Spatio-Temporal Network Anomaly Detection by Assessing Deviations of Empirical Measures*

Ioannis Ch. Paschalidis† Georgios Smaragdakis‡

***Abstract*—We introduce an Internet traffic anomaly detection mechanism based on large deviations results for empirical measures. Using past traffic traces we characterize network traffic during various time-of-day intervals, assuming that it is anomaly-free. We present two different approaches to characterize traffic: (*i*) a model-free approach based on the method of types and Sanov's theorem, and (*ii*) a model-based approach modeling traffic using a Markov modulated process. Using these characterizations as a reference we continuously monitor traffic and employ large deviations and decision theory results to "compare" the empirical measure of the monitored traffic with the corresponding reference characterization, thus, identifying traffic anomalies in real-time. Our experimental results show that applying our methodology (even short-lived) anomalies are identified within a small number of observations. Throughout, we compare the two approaches presenting their advantages and disadvantages to identify and classify temporal network anomalies. We also demonstrate how our framework can be used to monitor traffic from multiple network elements in order to identify both spatial and temporal anomalies. We validate our techniques by analyzing real traffic traces with time-stamped anomalies.**

***Index Terms*—Network security, statistical anomaly detection, method of types, Markov processes, large deviations.**

I. INTRODUCTION

ALTHOUGH significant progress has been made in network monitoring instrumentation, automated on-line traffic anomaly detection is still a missing component of modern network security and traffic engineering mechanisms. Network anomaly detection approaches can be broadly grouped into two classes: *signature-based anomaly detection* where known patterns of past anomalies are used to identify ongoing anomalies (e.g., see [1, 2] for intrusion detection), and *anomaly detection* which identifies patterns that substantially deviate from normal patterns of operation [3]. Earlier work has showed that systems based on pattern matching had detection rates below 70% [4, 5]. Furthermore, such systems need constant (and expensive) updating to keep up with new attack signatures. As a result, more attention has to be drawn to methods for traffic anomaly detection since they can identify even novel (unseen) types of anomalies.

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

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

‡ Deutsche Telekom Laboratories, Berlin, Germany, e-mail: georgios.smaragdakis@telekom.de.

In this work we focus on anomaly detection and in particular on *statistical anomaly detection*, where statistical methods are used to assess deviations from normal operation. Our main contribution is the introduction of a new statistical traffic anomaly detection framework that relies on identifying deviations of the empirical measure of some underlying stochastic process characterizing system behavior. In contrast with other approaches [1, 2, 6], we are not trying to characterize the abnormal operation, mainly because it is too complex to identify all the possible anomalous instances (especially those that have never been observed). Instead we observe past system behavior and, assuming that it is anomaly-free, we obtain a statistical characterization of "normal behavior." Then, using this knowledge we continuously monitor the system to identify time instances where system behavior does not appear to be normal. The novelty of our approach is in the way we characterize normal behavior and in how we assess deviations from it. More specifically, we propose two methods to characterize normal behavior: (*i*) a *model-free* approach employing the method of types [7] to characterize the type (i.e., empirical measure) of an independent and identically-distributed (i.i.d.) sequence of appropriately averaged system activity, and (*ii*) a *model-based* approach where system activity is modeled using a *Markov Modulated Process (MMP)*. Given these characterizations, we employ the theory of *Large Deviations (LD)* [7] and decision theory results to assess whether current system behavior *deviates* from normal. LD theory provides a powerful way of handling rare events and their associated probabilities with an asymptotically exact exponential approximation. The key technical results we rely upon are Sanov's theorem [7] in the model-free approach, a related result for the empirical measure of a Markov process for the model-based case, and Hoeffding's [8] composite hypothesis testing rule for assessing deviations from normal activity.

We note that the words "traffic" and "router" are purposefully absent from the previous paragraph. Rather, we use the generic term "system". This is to indicate that our approach can be easily adapted to identify anomalies in any trace of system activity we would like to monitor (e.g., access to various application ports, IP source-destination addresses, system calls, etc.). In this paper, however, we focus on two case studies: (*a*) three different representations (bytes, packets and flows) of sampled origin-destination flow data from a backbone network, and (*b*) the aggregate traffic that arrives to or originates from the border router of some local area network (LAN) we wish to monitor.

Traffic has diurnal variations which are primarily due to human activity. However, for relatively short time-scales (e.g., of about an hour), and especially during busy hours, stationary models can be appropriate. The model-free approach aggregates traffic over short time intervals to which we will refer to as *time buckets*. Although the correlation between samples in short time scales is significant, it reduces rapidly between aggregates over a time bucket. Hence, we consider the sequence of traffic aggregates over a time bucket as an i.i.d. sequence and employ the method of types to characterize its distribution. Our model-based approach uses an MMP process to model legitimate traffic during some time-of-day interval. Earlier work has shown that MMP models can accurately characterize network traffic [9, 10], at least for the purposes of estimating important quality-of-service metrics.

We should point out that in both the i.i.d. and MMP settings our goal is not to adopt the most sophisticated and accurate traffic model; the quality of the model should be judged based on whether it is useful in anomaly detection. There is no doubt that there are much more detailed and complex models that produce traces that are, statistically, very similar to actual traffic traces. On the other hand, it can be argued that low-dimensional models (which are bound to be more inaccurate) are more desirable for anomaly detection because they tend to be more robust to "normal" statistical fluctuations one expects in traffic. Our experimental results demonstrate that our characterizations, based on low-dimensional models, do a respectful job in identifying anomalies.

The proposed framework is general enough to also take into account spatial information. By combining observations at different locations of a network we are able to construct multi-dimensional stochastic processes to characterize system behavior. Both our model-free and our MMP approaches can work with vector characterizations and identify spatio-temporal anomalies (as both temporal and spatial information is preserved).

The methods we present are statistical; as a result, our approach has the potential of detecting novel anomalies, such as previously unseen attacks. This is crucial for network security as new types of attacks are constantly being engineered. A novel feature of our approach is that it compares subtle *distributional* differences between the reference traffic characterization and observed traffic traces. As we will see, this is critical as it enables us to detect attacks –including some short-lived ones or anomalies that traverse different locations of a network – that do not result in significant changes in traffic volume. First or second moments of traffic measurements would be too insensitive to these types of attacks. A distinctive feature of our approach is that it appears able to identify temporal and spatial anomalies in short time-scales as opposed to techniques working over much longer time-scales [3] or others that try to identify spatial anomalies [11–13] collapsing the temporal correlation of network feature samples. As we will also describe later, the only infrastructure requirement in order to deploy our method is a simple counter.

As is common in other statistical anomaly detection approaches, we rely upon observing the system during an anomaly-free period to learn what constitutes normal behavior. Of course, one can never ensure that a trace of system activity is anomaly-free. Yet, even in those cases that the reference trace is "tainted" it is useful to know that the current activity is statistically different. Non-stationarity in system activity can also cause problems to our approach as it may be responsible for legitimate distributional differences between past and current activity. However, as long as stationarity holds over relatively short periods of time one could often update the reference trace with more recent and relevant activity, thus, reducing the possibility of misdetections and false alarms. In our work, the time-scale over which measurements are available depends on the application and is of secondary concern. All that is needed is enough data to reliably characterize normal behavior. The latter is a reasonable assumption when network activity is continuously monitored. We report a number of experimental results from applying our approaches to two different network traces: (*a*) one week of sampled origin-destination flow data from the Abilene backbone network [12], and (*b*) the 1999 MIT Lincoln Lab (DARPA evaluation) trace [5]. We are able to detect a variety of anomalies such as attack and volume anomalies (even short ones) within a few samples.

The rest of the paper is organized as follows. In Section II we present our model-free method for anomaly detection. In Section III, we provide the basic theoretical background of our model-based method. In Section IV we extend the initial framework incorporating spatial information. In Sections V and VI we compare the two methods and validate our methodology using real measurements with time-stamped anomalies. In Section V we also report on the performance of our spatio-temporal framework in large scale networks. In Section VII we review related work and identify the major differences with our approaches. We conclude in Section VIII.

II. A Model-Free Approach

In this section we discuss our model-free approach and provide the structure of an algorithm to detect temporal network anomalies. As noted in the Introduction we focus on traffic at points of interest in the network, even though our approach is general enough to be applied to any trace of system activity. We assume that the *traffic trace* we monitor (in bits/bytes/packets/flows per time unit), corresponding to a specific time-of-day interval, can be characterized by a stationary model over a certain period (e.g., a month) if no technological changes (e.g., link bandwidth upgrades) have taken place.

Consider a time series $X_1, \ldots, X_n$ of traffic activity (say, in bits/bytes/packets/flows per sample). Let Y_t^b the *partial sum* (or aggregate traffic) over the time bucket starting at $(t-1)b$ and containing b samples, namely, $Y_t^b = \sum_{i=1}^{b} X_{(t-1)b+i}$. The crucial assumption we make is that $Y_1^{b^*}, \ldots, Y_{\lfloor n/b \rfloor}^{b^*}$ is an i.i.d. sequence for some appropriate bucket size b^*. This is a reasonable assumption in many settings as temporal correlations tend to become weaker over longer time intervals. In the Appendix we provide some discussion and a methodology to determine an appropriate b^*.

We quantize the values of the partial sums $Y_t^{b^*}$ mapping them to the finite set $\Sigma = \{\alpha_1, \ldots, \alpha_N\}$ of cardinality N. For the rest of the paper, we will be referring to Σ as the *underlying alphabet*. The quantization is done as follows: we let $[r_0, r_N]$ be the range of values $Y_t^{b^*}$ takes, divide it into N subintervals $[r_0, r_1], \ldots, [r_{N-1}, r_N]$ of equal length, and map $[r_{i-1}, r_i]$ to α_i for $i = 1, \ldots, N$. To select the appropriate size of the alphabet N we follow the approach of [10] and use the so called Akaike's Information Criterion (AIC) [14]. In particular, N is set to minimize:

$$Q(N, r_1, \ldots, r_{N-1}) = -\mathscr{L}(r_1, \ldots, r_{N-1}) + N(N-1),$$

where $\mathscr{L}(\cdot)$ is the *log*-likelihood of the model with respect to a process realization. The key observation motivating the AIC is that $\mathscr{L}(\cdot)$ tends to favor models with a larger number of free parameters. The AIC removes this bias by introducing a penalty for the number of free parameters; thus, the resulting N is considered the most appropriate for the given trace (minimizing modeling and estimation error). Once we have N, elements of the alphabet that are not observed in the trace are merged with neighboring ones to obtain N' which is the final size of the alphabet.

A. Large deviations of the empirical measure

Combinatorial methods can be applied for the empirical measures of Σ-valued process. Let $\mathbf{Y}_t^{b^*} = (Y_{t-w+1}^{b^*}, \ldots, Y_t^{b^*})$ be the trace of the w most recent partial sums using a bucket size b^*. We assume that the elements of $\mathbf{Y}_t^{b^*}$ are i.i.d., following a law $\boldsymbol{\mu} \in M_1(\Sigma)$, where $M_1(\Sigma)$ denotes the space of all probability measures on the alphabet Σ. Let also, $\Sigma_{\boldsymbol{\mu}}$ denote the support of the law $\boldsymbol{\mu}$, i.e., $\Sigma_{\boldsymbol{\mu}} = \{\alpha_i : \boldsymbol{\mu}(\alpha_i) > 0\}$.

Define the *type* (empirical measure) of $\mathbf{Y}_t^{b^*}$ as

$$\mathscr{E}_w^{\mathbf{Y}_t^{b^*}}(\alpha_i) = \frac{1}{w} \sum_{j=1}^{w} \mathbf{1}_{\alpha_i}(Y_{t-w+j}^{b^*}), \qquad i = 1, \ldots, N,$$

where $\mathbf{1}_{\alpha_i}$ is the indicator function of $Y_{t-w+j}^{b^*}$ being of type α_i. Namely, $\mathscr{E}_w^{\mathbf{Y}_t^{b^*}}(\alpha_i)$ is the fraction of occurrences of α_i in the sequence $\mathbf{Y}_t^{b^*}$. Let $\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}} = (\mathscr{E}_w^{\mathbf{Y}_t^{b^*}}(\alpha_1), \ldots, \mathscr{E}_w^{\mathbf{Y}_t^{b^*}}(\alpha_N))$.

The next theorem, which is due to Sanov, establishes a large deviations result for $\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}}$ (see [7, Sec. 2.1.10]).

Theorem II.1 *For every $\boldsymbol{\nu} \in M_1(\Sigma)$ let*

$$I_1(\boldsymbol{\nu}) = H(\boldsymbol{\nu}|\boldsymbol{\mu}),$$

where $H(\boldsymbol{\nu}|\boldsymbol{\mu})$ is the relative entropy of the probability vector $\boldsymbol{\nu}$ with respect to $\boldsymbol{\mu}$:

$$H(\boldsymbol{\nu}|\boldsymbol{\mu}) \triangleq \sum_{i=1}^{N} \boldsymbol{\nu}(\alpha_i) \log \frac{\boldsymbol{\nu}(\alpha_i)}{\boldsymbol{\mu}(\alpha_i)}.$$

Then, for any set Γ of probability vectors in $M_1(\Sigma)$

$$\begin{aligned} -\inf_{\boldsymbol{\nu} \in \Gamma^\circ} I_1(\boldsymbol{\nu}) &\leq \liminf_{w \to \infty} \frac{1}{w} \log \mathbf{P}[\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}} \in \Gamma] \\ &\leq \limsup_{w \to \infty} \frac{1}{w} \log \mathbf{P}[\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}} \in \Gamma] \leq -\inf_{\boldsymbol{\nu} \in \Gamma} I_1(\boldsymbol{\nu}), \end{aligned}$$

where Γ° denotes the interior of Γ.

More intuitively, Theorem II.1 states that for a long trace $\mathbf{Y}_t^{b^*}$ (i.e., large w) its empirical measure is "close to" $\boldsymbol{\nu}$ with probability that behaves as

$$\mathbf{P}[\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}} \approx \boldsymbol{\nu}] \asymp e^{-wI_1(\boldsymbol{\nu})}.$$

We will be referring to exponents like $I_1(\boldsymbol{\nu})$ as the *exponential decay rate* of the corresponding probability – in this case $\mathbf{P}[\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}} \approx \boldsymbol{\nu}]$.

B. Anomaly detection

Theorem II.1 can be used to identify anomalies. Specifically:

1) From an anomaly-free trace construct the alphabet $\Sigma = \{\alpha_1, \ldots, \alpha_N\}$ and the empirical measure (law) $\boldsymbol{\mu}$ induced by this sequence.
2) For each time t let $\mathbf{Y}_t^{b^*} = (Y_{t-w+1}^{b^*}, \ldots, Y_t^{b^*})$ be the trace of the w most recent partial sums using a bucket size b^*. Compute its empirical measure and let $\boldsymbol{\mathscr{E}}_w^{\mathbf{Y}_t^{b^*}} = \boldsymbol{\nu}_{t,w}$ be the result.

Based on Thm. II.1, $\rho_{t,w} \triangleq e^{-wI_1(\boldsymbol{\nu}_{t,w})}$ approximates the probability that the trace $\mathbf{Y}_t^{b^*}$ is drawn from the probability law $\boldsymbol{\mu}$. Thus, if $\rho_{t,w}$ is consistently low over some observed time interval, we can conclude that the observed trace deviates from the anomaly-free trace, which indicates an anomaly. In particular, we can identify an anomaly at time t if

$$\rho_{\tau,w} < \delta, \qquad \forall \tau = t-k+1, \ldots, t, \tag{1}$$

where n is the length of the traffic trace we process, $w = \lfloor n/b^* \rfloor$ is the number of partial sums we generate from this trace, and δ is the detection threshold we use. The parameters n, b^*, k, δ affect the performance of the above rule and can be tuned experimentally. Although, this is a valid approach and yields good results in all experiments we report later, tuning the rule's parameters can be costly. Clearly, the smaller the k and the larger the δ the smaller the misdetection probability (i.e., one minus the probability of successfully identifying an anomaly) and the larger the false alarm probability. Next, we present a more formal anomaly detection rule that optimally resolves this trade-off.

C. A formal anomaly detection test

Theorem II.1 rigorously identifies a distance metric – the exponent $I_1(\boldsymbol{\nu})$ – between the two measures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, constructed as specified in Steps (1) and (2) of Subsection II-B. The key question we wish to answer is whether $\mathbf{Y}_t^{b^*}$ is generated from $\boldsymbol{\mu}$ or from some other law. This is known as a *composite hypothesis* problem as one hypothesis (no anomaly) has a known law $\boldsymbol{\mu}$ while the alternative hypothesis is characterized by a family of laws (all laws other than $\boldsymbol{\mu}$). Hoeffding ([8]; see also [7]) has suggested an optimality criterion for these problems and a rule that is optimal. It is also well known, that the empirical measure $\boldsymbol{\nu}_{t,w}$ of $\mathbf{Y}_t^{b^*}$ is a sufficient statistic. In particular, Hoeffding's rule is

$$\mathscr{S}_1^{*w}(\mathbf{Y}_t^{b^*}) = \begin{cases} \text{no anomaly}, & \text{if } I_1(\boldsymbol{\nu}_{t,w}) < \eta, \\ \text{anomaly}, & \text{otherwise}, \end{cases} \tag{2}$$

and is optimal in the sense that it maximizes the exponential decay rate of the misdetection probability over all tests with a false alarm probability with exponential decay rate larger than η. That is, the parameter η used in the test controls the false alarm probability and can be determined as $\eta = -\log \epsilon / w$ where ϵ is our tolerable false alarm rate. The remaining parameters w, b^* affect the rule's performance and can be tuned experimentally. Notice that we can compute consecutive $\boldsymbol{\nu}_{t,w}$ by using a sliding window of length w. Thus, we generate a new decision with every traffic sample based on the most recent w-long trace. As we will see, this enables us to detect anomalies very fast.

We proceed by presenting a model-based method, where the i.i.d assumption is not a requirement, thus, it can be directly applied to the time-series and not to the partial sums.

III. A Model-Based Approach

The approach of Section II aggregated traffic over a time bucket to yield an i.i.d. sequence. One potential disadvantage of this aggregation is that it increases the response time to an anomaly since data is being processed on the slower time-scale of time buckets. In this section, the question we are seeking to answer is whether it is possible to process data on the timescale we collect them. To that end, and because the i.i.d. assumption will no longer hold, we will impose some more structure on the stochastic nature of the traffic time-series. In particular, we will assume a Markovian structure as it is tractable and has been shown to represent traffic well [9, 10], at least for the purpose of estimating distribution-dependent metrics like loss probabilities.

A. An MMP model

We start again with a time series $X_1, \ldots, X_n$ of traffic activity during a small time interval (several hours) which we will model as an MMP process. Such a process is characterized by an underlying Markov chain with transition probability matrix $\boldsymbol{\Xi} = \{p(i,j)\}_{i,j=1}^{M}$. To each state i, $i = 1, \ldots, M$, we associate an interval $[r_{i-1}, r_i]$ of real numbers from which traffic activity observations are drawn. That is, when the MMP is in state i at time t then X_t takes values in $[r_{i-1}, r_i]$. (For the application we are considering we do not need to specify how observations are drawn from $[r_{i-1}, r_i]$; in general they can follow some probability distribution.) MMPs, when the state is "hidden", are also known in the literature as Hidden Markov Models (HMMs) [15]. We restrict ourselves to models in which the ranges of possible observations corresponding to different states are disjoint. Thus, an observation can be uniquely associated to an MMP state and the state is no longer hidden.

To model the traffic trace as a MMP we let $[r_0, r_M]$ be the range of all observations we make, split $[r_0, r_M]$ into M subintervals of equal length, and assign state i, $i = 1, \ldots, M$, to interval $[r_{i-1}, r_i]$. To select the appropriate number of states M we use the AIC as in Section II. Given M, the transition probabilities $\boldsymbol{\Xi}$ are obtained via maximum likelihood estimation. Specifically, let $\mathbf{Y}$ denote a sequence $Y_1, Y_2, \ldots, Y_n$ of states that the Markov chain visits. A maximum likelihood estimator of the transition probabilities is given by

$$\hat{p}_n(i,j) = q_{nf}(j \mid i) \triangleq \frac{q_n(i,j)}{q_{n1}(i)}, \; i,j = 1, \ldots, M, \quad (3)$$

where $q_n(i,j)$ denotes the fraction of transitions from i to j in the sequence $\mathbf{Y}$ and $q_{n1}(i)$ the fraction of transitions out of i. We assume that n is large enough to have $q_{n1}(i) > 0$ for all i. As $n \to \infty$, $\hat{p}_n(i,j) \to p(i,j)$ with probability one (w.p.1). We consider the constructed model to be reliable since it is the outcome of a long period of anomaly-free observations. Different models can be constructed for different time-of-day intervals (business hours, evening hours, overnight, etc.).

B. Large deviations of the empirical measure

Once we obtain the MMP model from an anomaly-free trace we are interested in comparing ongoing traffic activity to the model in order to identify potential deviations.

Assume that the MMP has an irreducible underlying Markov chain with transition probability matrix $\boldsymbol{\Xi} = \{p(i,j)\}_{i,j=1}^{M}$. Let $\mathbf{p}$ denote the vector consisting of the rows of $\boldsymbol{\Xi}$. As before, $\mathbf{Y}$ denotes a sequence $Y_1, Y_2, \ldots, Y_n$ of states that the Markov chain visits with the initial state being $Y_0 = \sigma$. Consider the empirical measures

$$\mathscr{E}_{n,2}^{\mathbf{Y}}(\mathbf{y}) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\mathbf{y}}(Y_{k-1} Y_k),$$

where $\mathbf{y} \in \mathscr{A}^2 \triangleq \{1, \ldots, M\} \times \{1, \ldots, M\}$ and $\mathbf{1}_{\mathbf{y}}$ is the indicator function for the subset $\mathbf{y}$. Note that when $\mathbf{y} = (i,j) \in \mathscr{A}^2$ the empirical measure $\mathscr{E}_{n,2}^{\mathbf{Y}}(\mathbf{y})$ denotes the fraction of times that the Markov chain makes transitions from i to j in the sequence $\mathbf{Y}$. Let now $\mathscr{A}_{\mathbf{p}}^2 \triangleq \{(i,j) \in \mathscr{A}^2 \mid p(i,j) > 0\}$ denote the set of pairs of states that can appear in the sequence $Y_1, Y_2, \ldots, Y_n$ and denote by $M_1(\mathscr{A}_{\mathbf{p}}^2)$ the standard $|\mathscr{A}_{\mathbf{p}}^2|$-dimensional probability simplex, where $|\mathscr{A}_{\mathbf{p}}^2|$ denotes the cardinality of $\mathscr{A}_{\mathbf{p}}^2$. Note that the vector of $\mathscr{E}_{n,2}^{\mathbf{Y}}(\mathbf{y})$'s, denoted by $\boldsymbol{\mathscr{E}}_{n,2}^{\mathbf{Y}} = (\mathscr{E}_{n,2}^{\mathbf{Y}}(\mathbf{y});\; y \in \mathscr{A}_{\mathbf{p}}^2)$, is an element of $M_1(\mathscr{A}_{\mathbf{p}}^2)$. For any $\mathbf{q} \in M_1(\mathscr{A}_{\mathbf{p}}^2)$, let

$$q_1(i) \triangleq \sum_{j=1}^{M} q(i,j) \qquad \text{and} \qquad q_2(i) \triangleq \sum_{j=1}^{M} q(j,i) \quad (4)$$

be its marginals. Whenever $q_1(i) > 0$, let $q_f(j \mid i) \triangleq q(i,j)/q_1(i)$. As before, we will be using the notation $\mathbf{q}_f = (q_f(1 \mid 1), \ldots, q_f(M \mid 1), q_f(1 \mid 2), \ldots, q_f(M \mid M))$. We say that a probability measure $\mathbf{q} \in M_1(\mathscr{A}_{\mathbf{p}}^2)$ is *shift invariant* if both its marginals are identical, i.e., $q_1(i) = q_2(i)$ for all i. A large deviations result for $\boldsymbol{\mathscr{E}}_{n,2}^{\mathbf{Y}}$ is established in the next theorem and is proven in [7, Sec. 3.1.3].

Theorem III.1 ([7]) *For every $\mathbf{q} \in M_1(\mathscr{A}_{\mathbf{p}}^2)$ let*

$$I_2(\mathbf{q}) = \begin{cases} \displaystyle\sum_{i=1}^{M} q_1(i) H(q_f(\cdot \mid i) \mid p(i,\cdot)), & \textit{if } \mathbf{q} \textit{ is shift invariant,} \\ \infty, & \textit{otherwise,} \end{cases}$$

where $H(q_f(\cdot \mid i) \mid p(i,\cdot))$ is the relative entropy, that is,

$$H(q_f(\cdot \mid i) \mid p(i,\cdot)) = \sum_{j=1}^{M} q_f(j \mid i) \log \frac{q_f(j \mid i)}{p(i,j)}.$$

Then, for any set Γ of probability vectors in $M_1(\mathscr{A}^2_{\mathbf{p}})$,

$$-\inf_{\mathbf{q}\in\Gamma^\circ} I_2(\mathbf{q}) \le \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[\mathscr{E}^{\mathbf{Y}}_{n,2} \in \Gamma] \le \limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[\mathscr{E}^{\mathbf{Y}}_{n,2} \in \Gamma] \le -\inf_{\mathbf{q}\in\Gamma} I_2(\mathbf{q}),$$

where Γ° denotes the interior of Γ.

More intuitively, Theorem III.1 states that for a long trace $\mathbf{Y}$ (i.e., large n) its empirical measure is "close to" $\mathbf{q}$ with probability that behaves as

$$\mathbf{P}[\mathscr{E}^{\mathbf{Y}}_{n,2} \approx \mathbf{q}] \asymp e^{-nI_2(\mathbf{q})}.$$

C. Anomaly detection

As in the model-free method, Theorem III.1 can be used to identify anomalies. Specifically:

1) From an anomaly-free trace obtain a MMP as outlined in Subsection III-A. Let $\mathbf{p}$ be the resulting transition probability vector.
2) For each time t let $\mathbf{Y}_t = (Y_{t-n+1}, \dots, Y_t)$ be the trace of current traffic activity consisting of n consecutive traffic measurements. Compute its empirical measure and let $\mathscr{E}^{\mathbf{Y}_t}_{n,2} = \mathbf{q}_{t,n}$ be the result.

Based on Thm. III.1, $\rho_{t,n} \triangleq e^{-nI_2(\mathbf{q}_{t,n})}$ approximates the probability that the trace $\mathbf{Y}_t$ is drawn from the MMP with transition probability vector $\mathbf{p}$. Thus, if $\rho_{t,n}$ is consistently low over some observed time interval, we conclude that the observed trace does not "appear drawn" from the reliable model, which indicates an anomaly. For an automated anomaly detection rule one can specify some parametric threshold rule and tune the parameters using a "training" data set. For instance, we can identify an anomaly at time t if

$$\rho_{\tau,n} < \delta, \qquad \forall \tau = t-k+1, \dots, t, \tag{5}$$

where n, k and δ are the parameters affecting the rule's performance (detection and false alarm rates). Although, this is a valid approach and yields good results in all experiments we report later, tuning the rule's parameters can be costly and can lead to an arbitrary balancing of the detection and false alarm rates. Next, we present a more formal anomaly detection rule that optimally resolves this trade-off.

D. A formal anomaly detection test

As before, Theorem III.1 identifies a distance metric – the exponent $I_2(\mathbf{q})$ – between the two Markov measures $\mathbf{p}$ and $\mathbf{q}$, constructed as specified in Steps (1) and (2) of Subsection III-C. A generalized (to the Markov case) Hoeffding rule can now be used to identify an anomaly. Such a generalization has been shown in [16]. As in Sec. II-C, the composite hypothesis testing problem is to determine whether $\mathbf{Y}_t$ is generated from $\mathbf{p}$ or from some other law. Hoeffding's rule is

$$\mathscr{S}^{*n}_2(\mathbf{Y}_t) = \begin{cases} \text{no anomaly,} & \text{if } I_2(\mathbf{q}_{t,n}) < \eta, \\ \text{anomaly,} & \text{otherwise.} \end{cases} \tag{6}$$

As before, η determines the desirable decay rate for the false alarm probability. The parameter n should be large enough for the asymptotics to be accurate, but as we will see, moderate values of n suffice.

We claim that both our approaches are general on-line methods to compare an observed traffic trace with a model that represents "typical" behavior. Many types of traffic anomalies can be detected such as attacks, worms, intrusion, port scanning, network failures, or flash crowds. We proceed by extending our framework, incorporating spatial information to identify network anomalies.

IV. Incorporating Spatial Information

The approaches we discussed so far only exploit temporal information. Yet, activity traces of interest can be collected in many locations and attacks at one place may be precursors or aftershocks of attacks elsewhere. Consider, for example, a worm spreading in the Internet or an orchestrated distributed attack of vulnerable computational resources in a single server.

We introduce spatial information in the previous models by considering vectors of traffic activity at various locations of interest in the network. More specifically, consider the traffic activity time series $\mathbf{X}_1, \dots, \mathbf{X}_n$, where $\mathbf{X}_i \in \mathbb{R}^d$ now represents the rate of the network feature of interest (e.g., number of bytes/packets/flows) during time slot i at all d network locations we would like to monitor. The methods we discussed earlier –method of types and MMP model– readily extend to this case as they do not depend on $\mathbf{X}_i$ being a scalar. Of course, and especially for large d, one would now require longer traces to estimate the set of parameters and the anomaly detection algorithm would need longer samples to identify an anomaly. On the other hand by selecting a small number of network elements as monitors it is tractable to incorporate spatial information, as we experimentally validate in the next sections. The only additional requirements in order to incorporate spatial information in the manner suggested is that the network elements must be synchronized and being able to exchange the time series of the network features they monitor. This is a reasonable assumption for currently manufactured routers, especially when time slots are on the order of 1 minute.

We proceed by presenting spatio-temporal anomaly detection algorithms for both the model-free and the model-based approach.

A. Spatio-temporal anomaly detection algorithm using the model-free method

We first select d network elements or features we want to monitor. As in Section II, we will work with $\mathbf{Y}^{b^*,j}_t = (Y^{b^*,j}_{t-w+1}, \dots, Y^{b^*,j}_t)$ which denotes the trace of the w most recent partial sums at network element j using a bucket size b^*, for $j = 1, \dots, d$. We use $\mathbf{Y}^{b^*}_t = (\mathbf{Y}^{b^*,1}_t, \dots, \mathbf{Y}^{b^*,d}_t)$ to denote the trace of system activity at all network elements d.

1) For each chosen network element or feature j, apply Step (1) of Subsection II-B in order to obtain the underlying alphabet and the empirical measure.
2) Create the underlying multi-dimensional alphabet $\Sigma^d = \{\alpha_1^{(1)}, \ldots, \alpha_{N_1}^{(1)}\} \times \cdots \times \{\alpha_1^{(d)}, \ldots, \alpha_{N_d}^{(d)}\}$ and compute the associated empirical measure (law) $\boldsymbol{\mu}^d$ from past (anomaly-free) observations.
3) For each time t consider the trace $\mathbf{Y}_t^{b^*}$ of the w most recent partial sums and compute its empirical measure which yields $\mathscr{E}_{w,b^*}^{\mathbf{Y}_t^{b^*}} = \boldsymbol{\nu}_{t,w}^d$, where the superscript d indicates that this represents the fraction of occurrences of Σ^d-letters in the trace $\mathbf{Y}_t^{b^*}$.

As in the scalar case, $\rho_{t,w} \stackrel{\triangle}{=} e^{-wI_1(\boldsymbol{\nu}_{t,w}^d)}$ approximates the probability that the trace $\mathbf{Y}_t^{b^*}$ is drawn from the probability law $\boldsymbol{\mu}^d$. We can use the exact same test of (2) to identify an anomaly while controlling the desirable false alarm rate. As before, the parameters w, b^* can be selected to improve the performance of the algorithm.

B. Spatio-temporal anomaly detection algorithm using the model-based method

We first select d network elements or features we want to monitor and consider the time series of traffic activity $\mathbf{X}_n^j = (X_1^j, \ldots, X_n^j)$ at each network element j, $j = 1, \ldots, d$. For each such time series we split the range of values it takes into M_j subintervals following the procedure described in Section III, thus, defining M_j states. M_j is selected using the AIC criterion. Let $s_1^j, \ldots, s_{M_j}^j$ denote the states for the jth network element. Putting together the traces from all d network elements or features we form the trace $\mathbf{X}_n = (\mathbf{X}_n^1, \ldots, \mathbf{X}_n^d)$ with corresponding states $(s^1, \ldots, s^d)$, where $s^j \in \{s_1^j, \ldots, s_{M_j}^j\}$ is the state of the jth network element or feature. We have now constructed a multi-dimensional process and we assume that states evolve according to a Markov chain. We will use $\mathbf{Y}_t$ to denote the state of this multi-dimensional Markov chain at time t. The anomaly detection algorithm outlined next is similar to the scalar case discussed in Section III.

1) Using past (anomaly free) information compute the transition probability vector $\mathbf{p}^d$ for the multi-dimensional Markov chain.
2) For each time t let $\mathbf{Y}_{t,n} = (\mathbf{Y}_{t-n+1}, \ldots, \mathbf{Y}_t)$ be the sequence of states the multi-dimensional Markov chain visits over n consecutive time slots. Compute its empirical measure and let $\mathscr{E}_{n,2}^{\mathbf{Y}_{t,n}} = \mathbf{q}_{t,n}^d$ be the result, where the superscript d indicates that this is the empirical measure of the d-dimensional Markov chain we have constructed.
3) Then, $\rho_{t,n} \stackrel{\triangle}{=} e^{-nI_2(\mathbf{q}_{t,n}^d)}$ approximates the probability that the trace $\mathbf{Y}_{t,n}$ is drawn from the MMP with transition probability matrix $\mathbf{p}^d$. Use the rule in (6) to identify an anomaly, where η bounds the asymptotic decay rate of the false alarm probability.

We remark that the proposed mechanism may not scale well with the number d of network elements. Nevertheless, it provides on-line anomaly detection for selected points of interest, especially when the size of the alphabet (model-free anomaly detection) or the number of states (model-based anomaly detection) is small; this turned out to be the case in our experiments, as we comment on in the following sections.

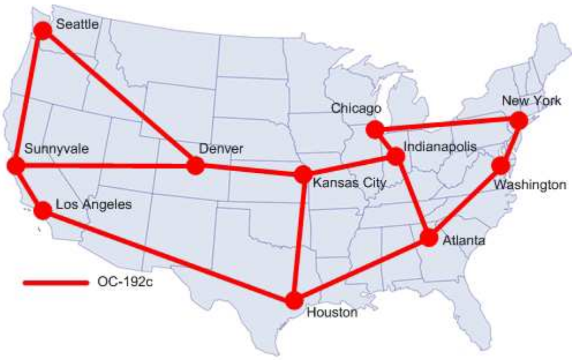


Fig. 1. The Abilene backbone network and PoPs.

V. Experimental Setup I: The Abilene Data Set

In this section, we validate our methodology against real traffic traces from a backbone network. Our source of data is the IP-level traffic flow measurements collected form every Point of Presence (PoPs) in the Abilene Internet2 backbone network. Abilene is the major academic network, connecting over 200 universities in the US, and peering with other research networks in Europe and Asia. Abilene has 11 PoPs resulting in 121 origin-destination flows (see Fig. 1).

The data we are using is sampled flow data from every router of Abilene for a period of one week (April 7 to 13, 2003). Sampling is random capturing 1% of all packets entering every router. Three different representations (features) of sampled flow data are used, a time-series of the number of bytes (B), of packets (P) and of flows (F). In order to avoid synchronization issues, the measurements are aggregated into 5 minutes bins. The issue of how packets are sampled is an important one but we do not consider it here because, in most cases, packet sampling is predetermined by the monitoring instrumentation. We comment later on the effect of the sampling frequency to our detection mechanisms.

A log with the anomalies that took place was also available. Three different types of anomalies are present: DoS: distributed Denial of service attack against a single victim; SCAN: scanning a host for a vulnerable port (port scan) or scanning the network for a target port (network scan); APLHA: unusually high rate point to point byte transfer. There are also some anomalies that are labeled as unknown (UNKN). In total there are 270 anomalies: 133 DoS, 81 SCAN, 32 ALPHA and the rest are unknown [12]. Origin-destination flows aggregate the traffic of thousands of connections (in a period of 5 minutes), thus, traffic anomalies of a destination may hide in the byte representation, but can appear in other representations like the packet or flow representations. DoS anomalies were always present in the packet (P) representation. This is expected as most DoS attacks bombard a single destination with a huge number of packets. Instances of DoS are not observed in the flow representation and may be observed in the byte (B) representation. The SCAN anomalies are observed only in the flow (F) representation. ALPHA anomalies are characterized by spikes in the byte representation only. Following the above observations we can even characterize anomalies that are denoted as unknown.

A. *Metrics of interest*

In order to validate the performance of our technique, we use two metrics, that are pretty standard for anomaly detection studies, namely the detection rate and the false alarm rate.

Detection Rate is the proportion of anomalies identified by the algorithm to the total number of existing anomalies. *False Alarms* are erroneous anomaly decisions. The *False Alarm Rate*, or the *False Alarm Probability*, is the proportion of anomaly decisions according to the rule (2) (or (6), depending on the model) in the set of all observation windows that were anomaly-free.

B. *Outline of the technique*

We apply both our methods to the different time-series (representations of B/P/F) for the 121 origin-destination flows. In order to avoid the effect of diurnal variation we consider 200 samples (each one representing the activity of 5 minutes) for every day of the week. We use as reference the activity that has been observed for the same time interval the previous day. For the first day of the week, as we do not have information from the previous day, we take as reference the network activity of the second day.

We apply the model-free approach following the algorithm described in Section II. We construct the alphabet of the three representations and the corresponding probability law for every day of the anomaly-free week. We then process the network activity for the next day according to the rule (2). Working with statistics of the autocorrelation function (see Appendix), we found that $b^* = 3$ and $w = 20$ are good values for our data set.

We also follow the approach of Section III to devise an appropriate MMP traffic model. Using this model, for every day of the week and for every time sample we use the rule (6) to identify an anomaly for an appropriately selected trace length n.

On a notational remark, we denote an anomaly as ORIG-DEST-xxxx, where ORIG is the ingress PoP, DEST is the egress PoP and xxxx is the time point in the time series (from $1 - 2016$) of the related representation where an anomaly occurs.

C. *Temporal anomaly detection examples*

In this subsection we discuss the performance of our framework and we compare the two proposed methods. Fig. 2 considers an ALPHA anomaly in the Washington-New York origin-destination flow (byte representation). In the top graph of that figure we plot the probability $\rho_{t,w} = e^{-wI_1(\boldsymbol{\nu}_{t,w})}$ that the trace is drawn from the anomaly free law when the model-free based method is applied. Notice that we can set an appropriate threshold so that when $\rho_{t,w}$ falls below that threshold we correctly identify an anomaly. The middle graph of Fig. 2 plots the exponent of $\rho_{t,w}$ and compares it to various thresholds η ($\eta_1, \eta_2, \eta_3 = 0.1\%, 1\%, 5\%$) that are set depending on our tolerance for false alarms. Fig. 3 illustrates detection of the same anomaly with the model-based method. Notice, that depending on the threshold η used in the middle graph we may or may not have a false alarm. It is worth noticing that spikes in the traffic volume are not necessarily identified as anomalies. In essence, both methods indicate an anomaly only when a substantial *distributional deviation* from the reference is identified. The same observations are valid for other types of attacks, e.g., DoS (Figure 4) and SCAN (Figure 5).

We note that except for Figs. 2 and 3, in all remaining figures we plot the exponents $I_1(\boldsymbol{\nu}_{t,w})$ and $I_2(\mathbf{q}_{t,n})$ (cf. (2) and (6)) to identify anomalies. Results using the probabilities $\rho_{t,w}$ and $\rho_{t,n}$ as anomaly indicators can be found in a preliminary study [17]. Figure 4 suggests that the model-free method may not always pinpoint the time of the anomaly due to aggregation effect of the partial sums and our use of a window of size w. Of course, smaller values of b^* lead to faster response times, but, also, larger modeling error due to increased correlation between consecutive elements of the time series $\mathbf{Y}_t^{b^*}$. In the absence of any information on when the anomaly occurs within this window we can estimate its time as $\frac{w}{2}$ time units after the first time $I_1(\boldsymbol{\nu}_{t,w})$ exceeds η. On the other hand, as seen in the figure, the model-based method can identify when an anomaly occurs more precisely. The disadvantage of the model-based method is that the false alarms rate is higher for the same η than that of the model-free method which benefits from the averaging over the time bucket b^*.

We summarize our results in Table I (see page 11). Notice that ϵ (hence, $\eta = -\log \epsilon / w$) can effectively control the false alarm rate; the actual false alarm rate is on the same order or magnitude as ϵ. We should also point out that the performance of our framework is related to the sampling frequency, i.e., if we increase the sampling frequency to 1 minute or even few seconds, we expect the sensitivity of our detection mechanism (and hence, the detection rate) to increase.

On the other hand dealing with more samples, will lead to a larger alphabet (if the model-free approach is applied) or a larger number of states (if the model-based approach is applied). In practice, an appropriate sampling design for anomaly detection applications using our proposed framework must take into account the sampling frequency and the dimensionality of the estimated alphabet or MMP respectively. Applying our framework in this data set we were able to construct alphabets with small cardinality (when the model-free approach was applied a typical size was 3 symbols) and small number of states (when the model-based approach was applied a typical size was 3 states). As we will argue in the next subsection this was beneficial for spatio-temporal network anomaly detection.

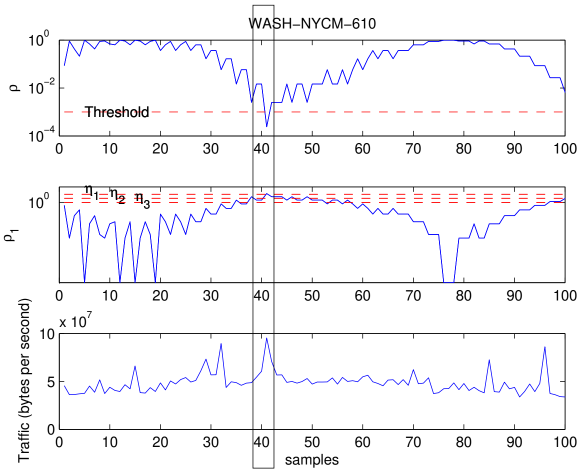


Fig. 2. Model-Free Method. (Top): Value of $\rho_{t,w}$. (Middle): Value of $I_1(\boldsymbol{\nu}_{t,w})$ used in (2). (Bottom): Byte representation for the Washington-New York origin-destination flow. The rectangle denotes an ALPHA anomaly.

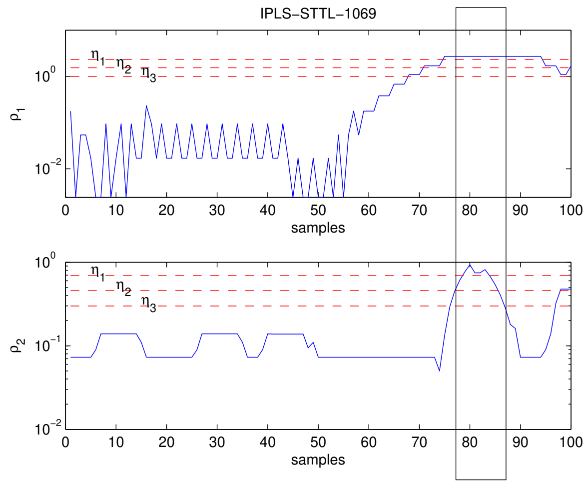


Fig. 4. Comparison of the two methods. (Top): Model-Free Method, (Bottom): Model-Based Method. The rectangle denotes a DoS anomaly.

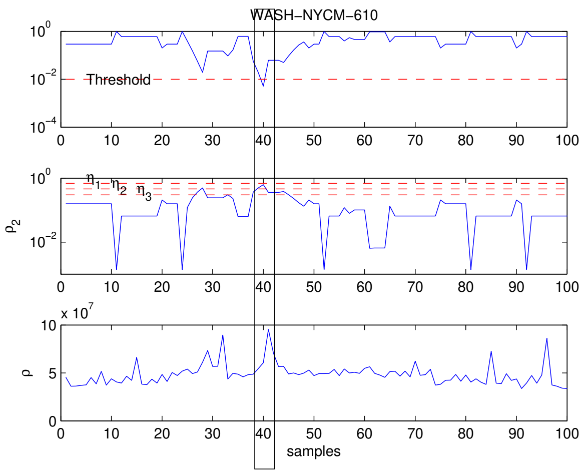


Fig. 3. Model-Based Method. (Top): Value of $\rho_{t,n}$. (Middle): Value of $I_2(\mathbf{q}_{t,n})$ used in (6). (Bottom): Byte representation for the Washington-New York origin-destination flow. The rectangle denotes an ALPHA anomaly.

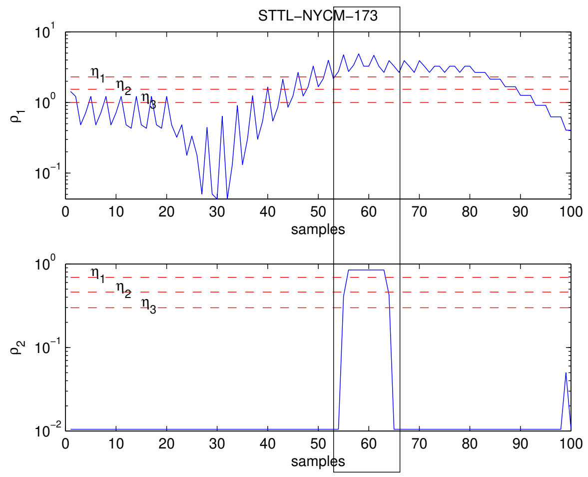


Fig. 5. Comparison of the two methods. (Top): Model-Free Method, (Bottom): Model-Based Method. The rectangle denotes a SCAN anomaly.

D. Spatio-temporal anomaly detection examples

We next turn our attention to spatio-temporal anomaly detection. As was mentioned before, the size of the alphabet and the number of states of the MMP for the Abilene data set is small when only temporal information is considered. Thus, it is easy to monitor subnets of PoPs (of low dimensionality d) by specifying the group of PoPs of interest and the role of each PoP (origin or destination). We present results for two case studies with different spatial characteristics. We apply our framework to: (a) flows that originate (end) from (at) PoPs that are 1-hop neighbors and (b) flows that originate (end) from (at) PoPs that are many hops away from each other.

In the first case study, the flows originate (end) at the Sunny Valley (SNVA) PoP with destination (originating from) the PoPs in its vicinity. In Figures 6 and 7, we illustrate instances of the identification of anomalies applying the model-free and the model-based method, respectively. The values of the parameters for the two methods are obtained from the temporal anomaly detection examples. Table II (in page 11) reports the detection and false alarm rates we achieved. It is worth noticing that the detection rate reached 100% and the false alarms rate was very low (lower than the values when only temporal anomalies were studied). This is due to two main reasons: (a) instantaneous high values in the time-series of observations that do not necessarily indicate attacks are smoothed due to time averaging, and (b) attacks may have temporal and/or spatial correlation, so the rare events are magnified when different locations are monitored.

In the second case study, we consider flows that originate from (end at) Los Angeles or Sunny Valley and end at (originate from) New York or Washington. In other words we examine flows that traverse a number of hops in the Abilene backbone network, traveling from the East (West) Coast to the West (East) Coast. Applying the same methodology as in the first case study, we were able again to achieve very high

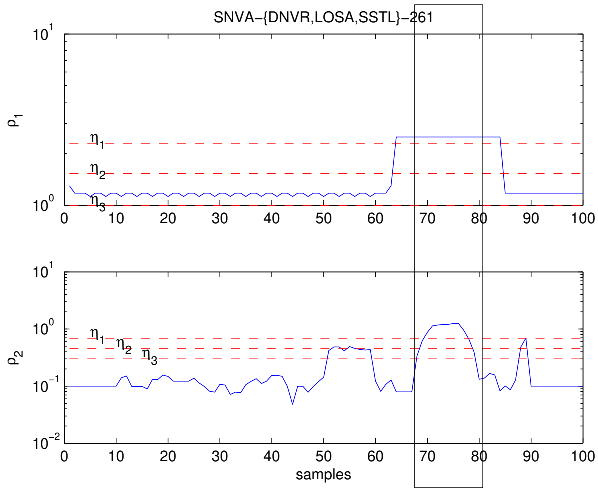


Fig. 6. Comparison of the two methods, incorporating both temporal and spatial information. (Top): model-free method, (Bottom): model-based method, when applied to the P representation of flows that originate from SNVA and end at the neighboring PoPs: DNVR, LOSA and STTL. The rectangle denotes a DoS anomaly.

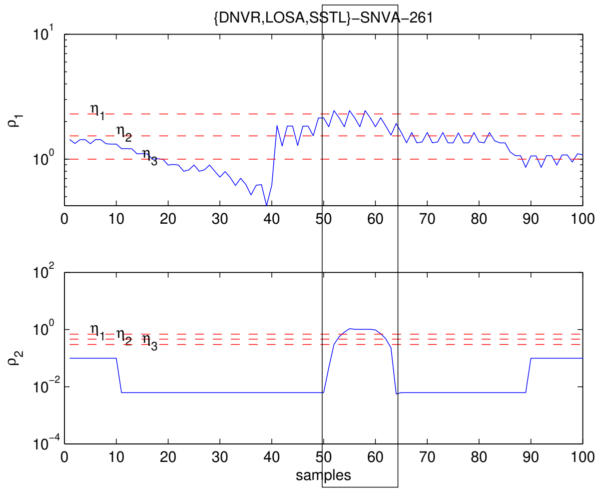


Fig. 7. Comparison of the two methods, incorporating both temporal and spatial information. (Top): model-free method, (Bottom): model-based method, when applied to the F representation of flows that end at SNVA originating from the neighboring PoPs: DNVR, LOSA and STTL. The rectangle denotes a SCAN anomaly.

detection rate with very low false alarms rate (see Table II). Although the PoPs are geographically apart it seems that the origin or the destination of attacks may be far away especially when the origin or the destination PoP (e.g., Los Angeles, New York) route a large proportion of traffic in the network. Although the space of possible combinations of origin or destination PoPs is very large, all the combinations (of a small number of PoPs - up to 4) we tested are in line with the results presented above.

VI. Experimental Setup II: The DARPA Evaluation Data Set

Next, we validate our method against the 1999 MIT Lincoln Lab (DARPA Evaluation) data set [5] which has been widely used for testing anomaly detection systems. The data set consists of tcpdump data collected at the border router of a Local Area Network (LAN). Using the tcpstat tool, we estimate the aggregated traffic per second from the packet headers of the packets served by the border router. Thus, we simulate a simple bit counter that counts the traffic volume served by the border router every one second. As was mentioned in the previous section, not all anomalies can be identified from the byte representations, because of the aggregation of end-to-end flows in origin-destination flows. For the case of one link though, the number of end-to-end flows are mush less, thus, by observing the aggregated traffic of the gateway link, we can more accurately identify network anomalies inside the LAN. Please also notice that only temporal information can be incorporated for this data set as only on router is monitored.

Three weeks of training data were provided in the DARPA Evaluation data set. The first and third weeks of the training data do not contain any attacks. This data was provided to facilitate the training of anomaly detection systems. The second week of the training data contains a selected subset of attacks from the 1998 DARPA evaluation data set in addition to several new attacks. In addition, two weeks of network based attacks in the midst of normal background data were also provided. The experimental setting includes different types of machines and operating systems.

There are 201 instances of about 56 types of attacks distributed throughout these two weeks. In particular, the attack events that either occurred or were attempted are the following: Denial of Service (DoS): unauthorized attempt to disrupt the normal functioning of a victim host or network; Remote to Local (R2L): obtaining user privileges on a local host by a remote user without proper authorization; User to Root (U2R): unauthorized access to local superuser or administrator privileges by a local unprivileged user; Surveillance or Probe (PROBE): unauthorized probing of a machine or network to look for vulnerabilities, explore configurations, or map the network's topology; and Data Compromise (DATA): unauthorized access or modification of data on a local or remote host. On a technical note, the aforementioned intrusion attacks (R2L, U2R), might be correlated with changes in traffic, e.g., ftp file transfer, or system reboot, resulting in distributional changes in traffic. A detailed taxonomy of the attacks is presented in [18].

Throughout our study, we also observed some anomalies that we could not classify using the DARPA Evaluation report. Trying to classify these anomalies, we found that some of them are correlated with unusually high traffic volume; hence, we will refer to them as *volume traffic anomalies*. The identification of these types of anomalies is very important for traffic engineering tasks such as network provisioning, monitoring, pricing and mitigation of high traffic volume. A detailed study including these types of anomalies appeared in [12] showing their significance. It should be mentioned

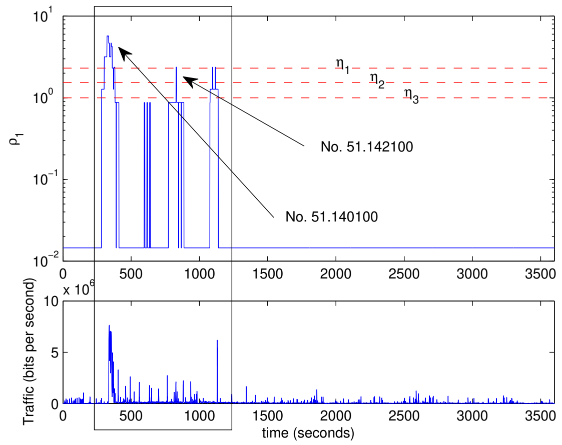


Fig. 8. Model-Free Method. (Top): Value of $I_1(\boldsymbol{\nu}_{t,w})$ – Monday, fifth week (with attacks) between 2:00pm - 3:00pm, (Bottom): Outbound traffic – Monday, fifth week (with attacks) between 2:00pm - 3:00pm. The rectangle identifies the time period there was an attack.

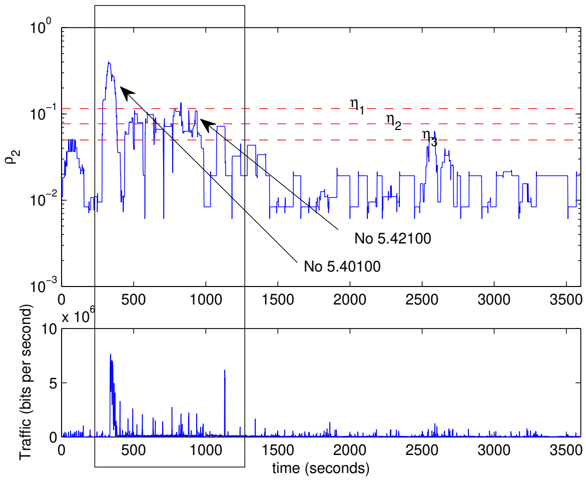


Fig. 9. Model-Based Method. (Top): Value of $I_2(\mathbf{q}_{t,n})$ – Monday, fifth week (with attacks) between 2:00pm - 3:00pm, (Bottom): Outbound traffic – Monday, fifth week (with attacks) between 2:00pm - 3:00pm. The rectangle identifies the time period there was an attack.

that persistent high-volume traffic, or more generally changes in the first and second moment can also be picked up by cruder methods (e.g., monitoring the mean and variance). The advantage of our approach is that it can identify all significant distributional changes including the ones that may only be reflected in higher moments.

We followed the same outline that was presented for the previous data set from Abilene. The first $36,000$ seconds (from 08:00-18:00) of the outbound traffic of each day of the first week were used to construct the alphabet and the MMP for the model-free and model-based model, respectively. We then observed the traffic of each day of the fifth week and we investigated how this deviates from the reference traffic of the same day of the first week. For the model-free method, the optimal values were found to be $b^* = 20$ and $w = 3$. For the model-based method, the optimal trace length was $n = 60$. The performance of both methods when applied to the DARPA data set is summarized in Table III. As there are no representations of different features in this data set, we provide the aggregate false alarms rate for each method.

Coming back to the last comment of Section V-C, the performance of both approaches has been improved, as more samples are used (see Figures 8 and 9), but this comes at the cost of creating larger alphabets or creating an MMP with more states.

VII. Related Work

In this section we refer to the work most closely related to our study. In [3], the authors used wavelet filters to detect anomalies in network traffic including outages, flash crowds, attacks and measurements failures. Our approach differs from that one in the sense that we try to detect short-lived network traffic anomalies within a few samples. Namely, our method, as we implemented it, does not investigate traffic anomalies occurring over long time-scales (hours or days); instead we focused on anomalies over relatively short time-scales. Moreover in our framework the time-series is not restricted to be a scalar, thus can be extended to investigate both spatial and temporal anomalies.

From a theoretical point of view, the authors in [19] studied a number of information-theoretic measures for anomaly detection. Their study was also performed using the DARPA Evaluation data set. Among other observations, they concluded that the relative entropy can better measure the similarity between two datasets. Both our approaches rigorously derive a rule on how to compare two datasets. It turns out that the relative entropy plays a critical role in both rules we derive.

The authors in [11–13] have introduced a framework to diagnose spatial anomalies, which is based on principal component analysis to partition the high dimensional space where a set of network traffic measurements live into disjoint subspaces corresponding to normal and anomalous conditions. Our methodology does not require whole network information and focuses on rapidly identifying both temporal and spatial anomalies in each origin-destination flow or link, by preserving both temporal and spatial correlation of samples.

Very recently the authors in [13] used data mining and information theory techniques to identify network anomalies. Their methods take into account more information than the traffic volume, including, the origin and destination address of each flow, as well as source and destination ports using results from netflow. As we commented in the Introduction our methods can be easily adapted to handle such traces of activity as well. Our methods are on-line, providing a rigorous way to identify anomalies using a fixed sliding window. All the other methods we surveyed are off-line.

In [20], the authors proposed on-line techniques, based on sketches, to identify change detection of individual IP-to-IP flows. The main focus of our framework is on identifying anomalies in the aggregated traffic, which requires a less detailed summary of the network activity.

TABLE I
SETUP I: DETECTION AND FALSE ALARMS RATES FOR EACH TYPE OF ANOMALY, USING THE MODEL-FREE METHOD (WITH $w = 20$ SAMPLES, $b^* = 3$ SAMPLES) AND THE MODEL-BASED METHOD (WITH $n = 10$ SAMPLES), FOR A DESIRABLE FALSE ALARM PROBABILITY $\epsilon = 0.1\%, 1\%, 5\%$.

	Model-Free Method		Model-Based Method	
Anomaly	Detection Rate	False Alarms Rate	Detection Rate	False Alarms Rate
DoS	$90\%, 92\%, 94\%$	$0.5\%, 3\%, 5\%$	$85\%, 87\%, 90\%$	$0.5\%, 3\%, 10\%$
SCAN	$90\%, 92\%, 94\%$	$0.5\%, 3\%, 7\%$	$83\%, 86\%, 90\%$	$0.5\%, 3\%, 10\%$
ALPHA	$91\%, 93\%, 94\%$	$0.5\%, 3\%, 6\%$	$82\%, 84\%, 87\%$	$0.5\%, 3\%, 11\%$
UNKN	$80\%, 85\%, 88\%$	$1\%, 4\%, 7\%$	$78\%, 80\%, 82\%$	$2\%, 4\%, 10\%$

TABLE II
SETUP I: DETECTION AND FALSE ALARMS RATES FOR EACH TYPE OF ANOMALY, USING THE MODEL-FREE METHOD (WITH $w = 20$ SAMPLES, $b^* = 3$ SAMPLES) AND THE MODEL-BASED METHOD (WITH $n = 10$ SAMPLES), FOR A DESIRABLE FALSE ALARM PROBABILITY $\epsilon = 0.1\%, 1\%, 5\%$.

	Model-Free Method		Model-Based Method	
	Detection Rate	False Alarms Rate	Detection Rate	False Alarms Rate
Scenario 1:				
SNVA–{DNVR,LOSA,WASH}	$93\%, 95\%, 97\%$	$1\%, 2\%, 6\%$	$90\%, 90\%, 92\%$	$1\%, 3\%, 8\%$
{DNVR,LOSA,WASH}–SNVA	$92\%, 94\%, 97\%$	$1\%, 2\%, 5\%$	$85\%, 87\%, 90\%$	$1\%, 5\%, 8\%$
Scenario 2:				
{LOSA,SNVA}–{NYCM,WASH}	$100\%, 100\%, 100\%$	$0.5\%, 1\%, 5\%$	$88\%, 90\%, 92\%$	$0.5\%, 3\%, 6\%$
{NYCM,WASH}–{LOSA,SNVA}	$100\%, 100\%, 100\%$	$0.5\%, 1\%, 5\%$	$86\%, 90\%, 92\%$	$0.5\%, 3\%, 6\%$

TABLE III
SETUP II: DETECTION AND FALSE ALARMS RATES FOR EACH TYPE OF ANOMALY, USING THE MODEL-FREE METHOD (WITH $w = 3$ SAMPLES, $b^* = 20$ SECONDS) AND THE MODEL-BASED METHOD (WITH $n = 60$ SECONDS), FOR A DESIRABLE FALSE ALARM PROBABILITY $\epsilon = 0.1\%, 1\%, 5\%$.

	Model-Free Method	Model-Based Method
Attack Category	Detection Rate	Detection Rate
DATA	$99\%, 100\%, 100\%$	$97\%, 100\%, 100\%$
DoS	$90\%, 92\%, 97\%$	$86\%, 91\%, 93\%$
PROBE	$89\%, 94\%, 100\%$	$84\%, 90\%, 92\%$
R2L	$80\%, 82\%, 86\%$	$71\%, 74\%, 76\%$
U2R	$80\%, 82\%, 88\%$	$78\%, 80\%, 83\%$
False Alarms Rate	$0.5\%, 2\%, 5\%$	$0.5\%, 3\%, 6\%$

VIII. CONCLUSIONS

We introduced a general distributional fault detection scheme able to identify a large spectrum of temporal anomalies from attacks and intrusions to various volume anomalies and problems in network resource availability. We then showed how this framework can be extended to incorporate spatial information, resulting in robust spatio-temporal anomaly detection in large scale operational networks. Although most of the proposed anomaly detection frameworks are able to identify temporal or spatial anomalies [21], we are able to identify both as we preserve both the temporal and spatial correlation of network feature samples.

We provided two different approaches, a model-free and a model-based one. The model-free method works on a longer time-scale processing traces of traffic aggregates over a small time interval. Using an anomaly-free trace it derives an associated probability law. Then it processes current traffic and quantifies whether it conforms to this probability law. The model-based method constructs a Markov modulated model of anomaly-free traffic measurements and relies on large deviations asymptotics and decision theory results to compare this model to ongoing traffic activity. We presented a rigorous framework to identify traffic anomalies providing asymptotic thresholds for anomaly detection. In our experimental results the model-free approach showed a somewhat better performance than the model-based one. This may be due to the fact that the former gains from the aggregation over a time-bucket in addition to the fact that the latter one requires the estimation of more parameters, hence, it may introduce a larger modeling error. For future work, it would be interesting to analyze the robustness of the anomaly detection mechanism to various model parameters.

Since we monitor the detailed distributional characteristics of traffic and do not rely on the mean or the first few moments we are confident that our approach can be successful against new types of (emerging) temporal and spatial anomalies.

Our method is of low implementation complexity (only an additional counter is required), and is based on first principles, so it would be interesting to investigate how it can be embedded on routers or other network devices.

REFERENCES

[1] M. Roesch, "Snort - lightweight intrusion detection for networks," in *LISA '99: Proceedings of the 13th USENIX conference on System administration*, Seattle, Washington, November 1999, pp. 229–238.

[2] V. Paxson, "Bro: a system for detecting network intruders in real-time," *Computer Networks*, vol. 31, no. 23–24, pp. 2435–2463, 1999.

[3] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in *Proceedings of the ACM SIGCOMM Workshop on Internet Measurement*, Marseille, France, November 2002, pp. 71–82.

[4] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, and M. Zissman, "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation," in *Proceedings of the DARPA Information Survivability Conference and Exposition*, Los Alamitos, CA, January 2000, pp. 12–26.

[5] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA off-line intrusion detection evaluation," *Computer Networks*, vol. 34, no. 4, pp. 579–595, 2000.

[6] V. Yegneswaran, J. T. Giffin, P. Barford, and S. Jha, "An Architecture for Generating Semantics-Aware Signatures," in *USENIX Security Symposium*, Baltimore, MD, USA, July 2005, pp. 97–112.

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

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

[9] I. Paschalidis and S. Vassilaras, "On the estimation of buffer overflow probabilities from measurements," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 178–191, 2001.

[10] ——, "Model-based estimation of buffer overflow probabilities from measurements," in *Proceedings of the ACM SIGMETRICS 2001/Performance 2001 conference*, Cambridge, Massachusetts, June 16-20th 2001, pp. 154–163.

[11] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proceedings of ACM SIGCOMM*, Portland, OR, August 2004, pp. 219–230.

[12] ——, "Characterization of Network-Wide Anomalies in Traffic Flows," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, Taormina, Italy, October 2004, pp. 201–206.

[13] ——, "Mining Anomalies Using Traffic Feature Distributions," in *Proceedings of ACM SIGCOMM*, Philadelphia, PA, August 2005, pp. 217 – 228.

[14] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, Budapest, Hungary, 1973, pp. 267–281.

[15] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.

[16] 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.

[17] I. C. Paschalidis and G. Smaragdakis, "A large deviations approach to statistical traffic anomaly detection," in *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, California, 2006, pp. 1900–1905.

[18] J. Mirkovic and P. Reiher, "A taxonomy of DDoS attack and DDoS defense mechanisms," *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 2, pp. 39–53, 2004.

[19] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection," in *Proceedings of the 2001 IEEE Symposium on Security and Privacy*, May 2001, pp. 130–143.

[20] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: methods, evaluation, and applications," in *ACM/SIGCOMM IMC'03*, 2003.

[21] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network Anomography," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, Berkeley, CA, October 2005, pp. 317–330.

[22] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[23] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, 1997.

[24] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 835–846, 1997.

[25] P. Abry and D. Veitch, "Wavelet analysis of long-range-dependent traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 2–15, 1998.

[26] G. Urvoy-Keller, "On the Stationarity of TCP Bulk Data Transfers," in *Proceedings of Passive and Active Network Measurement Workshop*, Boston, MA, March 2005, pp. 27–40.

APPENDIX

The goal of this Appendix is to provide a methodology to estimate the value of b^*, i.e., the optimal size of bucket that is sufficient to argue that the intra-bucket samples are correlated and the inter-bucket samples are independent, as well as present statistical tests to validate the key assumptions of the model-free approach, namely, that the sequence of partial sums $Y_1^{b^*}, \ldots, Y_{\lfloor n/b \rfloor}^{b^*}$ is an i.i.d. sequence for some appropriate bucket size b^*.

A. Data correlation

We start by empirically characterizing the correlation between elements $X_1, \ldots, X_n$ of a traffic trace over a short time-scale (this may depend on the sampling frequency - 5 minutes in the Abilene data set, 1 sec in the DARPA dataset). Early results in the networking literature advocate that traffic is *self-similar* [22–24]. We will use the notion of long-range and short-range dependence described below.

Let the *autocorrelation function (ACF)* be defined as

$$ACF(k) = \frac{\mathbf{E}[(X_t - \mu)(X_{t+k} - \mu)]}{\sigma^2},$$

where μ is the mean value and σ^2 is the variance of the time series. We say that a process exhibits *Long-Range Dependency (LRD)* if $ACF(k)$ decays following a power law, that is, if there exists a $\alpha \in (0, 1)$ and a constant $c > 0$ such that:

$$\lim_{k \to \infty} \frac{ACF(k)}{c \cdot k^{-\alpha}} = 1.$$

The quantity $\mathscr{H}$ satisfying $\alpha = 2(1 - \mathscr{H})$ is called the *Hurst* parameter and ranges in $(0.5, 1)$ for a LRD series. Slower decay of $ACF(k)$ occurs for $\mathscr{H}$ closer to 1 and implies longer-range dependency (a rigorous estimator of $\mathscr{H}$ is the so-called Abry-Veitch estimator [25]). *Short-Range Dependence (SRD)*, on the other hand, implies that $ACF(k)$ decays to zero exponentially fast.

We analyzed the Abilene Data set. We also analyzed the 1999 MIT Lincoln Lab (DARPA Evaluation) data set [5]. In each case, for each given b, we can form the sequence of partial sums $Y_1^b, Y_2^b, \ldots$. To form an i.i.d. sequence we would like to select a bucket size b^* such that

$$|ACF(k)| \le \frac{2}{\sqrt{n}}, \qquad \forall\ k > b^*.$$

We first attempted to find the b^* value for the period of one day in the DARPA data set. Our tests showed that this is not possible. This is to be expected since traffic over long periods

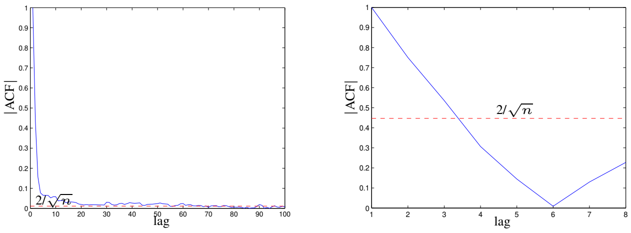


Fig. 10. (Left): the autocorrelation function for 10 hours of the DARPA dataset. (Right): typical autocorrelation function (for the origin-destination flow IPLS-STTL in the byte representation, with observation window n=20) from the Abilene dataset, for various time lags.

exhibits diurnal variation (non-stationarity). On the contrary, for smaller time periods (e.g., the first 10 hours of the same day), the ACF coefficient gets very close to the bound (albeit, not exactly below it) even for small lags of size close to 20 (see Figure 10 (left)). In fact, ACF decays even faster for smaller time periods – less than 6 hours – where stationarity is a reasonable approximation as the diurnal variation (24 hours cycle) is not strong. In the Abilene dataset, the value of $|ACF|$ dropped below the threshold mentioned above for $b^* = 3$ (see Figure 10 (right)) and sample sizes on the order of the observation window we use for anomaly detection in Sec. V. To conclude, we were able to find a value of b^* in both datasets so that the inter-bucket correlation is small, which suggests that the i.i.d assumption over buckets of size b^* is reasonable. As we noted in the Introduction, the key question is not so much the extent of inter-bucket correlation but whether the characterization we use is useful for anomaly detection or not. Clearly, there may be situations where dependencies are so strong that one needs to use a very large b^* resulting in very slow response to anomalies. In those situations, our Markovian characterization may be more appropriate.

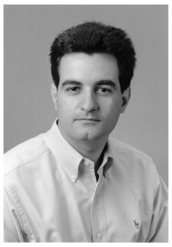
B. *Stationarity of partial sums*

In this section we study if the stationarity assumption is valid, that is, if the partial sums $Y_1^b, Y_2^b, \ldots$ have identical distribution in time intervals of 10 hours or less. To that end, we first use the well known Kolmogorov-Smirnov (K-S) goodness of fit test.

Given two independent and identical distributed series $\{X_t^1\}$ and $\{X_t^2\}$, $t \in \{1, 2, .., n\}$, the K-S test provides an answer to the question whether the two samples are drawn from the same distribution or not. The test starts with the initial hypothesis H_0, that the two samples are drawn from an identical distribution. It calculates the empirical cumulative distribution functions of both samples and evaluates the absolute maximum difference D_{max} between these two distribution functions. The test outputs the limit distribution of D_{max} under the hypothesis H_0. Given a threshold, on the limit distribution, the test outputs if the initial hypothesis is valid or not.

In our setting, we used the sequence of partial sums and constructed a subsequence (e.g., containing the odd elements of the sequence). We also constructed a second subsequence containing the elements of the original sequence that are consecutive (in time) to the elements of the first subsequence. This test is general and can be employed by using as a starting point any arbitrary point in the time series; it was able to identify stationarity regions for persistent connections in [26].

We start from the original hypothesis H_0 indicating that the two subsequences are identically distributed. We used the first 10 hours of the traffic for each day for the DARPA dataset. For the Abilene dataset the threshold was quite high, but in both cases, the K-S test validated H_0.



Ioannis Ch. Paschalidis (M'96, SM'06) received the Diploma in ECE from the National Technical University of Athens, Athens, Greece, in 1991, and the S.M. and Ph.D. degrees in EECS from the Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, in 1993 and 1996, respectively. He joined Boston University in 1996 where he is an Associate Professor in the Department of Electrical & Computer Engineering and in the Systems Engineering Division. He also serves as the Co-Director of the Center for Information and Systems Engineering (CISE) and as the Academic Director of the Sensor Network Consortium (SNC) – an industry consortium with companies active in sensor networking. He has held visiting appointments with MIT, and the Columbia University Business School. His current research interests lie in the fields of systems and control, optimization, networking, operations research, and computational biology. The main application areas he is targeting include communication and sensor networks, supply chains, and protein docking.

He has received an NSF CAREER award (2000), the second prize in the 1997 George E. Nicholson paper competition by INFORMS, and was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the National Academy of Engineering. He is an associate editor of the IEEE Transactions on Automatic Control and of Operations Research Letters.

Georgios Smaragdakis is a Senior Research Scientist at Deutsche Telekom Laboratories and the Technical University of Berlin. He earned his Ph.D. in Computer Science from Boston University, the Diploma in Electronic and Computer Engineering from the Technical University of Crete, and he interned at Telefónica Research, Barcelona. His research interests include the design and analysis of computer networks and content distribution systems with main applications in overlay network creation and maintenance, resource allocation and sharing, content routing, and network security.