A Large Deviations Approach to Statistical Traffic Anomaly Detection[*]

Ioannis Ch. Paschalidis[†] Georgios Smaragdakis[‡]

***Abstract*— We introduce an Internet traffic anomaly detection mechanism based on large deviations asymptotic results. 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 results to compute the probability that the monitored traffic is "consistent" with the corresponding reference characterization. Low values of this probability identify, in real-time, traffic anomalies. 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. We validate our techniques by analyzing real traffic traces with time-stamped anomalies.**

***Index Terms*— Network security, intrusion detection, statistical anomaly detection, method of types, 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. Previous studies showed that many types of traffic anomalies, such as attacks, worms, misconfiguration, network failures and a rapid increase of traffic volume (flash crowds), can be detected by monitoring the aggregate traffic at a border router. These approaches are typically grouped in two categories: *signature based anomaly detection* where known patterns of past anomalies are used to identify ongoing anomalies [1, 2], and *statistical anomaly detection* which identifies patterns that substantially deviate from normal patterns of operation [3]. Recent research studies showed that systems based on pattern matching had detection rates below 70% [4]. 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 statistical methods for traffic anomaly detection since they can identify even novel (unseen) types of anomalies.

In this paper we introduce a new statistical traffic anomaly detection framework that employs rigorous fault detection methodologies at short time scales. Our approach is fairly standard for statistical anomaly detection. First, we learn what constitutes "normal behavior" by observing past system behavior, assuming that it is anomaly-free. 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 [5] to characterize the type (i.e., empirical measure) of an 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 of normal behavior obtained from past system history, we compute the probability that current system behavior *deviates* from normal. Naturally, we employ the theory of *Large Deviations (LD)* [5]. LD theory provides a powerful way of handling rare events and their associated probabilities. The two key technical results we rely upon are Sanov's theorem [5] in the model-free approach and a related result for the empirical measure of a Markov process for the model-based case.

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 [6]. 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 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 [7, 8].

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

[*] Research partially supported by the NSF under grants DMI-0330171, DMI-0300359, CNS-0435312, ECS-0426453 and by the ARO under the ODDR&E MURI2001 Program Grant DAAD19-01-1-0465 to the Center for Networked Communicating Control Systems.

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

[‡] Computer Science Department, Boston University, e-mail: gsmaragd@cs.bu.edu.

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– 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. To the best of our knowledge, this is the first attempt to develop a framework that provides a statistical on-line methodology for anomaly detection in short time scales identifying temporal anomalies as opposed to techniques working over much longer time-scales [3]. The only infrastructure requirement in order to deploy our method is a simple counter.

As mentioned earlier, 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. Moreover, one would often update the reference trace with more recent activity, thus, almost eliminating the possibility that non-typical behavior (hence, relatively short-lived) will be classified as typical. 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, and (*b*) the 1999 MIT Lincoln Lab (DARPA evaluation) trace [4]. We are able to detect a variety of anomalies such as attacks and volume anomalies (even short ones) within a few samples, with very high success rate and low false alarm rate.

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 Sections IV and V we compare the two methods and validate our methodology using real measurements with time-stamped anomalies. In Section VI we review related work and identify the major differences with our approaches. We conclude in Section VII.

II. A Model-Free Approach

In this section we discuss our model-free approach. 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^*. (It is possible to select such a b^* by finding the value of b^* so that $|ACF(k)| \le \frac{2}{\sqrt{n}}, \ \forall\ k > b^*$, where $ACF(k) = \mathbf{E}[(X_t - \mu)(X_{t+k} - \mu)]/\sigma^2$ is the autocorrelation function with μ denoting the mean and σ^2 the variance of the timeseries.)

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 $[r_0, r_N]$ 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 [8] and use the so called Akaike’s Information Criterion (AIC) [9]. 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 $\mathbf{Y}_t = (Y_{t-w+1}^{b^*}, \ldots, Y_t^{b^*})$ be the trace of the w most recent partial sums using a bucket size b^*. We assume that these random variables 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 $\boldsymbol{\mu}$, i.e., $\Sigma_{\boldsymbol{\mu}} = \{\alpha_i : \boldsymbol{\mu}(\alpha_i) > 0\}$.

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

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

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

The next theorem, which is due to Sanov, establishes a large deviations result for $\boldsymbol{\mathscr{E}}_{w,b^*}^{\mathbf{Y}_t}$ (see [5, 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 $M_1(\Sigma)$

$$-\inf_{\boldsymbol{\nu}\in\Gamma^\circ} I_1(\boldsymbol{\nu}) \le \lim\inf_{w\to\infty} \frac{1}{w} \log \mathbf{P}[\boldsymbol{\mathscr{E}}_{w,b^*}^{\mathbf{Y}_t} \in \Gamma]$$

$$\le \lim\sup_{w\to\infty} \frac{1}{w} \log \mathbf{P}[\boldsymbol{\mathscr{E}}_{w,b^*}^{\mathbf{Y}_t} \in \Gamma] \le -\inf_{\boldsymbol{\nu}\in\Gamma} I_1(\boldsymbol{\nu}),$$

where Γ° denotes the interior of Γ.

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

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

B. Anomaly Detection Algorithm

Theorem II.1 can be used to identify anomalies. The proposed algorithm is summarized as follows:

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 $\mathbf{Y}_t = (Y_{t-w+1}^{b^*}, \ldots, Y_t^{b^*})$ be the trace of the w most recent partial sums using a

bucket size b^*. Compute its empirical measure and let $\boldsymbol{\mathscr{E}}^{\mathbf{Y}_t}_{w,b^*} = \boldsymbol{\nu}_{t,w}$ be the result.

3) Then, $\rho_{t,w} \stackrel{\triangle}{=} e^{-wI_1(\boldsymbol{\nu}_{t,w})}$ approximates the probability that the trace $\mathbf{Y}_t$ is drawn from the probability law $\boldsymbol{\mu}$.

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. $\rho_{t,w}$ can be better observed in the logarithmic scale. Formally, we identify an anomaly at time t if

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

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 ϵ is the detection threshold we use. The parameters n, b^*, k, ϵ affect the rule's performance and can be tuned experimentally. Notice that we can compute consecutive $\rho_{\tau,w}$ by using a sliding window of length n. Thus, we generate a new value for $\rho_{\tau,w}$ with every traffic sample. 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 timeseries and not to the partial sums.

III. A Model-Based Approach

We start with devising an MMP model for representing traffic activity. We should point out that 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.

A. An MMP model

We assume that the origin-destination traffic (in bits/bytes/packets/flows per time unit) or the (inbound or outbound) *traffic trace* observed at a border router, 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. A traffic trace can be a sequence of bits/bytes/packets/flows per time unit, where time units are defined depending on the available data or as we see fit for detecting anomalies. We propose the use of an MMP to model the traffic activity during a small time interval (several hours). Such a process is characterized by an underlying Markov chain with transition probability matrix $\boldsymbol{\Xi}$. To each state i, $i = 1,\dots,M$, we associate an interval $[r_{i-1}, r_i]$ of real numbers from which observations are drawn. That is, when the MMP is in state i traffic activity observations range in $[r_{i-1}, r_i]$. (For the application we are considering we do 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) [10]. 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 (the state is no longer hidden) and therefore the term disjoint Markov Modulated Process (d-MMP) is more appropriate.

To model the traffic trace as a d-MMP we let $[r_0, r_M]$ be the the range of all observations we make, split $[r_0, r_M]$ into M subintervals of equal length, and assign state i, $i = 1,\dots,M$, 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. 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 d-MMP model from an anomaly-free trace we are interested in comparing ongoing traffic activity to the model in order to identify potential deviations that would indicate an anomaly. To that end, given any trace we need to determine the probability that the trace is "explained" by the model.

Assume that the d-MMP has an irreducible underlying Markov chain with M states $1, 2, \dots, M$ and transition probability matrix $\boldsymbol{\Xi} = \{p(i,j)\}_{i,j=1}^{M}$. Let $\mathbf{p}$ denote the vector consisting of the rows of $\boldsymbol{\Xi}$. Let $\mathbf{Y}$ denote a sequence $Y_1, Y_2, \dots, Y_n$ of states that the Markov chain visits with the initial state being $Y_0 = \sigma$, and consider the empirical measures

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

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

$$q_1(i) \stackrel{\triangle}{=} \sum_{j=1}^{M} q(i,j) \qquad \text{and} \qquad q_2(i) \stackrel{\triangle}{=} \sum_{j=1}^{M} q(j,i) \tag{2}$$

be its marginals. Whenever $q_1(i) > 0$, let $q_f(j \mid i) \stackrel{\triangle}{=} q(i,j)/q_1(i)$. We will be using the notation $\mathbf{q}_f = (q_f(1 \mid 1), \dots, q_f(M \mid 1), q_f(1 \mid 2), \dots, q_f(M \mid M))$. We say that a probability measure $\mathbf{q} \in M_1(\mathscr{A}^2_{\mathbf{p}})$ 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 $\boldsymbol{\mathscr{E}}^{\mathbf{Y}}_{n,2}$ is established in the next theorem and is proven in [5, Sec. 3.1.3].

Theorem III.1 ([5]) *For every* $\mathbf{q} \in M_1(\mathscr{A}^2_{\mathbf{p}})$ *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 $M_1(\mathscr{A}_{\mathbf{p}}^2)$,

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

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

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

C. Anomaly detection algorithm

Theorem III.1 can be used to identify anomalies. The proposed algorithm is summarized as follows.

1) From an anomaly-free trace obtain a d-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 $\mathscr{E}_{n,2}^{\mathbf{Y}_t} = \mathbf{q}_{t,n}$ be the result.
3) Then, $\rho_{t,n} \overset{\triangle}{=} e^{-nI_2(\mathbf{q}_{t,n})}$ approximates the probability that the trace $\mathbf{Y}_t$ is drawn from the d-MMP with transition probability vector $\mathbf{p}$.

If $\rho_{t,n}$ is consistently low over some observed time interval, we conclude that the observed trace is not "consistent" with the reliable model, which indicates an anomaly. As we will see, $\rho_{t,n}$ can be better observed in a logarithmic scale. For an automated anomaly detection rule one can use the rule in 1, i.e., identify an anomaly at time t if

$$\rho_{\tau,n} < \epsilon, \qquad \forall \tau = t-k+1, \dots, t, \tag{3}$$

where n, k and ϵ are the parameters affecting the rule's performance (success and false alarm rates) and which have to be tuned. Obviously, other variations of this rule are possible, for instance, setting a threshold for the time-average of $\rho_{\tau,n}$ over the window $[t-k+1, \dots, t]$.

IV. Experimental Setup I: The Abilene Data Set

In this section, we validate our methodology against real traffic 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.

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 of 1% of all packets entering every router. Three different representations (features) of sampled flow data are used, as timeseries 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.

A log with the anomalies that took place is also provided. 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 271 anomalies: 133 DoS, 81 SCAN, 32 ALPHA and the rest are unknown [11]. 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. Outline of the Technique

We apply both our methods to the different timeseries (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) every day. 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 Subsection II-B. 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. We compute $\rho_{t,w}$ – the probability that the observed traffic follows the same (anomaly-free) law – using the procedure outlined in Subsection II-B. Working with statistics of the autocorrelation function, we found that $b^* = 3$ and $w = 10$ are good values for our data set. In order to identify the threshold ϵ in (1) we run our algorithm and compute $\rho_{t,w}$ for the reference trace, which should yield $O(1)$ probabilities. We found that $\rho_{t,w}$ is always below 10^{-3} in the reference trace, thus we set $\epsilon = 10^{-3}$.

We also follow the approach of Section III-A to devise an appropriate d-MMP traffic model. Using this model, for every day of the week and for every time sample we compute $\rho_{t,n}$ – the probability that the observed traffic trace is consistent with the d-MMP model – for an appropriately selected trace length n (cf. Step 2 of the algorithm in Subsection III-C). By following the same procedure as above, we select $\epsilon = 10^{-2}$ in (3).

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.

B. Anomaly Detection Examples

In this subsection we discuss the performance of our framework and we compare the two proposed methods. Fig. 1, illustrates a DoS attack in the Indianapolis-Seattle origin-destination flow and the associated probability $\rho_{t,w}$

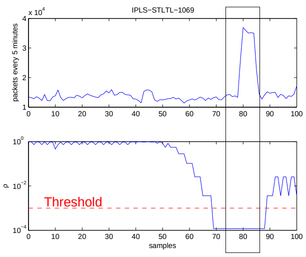

Fig. 1. Model-Free Method. (Top): Packet representation for the Indianapolis-Seattle flow. (Bottom): Value of $\rho_{t,w}$. The rectangle denotes a DoS attack.

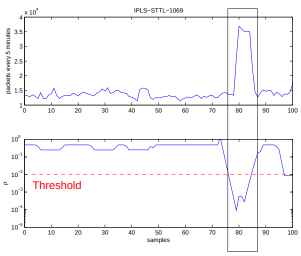

Fig. 2. Model-Based Method. (Top): Packet representation for the Indianapolis-Seattle flow. (Bottom): Value of $\rho_{t,n}$. The rectangle denotes a DoS anomaly.

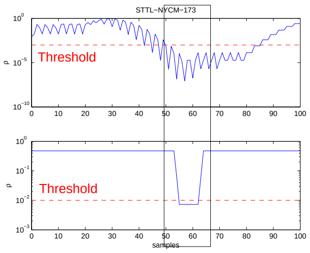

Fig. 3. Comparison of the two methods. (Top): Model-Free Method, (Bottom): Model-Based Method. The rectangle denotes a SCAN anomaly.

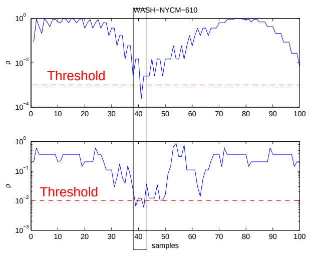

Fig. 4. Comparison of the two methods. (Top): Model-Free Method, (Bottom): Model-Based Method. The rectangle denotes an ALPHA anomaly.

when the model-free based method is applied. This probability is significantly low in the vicinity of the anomaly. It is to be expected that it can not pinpoint the time of the anomaly due to aggregation effect of the partial sums and our use of a window of size w. 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 $\rho_{t,w}$ drops below ϵ. On the other hand, applying the model-based method (Fig. 2), we can more precisely identify when an anomaly occurs. The disadvantage of the model-based method is that the false alarm rate is larger than that of the model-free method which benefits from the averaging over the time bucket b^*. The same observations are valid for other types of attacks, namely SCAN (Fig. 3) and ALPHA (Fig. 4).

We summarize our results in Table I. We should 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 success rate) to increase.

V. Experimental Setup II: The DARPA Evaluation Data Set

Next, we validate our method against the 1999 MIT Lincoln Lab (DARPA Evaluation) data set [4]. The data set consists of tcpdump data collected at the border router of a local network.

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. A detailed taxonomy of the attacks is presented in [12].

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 [11] showing their significance.

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 d-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 $b^* = 20$, $w = 3$, $k = 10$ and the threshold was set to 10^{-3}. For the model-based method optimal the optimal values were $n = 60$, $k = 3$ and the threshold was set to 10^{-5}. The performance of both methods when applied to the DARPA data set is summarized in Table II. As there are no representations of different features in this data set, we give the aggregate false alarm rate for each method.

VI. Related Work

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.

Recently, a number of intrusion detection tools such as Snort [1] or Bro [2] have been developed. Their aim is to identify application specific intrusions and attacks. Our approach is clearly advantageous as it is not application specific and can identify most types of traffic anomalies.

From a theoretical point of view, the authors in [13] 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

	Model-Free Method		Model-Based Method	
Anomaly	Success Rate	False Alarm Rate	Success Rate	False Alarm Rate
DoS	92%	5%	87%	10%
SCAN	92%	9%	86%	11%
ALPHA	93%	3%	87%	10%
UNKN	88%	10%	80%	12%
Overall	92%	7%	86%	11%

TABLE I

SETUP I: SUCCESS AND FALSE ALARM RATES FOR EACH TYPE OF ANOMALY, USING THE MODEL-FREE METHOD (WITH $\epsilon_1 = 10^{-3}$, $w = 20$, $b^* = 3$ SAMPLES, AND $k = 3$ SAMPLES) AND THE MODEL-BASED METHOD (WITH $\epsilon_2 = 10^{-2}$, $n = 10$ SAMPLES, AND $k = 3$ SAMPLES).

	Model-Free Method	Model-Based Method
Attack Category	Success Rate	Success Rate
DATA	100%	100%
DoS	90%	86%
PROBE	89%	84%
R2L	86%	76%
U2R	88%	83%
Overall	90%	83%
False Alarm Rate	7%	12%

TABLE II

SETUP II: SUCCESS AND FALSE ALARM RATES FOR EACH TYPE OF ANOMALY, USING THE MODEL-FREE METHOD (WITH $\epsilon_1 = 10^{-3}$, $w = 3$, $b^* = 20$ SECONDS AND $k = 10$ SECONDS) AND THE MODEL-BASED METHOD (WITH $\epsilon_2 = 10^{-5}$, $n = 60$ SECONDS, AND $k = 10$ SECONDS).

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, 14, 15] 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 temporal anomalies in each origin-destination flow or link.

Very recently the authors in [15] 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.

VII. CONCLUSIONS

We introduced a general distributional fault detection scheme able to identify all sorts of temporal anomalies anomalies from attacks and intrusions to various volume anomalies and problems in network resource availability. 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 computes the probability that 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 asymptotic results to compare this model to ongoing traffic activity. In particular, our results compute the probability that ongoing traffic activity is "consistent" with the model. In both methods, deviations in distributional characteristics of the traffic result in low values of the "conformance" probability, which identifies an anomaly. To the best of our knowledge, this is the first work for online anomaly detection based on large deviations results and distributional characteristics of empirical measures. We present a rigorous framework to identify traffic anomalies providing asymptotic thresholds for anomaly detection. We also estimate the size of the sliding window which provides reliable results for traffic anomaly detection in real-time.

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

Acknowledgments: We are grateful to Anukool Lakhina (Boston University) for providing us with the Abilene data set and for very helpful discussions.

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, 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.

[5] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer-Verlag, New York, 1998.

[6] M. Crovella, "Traffic modeling 101: Methods and results for single links and whole networks," Tutorial presented at ACM SIGCOMM'04, Portland, OR, August 2004.

[7] I. C. Paschalidis and S. Vassilaras, "On the Estimation of Buffer Overflow Probabilities from Measurements," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 178–191, 2001.

[8] ——, "Model-Based Estimation of Buffer Overflow Probabilities from Measurements," in *Proceedings of the ACM SIGMETRICS 2001/Performance 2001 conference*, Cambridge, MA, June 2001, pp. 154–163.

[9] 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.

[10] 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.

[11] A. Lakhina, M. Crovella, and C. Diot, "Characterization of Network-Wide Anomalies in Traffic Flows," in *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, Taormina, Italy, October 2004, pp. 201–206.

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

[13] 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.

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

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