Volume Anomaly Detection in Data Networks: an Optimal Detection Algorithm vs. the PCA Approach

Pedro Casas[1,3], Lionel Fillatre[2], Sandrine Vaton[1], and Igor Nikiforov[2]

[1] TELECOM Bretagne, Computer Science Department,
Technopôle Brest-Iroise, 29238 Brest, France
{pedro.casas,sandrine.vaton}@telecom-bretagne.eu

[2] Charles Delaunay Institute/LM2S, FRE CNRS 2848,
Université de Technologie de Troyes,
12 rue Marie Curie, 10010 Troyes, France
{lionel.fillatre,igor.nikiforov}@utt.fr

[3] Universidad de la República, Faculty of Engineering,
Julio Herrera y Reissig 565, 11300 Montevideo, Uruguay

Abstract. The crucial future role of Internet in society makes of network monitoring a critical issue for network operators in future network scenarios. The Future Internet will have to cope with new and different anomalies, motivating the development of accurate detection algorithms. This paper presents a novel approach to detect unexpected and large traffic variations in data networks. We introduce an optimal volume anomaly detection algorithm in which the anomaly-free traffic is treated as a nuisance parameter. The algorithm relies on an original parsimonious model for traffic demands which allows detecting anomalies from link traffic measurements, reducing the overhead of data collection. The performance of the method is compared to that obtained with the Principal Components Analysis (PCA) approach. We choose this method as benchmark given its relevance in the anomaly detection literature. Our proposal is validated using data from an operational network, showing how the method outperforms the PCA approach.

Key words: Network Monitoring and Traffic Analysis, Network Traffic Modeling, Optimal Volume Anomaly Detection.

1 Introduction

Bandwidth availability in nowadays backbone networks is large compared with today's traffic demands. The core backbone network is largely over-provisioned, with typical values of bandwidth utilization lower than 30%. The limiting factor in terms of bandwidth is not the backbone network but definitely the access network, whatever the access technology considered (ADSL, GPRS/EDGE/UMTS, WIFI, WIMAX, etc.). However, the evolution of future access technologies and the development of optical access networks (Fiber To The Home technology) will dramatically increase the bandwidth for each end-user, stimulating the proliferation of new "bandwidth aggressive" services (High Definition Video on Demand, interactive gaming, meeting virtualization, etc.). Some authors forecast a value of bandwidth demand per user as high as 50 Gb/sec in 2030. In this future scenario, the assumption of "infinite" bandwidth at the backbone network will no longer be applicable. ISPs will need efficient methods to engineer traffic demands at the backbone network. This will notably include a constant monitoring of the traffic demand in order to react as soon as possible to abrupt changes in the traffic pattern. In other words, network and traffic anomalies in the core network will represent a crucial and challenging problem in the near future.

Traffic anomalies are unexpected events in traffic flows that deviate from what is considered as normal. Traffic flows within a network are typically described by a traffic matrix (TM) that captures the amount of traffic transmitted between every pair of ingress and egress nodes of the network, also called the Origin Destination (OD) traffic flows. These traffic flows present two different properties or behaviors: on one hand, a stable and predictable behavior due to usual traffic usage patterns (e.g. daily demand fluctuation); on the other hand, an abrupt and unpredictable behavior due to unexpected events, such as network equipment failures, flash crowd occurrences, security threats (e.g. denial of service attacks), external routing changes (e.g. inter-AS routing through BGP) and new spontaneous overlay services (e.g. P2P applications). We use the term "volume anomaly" [18] to describe these unexpected network events (large and sudden link load changes). Figure 1 depicts the daily usual traffic pattern together with sporadic volume anomalies in four monitored links from a private international Tier-2 network. As each OD flow typically spans multiple network links, a volume anomaly in an OD flow is simultaneously visible on several links. This multiple evidence can be exploited to improve the detection of the anomalous OD flows. Volume anomalies have an important impact on network performance, causing sudden situations of strong congestion that reduce the network throughput and increase network delay. Even more, in the case of volume network attacks, the cost associated with damages and side effects can be excessively high to the network operator. The early and accurate detection of these anomalies allows to rapidly take precise countermeasures, such as routing reconfiguration in order to mitigate the impact of traffic demands variation, or more precise anomaly diagnosis by deeper inspection of other types of traffic statistics.

There are at least two major problems regarding current anomaly detection in OD traffic flows: (i) most detection methods rely on highly tuned data-driven traffic models that are not stable in time [18,22] and so are not appropriate for the task, causing lots of false alarms and missing real anomalies; (ii) current detection methods present a lack of theoretical support for their optimality properties (in terms of detection rate, false alarm generation, delay of detection, etc.), making it almost impossible to compare their performances. In this context, there are many papers with lots of new anomaly detection algorithms that claim

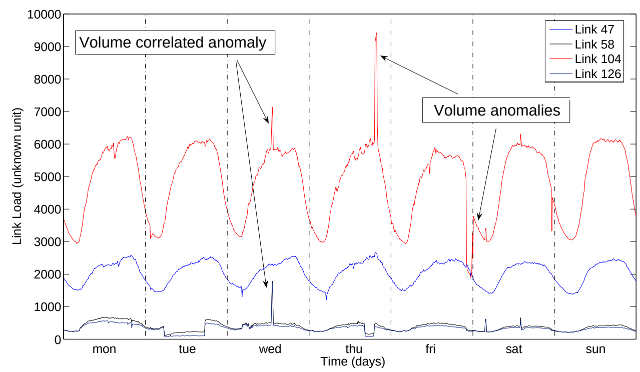


Fig. 1. Network anomalies in a large Tier-2 backbone network.

to have the best performance so far, but the generalization of these results is not plausible without the appropriate theoretical support. In this paper we focus on the optimal detection of volume traffic anomalies in the TM. We present a new linear and parsimonious model to describe the TM. This model remains stable in time, making it possible to overcome the stability problems of different current approaches. At the same time, it allows to monitor traffic flows from simple link load measurements, reducing the overhead of direct flow measurements. Based on this model, we introduce a simple yet effective anomaly detection algorithm. The main advantages of this algorithm rest on its optimality properties in terms of detection rate and false alarm generation.

1.1 Related Work

The problem of anomaly detection in data networks has been extensively studied. Anomaly detection consists of identifying patterns that deviate from the normal traffic behavior, so it is closely related to traffic modeling. This section overviews just those works that have motivated the traffic model and the detection algorithm proposed in this work. The anomaly detection literature treats the detection of different kinds of anomalous behaviors: network failures [8–10], flash crowd events [11,12] and network attacks [15–17,19,25]. The detection is usually performed by analyzing either single [13–15,23] or multiple time-series [18,19,26], considering different levels of data aggregation: IP flow level data (IP address, packet size, n^{o} of packets, inter-packet time), router level data (from router's management information), link traffic data (SNMP measurements from now on) and OD flow data (i.e a traffic matrix). The usual behavior of traffic data is modeled by several approaches: spectral analysis, Principal Components Analysis (PCA), wavelets decomposition, autoregressive integrated moving average models (ARIMA), etc. [13] analyzes frequency characteristics of network traffic at the IP flow level, using wavelets filtering techniques. [19] analyses the distribution of IP flow data (IP addresses and ports) to detect and classify network attacks. [15] uses spectral analysis techniques over TCP traffic for denial of service detection. [14] detects anomalies from SNMP measurements, applying exponential smoothing and Holt-Winters forecasting techniques. In [18], the authors use the Principal Components Analysis technique to separate the SNMP measurements in anomalous and anomaly-free traffic. These methods can detect anomalies by monitoring links traffic but they do not appropriately exploit the spatial correlation induced by the routing process. This correlation represents a key feature that can be used to provide more robust results. Moreover, the great majority of them cannot be applied when the routing matrix varies in time (because links traffic distribution changes without a necessary modification in OD flows), and many of the developed techniques are so data-driven that they are not applicable in a general scenario (notably the PCA approach in [18], as mentioned in [22]).

The authors in [26] analyze traffic at a higher aggregation level (SNMP measurements and OD flow data), using ARIMA modeling, Fourier transforms, wavelets and PCA to model traffic evolution. They extend the anomaly detection field to handle routing changes, an important advantage with respect to previous works. Unfortunately, all these methods present a lack of theoretical results on their optimality properties, limiting the generalization of the obtained results. [24] considers the temporal evolution of the TM as the evolution of the state of a dynamic system, using prediction techniques to detect anomalies, based on the variance of the prediction error. The authors use the Kalman filter technique to achieve this goal, using SNMP measurements as the observation process and a linear state space model to capture the evolution of OD flows in time. Even though the approach is quite appealing, it presents a major drawback: it depends on long-time periods of direct OD flow measurements for calibration purposes, an assumption which can be too restrictive in a real application, or directly infeasible for networks without OD flow measurement technology.

Our work deals with volume anomalies, i.e. large and sudden changes in OD flows traffic, independently of their nature. As direct OD flow measurements are rarely available, the proposed algorithm detects anomalies in the TM from links traffic data and routing information. This represents quite a challenging task: as the number of links is generally much smaller than the number of OD flows, the TM process is not directly observable from SNMP measurements. To solve this observability problem, a novel linear parsimonious model for anomaly-free OD flows is developed. This model makes it possible to treat the anomaly-free traffic as a nuisance parameter, to remove it from the detection problem and to detect the anomalies in the residuals.

1.2 Contributions of the Paper

This paper proposes an optimal anomaly detection algorithm to deal with abrupt and large changes in the traffic matrix. In [1] we present an optimal "sequential" algorithm to treat this problem, minimizing the anomaly detection delay (i.e. the time elapsed between the occurrence of the anomaly and the rise of an alarm). In this work, we draw the attention towards a "non-sequential" detection algorithm. This algorithm is optimal in the sense that it maximizes the correct detection probability for a bounded false alarm rate. To overcome the stability problems of previous approaches, a novel linear, parsimonious and non data-driven traffic model is proposed. This model remains stable in time and renders the process of traffic demand observable from SNMP measurements. The model can be used in two ways, either to estimate the anomaly-free OD flow volumes or to eliminate the anomaly-free traffic from the SNMP measurements in order to provide residuals sensitive to anomalies. Since a few anomaly-free SNMP measurements (at most one hour of measurements) is sufficient to obtain a reliable model of the OD flows, the proposed method is well adapted to highly non-stationary in time traffic and to dynamic routing. Using real traffic data from the Internet2 Abilene backbone network [32], we present an empirical comparison between our anomaly detection algorithm and the well known Principal Components Analysis (PCA) method introduced in [18]. The PCA approach has an important relevance in the anomaly detection field [18, 20, 22] but presents some important conception problems that we detect and analyze in our study. Through this analysis we verify the optimality properties of our detection algorithm and the stability of our traffic model, and show how our method outperforms the PCA approach in the considered dataset.

The remainder of this paper is organized as follows. The linear parsimonious OD flow model is introduced and validated in section 2. Section 3 describes the two different algorithms for anomaly detection that we compare in this work: our optimal detection algorithm and the previously introduced PCA approach. The evaluation and validation of our algorithm as well as a deep analysis of the PCA approach performance over real traffic data is conducted in section 4. Finally, section 5 concludes this work.

2 Handling Abrupt Traffic Changes

The anomaly detection algorithm that we present in this work consists of a non-sequential method. This algorithm presents optimality properties in terms of maximization of the detection probability for a bounded false alarm rate. To avoid direct OD flow measurements, the algorithm uses SNMP measurements $\mathbf{y}_t = \{y_t(1), \ldots, y_t(r)\}$ as input data; $y_t(i)$ represents the traffic volume (i.e. the amount of traffic) at link i in time interval t. High hardware requirements are necessary to network-wide collect and process direct OD flow measurements [7], so traffic models are generally developed using link SNMP measurements $\mathbf{y}_t$ and a routing matrix R to "reconstruct" OD flows. This reconstruction represents an ill-posed problem, as the number of unknown OD flows is much larger than the number of links [7]; in other words, it is not possible to directly retrieve the traffic demands $\mathbf{d}_t = \{d_t(1), \ldots, d_t(m)\}$ from $\mathbf{y}_t = R.\mathbf{d}_t$ given the ill-posed nature of the observation problem: $r << m$. Each traffic demand $d_t(i)$ represents the amount of traffic for OD couple i at time t. To overcome this difficulty, a parsimonious linear model for anomaly-free traffic is proposed. The idea of this model is that the anomaly-free traffic $\mathbf{d}_t$, sorted by OD flow volume can be decomposed at each time t over a known family of q basis functions $S = \{\mathbf{s}(1), \mathbf{s}(2), \ldots, \mathbf{s}(q)\}$ such that $q << m$. Therefore, the anomaly-free traffic can be expressed as $\mathbf{d}_t \approx S\boldsymbol{\mu}_t$ where the $m \times q$ matrix S is assumed to be known and $\boldsymbol{\mu}_t \in \mathbb{R}^q$ is a vector of unknown coefficients which describes the OD flows decomposition w.r.t. the set of vectors $\mathbf{s}(i)$. In this work, the traffic model is used to treat the anomaly-free traffic as a nuisance parameter, performing the anomaly detection in the traffic "residuals" that are obtained after removing the anomaly-free traffic. The anomaly-free traffic is removed by projection of the measured traffic on some space which is orthogonal to the space generated by the basis S. This transformation is based on the theory of invariance in statistics.

The parsimonious linear traffic model can be used to solve other problems than the anomaly detection one: TM estimation, using a least mean squares approach as it is shown in section 2.3, filtering and prediction with a Kalman approach, etc.

2.1 Stochastic Traffic Model for Anomaly Detection

It is assumed that the stochastic process of the anomaly-free OD traffic demand $\mathbf{d}_t$ obeys the following linear expression:

$$\mathbf{d}_t = \boldsymbol{\lambda}_t + \boldsymbol{\xi}_t \tag{1}$$

where $\boldsymbol{\lambda}_t \in \mathbb{R}^m$ is the mean traffic demand and $\boldsymbol{\xi}_t$ is a white Gaussian noise with covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. The process $\boldsymbol{\lambda}_t$ represents the "regular" part of the OD TM which can be correctly modeled when the behavior of the network is anomaly-free. The white Gaussian noise $\boldsymbol{\xi}_t$ models the natural variability of the OD TM together with the modeling errors. In order to describe the anomaly-free traffic $\boldsymbol{\lambda}_t$ with a small number of coefficients, a key feature of the TM is employed: its spatial stationarity; many classical TM models make use of this assumption, e.g. the gravity model [3,4,6]. The other key observation for this model is the "mice and elephants phenomenon": a small percentage of OD flows contribute to a large proportion of the total traffic [2,3]. The existence of such dominant flows together with the spatial stationarity of flows makes it reasonable to assume that, in the absence of an anomaly, the largest OD flows in a network remain the largest and the smallest flows remain the smallest during long periods of time; this assumption is confirmed in the empirical validation of the model, at least for several days, see section 2.3. Therefore, regarding the order of increasing OD flows (w.r.t. their traffic volume), it seems quite logical to accept that this order remains stable in time. It should be clear to the reader that this assumption can not be generalized to all network topologies and scenarios, but that holds for networks with a high level of aggregation (e.g. a backbone network or a large international VPN). The sorted OD flows can be interpreted as a discrete non-decreasing signal with certain smoothness. The curve obtained by interpolating this discrete signal is assumed to be a continuous curve, hence it can be parameterized by using a polynomial splines approximation.

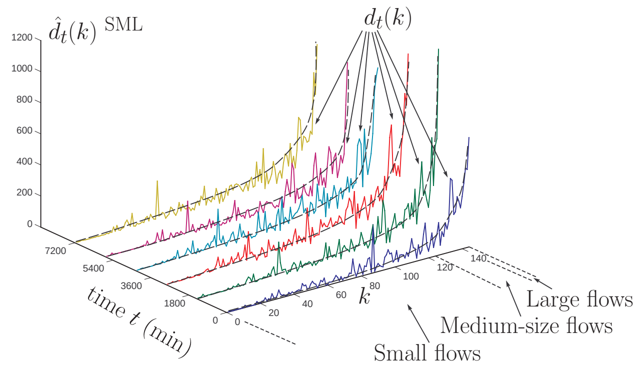


Fig. 2. Approximation of OD flows (full lines) by the spline-based model (dashed lines) for the Abilene network.

Figure 2 shows the anomaly-free OD flows for the Abilene network, sorted in the increasing order of their volume of traffic, for different time instants t. The full lines depict the value of each sorted OD flow $d_t(k), k = 1..m$, the dashed lines represent the polynomial approximation of the sorted flows. In order to appreciate the time stability of this approximation, the curves are plotted for 6 consecutive days (from Sunday to Friday). Given the shape of the curve formed by the sorted OD flows, a cubic splines' approximation is applied; basic definitions and results on polynomial splines can be found in [27]. A discrete spline basis is designed, discretizing the continuous splines according to m points uniformly chosen in the interval $[1; m]$ and rearranging them according to the OD flows sorting order. The obtained linear parsimonious model for the anomaly-free traffic demand can be expressed as:

$$\mathbf{d}_t = S\boldsymbol{\mu}_t + \boldsymbol{\xi}_t \tag{2}$$

where $S = \{\mathbf{s}(i), i = 1..q\}$ is a $m \times q$ known matrix with a small number of columns w.r.t. m ($q << m$). The vectors $\mathbf{s}(i)$, which correspond to the rearranged discrete splines, form a set of known basis vectors describing the spatial distribution of the traffic; $\boldsymbol{\mu}_t = \{\mu_t(1) \ldots \mu_t(q)\}^T$ is the unknown time varying parameter vector which describes the OD flow intensity distribution with respect to the set of vectors $\mathbf{s}(i)$. The model for the anomaly-free link traffic is given by:

$$\mathbf{y}_t = G\boldsymbol{\mu}_t + \boldsymbol{\zeta}_t, \tag{3}$$

where $G = RS$ and $\boldsymbol{\zeta}_t \sim \mathcal{N}(0, \Phi)$, with $\Phi = R\Sigma R^T$. The computation of the rank of G is not simple since it depends on the routing matrix R. In practice, since the number of columns of G is very small, the product RS and its rank can be computed very fast. Therefore, it will be assumed that G is full column rank. To simplify notation and computations, the whitened measurements vector is introduced:

$$\mathbf{z}_t = \Phi^{-\frac{1}{2}}\mathbf{y}_t = H\boldsymbol{\mu}_t + \boldsymbol{\varsigma}_t, \tag{4}$$

where $H = \Phi^{-\frac{1}{2}}G$ and $\boldsymbol{\varsigma}_t \sim \mathcal{N}(0, I_r)$ (I_r is the $r \times r$ identity matrix). The purpose of this transformation is simply to whiten the Gaussian noise. Finally, the covariance matrix Σ is unknown. The solution consists of computing an estimate $\widehat{\Sigma}$ from a few anomaly-free measurements. Results on the estimation of $\widehat{\Sigma}$ can be found in [28].

2.2 Validation of the model - the dataset

The validation of the proposed traffic model is conducted using real data from the Abilene network, an Internet2 backbone network. Abilene consists of 12 router-level nodes and 30 OC192 links (2 OC48). The used router-level network topology and traffic demands are available at [33]. Traffic data consists of 6-months traffic matrices collected via Netflow from the Abilene Observatory [32]. The Abilene network is mainly a research experimental network; for this reason and as a particular case, the available dataset [33] consists of complete direct OD flow measurements $\mathbf{d}_t$. In order to reduce the overhead introduced by the direct measurement and process of flow-level data, our traffic model relies on SNMP links' load measurements $\mathbf{y}_t$. For the purpose of validation, we use the Abilene routing matrix R_o (available at [33]) to retrieve $\mathbf{y}_t$ from the OD flow measurements: $\mathbf{y}_t = R_o.\mathbf{d}_t$. In the following evaluations, we assume that traffic demands $\mathbf{d}_t$ are unknown and just consider the link load values $\mathbf{y}_t$ as the input known data.

The number of links is $r = 30$ and the number of OD flows is $m = 144$. The sampling rate is one measurement each 10 minutes. In order to verify the stability properties of the model, two sets of measurements are used: the first one, the "learning" anomaly-free dataset, is composed of one hour of anomaly-free SNMP measurements and it is used to construct the spline basis S; the second one, the "testing" dataset, is composed of 720 SNMP measurements (five days measurement period) and it is used to validate the model. Let T_{learning} (T_{testing} respectively) be the set of time indexes associated with SNMP measurements from the learning anomaly-free dataset (testing dataset respectively). The learning anomaly-free dataset is measured one hour before the testing dataset.

The same dataset is further used for the evaluation of the anomaly detection algorithms; therefore, the set of "true" anomalies is manually identified in the testing dataset. Manual inspection declares an anomaly in an OD flow if the unusual deviation intensity of the guilty OD flow leads to an increase of traffic (i) larger than 1.5% of the total amount of traffic on the network and (ii) larger than 1% of the amount of traffic carried by the links routing this guilty OD flow, for each of these links. Hence, only significant volume anomalies are considered as "true anomalies" (small volume anomalies have little influence on link utilization). Let $T_{\text{testing}}^{\text{free}} \subset T_{\text{testing}}$ be the set of time indexes associated with the 680 non-consecutive SNMP measurements of the testing dataset manually declared as anomaly-free (40 measurements of the testing dataset are affected by at least one significant volume anomaly).

2.3 Numerical validation of the model

Although many aspects could potentially be included in the evaluation, the size of the estimation error is considered as the quality indicator, using the root mean squared error (RMSE) as a measure of this size:

$$\mathrm{RMSE}^{\mathrm{label}}(t) = \sqrt{\sum_{k=1}^{m} \left(\hat{d}_t^{\,\mathrm{label}}(k) - d_t(k)\right)^2}, \quad \forall t \in T_{\mathrm{testing}}^{\mathrm{free}} \tag{5}$$

where $d_t(k)$ is the true traffic volume of the anomaly-free OD flow k at time t and $\hat{d}_t^{\,\mathrm{label}}(k)$ denotes the corresponding estimate for the method entitled 'label'. Three estimates are compared: (i) simple gravity estimate [5] with label 'SG', (ii) tomogravity estimate [4, 5] with label 'TG' and (iii) spline-based Maximum Likelihood (ML) estimate with the label 'SML'. Since the traffic linear model is a Gaussian model, the Maximum Likelihood estimate of $\mathbf{d}_t$, namely $\hat{\mathbf{d}}_t^{\,\mathrm{SML}}$ corresponds to the least mean squares estimate, given by $\hat{\mathbf{d}}_t^{\,\mathrm{SML}} = S(H^T H)^{-1} H^T \mathbf{z}_t$. The statistical properties of the ML estimate are well known [28] contrary to the simple gravity and tomogravity estimates. The spline-based model is computed using the learning dataset, following these steps: (i) the tomogravity estimate $\hat{d}_t^{\,\mathrm{TG}}(k)$ is computed for all OD flows k and all $t \in T_{\mathrm{learning}}$, (ii) the mean flow values $\bar{d}^{\,\mathrm{TG}}(k) = \frac{1}{\mathrm{card}(T_{\mathrm{learning}})} \sum_{t \in T_{\mathrm{learning}}} \hat{d}_t^{\,\mathrm{TG}}(k)$ are computed, where $\mathrm{card}\,(T_{\mathrm{learning}})$ is the number of time indexes in the learning dataset and (iii) sorted in ascending order to obtain a rough estimate of the OD flows traffic volume. The spline-based model is designed with cubic splines and 2 knots (representing small, medium-size and large OD flows). The mean value $\bar{d}^{\,TG}(k)$ is also used to compute an estimate $\hat{\sigma}_k^2$ of σ_k^2, which leads to an estimate $\hat{\Phi}$ of Φ.

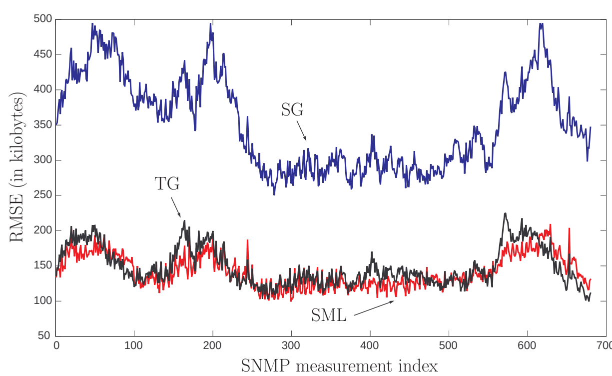


Fig. 3. Comparison between the SG, TG and SML RMSE for 680 anomaly-free measurements.

Figure 3 depicts the error $\mathrm{RMSE}^{\mathrm{label}}(t)$ over the set $T^{\mathrm{free}}_{\mathrm{testing}}$. The total error in $T^{\mathrm{free}}_{\mathrm{testing}}$, $\mathrm{TRMSE}^{\mathrm{label}} = \sum_{t \in T^{\mathrm{free}}_{\mathrm{testing}}} \mathrm{RMSE}^{\mathrm{label}}(t)$ is presented in table 1 as a global indicator of methods' performance. The spline-based estimate outperforms the other estimates, as it produces the smallest total estimation error. The TRMSE corresponding to the tomogravity estimate TG is quite close to the error produced with our model. However, the SML estimate presents a major advantage w.r.t. the TG estimate: as it was previously said, the ML estimate presents well established statistical properties, which is not the case for the TG estimate. The SML estimate is asymptotically optimal, i.e. it is asymptotically unbiased and efficient. Moreover, the spline-based model can be used in order to design anomaly detection algorithms with optimality properties, which is not the case for the tomogravity estimate. As a final validation, the Gaussian assumption of the model is studied. The "residuals" of measurements are analyzed, i.e. the obtained traffic after filtering the "regular" part, $G\boldsymbol{\mu}_t$. The residuals are obtained by projection of the whitened measurements vector $\mathbf{z}_t = \Phi^{-\frac{1}{2}}\mathbf{y}_t$ onto the left null space of H, using a linear transformation into a set of $r-q$ linearly independent variables $\mathbf{u}_t = W\mathbf{z}_t \sim \mathcal{N}(0, I_{r-q})$. The matrix W is the linear rejector that eliminates the anomaly-free traffic, built from the first $r-q$ eigenvectors of the projection matrix $P_H^{\perp} = I_r - H(H^T H)^{-1} H^T$ corresponding to eigenvalue 1.

Table 1. TRMSE (in kilobytes) for 680 anomaly-free measurements for gravity (SG), tomogravity (TG) and spline-based (SML) models.

Method	SG	TG	SML
TRMSE (kB)	9337	3935	3766

The rejector verifies the following relations: $WH = 0$, $W^T W = P_H^{\perp}$ and $WW^T = I_{r-q}$. $P_H^{\perp}$ represents the projection matrix onto the left null space of H. The Kolmogorov-Smirnov test [29] at the level 5% accepts the Gaussian hypothesis for 670 of the 680 measurements with time indexes in $T^{\mathrm{free}}_{\mathrm{testing}}$ (acceptation ratio of 98.5%), which confirms the Gaussian assumption.

3 Optimal Anomaly Detection and the PCA Approach

In this section we introduce the optimal volume anomaly detection algorithm. The goal of the proposed method is to detect an additive change θ in one or more OD flows of the traffic demand time series $\mathbf{d}_t$ from a sequence of SNMP measurements $\mathbf{y}_t = R.\mathbf{d}_t$. For technical reasons, it will be assumed that the amplitude of the change θ is constant; however, as it is shown in the results from section 4, this technical assumption does not restrict in practice the applicability of the proposed approach. Our detection algorithm consists of a non-sequential approach, also known as a "snapshot" method. Non-sequential approaches allow to define optimal algorithms, regarding the maximization of the probability of anomaly detection and the minimization of false alarms (i.e. raising an alarm in the absence of an anomaly). In this work, a simple snapshot approach is presented, which allows to detect an anomaly with the highest probability of detection for a given probability of false alarm. The previously introduced anomaly-free traffic model is slightly modified, in order to explicitly consider the temporal variation of the covariance matrix Σ. The Gaussian noise $\boldsymbol{\xi}_t$ is now assumed to have a covariance matrix $\gamma_t^2 \Sigma$; $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ is assumed to be known and stable in time. The scalar γ_t is unknown and serves to model the mean level of OD flows' volume variance. As it is explained in [28], this distinction between Σ and γ_t was not necessary in section 3. However, in the detection problem, this separation allows to accurately define the detection thresholds.

3.1 Optimal volume anomaly detection

Typically, when an anomaly occurs in one or several OD flows, the measurement vector $\mathbf{y}$ presents an abrupt change in those links where the OD flows are routed. The detection of this anomalous increase can be treated as a hypothesis testing problem, considering two alternatives: the null hypothesis $\mathcal{H}_0$, where OD flows are anomaly-free and the alternative hypothesis $\mathcal{H}_1$, where OD flows present an anomaly:

$$\mathcal{H}_0 = \{\mathbf{z} \sim \mathcal{N}(\boldsymbol{\varphi} + H\boldsymbol{\mu}, \gamma_t^2 I_r);\ \boldsymbol{\varphi} = 0,\ \boldsymbol{\mu} \in \mathbb{R}^q\}, \tag{6}$$

$$\mathcal{H}_1 = \{\mathbf{z} \sim \mathcal{N}(\boldsymbol{\varphi} + H\boldsymbol{\mu}, \gamma_t^2 I_r);\ \boldsymbol{\varphi} \neq 0,\ \boldsymbol{\mu} \in \mathbb{R}^q\}. \tag{7}$$

Here $\boldsymbol{\varphi}$ represents the evidence of an anomaly. In the anomaly detection problem, $\boldsymbol{\mu}$ is considered as a nuisance parameter since (i) it is completely unknown, (ii) it is not necessary for the detection and (iii) it can mask the anomalies. It is possible to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$ if, in the case of an anomaly, $\boldsymbol{\varphi}$ has a non-null component in the left null space of H. This verifies for any value of $\boldsymbol{\varphi}$ in the form of $\boldsymbol{\varphi} = \theta\, \Phi^{-\frac{1}{2}}\, \mathbf{r}^*$, where $\mathbf{r}^*$ stands for the sum of the normalized columns of the routing matrix R with indexes corresponding to the anomalous demands.

The quality of a statistical test is defined by the false alarm rate and the power of the test. The above mentioned testing problem is difficult because (i) $\mathcal{H}_0$ and $\mathcal{H}_1$ are composite hypotheses and (ii) there is an unknown nuisance parameter $\boldsymbol{\mu}$. There is no general way to test between composite hypotheses with a nuisance parameter. In this paper we use the statistical test $\phi^* : \mathbb{R}^r \mapsto \{\mathcal{H}_0, \mathcal{H}_1\}$ of [31], inspired by the fundamental paper of Wald [30]. The test is designed as:

$$\phi^*(\mathbf{z}) = \begin{cases} \mathcal{H}_0 & \text{if } \ \Lambda(\mathbf{z}) = \|P_H^{\perp}\mathbf{z}\|^2/\gamma_t^2 < \lambda_\alpha \\ \mathcal{H}_1 & \text{else} \end{cases} \tag{8}$$

where $\|\cdot\|$ represents the Euclidean norm.

Let K_α be the class of tests with an upper bounded maximum false alarm probability, $K_\alpha = \{\phi\ : \sup_{\boldsymbol{\mu}} \Pr_{\boldsymbol{\varphi}=0,\boldsymbol{\mu}}(\phi(\mathbf{z}) = H_1) \leqslant \alpha\}$, $0 < \alpha < 1$; here $\Pr_{\boldsymbol{\varphi}=0,\boldsymbol{\mu}}$ stands for the probability when $\mathbf{z} \sim \mathcal{N}(H\boldsymbol{\mu}, \gamma_t^2 I_r)$. The power function or hit rate is defined as $\beta_\phi(\boldsymbol{\varphi}, \boldsymbol{\mu}) = \Pr_{\boldsymbol{\varphi} \neq 0, \boldsymbol{\mu}}(\phi(\mathbf{z}) = \mathcal{H}_1)$. A priori, this probability depends on the nuisance parameter $\boldsymbol{\mu}$ as well as on the parameter $\boldsymbol{\varphi}$ which is highly undesirable. However, the test $\phi^*(\mathbf{z})$ defined by equation (8) has uniformly best constant power (UBCP) in the class K_α over the family of surfaces $S = \{S_c : c \geq 0\}$ defined by $S_c = \{\boldsymbol{\varphi} : \|P_H^\perp \boldsymbol{\varphi}\|^2 = c^2\}$. UBCP means that $\beta_{\phi^*}(\boldsymbol{\varphi}, \boldsymbol{\mu}) = \beta_{\phi^*}(\boldsymbol{\varphi}', \boldsymbol{\mu}), \forall \boldsymbol{\varphi}, \boldsymbol{\varphi}' \in S_c$ and $\beta_{\phi^*}(\boldsymbol{\varphi}, \boldsymbol{\mu}) \geqslant \beta_\phi(\boldsymbol{\varphi}, \boldsymbol{\mu})$ for any test $\phi \in K_\alpha$. The threshold λ_α is chosen to satisfy the false alarm bound α, $\Pr_{\boldsymbol{\varphi}=0,\boldsymbol{\mu}}(\Lambda(\mathbf{z}) \geqslant \lambda_\alpha) = \alpha$. The UBCP property of this test represents the optimality condition of the detection algorithm.

3.2 Principal Components Analysis for Anomaly Detection

The Principal Components Analysis (PCA) approach for anomaly detection [18, 20, 22] consists of a two steps methodology: (i) parsimonious anomaly-free traffic modeling, using a decomposition of traffic measurements into a principal components basis and (ii) anomaly detection in the traffic residuals, i.e. the traffic not described by the PCA decomposition. PCA is a linear coordinate transformation that maps a given set of data points to a new coordinate system, such that the greatest variance of any projection lies on the first coordinate $\mathbf{w}_1$ (called the first principal component or first PC), the second greatest variance on the second coordinate $\mathbf{w}_2$, and so on. Given a traffic measurement matrix $\mathbf{Y} \in \mathbb{R}^{p \times r}$, where each column represents a time series of p samples of SNMP measurements for each link, the PCA traffic modeling consists of computing the r principal components of $\mathbf{Y}$, $\mathbf{w}_{i=1..r}$, using the first k principal components to capture the anomaly-free behavior of traffic and the remaining $r-k$ components to construct residuals sensitive to anomalies. The first k principal components are the "normal components" and the remaining $r-k$ are the "anomalous components". Each of the principal components can be computed as follows:

$$\mathbf{w}_1 = \arg \max_{||\mathbf{w}||=1} ||\mathbf{Y}\mathbf{w}||$$

$$\mathbf{w}_k = \arg \max_{||\mathbf{w}||=1} ||(\mathbf{Y} - \sum_{i=1}^{k-1} \mathbf{Y}\mathbf{w}_i\mathbf{w}_i^T)\mathbf{w}||$$

The idea behind this approach is that traffic anomalies are sparse in $\mathbf{Y}$, and so the first components of the transformation will correctly describe the anomaly-free behavior. The space spanned by the set of normal components is the "normal sub-space" S and the space spanned by the anomalous components is the "anomalous sub-space" $\hat{S}$. After the construction of the normal and anomalous sub-spaces, the links' traffic $\mathbf{y}$ can be separated at each time t in the modeled traffic $\mathbf{y}_{\text{model}}$ and the residual traffic $\mathbf{y}_{\text{residual}}$ by simple projection onto S and $\hat{S}$:

$$\mathbf{y} = \mathbf{y}_{\text{model}} + \mathbf{y}_{\text{residual}}$$

$$\mathbf{y}_{\text{model}} = \mathbf{P}\mathbf{P}^T\mathbf{y}$$

$$\mathbf{y}_{\text{residual}} = \left(\mathbf{I} - \mathbf{P}\mathbf{P}^T\right)\mathbf{y}$$

where $\mathbf{P} \in \mathbb{R}^{r \times k}$ stands for the matrix with the first k PCs as column vectors and $\mathbf{P}\mathbf{P}^T$ represents the projection matrix onto the normal sub-space. The anomaly detection is then performed in the residual traffic, looking for large changes in the squared norm of residuals, $||\mathbf{y}_{\text{residual}}||^2$.

4 Validation of the detection algorithm and PCA evaluation

The detection algorithm is applied to the SNMP measurements of the testing dataset. The false alarm probability is fixed to $\alpha = 0.01$. For the detection purpose, it is crucially important to have a good estimate of γ_t. This parameter is estimated from the learning dataset by using the ML estimate of noise variance [28] in residuals $\mathbf{u}_t$. Since this parameter can slowly vary in time, its value is updated during the test: at time t, if no anomaly has been declared in the last hour, γ_t is estimated by its value one hour before. The performance of our method is compared to the performance obtained with the PCA approach. This method is chosen as benchmark given its relevance in the anomaly detection literature [18,20,22]. The obtained results are presented in table 2. The column *Spline-based* shows that the proposed test (8) obtains a false alarm rate of 1.18%, close to the prescribed value $\alpha = 0.01$. The probability to detect a volume anomaly is about 77.5%. The column *PCA* presents the results obtained with the PCA approach. The best performance that can be attained with the PCA test is considered in this evaluation, using just the first PC to model the normal sub-space; the following discussion about results in figure 4 clarifies this election. The detection

Table 2. Results of the detection for 720 measurements composed of 680 anomaly-free measurements and 40 anomalous measurements for the spline-based and PCA tests.

Situation	Spline-based	PCA (1 PC)
Normal operation	672 (98.82 %)	671 (98.68 %)
False alarms	8 (1.18 %)	9 (1.32 %)
Missed detections	9 (22.50 %)	25 (62.50 %)
Correct detections	31 (77.50 %)	15 (37.50 %)

threshold of this test is chosen to obtain a similar false alarm rate of 1.32%. The PCA test presents a very low correct detection rate for this level of false alarm, about 37.50%. Figure 4 illustrates the ROC curves for the Spline-based and the PCA tests for different number of first PCs to model the normal sub-space. The figure presents the correct detection rate β for different values of the false alarm rate α. The ROC curves allow to compare the accuracy of both tests and the sensitivity of each detection method w.r.t. the variation of the detection thresholds, showing the existing trade-off between the correct detection and the false alarm rates. Results obtained with the PCA approach in the Abilene dataset

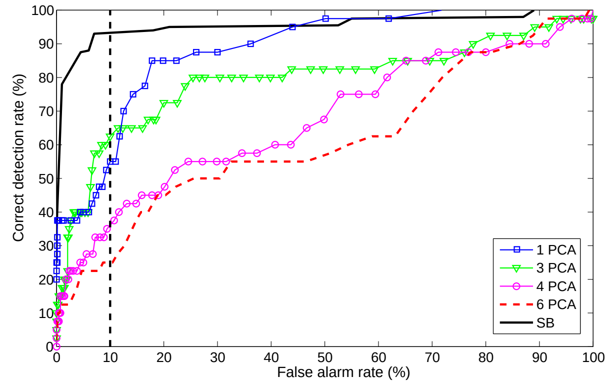
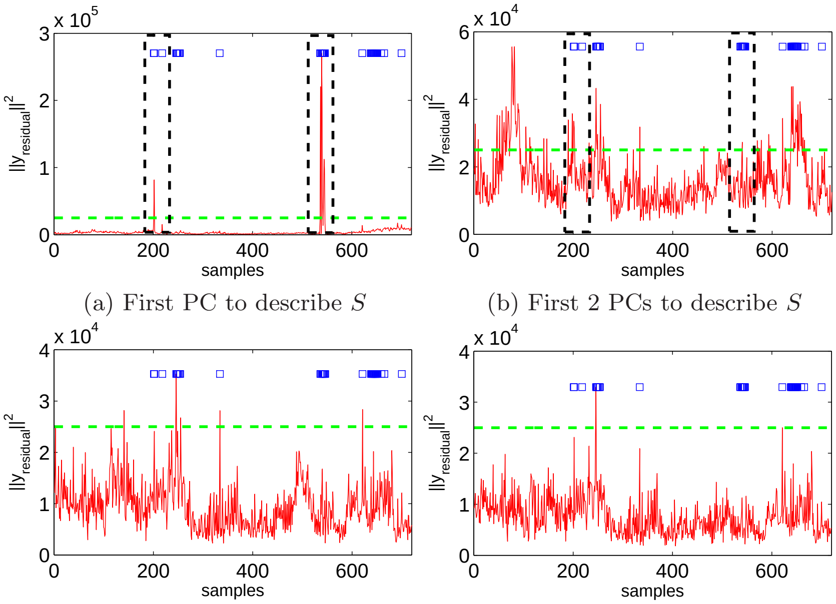


Fig. 4. Correct detection rate vs false alarm rate for the spline-based test (SB - solid line) and the PCA test, considering different number of first PCs to model the normal sub-space.

are far from those obtained with our method; the PCA test presents more than 2 times lower detection rates for a reasonable false alarm rate, below 5%. There are at least three major problems regarding the PCA approach: (i) its performance strongly depends on the number of components selected to describe the normal space; (ii) the traffic modeling procedure is data-driven, posing serious stability problems and (iii) the learning step is unsupervised but very time-consuming, becoming prone to bad-learning effects. Similar problems were also analyzed and verified by the authors of the original PCA for anomaly detection approach [18] in [21, 22]. Let us begin by the first issue; in [18], the separation between the normal and anomalous principal components is performed using a simple ad-hoc threshold-based separation method that is highly tuned for each dataset and cannot therefore be generalized, making the PCA approach inapplicable in a general scenario. Figure 5 depicts the temporal evolution of $||\mathbf{y}_{\text{residual}}||^2$, using a different number of PCs to describe the normal sub-space (1, 2, 4 and 5 first PCs are used to model the anomaly-free traffic). The dotted line represents the detection threshold; the squares indicate the times when an anomaly truly occurs, according to the manual inspection performed in section 2.2. It can be appreciated that the false positive rate is very sensitive to small differences in the number of principal components used to describe the normal sub-space. The ROC curves in figure 4 show that there is no single PCA representation for the Abilene dataset that offers a good balance between correct detection and false alarm rates.

Regarding the second issue, the traffic modeling in the PCA approach is data-driven, i.e. the PCA decomposition strongly depends on the considered SNMP measurements matrix $\mathbf{Y}$. In [18], the normal and anomalous sub-spaces are constructed from a given matrix $\mathbf{Y}^o$ at a certain time window t_o, and the representation is assumed to be stable during long-time periods, from week to week.



(a) First PC to describe S

(b) First 2 PCs to describe S

(c) First 4 PCs to describe S

(c) First 5 PCs to describe S

Fig. 5. Temporal evolution of $||\mathrm{y}_{\mathrm{residual}}||^2$, using a different number of first PCs to model the normal sub-space S. The squares indicate when an anomaly truly occurs. The dotted line depicts the detection threshold. Large anomalies pollute the normal sub-space and are not detected with the PCA approach. (a) Both large anomalies at samples 200 and 540 are correctly detected using 1 PC to describe S. (b) Large anomalies are not detected using a 2 PCs representation of S.

However, it is easy to see that this approach is highly unstable, even from one time window to the other. Let us consider an extreme-case example that will also illustrate the learning problems of the approach. The PCA approach assumes that the normal sub-space can be correctly described by the first principal components of $\mathbf{Y}^o$ as they capture the highest level of "energy". Figure 6 depicts the temporal evolution of the variance captured by each principal component $\mathbf{w}_i$, $||\mathbf{Y}\mathbf{w}_i||^2$, considering time windows of 12hs (i.e. the set of PCs is recomputed every 12hs). In almost every time window, the first principal component captures the highest energy, justifying the use of one single PC to describe the normal traffic behavior. However, large anomalies at time windows t_3 and t_8, also visible in figure 5.(a) contribute to a large proportion of the captured energy; in this case, a second principal component may be added as a descriptor of the normal traffic. Since this second component corresponds in fact to an anomaly, the normal sub-space is inadvertently polluted, turning useless the learning step. In figure 5.(b), both large anomalies at t_3 and t_8 are not detected due to this effect.

This brings us to the last but not least problem; the learning step of the PCA approach is very "time-consuming": the number of samples p must be greater

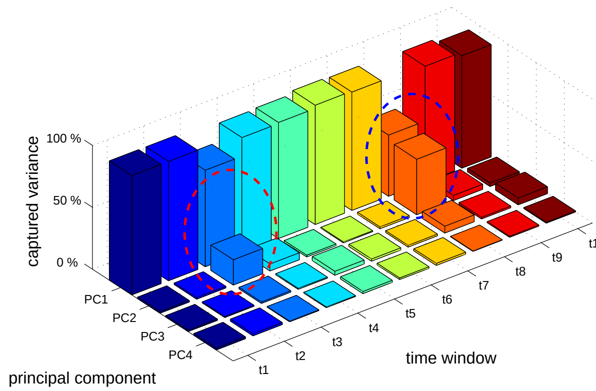


Fig. 6. Temporal evolution of the total variance captured by each PC $\mathbf{w}_i$, $||\mathbf{Y}\mathbf{w}_i||^2$. Each time window $t_{j=1..10}$ consists of 12hs of SNMP data. Large anomalies may inadvertently pollute the normal sub-space at t_3 and t_8.

than the number of links r, in order to obtain at least r independent PCs [22]. In this sense, the approach is more prone to suffer from this kind of polluting effect, since it is likely that an anomaly occurs on longer time periods. Our algorithm is not data-driven and has a very short learning-step: as we show in section 2.3, at most one hour of measurements is sufficient to obtain a reliable model of the OD flows. The effect of a training step over polluted data does not represent a problem to our short-learning approach, as it is quite simple to assure or look for a 1-hour anomaly-free time period.

5 Conclusions and Some Extensions

In this paper, we have presented and evaluated a new statistical algorithm for volume anomaly detection in data networks. This algorithm presents well-established optimality properties in terms of detection probability and false alarm generation, unavailable in previous proposals in the field and extremely important in order to provide solid results. For the purpose of anomaly detection, we have introduced an original linear parsimonious spline-based traffic model which allows to treat the anomaly-free traffic as a nuisance parameter. This model parameterizes traffic flows from simple link load measurements, reducing the overhead of direct flow measurements. Compared to other different traffic models, this model is not data-driven and remains stable in time, a necessary property to achieve reliable results. We have also applied this traffic model to the traffic matrix estimation problem, achieving better results than those obtained with classical models (e.g. tomogravity model). We have analyzed the performance of a very well known anomaly detection method, the so called PCA for anomaly detection approach into a real traffic dataset and studied in depth some of the weaknesses of this approach. We have finally compared our algorithm to the PCA approach and showed that the spline-based anomaly detection test outperforms the PCA based approach as predicted by the optimality properties of the test, which highlights the impact of our proposal for volume anomaly detection. In this work we have only treated the anomaly detection problem. In [1] we present some interesting countermeasures to react against anomalies, based on routing reconfiguration.

References

1. Casas, P., Fillatre, L., Vaton, S.: Robust and Reactive Traffic Engineering for Dynamic Traffic Demands. In: Proc. EuroNGI Conference on Next Generation Networks (2008)
2. Johansson. C., Gunnar, A.: Data-driven Traffic Engineering: techniques, experiences and challenges. In: Proc. IEEE BROADNETS (2006)
3. Medina, A., Salamatian, K., Bhattacharyya, S., Diot, C.: Traffic Matrix Estimation: Existing Techniques and New Directions. In: Proc. ACM SIGCOMM (2002)
4. Zhang, Y., Roughan, M., Lund, C., Donoho, D.: Estimating Point-to-Point and Point-to-Multipoint Traffic Matrices: an Information-Theoretic Approach. IEEE/ACM Trans. Networking, vol. 13, no. 5, pp. 947–960 (2005)
5. Zhang, Y., Roughan, M., Duffield, N., Greenberg, A.: Fast Accurate Computation of Large-Scale IP Traffic Matrices from Link Load Measurements. In: Proc. ACM SIGMETRICS (2003)
6. Gunnar, A,. Johansson, M., Telkamp, T.: Traffic Matrix Estimation on a Large IP Backbone - A Comparison on Real Data. In: Proc. USENIX/ACM IMC (2004)
7. Coates, M., Hero, A., Nowak, R., Yu, B.: Internet Tomography. IEEE Signal Processing Magazine, vol. 19, no. 3, pp 47–65 (2002)
8. Hood, C., Ji, C.: Proactive network fault detection In: Proc. IEEE INFOCOM (1997)
9. Katzela, I., Schwartz, M.: Schemes for fault identification in communications networks. IEEE/ACM Trans. Networking, vol. 3, no. 6, pp. 753–764 (1995)
10. Ward, A., Glynn, P., Richardson, K.: Internet service performance failure detection. Performance Evaluation Review (1998)
11. J. Jung, B. Krishnamurthy and M. Rabinovich Flash crowds and denial of service attacks: Characterization and implications for CDNs and webs. In: Proc. ACM WWW-02 (2002)
12. Xie, L. et al.: From Detection to Remediation: A Self-Organized System for Addressing Flash Crowd Problems. In: Proc. IEEE ICC (2008)
13. Barford, P., Kline, J., Plonka, D., Ron, A.: A Signal Analysis of Network Traffic Anomalies. In: ACM SIGCOMM Internet Measurement Workshop (2002)
14. Brutlag, J.D.: Aberrant Behavior Detection in Time Series for Network Monitoring In: Proc. 14th Systems Administration Conference (2000)
15. Cheng, C.M., Kung, H., Tan, K.S.: Use of Spectral Analysis in Defense Against DoS Attacks. In: Proc. IEEE GLOBECOM (2002)
16. Zou, C.C., Gong, W., Towsley, D., Gao, L.: The Monitoring and Early Detection of Internet Worms. IEEE/ACM Trans. Networking, vol. 13, no. 5, pp. 961–974 (2005)

17. Wang, H., Zhang, D., Shin, K.: Detecting SYN flooding attacks. In: Proc. IEEE INFOCOM (2002)
18. Lakhina, A., Crovella, M., Diot, C.: Diagnosing Network-Wide Traffic Anomalies. In: Proc. ACM SIGCOMM (2004)
19. Lakhina, A., Crovella, M., Diot, C.: Mining Anomalies Using Traffic Feature Distributions. In: Proc. ACM SIGCOMM (2005)
20. Li, X., Bian, F., Crovella, M., Diot, C., Govindan, R., Iannaccone, G., Lakhina, A.: Detection and Identification of Network Anomalies Using Sketch Subspaces. In: Proc. USENIX/ACM IMC (2006)
21. Ahmed, T., Coates, M., Lakhina, A.: Multivariate Online Anomaly Detection Using Kernel Recursive Least Squares. In: Proc. IEEE INFOCOM (2007)
22. Ringberg, H., Soule, A., Rexford, J., Diot, C.: Sensitivity of PCA for Traffic Anomaly Detection. In: Proc. ACM SIGMETRICS (2007)
23. Thottan, M., Ji, C.: Anomaly Detection in IP Networks. IEEE Trans. Signal Processing, vol. 51, no. 8, pp. 2191–2204 (2003)
24. Soule, A., Salamatian, K., Taft, N.: Combining Filtering and Statistical Methods for Anomaly Detection. In: Proc. USENIX/ACM IMC (2005)
25. Tartakovsky, A. et al.: A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. IEEE Trans. Signal Processing, vol. 54, no. 9, pp. 3372–3382 (2006)
26. Zhang, Y., Ge, Z., Greenberg, A., Roughan, M.: Network Anomography. In: Proc. USENIX/ACM IMC (2005)
27. Nürnberger, G.: Approximation by Spline Functions. Springer-Verlag (1989)
28. Rao, C.: Linear Statistical Inference and its Applications. J. Wiley & Sons (1973)
29. Lehman, E.: Testing Statistical Hypotheses, 2nd Edition. Chapman & Hall (1986)
30. Wald, A.: Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. American Math. Soc., vol. 54, pp. 426–482 (1943)
31. Fillatre, L., Nikiforov, I.: Non-bayesian detection and detectability of anomalies from a few noisy tomographic projections IEEE Trans. Signal Processing, vol 55, no. 2, pp. 401–413 (2007)
32. The Abilene Observatory, `http://abilene.internet2.edu/observatory/`
33. Zhang, Y.: Abilene Dataset 04, `http://www.cs.utexas.edu/yzhang/`