OPTIMAL VOLUME ANOMALY DETECTION IN NETWORK TRAFFIC FLOWS

Lionel Fillatre[★]*, Igor Nikiforov*[★]*, Pedro Casas*[♦] *and Sandrine Vaton*[♦]

[★] ICD/LM2S, Université de Technologie de Troyes
12, rue Marie Curie - BP 2060 - 10010, Troyes, France
phone: + (33) 3 25 75 96 73, fax: + (33) 3 25 71 56 99,
email: firstname.lastname@utt.fr

[♦] Computer Science Department, TELECOM Bretagne
Technople Brest-Iroise - CS 83818 - 29238, Brest, France
phone: + (33) 2 29 00 10 29, fax: + (33) 2 29 00 10 00,
email: firstname.lastname@telecom-bretagne.eu

ABSTRACT

Optimal detection of unusual and significant changes in network Origin-Destination (OD) traffic volumes from simple link load measurements is considered in the paper. The ambient traffic, i.e. the OD traffic matrix corresponding to the non-anomalous network state, is unknown and it is considered here as a nuisance parameter because it can mask the anomalies. Since the OD traffic matrix is not recoverable from simple link load measurements, the anomaly detection is an ill-posed decision-making problem. The method proposed in this paper consists of finding a linear parsimonious model of ambient traffic (nuisance parameter) and detecting anomalies by using an invariant detection algorithm based on a separation of the measurement space into disjoint subspaces corresponding to normal and anomalous network traffic. The method's ability to detect anomalies is evaluated in real traffic from Abilene, a United States backbone network. The theoretically expected results are confirmed.

1. INTRODUCTION

Network management becomes very complex as networks increase in size and complexity. The traffic demand in a network is typically described by a traffic matrix that captures the amount of traffic transmitted between every pair of ingress and egress nodes in a network, also called the Origin-Destination (OD) flows. A volume anomaly is a sudden change in an OD flow's traffic (for example, due to a denial-of-service attack, a flash crowd event, a virus/worm propagation, etc.) that spans multiple physical links of the network. The reliable detection of these unusual and significant changes in the OD traffic matrix is an important issue for network operation.

High hardware requirements are necessary to network-wide collect and process the direct OD flow measurements [1]. Consequently, the Simple Network Management Protocol (SNMP), which is a widely deployed standardized protocol, is preferred in practice to measure link loads and obtain some information on the traffic matrix. The challenge lies in the ill-posed nature of the problem: the number of unknown OD flows is much larger than the number of SNMP measurements. For this reason detecting an anomaly in the traffic matrix from SNMP measurements is a difficult task.

Several approaches are proposed to remedy this problem. The first group of methods consists in detecting anomalies in SNMP measurements without taking into account the traffic matrix. Such methods typically use time series (AutoRegressive Integrated Moving Average or ARIMA models among others) [2, 3, 4, 5] to model the SNMP measurements' evolution in time and detect deviations. In [6], the authors propose to decompose the SNMP measurements on a Principal Component Analysis (PCA) basis. These methods can detect anomalies by monitoring each link but they do not exploit the linear mathematical relation between the OD traffic matrix and the SNMP measurements, represented by the routing matrix. Moreover, they cannot be applied when the routing matrix varies in time (dynamic routing) and they cannot be used to estimate the OD traffic matrix. The second group of methods [1, 7, 8] exploits this linear mathematical relation. These approaches typically assume that the traffic matrix is well approximated by a known statistical model. Such a method requires a well known prior to be efficient, which is not always feasible in practice. In [9], the authors study a large number of methods based on different models for SNMP measurements (wavelets, PCA among others) and OD flows (ARIMA time series) to detect anomalies. A major drawback of these methods is the lack of theoretical results on their optimality properties (maximization of the probability to detect an anomaly for example) of the studied methods. Finally, the last group of methods consists in using the Kalman filtering technique [10] to model the time evolution of the traffic matrix and to detect changes in the OD flows. Unfortunately, strictly speaking, the ill-posed nature of the measurement model makes the Kalman filter not observable and the Kalman filtering efficiency strongly depends on the initialization, which is a serious limitation in practice.

The main contributions of this study are the following: firstly, a parsimonious linear model of non-anomalous OD flow volumes ("ambient" traffic) is proposed. This model can be used in two ways, either to estimate the OD flow volumes or to eliminate the non-anomalous "ambient" traffic from the SNMP measurements in order to provide residuals sensitive to anomalies. Secondly, 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 measurements and to dynamic routing. Finally, an optimal invariant detection algorithm is proposed to detect anomalies directly from SNMP measurements (no need of direct OD flow measurements). This algorithm is optimal in the sense that it maximises the probability of detecting the anomalies under a constrained false alarm probability.

2. PROBLEM STATEMENT

This section briefly presents the SNMP measurement model and the anomaly detection problem.

2.1 SNMP Measurements

Let us consider a network composed of r nodes and n monodirectional links [6, 9]. The volume of traffic $y(\ell)$,

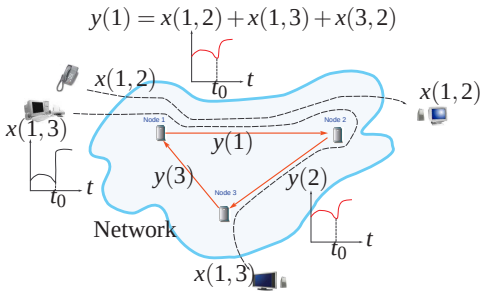

Figure 1: Detection of unusual changes in OD traffic volumes. A step-wise anomaly appears at time t_0 in the OD flow $x(1,3)$ and it is routed on links $y(1)$ and $y(2)$.

typically in bytes, on the link ℓ at time t (to simplify the notations, the subscript t is omitted) is provided by SNMP link load measurements [9]. Let $x(i,j)$ be the OD traffic demand from node i to node j at time t. This situation is shown in Fig. 1. The link loads and the traffic matrix are simply related by a linear equation $Y = AX$ where $Y = (y(1),\ldots,y(n))^T$, $X = (x(1),\ldots,x(m))^T$ contains the m ($m \gg n$) unknown traffic matrix elements $x(k) = x(i_k, j_k)$ written as a vector and $A = \{a(\ell,k)\}$ is the $n \times m$ routing matrix where $0 \leq a(\ell,k) \leq 1$ represents the fraction of OD flow k volume routed through link ℓ. Here, X^T denotes the transpose of the matrix X. Without loss of generality, the known matrix A is assumed to be full row rank, $\text{rank}(A) = n$. Clearly, since $m \gg n$, it is impossible to infer X from Y.

2.2 Volume anomaly detection problem

The detection problem consists in detecting a significant volume anomaly in an OD flow $x(i,j)$ by using only SNMP measurements $y(1),\ldots,y(n)$. For example, in Fig. 1, it is necessary to detect a sudden increase of the traffic volume $x(1,3)$ by using $y(1)$, $y(2)$ and $y(3)$ (typical anomalies to be detected in SNMP measurements are presented in [6]). As it was mentioned above, the main problem with gathering the traffic matrix from SNMP measurements is that $n \ll m$. To overcome this problem a parsimonious linear model of non-anomalous traffic has been used. The idea of this model is that the non-anomalous (ambient) traffic X can be represented at each time t by using a known family of q basis functions $B = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_q)$ such that $q < n$. Therefore, the ambient traffic can be expressed as $X \approx B\mu$ where the $m \times q$ matrix B is assumed to be known and $\mu \in \mathbb{R}^q$ is a vector of unknown coefficients which describes the OD flows' variation with respect to the set of vectors $\mathbf{b}_i$. The advantages of such a parametric model are the following: firstly, a non-parametric basis, typically the PCA basis, can be used to generate the matrix B but this solution needs direct OD flow measurements (infeasible in practice) and the PCA basis depends on the period when the measurements are made [11]. Secondly, the parametric detection methods' performance is better than the non-parametric one provided the adopted model is accurate enough. Therefore, the parametric approach is used to design a reliable detection algorithm derived from a spline-based OD flow model.

3. OD FLOW MODEL

The derivation of the ambient traffic matrix model consists of two different steps: i) description of the ambient traffic by using a spatial stationary gravity model and ii) linear approximation of the gravity model by using polynomial splines.

3.1 General gravity model

Gravity models [12, 13] assume that the traffic volume between two nodes is directly proportional to the relative attraction of each node and inversely proportional to some function of the separation (typically the distance) between these nodes. Let us assume that, at time t, the node i of the network is characterized by an attractive factor $A(i)$ related to the incoming flow at node i and a repulsion factor $R(i)$ associated to the outgoing flow from node i. A network-adapted formulation of the gravity model [12] may be given by the following equation:

$$x(i,j) \approx h\left(R(i) \cdot A(j)\right) \cdot d(i,j)^{-1}, \tag{1}$$

where $d(i,j)$ is the deterrence function representing the separation between nodes i and j at time t and $h : \mathbb{R} \mapsto \mathbb{R}$ is a weighting function relating the repulsion and attraction factors. Typically, the function h is a non-decreasing smooth function, which is a reasonable assumption in the absence of another physical model on the network. The function h is unknown and makes it possible to model more complex interactions between nodes and to give more flexibility to the model as suggested (but not used) in [12, 13]. Following [12], it is assumed that $d(i,j) = d$ for all $1 \leq i,j \leq r$ where d is an unknown real value. This assumption is quite reasonable in the case of an autonomous system which is geographically limited, as the Abilene network [14] for example, since the distance between nodes are not significant for end-users. Hence, it follows that (1) can be rewritten as:

$$x(k) = x(i_k, j_k) \approx d^{-1} \cdot h(\omega(k)) \tag{2}$$

where $\omega(k) = R(i_k) \cdot A(j_k)$ represents the importance of OD flow k in terms of traffic volume outgoing from i_k and incoming to j_k. Without any loss of generality, it is assumed that $0 \leq \omega(k) \leq 1$ for all k.

3.2 Spatial stationarity

Firstly, it is assumed that the order of importance between OD flows remains stable in time, i.e.

$$\omega(k_1) \leq \omega(k_2) \leq \ldots \leq \omega(k_m), \tag{3}$$

for all time t, where k_j is the index of the OD flow with the j-th rank of importance (the relevance of this assumption is shown in section 5). Secondly, one of the few invariants of Internet traffic is "the elephants and mice phenomenon" [15]: a few percentage of flows contributes to a large proportion of total traffic. Hence, OD flows can typically be classified in three classes of traffic depending of their volume of traffic: dominating OD flows, negligible OD flows and medium-size OD flows. This last class of traffic is defined to add more flexibility to the classification of OD flows. These classes induce a segmentation of the interval $[0;1]$ in three parts $[0;\pi_1]$, $[\pi_1;\pi_2]$ and $[\pi_2;1]$, with $0 < \pi_1 < \pi_2 < 1$, corresponding respectively to negligible, medium and dominating OD flows. As it has been experimentally confirmed, the segmentation is time-stable, at least during several days (see section 5).

3.3 OD flow spline-based model

The function h is assumed to be non-decreasing with a certain smoothness and defined piecewise, with respect to knots π_1 and π_2, on the interval $[0;1]$. Therefore, it is linearly approximated by using polynomial splines (basic definitions and results on polynomial splines can be found in [16]). Let $\{\omega \mapsto b_1(\omega), \ldots, \omega \mapsto b_q(\omega)\}$ be a basis of $q = p+3$ functions for the space of splines of degree p with $p-1$ continuous derivatives and 2 knots. By using (2), there exists a unique time-dependent vector $\mu = (\mu(1), \ldots, \mu(q))^T$ such as $X \approx B\mu$ where B is the $m \times q$ matrix whose element at position (i, j) is $b_j(\omega(i))$. Here, since the function h is unknown, sampling points $\omega(i)$ can be arbitrarily chosen in the interval $[0;1]$ provided condition (3) is verified. The existence of dominant OD flows is sometimes modelled by using α-stable laws [17]. In such a context, the spline-based approximation is naturally justified by the necessity to have a piecewise approximation of such a heavy-tailed power law distribution. Finally, it is assumed that model errors together with the natural variability of the OD flows follow a spatial Gaussian distribution [1], which leads to the model:

$$X = B\mu + \xi \tag{4}$$

where $\xi \sim \mathcal{N}(0, \gamma^2 \Sigma)$ is a Gaussian noise with the $m \times m$ spatial diagonal covariance matrix $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. The matrix Σ is assumed to be known and stable in time. On the contrary, the scalar γ^2 serves to model the mean level of the variance (due to the natural OD flow time variability) and it may depend on the time. In practice, Σ and γ^2 are estimated from a few anomaly-free SNMP measurements.

4. ANOMALY DETECTION PROBLEM

The goal is to detect an anomaly with the highest probability of detection for a given probability of false alarm, i.e. activating an alarm in absence of anomaly, which is an undesirable event.

4.1 Hypotheses testing: problem statement

According to the previous section, the non-anomalous link load measurement model is given by the following linear model :

$$Y = AB\mu + A\xi = G\mu + \zeta, \tag{5}$$

where $Y = (y(1), \ldots, y(n))^T$ and $\zeta \sim \mathcal{N}(0, \gamma^2 A\Sigma A^T)$. Without any loss of generality, the resulting matrix $G = AB$ is assumed to be full column rank. Since the matrix $\Phi = A\Sigma A^T$ is known, the testing problem consists of choosing between the two alternatives:

$$\mathscr{H}_0 = \{Z \sim \mathcal{N}(\theta + H\mu, \gamma^2 I_n);\ \theta = 0,\ \mu \in \mathbb{R}^q\} \tag{6}$$

$$\mathscr{H}_1 = \{Z \sim \mathcal{N}(\theta + H\mu, \gamma^2 I_n);\ \theta \neq 0,\ \mu \in \mathbb{R}^q\}, \tag{7}$$

with $Z = \Phi^{-\frac{1}{2}} Y$, $H = \Phi^{-\frac{1}{2}} G$, $\Phi^{-\frac{1}{2}}$ is the square-root matrix of Φ^{-1}, Φ^{-1} is the inverse of Φ and I_n is the identity matrix of size n. Here μ is considered as a nuisance vector parameter since i) it is completely unknown, ii) it is not necessary for the anomaly detection and iii) it can mask the anomalies. Typically, when an anomaly occurs in OD flow j, the vector θ has the form $\theta = \varepsilon \Phi^{-\frac{1}{2}} \mathbf{a}(j)$ where $\mathbf{a}(j)$ is the j-th normalized column of A and ε is the intensity of the anomaly. The aim is to detect the presence of an anomalous vector θ not explicable by the ambient traffic model (4).

Let $\mathscr{K}_\alpha = \{\phi : \sup_{\mu \in \mathbb{R}^q} \Pr_{\theta=0,\mu}(\phi(Z) = \mathscr{H}_1) \leq \alpha\}$ be the class of tests $\phi : \mathbb{R}^n \mapsto \{\mathscr{H}_0, \mathscr{H}_1\}$ with upper-bounded maximum false alarm probability, where the probability $\Pr_{\theta,\mu}$ stands for the vector of observations Z being generated by the distribution $\mathcal{N}(\theta + H\mu, \gamma^2 I_n)$ and α is the prescribed probability of false alarm. The power function β is defined with the probability of correct detection: $\beta(\theta;\mu) = \Pr_{\theta \neq 0,\mu}(\phi(Z) = \mathscr{H}_1)$. The subtlety of the above mentioned hypotheses testing problem consists of choosing between $\mathscr{H}_0$ and $\mathscr{H}_1$ with the best possible performance indexes (α, β) while considering μ as a nuisance parameter.

4.2 Anomaly detection methodology

It is easy to see that the problem remains invariant under the group of translations $G = \{g : g(Z) = Z + H\mathbf{c},\ \mathbf{c} \in \mathbb{R}^q\}$ (see an introduction to the principle of invariance in [18]). The maximal invariant statistics (also called "parity vector") $U = WZ$ is the transformation of the measured output Z into a set of $n-q$ linearly independent variables by projection onto the left null space of the matrix H. The matrix $W^T = (w_1, \ldots, w_{n-q})$ of size $n \times (n-q)$ is composed of the eigenvectors $w_1, \ldots, w_{n-q}$ of the projection matrix $P_H^\perp = I_n - H(H^T H)^{-1} H^T$ corresponding to eigenvalue 1. The matrix W satisfies the following conditions: $WH = 0$, $W^T W = P_H^\perp$, $WW^T = I_{n-q}$. Let $\mathscr{S}$ be the family of surfaces $\mathscr{S} = \{S_c : c > 0\}$ with $S_c = \{\theta : \|P_H^\perp \theta\|_2^2 / \gamma^2 = c^2\}$. Then, it is shown [19] that the test

$$\phi^*(Z) = \begin{cases} \mathscr{H}_0 \text{ if } \Lambda(Z) = \|P_H^\perp Z\|_2^2 / \gamma^2 < \lambda_\alpha \\ \mathscr{H}_1 \text{ else} \end{cases}, \tag{8}$$

where the threshold λ_α is chosen to satisfy the false alarm bound α, $\Pr_{\theta=0,\mu}(\Lambda(Z) \geq \lambda_\alpha) = \alpha$, is Uniformly Best Constantly Powerful (UBCP)[1] in the class $\mathscr{K}_\alpha$ over the family of surfaces $\mathscr{S}$. The statistics Λ is distributed according to the χ^2 law with $n-q$ degrees of freedom. This law is central under $\mathscr{H}_0$ and non-central under $\mathscr{H}_1$ with the non-centrality parameter $\theta^T P_H^\perp \theta / \gamma^2$.

5. NUMERICAL RESULTS

This section shows the relevance of the model and the performance of the detection algorithm.

5.1 Description of the data set

The evaluation of the proposed methods requires the knowledge of the real OD traffic flows. Such measurements are quite difficult to obtain in a commercial network but are available for the Abilene network. The Abilene backbone is composed of $r = 12$ core routers and $m = 144$ OD flows. For these numerical experiments, $n = 42$ backbone links are measured. More details on this network are given in [14] and real data are available in [20]. The primary data inputs are the time series of link loads (bytes across interfaces) gathered through SNMP. The sampling rate is one measurement per 10 minutes, i.e. each measurement corresponds to the total of volume of traffic (in bytes) which has passed through a

[1] A test $\phi^* \in \mathscr{K}_\alpha$ is UBCP on $\mathscr{S}$ if 1) $\beta_{\phi^*}(\theta') = \beta_{\phi^*}(\theta'')$, $\forall \theta', \theta'' \in S_c$; 2) $\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$, $\forall \theta \in S_c$, $\forall c > 0$ for any test $\phi \in \mathscr{K}_\alpha$ which satisfies 1).

given link during 10 minutes. Two sets of measurements are used: the first one, the anomaly-free data set, is composed of 6 anomaly-free SNMP measurements (one hour measurement period) and the second one, the testing data set, is composed of 720 SNMP measurements (five days measurement period). Let $\mathbb{T}_a$ (respectively $\mathbb{T}_b$) be the set of time index associated to SNMP measurements of the anomaly-free (resp. testing) data set. The anomaly-free data set is measured one hour before the testing one.

To identify the set of "true" anomalies in the testing data set (as a precursor to the validation step), unusual deviations from the mean in each OD flow are manually detected. Manual inspection declares an anomaly if the unusual deviation intensity of the guilty OD flow leads to an augmentation of traffic 1) larger than 1.5% of the total amount of traffic on the network and 2) larger than 1% of the amount of the traffic carried by the links routing this guilty OD flow. Hence, only significant volume anomalies are considered as "true anomalies" (small volume anomalies have little influence on link utilisations). Let $\mathbb{T}_b^\circ \subset \mathbb{T}_b$ be the set of time index t associated to the 680 non consecutive SNMP measurements of the testing data set manually declared as anomaly-free (40 measurements of the testing data set are affected by at least one significant volume anomaly).

5.2 Numerical validation of the model

Although many aspects could potentially be included in the evaluation, the focus is on the potential impact of performance errors on traffic engineering tasks. Hence, the root mean square error (RMSE) is used:

$$\mathrm{RMSE}^{label}(t) = \sqrt{\sum_{i=1}^{m} \left(\hat{x}_t^{\,label}(i) - x_t(i)\right)^2}, \ \forall t \in \mathbb{T}_b^\circ.$$

Here, $x_t(i)$ denotes the true traffic volume of OD flow i at time $t \in \mathbb{T}_b^\circ$ and $\hat{x}_t^{\,label}(i)$ denotes the corresponding estimate for the method entitled '*label*'. Three estimates are compared : 1) simple gravity estimate [14] with label 'SG', 2) tomogravity estimate [12] with label 'TG' and 3) spline-based Maximum Likelihood (ML) estimate with the label 'SML'. Since the measurement model (5) is a Gaussian linear one, the optimal estimate of X_t is the ML estimate [21] $\hat{X}_t^{\,SML}$ given by:

$$\hat{X}_t^{\,SML} = B(H^T H)^{-1} H^T Z_t.$$

The statistical properties of the ML estimate are well known [21] contrary to the simple gravity and tomogravity estimates [14, 12]. The temporal correlation of the noise sequence $(\zeta_t)_{t\geq 1}$ is ignored for the following reason: it can be theoretically shown that the integration of the ARMA (AutoRegressive and Moving Average) model of $(\zeta_t)_{t\geq 1}$ does not change the covariance matrix (and the bias) of $\hat{X}_t^{\,SML}$.

The spline-based model is computed by using SNMP measurements of the short anomaly-free data set : 1) the tomogravity estimate $\hat{x}_t^{\,TG}(k)$ is computed for all OD flow k and all $t \in \mathbb{T}_a$, 2) the mean flow values $\bar{x}^{TG}(k) = \sum_{t\in\mathbb{T}_a} \hat{x}_t^{\,TG}(k)$ are computed and 3) sorted in ascending order to obtain a rough estimate of the OD flow ranks. The spline-based model is designed with cubic splines ($p = 3$) and knots $\pi_1 = 0.8507$ and $\pi_2 = 0.9830$ with sampling points $\omega(k)$ uniformly distributed in the interval $[0;1]$. Small variations on the values π_1 and π_2 have no serious effect on the results.

Method	SG	TG	SML
Total RMSE	9336.9	3934.9	3765.6

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

The mean value $\bar{x}^{TG}(k)$ is also used as an estimate $\hat{\sigma}_k^2$ of σ_k^2, which leads to an estimate $\hat{\Phi}$ of Φ (quite efficient and sufficient in practice). In this estimation step, it is not necessary to know γ_t^2 as explained in [21].

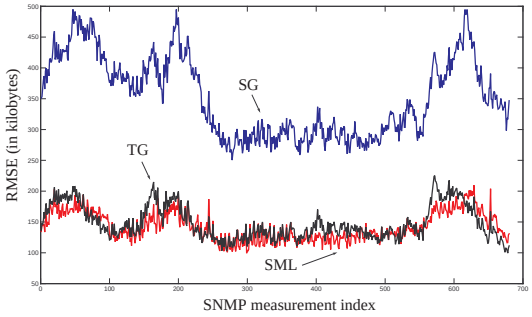

Figure 2: Comparison between the SG, TG and SML RMSE for 680 anomaly-free measurements.

Fig. 2 shows the error $\mathrm{RMSE}(t)$ over the set $\mathbb{T}_b^\circ$. The x-axis of the figure has no time meaning and it corresponds to the index of each measurement. The sum $\sqrt{\sum_{t\in\mathbb{T}_b^\circ} \mathrm{RMSE}(t)^2}$ on this time period is computed in Table 1 as a global indicator to compare the methods. Clearly, the spline-based estimate gives better results than the others. To verify the spatial Gaussian assumption, residuals U_t are computed for each $t \in \mathbb{T}_b^\circ$. The Kolmogorov-Smirnov test [18] at the level 5% accepts the Gaussian hypothesis for 670 of these measurements (acceptation 98.5% of the time).

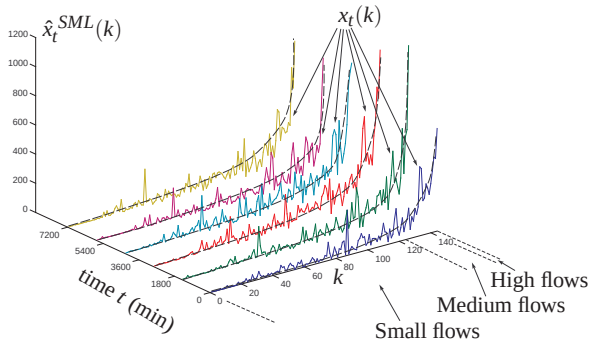

Figure 3: Approximation of real OD flows (full lines) by the spline-based model (dashed lines).

Typical non-anomalous OD flows, sorted in the increasing order of their volume of traffic, are shown as functions of time t in Fig. 3. Since the "shape" of sorted OD flows is almost constant over the time, only a few sorted OD flows are plotted. The SML estimated profiles of the OD flows are

Type of situation	Spline-based	PCA
Normal working	672 (98.82 %)	673 (98.97 %)
False alarms	8 (1.18 %)	7 (1.03 %)
Missed detections	9 (22.50 %)	35 (87.50 %)
Correct detections	31 (77.50 %)	5 (12.50 %)

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

also plotted on the same figure with dashed lines. It shows that the "shape" of the OD flows is well modeled by the proposed spline-based model and is stable over the time. Hence, the spatial stationarity assumption is quite reasonable.

5.3 Numerical validation of the detection algorithm

The detection algorithm is applied to the SNMP measurements of the testing data set. For the detection purpose, it is crucially important to have a good estimate of the noise level γ_t^2. This parameter is estimated from the short anomaly-free data set by using the ML estimate of noise variance [21] in residuals U_t. Since this parameter can slowly vary in time, its value is updated during the test. During the test, at time t, if no anomaly has been declared one hour before, γ_t^2 is estimated by its value one hour before.

The results are presented in Table 2. The second column shows that the proposed test obtains a false alarm rate of 1.18% comparable with the prescribed level of false alarm $\alpha = 0.01$. The probability to detect a volume anomaly is about 77.5%. The third column presents the results obtained by the PCA test described in [6]. The threshold of this test is chosen to obtain a similar false alarm rate of 1.03%. Clearly, the PCA test is not as sensitive (correct detection rate about 12.50%) as the proposed test. Indeed, the PCA decomposition of SNMP measurements is too rough to detect small (but significant) anomalies. Finally, Fig. 4 shows the correct

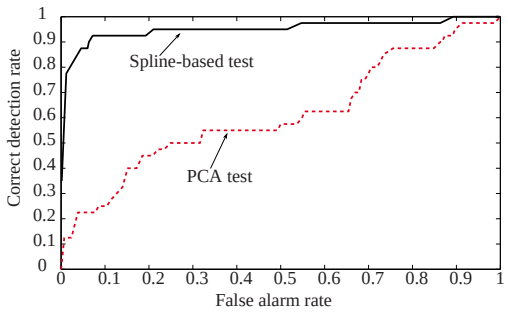

Figure 4: Correct detection rate versus false alarm rate for the spline-based test (solid line) and the PCA test (dotted line).

detection rates of the spline-based test and the PCA test for different false alarm rates varying between 0 and 1. For example, for a correct detection rate of 80%, the false alarm rate of the spline-based test ($\approx 2\%$) is clearly better than that of the PCA test ($\approx 70\%$). Hence it can be concluded that the numerical results confirm the theoretical properties of the detection method and shows that the proposed test outperforms the PCA approach, at least for the Abilene data set.

6. CONCLUSION

The problem of anomaly detection in OD traffic volume from SNMP measurements has been considered as a statistical hypotheses testing problem with nuisance parameters (non-anomalous traffic). Since the number of SNMP measurements is significantly lower than the number of OD flows, an original linear spline-based parsimonious model is proposed to describe the non-anomalous traffic and to overcome the ill-posed nature of the SNMP measurement model. Results obtained with real data traffic from a United States backbone network show that both the OD traffic matrix estimation and the volume anomaly detection approaches outperform the methods previously applied in the field.

REFERENCES

[1] M. Coates, A. Hero, R. Nowak, and B. Yu, "Internet tomography," *IEEE Signal Processing Mag.*, May 2002.

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

[3] A. Tartakovsky *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.

[4] B. Krishnamurthy *et al.*, "Sketch-based change detection: methods, evaluation, and applications," in *IMC*, 2003.

[5] T. Ahmed, M. Coates, and A. Lakhina, "Multivariate online anomaly detection using kernel recursive least squares," in *Infocom*, 2007.

[6] A. Lakhina *et al.*, "Diagnosing network-wide traffic anomalies," in *SIGCOMM*, 2004.

[7] C. Tebaldi *et al.*, "Bayesian inference on network traffic using link count data," *J. Amer. Statist. Assoc.*, vol. 93, no. 442, pp. 557–576, 1998.

[8] J. Cao *et al.*, "Time-varying network tomography: router link data," *Journal of Americal Statistical Association*, vol. 95, no. 452, pp. 1063–1075, 2000.

[9] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *IMC'05*, 2005.

[10] A. Soule *et al.*, "Traffic matrices: balancing measurements, inference and modelling," in *SIGMETRICS*, 2005.

[11] H. Ringberg *et al.*, "Sensitivity of PCA for traffic anomaly detection," in *SIGMETRICS*, 2007.

[12] Y. Zhang *et al.*, "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.

[13] A. Gunnar *et al.*, "Traffic matrix estimation on a large IP backbone - A comparison on real data," in *IMC*, 2004.

[14] Y. Zhang *et al.*, "Fast accurate computation of large-scale IP traffic matrices from link loads," in *SIGMETRICS*, 2003.

[15] A. Medina *et al.*, "Traffic matrix estimation : existing techniques and new directions," in *SIGCOMM*, 2002.

[16] G. Nürnberger, *Approximation by spline functions*. Springer-Verlag, 1989.

[17] P. Doukhan *et al.*, *Theory and Applications of Long-Range Dependence*. A Birkhäuser book, 2003.

[18] E. Lehman, *Testing Statistical Hypotheses, Second Edition*. Chapman & Hall, 1986.

[19] L. Fillatre and I. Nikiforov, "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.

[20] Y. Zhang, "6 months of Abilene traffic matrices," 2003, http://www.cs.utexas.edu/∼yzhang/research/AbileneTM/.

[21] C. Rao, *Linear statistical inference and its applications (second edition)*. John Wiley & Sons, 1973.