PCA-based Network-wide Correlated Anomaly Event Detection and Diagnosis

Yuanxun Zhang, Prasad Calyam, Saptarshi Debroy
University of Missouri-Columbia
Email: yzd3b@mail.missouri.edu, {calyamp, debroysa}@missouri.edu

Mukundan Sridharan
The Samraksh Company
Email: mukundan.sridharan@samraksh.com

***Abstract*—High-performance computing environments supporting large-scale distributed computing applications need multi-domain network performance measurements from open frameworks such as perfSONAR. Network-wide correlated anomaly events that can potentially impact data throughput performance need to be quickly and accurately notified for smooth computing environment operations. Since network topology is not always available along with the measurements data, it is challenging to identify and locate network-wide correlated anomaly events that impact data throughput performance. In this paper, we present a novel PCA-based correlated anomaly event detection scheme that can fuse multiple time-series of measurements and transform them using principal component analysis. We demonstrate using actual perfSONAR one-way delay measurement datasets that our scheme can: (a) effectively distinguish between correlated and uncorrelated anomalies, (b) leverage a source-side vantage point to diagnose whether a correlated anomaly event location is local or in an external domain, (c) act as a "black-box" correlation analysis tool for key insights in eventual root-cause identification.**

I. Introduction

Distributed computing applications are increasingly being developed in scientific communities in areas such as biology, geography and high-energy physics. These communities transfer data on a regular basis between computing and collaborator sites at high-speeds on multi-domain networks that span across continents. Given the real-time data consumption demands of users that require ensuring high data throughputs through effective network monitoring, there is a rapidly increasing trend to deploy multi-domain, open measurement frameworks such as perfSONAR [1]. The perfSONAR framework has been developed over the span of several years by worldwide-teams to assist in creating "measurement federations" for: measurement data collection, storage/archival and dissemination for monitoring and diagnosing bottlenecks that hinder end-to-end data transfer speeds. In addition, perfSONAR can be used to understand network performance changes due to network fault events (e.g., misconfigurations, outages) and cross-traffic congestion that impact application and protocol behavior.

Over 1000 perfSONAR measurement points have been deployed all over the world, and are sampling both active and passive measurements of various metrics several times a day [2]. They are exposing collected measurements via web-services in the form of vast data archives of current and historic measurements on national and international backbones (e.g., ESnet, Internet2, GEANT, SWITCH, RNP). The consumers of these measurements (e.g., network operators, researchers in scientific disciplines) are faced with the challenge to analyze and interpret the vast measurement archives across end-to-end multi-domain network paths with minimal human inspection.

Consequently, there is a dire need for automated techniques to query, analyze, detect and diagnose prominent network performance anomalies that hinder data transfer speeds, and works such [3] - [6], [15] address specific aspects of these needs. In particular, our earlier work on adaptive plateau event detection [5] (APD) proposed a scheme to detect uncorrelated network anomaly events (change-points from statistical norm) at the network-path level by analyzing for e.g., end-to-end one-way delay and throughput measurement time series from OWAMP and BWCTL active measurement tools used in perfSONAR deployments, respectively. Our APD scheme avoids manual calibration of 'sensitivity' and 'trigger elevation threshold' parameters used in earlier static plateau detector (SPD) schemes [7] [8] for diverse profiles of measurement samples on network paths. It uses principles of reinforcement learning in order to achieve low false alarm rates, at the cost of a fractional increase in online detection time.

The general limitation of any scheme that detects uncorrelated anomaly events is that the root-cause location cannot be correlated and isolated. Consequently, in a follow-up work [6], we developed a correlated topology-aware network anomaly event detection scheme that analyzes several network-path level (uncorrelated) anomaly events in order to localize the change-cause to a particular network segment. However, one of the major challenges in realizing such a scheme is the general unavailability of network topology information along with the multi-domain measurements data due to lack of topology publication services within ISPs or for other policy reasons. Without publicly accessible topology information for measurement points, it is even more difficult to identify and locate network-wide correlated anomaly events that impact data throughput performance. The identification and location diagnosis of correlated anomaly events is especially true in case of fault isolation analysis with high-dimensional measurement data spanning multiple network paths.

In this paper, we present a *novel scheme that can fuse time-series of perfSONAR path measurements from multiple domains with common intermediate hops, and transform them using principal component analysis (PCA) [9] for correlated anomaly event detection*. The proposed scheme involves fusion of multiple time-series periodically collected from perfSONAR dashboard queries by network operators in order to transform specific source-related perfSONAR measurements onto new axes (i.e., principal components). The transformation extracts common features upon which our earlier APD scheme

This material is based upon work supported by the Department of Energy under Award Numbers: DE-SC0001331 and DE-SC0007531. The views and opinions of authors expressed herein do not necessarily state or reflect those of the united states government or any agency thereof.

/978-1-4799-7795-6/15/$31.00 ©2015 IEEE

is applied on the transformed data to detect anomaly events at a network-wide level. This approach leverages the fact that PCA technique is best suited to be configured by network operators as a "black-box" [10] [11] for correlation analysis.

Through a case study with perfSONAR one-way delay measurements analysis, we demonstrate that our novel PCA-APD scheme can automatically distinguish between correlated and uncorrelated anomalies in the absence of complete network topology information with high detection accuracy, and low false alarm rate. More specifically, we show how our PCA-APD scheme can leverage source-side vantage points in measurement traces that contain source and destination sites information, in order to diagnose whether a correlated bottleneck anomaly event location is local or in an external domain. Thus, our proposed scheme is helpful for detection and diagnosis of correlated network anomaly events that relate to network faults or bottlenecks, and ultimately can save time, cost and effort in multi-domain network paths management.

The remainder paper organization is as follows: Section II describes the related work. Section III presents background on plateau anomaly detection and the PCA technique. Section IV presents details of our novel PCA-APD scheme. In Section V, we apply our PCA-APD scheme in a case study with short-term and long-term measurements to isolate bottleneck anomaly event locations within actual perfSONAR measurement traces. Section VI concludes the paper.

II. Related Work

To assist network operators in troubleshooting bottlenecks (e.g., prolonged congestion events or device misconfigurations) in high-speed networks, a number of smart and effective network monitoring tools based on statistical measurement data analysis techniques have been developed. In particular, there have been studies on correlated anomaly detection such as [10] - [13]. Authors in [10] use PCA technique on passive measurements for network anomaly detection on a network link basis. In [13], the authors address limitations of PCA's failure in detecting strong correlations in distributed network traffic anomalies. Both [10] and [13] do not use topology information and thus propose black-box techniques, in comparison to other topology-aware works such as [6, 11, 12, 14–16].

In our recent work [6], we used spatial and temporal analysis after combining topology and uncorrelated anomaly events information corresponding to multiple measurement time series for location diagnosis of correlated anomaly events. However, such analysis has a strict requirement for topology information, which is generally not made publicly available by domains that share perfSONAR measurement data. The authors in [11] use Kalman-filter for anomaly detection and build a traffic matrix of an enterprise network to overcome link basis limitations. In [12], the authors present a general framework called NICE (Network-wide Information Correlation and Exploration) for analyzing data through correlations and present a qualitative as well as quantitative analysis approach with network related data such as router logs and topology information. Routing connection relationships are used in [14] for network-wide anomaly detection in backbone networks; relationships are established based on features such as packet sizes, IP addresses and ports.

Authors in [15] use perfSONAR measurements for root-cause analysis and localizations of performance problems, however their analysis also has strict requirement for topology information. Our work on using measurement data for anomaly detection is closest to the work by authors in [16]. Therein, an anomaly detection system is developed based on prediction of upper and lower dynamic thresholds of various time-varying data trends that include sparse and transient data.

III. Background

In this section, we first define anomaly events that are of interest to network operators, and give an overview of adaptive plateau event detection (APD). Following this, we formally introduce the PCA technique which we will leverage in our proposed technique.

A. Adaptive Plateau Detector

One of the significant challenges in dealing with measurement data sets is to decide which kind of network events need to be labeled and notified as anomaly events that may indicate potential performance bottlenecks. Various traffic related anomaly events are caused due to IP route/AS path change events that involve traffic re-routing on backup paths due to ISP traffic migration for maintenance reasons. These events manifest in the form of spikes, dips, bursts, persistent variations and plateau trends in network performance metrics such as round-trip delay, available bandwidth and packet loss obtained through end-to-end active measurements. Based on documented experiences from network operators and application users [8] and based on our own discussions with other network operators (e.g., ESnet, Internet2, GEANT), the notification of 'plateau anomalies' shown in Fig. 1 are the most worthy to be notified. These anomaly events are commonly known to impact data transfer speeds at the application-level on high-speed network paths. Network operators, when analyzing a measurement time-series of network performance metrics, typically look for plateau event trends through visual inspections and seek for automated notification of such network-wide detected anomaly events.

Variants of plateau anomaly event detectors have been developed and adopted in large scale monitoring infrastructures such as NLANR AMP [7] and SLAC IEPM-BW [8], which are predecessors to the perfSONAR deployments. These detectors use static configurations of 'sensitivity' and 'trigger elevation threshold' parameters to detect that a plateau event or a 'change event' has occurred. Note that a small sensitivity value results in triggers of the slight variations in network performance magnitudes, whereas a large value could overlook actual anomalies that should be detected. Trigger elevation prevents repeated triggers of an already detected plateau event. A plateau event is detected if the most recent measurement sample value crosses the upper or lower thresholds of the summary (i.e., T_{SU}, T_{SL}) and quarantine (i.e., T_{QU}, T_{QL}) buffers as determined by the settings of sensitivity and trigger elevation parameters. The summary buffer is used to maintain sample history that indicates the normal state (before anomaly event occurs), and a quarantine buffer is used to store outlier data samples that are twice the normal state sample values.

The sample counts in above buffers are used to maintain trigger count values over a pre-configured trigger duration before an alarm of anomaly event occurrence (indicated by the cross mark in Fig. 1) is notified. The trigger duration before samples are marked for impending anomaly states (triangle symbols shown in Fig. 1) should be chosen long

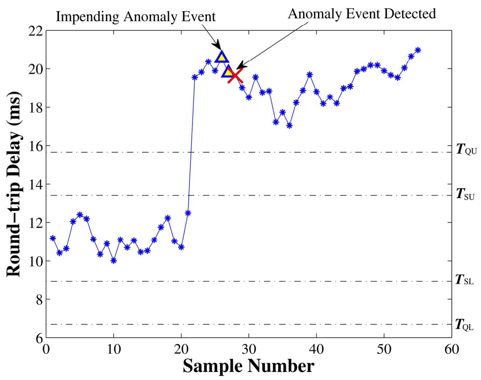

Fig. 1: Plateau-detector thresholds illustration

enough to avoid false alarms due to noise events corresponding to intermittent spikes, dips, or bursts. Our earlier adaptive plateau-detector (APD) algorithm [5] scheme avoids manual calibration of 'sensitivity' and 'trigger elevation threshold' parameters and has been shown to be more accurate than earlier static plateau detection schemes [7] [8] over a diverse profiles of measurement samples on network paths.

B. *Principal Component Analysis*

Principal Component Analysis (PCA) is a dimensionality-reduction approach that involves mapping a set of data points within time-series onto new coordinates. The new coordinates are called the principal axes or principal components that help to extract common features in the data points of multiple time-series, and thus visually separate the normal behavior from anomalous behavior.

Let $\mathbf{Y}$ be the $n \times m$ time-series measurement matrix, which denotes the time-series of all links and centered to have zero mean, with n being the number of rows and m being the number of columns. Thus, each column denotes the time-series of the i-th link and each row j represents an instance of all the links. And let $\mathbf{y} = \mathbf{y}(t)$ denote a n-dimensional vector of measurements (for all links) from a single time step t. Formally, PCA is a projection method that maps a given set of data points onto principal components ordered by the amount of data variance they capture. Applying PCA to $\mathbf{Y}$ yields a set of m principal components, $\{\mathbf{v}_i\}_{i=1}^{m}$. The first principal $\mathbf{v}_1$ is the vector that points in the direction of maximum variance in $\mathbf{Y}$:

$$\mathbf{v}_1 = \arg \max_{\|\mathrm{v}\|=1} \|\mathbf{Y}\mathbf{v}\| \tag{1}$$

Where $\|\mathbf{Y}\mathbf{v}\|$ is proportional to the variance of the data measured along $\mathbf{v}$. Proceeding iteratively, once the first $k-1$ principal components have been determined, the k-th principal component corresponds to the maximum variance of the residual. The residual is the difference between the original data and the data mapped onto the first $k-1$ principal axes. Thus, we can write the k-th principal component $\mathbf{v}_k$ as:

$$\mathbf{v}_k = \arg \max_{\|\mathrm{v}\|=1} \|(\mathbf{Y} - \sum_{i=1}^{k-1} \mathbf{Y}\mathbf{v}_i\mathbf{v}_i^{\mathbf{T}})\mathbf{v}\| \tag{2}$$

As shown in [10], PCA is useful to explore the intrinsic dimensionality of a set of data points. Most data variance can be captured by the first $k = 4$ principal components. In this way, all possible link measurements could be separated onto *normal* measurements subspace S_{no} and *abnormal* measurements subspace S_{ab}. Consequently, the *normal* measurements reside in a low k-dimensional subspace S_{no}. The remaining $(n-k)$ principal components constitute the *abnormal* measurements subspace S_{ab}.

Detection of anomalies relies on the decomposition of link measurements $\mathbf{y} = \mathbf{y}(t)$ at any step into normal and abnormal components, $\mathbf{y} = \mathbf{y}_{no} + \mathbf{y}_{ab}$, the $\mathbf{y}_{no}$ corresponds to modeled normal measurements (the projections of $\mathbf{y}$ onto S_{no}), and the $\mathbf{y}_{ab}$ corresponds to residual measurements (the projections of $\mathbf{y}$ onto S_{ab}), and can be computed as:

$$\mathbf{y}_{no} = \mathbf{P}\mathbf{P}^T\mathbf{y} = \mathbf{C}_{no}\mathbf{y} \;\text{ and }\; \mathbf{y}_{ab} = (\mathbf{I} - \mathbf{P}\mathbf{P}^T)\mathbf{y} = \mathbf{C}_{ab}\mathbf{y} \tag{3}$$

where $\mathbf{P} = [\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, ..., \mathbf{v_k}]$ is formed by the first k principal components which capture the dominant variance in data. The matrix $\mathbf{C}_{no} = \mathbf{P}\mathbf{P}^T$ represents the linear operator that performs projection onto normal subspace S_{no}, and the $\mathbf{C}_{ab}$ represents the projection onto the abnormal subspace S_{ab}.

As described in [10], a volume anomaly event typically results in a large change to $\mathbf{y}_{ab}$; thus, a useful metric for detecting abnormal measurements pattern is squared prediction error (**SPE**):

$$\mathbf{SPE} \equiv \|\mathbf{y}_{ab}\|^2 = \|\mathbf{C}_{ab}\mathbf{y}\|^2 \tag{4}$$

We consider network measurements to be normal if $\mathbf{SPE} \leq \delta_\alpha^2$, where δ_α^2 denotes the threshold for the **SPE** at the $1-\alpha$ confidence level. Such a statistic test for the **SPE** residual function is known as Q-statistic, which was developed in [9] to deal with residuals related to principal component analysis. The Q-statistic enables us to analyze the significance of the differences among the data sets. The residual measurements $\mathbf{y}_{ab}$ are valid when used to detect a volume anomaly event. However, we found that the $\mathbf{y}_{ab}$ is not valid to detect correlated anomaly events. In the measurement matrix $\mathbf{Y}$, the correlated anomalies will affect almost all the links at time t, which means the row t's data will be affected by correlated anomalies. Thus, on the contrary, the correlated anomalies will be captured in the first k principal components and reside in the *normal* measurements subspace S_{no}.

IV. PCA-APD Detection and Diagnosis

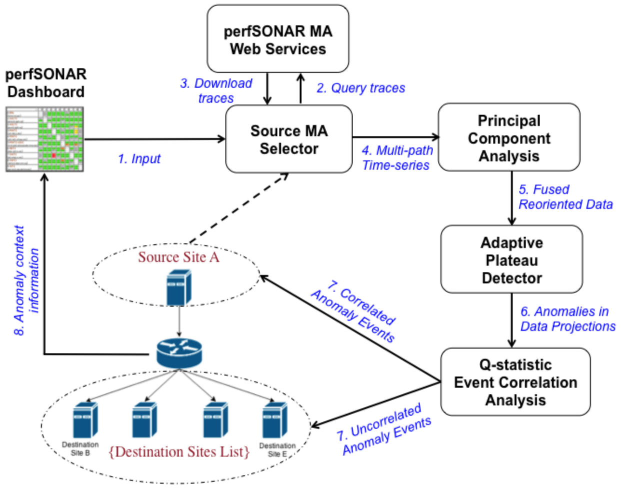
perfSONAR MA Web Services
perfSONAR Dashboard
3. Download traces
2. Query traces
Source MA Selector
Principal Component Analysis
1. Input
4. Multi-path Time-series
5. Fused Reoriented Data
Source Site A
Adaptive Plateau Detector
8. Anomaly context information
7. Correlated Anomaly Events
6. Anomalies in Data Projections
Q-statistic Event Correlation Analysis
Destination Site B
{Destination Sites List}
Destination Site E
7. Uncorrelated Anomaly Events

Fig. 2: PCA-APD scheme sequence diagram

A. Scheme Overview

Fig. 2 shows the sequence of steps involved in our PCA-APD detection and diagnosis scheme. The first step involves finding potential anomalous sources from source-specific data collection in perfSONAR dashboard [6] through querying of the openly-accessible, distributed measurement archives (accessible at an address e.g., http://testproject.exampleuniversity.edu:8085) by using perfSONAR-compliant web service clients. The site list of measurement archives (MAs) that are available for query can be selected using a global lookup service hosted by the perfSONAR community. This service registers the addresses of all openly-accessible measurement archives within individual domains. Through standardized request/response messages, active measurement time series data relating to end-to-end performance measurement tools such as OWAMP (one-way delay as specified in IETF RFC 4656) are downloaded for any given site (i.e., Source Site A). The downloaded multipath time series datasets are in the form of XML files, which are then processed using parsing for applying PCA technique in the subsequent step.

The output of PCA is the fused reoriented data comprising of eigen vectors, where the first eigen vector captures maximum variability and the last is left with minimum variability. What this translates into in-reality is that - the data projection using the first eigen vector has variability that is common to most of datasets and the last eigen vectors have the variability that is least common in the dataset (e.g., variability present in only one dataset amongst all). Next, data dimension selection is performed on the fused reoriented data. For example, if we are interested only in the common anomalies, we will select only the first principal component as described in the previous section. After the data dimension (number of eigen vectors) is selected, the data is projected using the principal components, and is passed as input for APD algorithm to detect anomalies.

B. Anomaly detection using PCA-APD

We remark that - although most of the correlated anomalies subspace are captured in the first, or first and second principal components, it is likely that the normal subspace is also located in the lower k-components. In order to accurately capture all of the anomaly events within measurement time-series, we leverage our APD scheme on the PCA transformed (or fused reoriented) measurement data. To further classify the correlated and uncorrelated anomaly events, we employ the Q-statistic test described earlier in Section III-B. Moreover, if we find that the site-of-interest (i.e., Source Site A) is featured in many or all of the correlated anomaly event paths, we can conclude that the anomaly event root-cause is local. If otherwise, we can conclude that the anomaly event root-cause is in an external domain, and above sequence of diagnosis steps can be applied to other domains whose measurement data is accessible with the hope of localizing the root-cause in one of the external domains.

To substantiate the above rationale for correlated and uncorrelated anomalies, we use synthetic measurement time-series for study purposes that comprise of 16 traces of one-way delay measurements collected from perfSONAR archives that do not have any anomaly events. Into these traces, we inject 5 anomaly events within a common time period window to create a correlated anomaly event, and also inject 16 uncorrelated anomaly events in other time period windows.

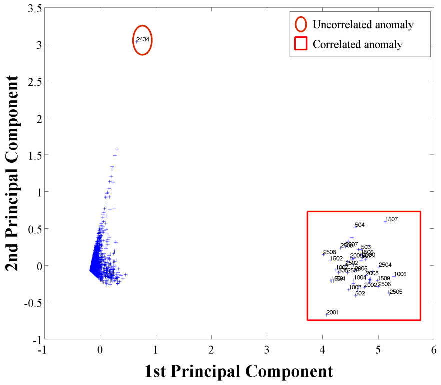

Fig. 3: Correlated and uncorrelated anomaly subspace separation with PCA application

As shown in Fig. 3, all the correlated anomaly events are captured in the first principal component, and an uncorrelated anomaly event is captured in the second principal component. In repeated studies with different synthetic measurement time-series, we found that all the correlated anomaly events are captured mostly in the first principal component, and at worst in the second principal component in a very few number of instances.

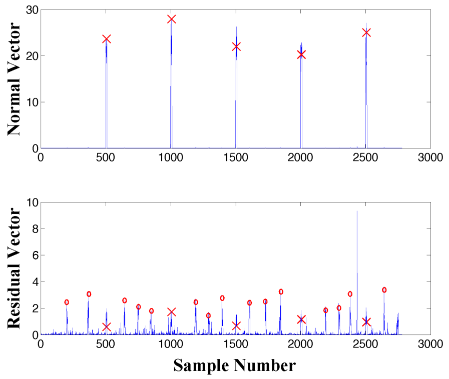

Fig. 4: Measurements of normal space vector squared magnitude ($\|\mathbf{y}_{no}\|^2$, upper), and residual space vector squared magnitude ($\|\mathbf{y}_{ab}\|^2$, lower) for the synthetic data

As shown in Fig. 4, we separate the link measurements $\mathbf{y}$ into normal subspace and residual subspace. The lower part of the figure shows the **SPE** of $\mathbf{y}$**'s** projection in the residual subspace $\mathbf{y}_{ab}$, and the upper part shows $\mathbf{y}$**'s** projection in the normal subspace $\mathbf{y}_{no}$. On these plots, we have marked the correlated anomalies with crosses (x) and uncorrelated anomalies with circles (o). In the lower part of the figure, it is clear that the magnitude of the residual vector $\mathbf{y}_{ab}$ is dominated by uncorrelated anomalies rather than correlated anomalies. As a result, it is difficult to discern the correlated and uncorrelated anomalies in the residual vector $\mathbf{y}_{ab}$. However, in the upper part of the figure, only correlated anomalies along with normal measurement data are captured in the projection. Thus, the magnitude of normal measurement data is obviously different from the correlated anomaly measurement data, which makes the detection of anomalies much easier to distinguish.

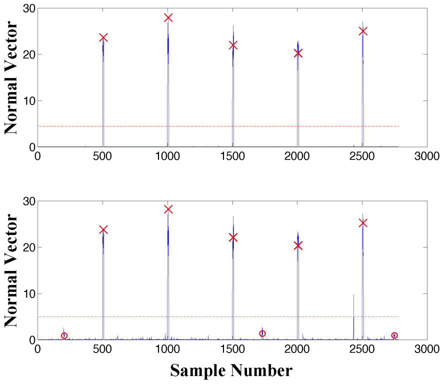

Fig. 5: Measurements of normal space vector squared magnitude $\|\mathbf{y}_{no}\|^2$ ($\mathbf{P} = [\mathbf{v_1}]$, upper) and ($\mathbf{P} = [\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, \mathbf{v_4}]$, lower) for the synthetic data

Above observation shows that the normal vector $\mathbf{y}_{no}$ is suitable to detect correlated anomalies at a network-wide level. However, we still want to find uncorrelated anomalies. In Fig. 4, only correlated anomalies are captured in the normal vector. Although residual vector can capture all the correlated and uncorrelated anomalies, it is difficult to discern them because only the first principal axis is selected in the Fig. 4 to capture normal traffic and correlated anomalies. Hence, we need to increase principal axes to capture uncorrelated anomalies.

In the Fig. 5, only correlated anomalies are captured in the first principal component projection. However, in the lower plot of Fig. 5, correlated anomalies and some of the uncorrelated anomalies are captured in the first 4 principal components projection. The Q-statistics (δ^2_α) are also shown in these plots. From the lower plot, we found Q-statistic (δ^2_α) is sensitive to the detected correlated anomalies but not the uncorrelated anomalies. Based on these characteristics of correlated and uncorrelated anomalies in the normal subspace, and the drawbacks of Q-statistic, we apply the APD scheme to detect anomalies.

The link measurements $\mathbf{y}$'s projection onto normal subspace in Eqn.(4) can be written as:

$$\mathbf{SPE} \equiv \|\mathbf{y}_{no}\|^2 = \|\mathbf{PP}^T\mathbf{y}\|^2,\ \mathbf{P} = [\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3}, ..., \mathbf{v_k}] \quad (5)$$

In APD [5], we use $\mu \pm s * \sigma$ as a threshold to define the health norm of network measurements, where μ denotes the mean of measurement samples, σ corresponds to the standard deviation of the measurements samples, and s specifies the magnitude of deviation. Combined with APD scheme, we may consider the network measurements to be normal if,

$$\mu - s * \sigma \leq \|\mathbf{PP}^T\mathbf{y}\|^2 \leq \mu + s * \sigma \quad (6)$$

Now if we combine Eqn.(6) with Q-statistic, we formalize conditions for correlated and uncorrelated anomalies. We say correlated anomalies have occurred in the network measurements if,

$$\begin{cases} \|\mathbf{PP}^T\mathbf{y}\|^2 > \mu + s * \sigma \quad \text{and} \quad \|\mathbf{PP}^T\mathbf{y}\|^2 > \delta^2_\alpha \\ \delta^2_\alpha < \|\mathbf{PP}^T\mathbf{y}\|^2 < \mu - s * \sigma \end{cases} \quad (7)$$

And similarly, we conclude that uncorrelated anomalies have occurred in the network measurements if,

$$\begin{cases} \|\mathbf{PP}^T\mathbf{y}\|^2 < \mu + s * \sigma \quad \text{and} \quad \|\mathbf{PP}^T\mathbf{y}\|^2 < \delta^2_\alpha \\ \delta^2_\alpha > \|\mathbf{PP}^T\mathbf{y}\|^2 > \mu - s * \sigma \end{cases} \quad (8)$$

With the correlated and uncorrelated anomaly detection conditions formalized, we next analyze the accuracy of our proposed anomaly detection scheme.

C. Detection Accuracy Analysis

We now illustrate how the APD scheme can accurately detect anomalies in the output of PCA with a low number of false alarms. For this, we use the synthetic trace data described earlier in Section IV-B which closely resembles actual perfSONAR traces from DOE sites (Discussed in Section V). We show the advantage of using APD in terms of detecting correlated anomalies in the dataset. Alternately, for plateau detectors using static thresholds, the detection of uncorrelated anomalies of smaller magnitudes would result in a large number of false alarms. To demonstrate these facts, we plot the anomaly detection performance of the APD, SPD, and Q-statistic schemes with increasing number of principal components used for projection of the data. Recall that the Q-statistic is a statistic test to detect threshold-crossing samples. To adapt the Q-statistic into a plateau detector, we look for 7 (same as the trigger count in APD and SPD schemes) consecutive threshold-crossing to classify it as a plateau event. Fig. 6 shows the detection performance of the three schemes for correlated anomaly event cases. APD, SPD and Q-statistic detect all the 5 correlated anomalies magnified by PCA, however the SPD scheme has a false alarm in the low normal space, because of its static sensitivity parameter setting limitations.

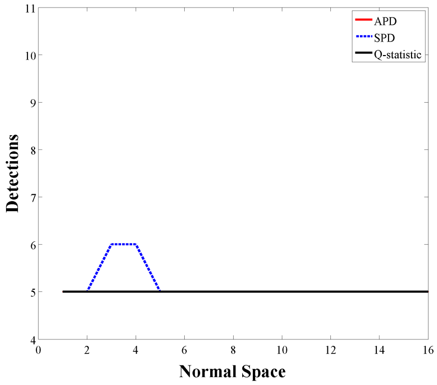

Fig. 6: Detection of the 5 correlated anomalies by APD, SPD and Q-statistic schemes on PCA output

Fig. 7 shows the detection performance for the three schemes as the number of principal components in the data projection increases for uncorrelated anomalies. As we increase the principal components, we find more correlated anomalies with our APD scheme. From this result, we can find that the Q-statistic scheme completely misses all the uncorrelated anomalies since the magnitude of the uncorrelated anomalies is much smaller compared to the correlated anomalies, which shifts the static threshold for detection to a higher value. With increasing number of principal components used for projection of the data, the SPD scheme detects all of the uncorrelated anomalies, but produces 4 false positive alarms. This is because it does not account for the change in the variance in the data. However, the APD scheme adapts to both the changes in mean and variance, and also correctly detects all of the uncorrelated anomalies with only 1 false negative alarm in this case.

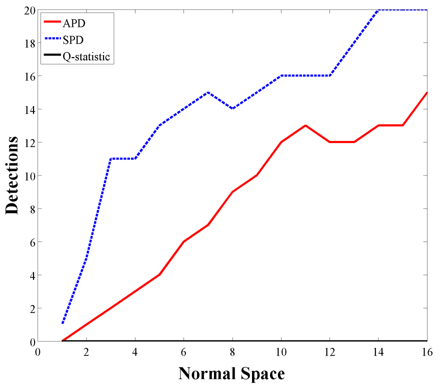

Fig. 7: Detection of the 15 uncorrelated anomalies by APD, SPD and Q-statistic schemes on PCA output

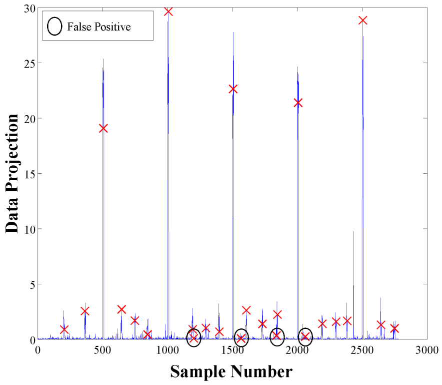

Fig. 8: Detection accuracy of correlated anomalies and uncorrelated anomalies by SPD on PCA output

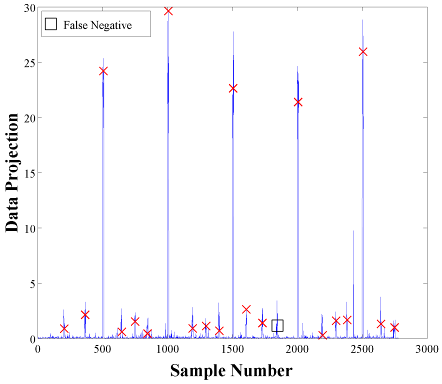

Fig. 9: Detection accuracy of correlated anomalies and uncorrelated anomalies by APD on PCA output

Figs. 8 and 9 show the number of detections of the SPD and APD schemes, respectively for the data projections using all the principal components (i.e., all anomalies in the dataset). As shown in Fig. 9, the APD scheme detects all the 5 correlated anomalies and detects 15 uncorrelated anomalies correctly with 1 false negative alarm. In contrast, as shown in Fig. 8, the SPD scheme detects all the 5 correlated anomalies and detects 16 uncorrelated anomalies with 4 false positive alarms, thus performs relatively poorly. Hence, from the experiments and analysis above, we can conclude that when APD scheme is applied to the data projections of principal components, it shows highly accurate anomaly detection in both correlated anomaly and uncorrelated anomaly cases with low false alarms. However, when SPD scheme is applied to the data projections of principal components, it similarly shows its ability in detecting correlated anomalies and uncorrelated anomalies but with a relatively higher false alarm rate. Lastly, when the Q-statistic scheme is applied, it shows high accuracy for detection in correlated anomalies, but completely misses all the uncorrelated anomalies.

V. Case Study: Source-side diagnosis

In this section, we validate the use of our proposed PCA-APD scheme to analyze correlated anomaly events at the network-wide level using source-side information within actual perfSONAR traces. Single day and month long traces are collected (as shown in tables I and II) to validate the effectiveness of the proposed scheme for both short-term and long-term network behavior. The datasets in this case study consist of plateau anomalies such as persistent increase and other anomaly events such as intermittent bursts and dips. We consciously ignore intermittent burst and dip events because these types of anomalies are generally caused by user application behavior, and are not of interest to network operators for routine monitoring and bottleneck troubleshooting. All of the actual perfSONAR traces correspond to one-way delay measurements collected between US Department of Energy (DOE) lab sites such as FNAL (Fermi National Accelerator Laboratory), and ANL (Argonne National Laboratory).

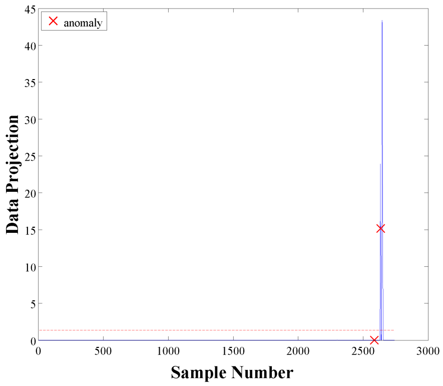

Fig. 10: Short-term traces: Detection of anomalies by APD on first principal component data projection

As discussed in Section IV, we know PCA-APD scheme can accurately detect synthetic correlated and uncorrelated anomalies with low false alarm rates, and the Q-statistic scheme specifically helps detect correlated anomalies fully accurately. To show evidence in practice, we used actual perfSONAR traces collected over a one-day period shown in Table I. Fig. 10 shows the two anomalies that are detected in this set of traces by the PCA-APD scheme using the first principal component. If we assume all correlated anomalies are captured in the first principal component, and the uncorrelated anomalies are captured in the rest of principal components, we may misidentify correlated anomaly events in certain situations. Consequently,

TABLE I: Short-term perfSONAR traces description

Trace ID	Source ↔ Destination	Time Range (Start - End)	Time Series Characteristics
1	atla-owamp.es.net ↔ anl-owamp.es.net	2014-03-04 00:00:22 - 2014-03-04 23:58:32	Intermittent Bursts, Intermittent Dips
2	atla-owamp.es.net ↔ bnl-owamp.es.net	2014-03-04 00:00:25 - 2014-03-04 23:58:36	Intermittent Bursts, Intermittent Dips
3	atla-owamp.es.net ↔ bois-owamp.es.net	2014-03-04 00:00:11 - 2014-03-04 23:58:36	Intermittent Bursts, Intermittent Dips
4	atla-owamp.es.net ↔ denv-owamp.es.net	2014-03-04 00:00:54 - 2014-03-04 23:58:44	Intermittent Bursts, Intermittent Dips
5	atla-owamp.es.net ↔ elpa-owamp.es.net	2014-03-04 00:00:07 - 2014-03-04 23:58:05	Intermittent Bursts, Intermittent Dips
6	atla-owamp.es.net ↔ fnal-owamp.es.net	2014-03-04 00:00:02 - 2014-03-04 23:58:10	Intermittent Bursts, Intermittent Dips
7	atla-owamp.es.net ↔ hous-owamp.es.net	2014-03-04 00:00:16 - 2014-03-04 23:58:42	Intermittent Bursts, Intermittent Dips
8	atla-owamp.es.net ↔ kans-owamp.es.net	2014-03-04 00:00:05 - 2014-03-04 23:58:44	Intermittent Bursts, Intermittent Dips
9	atla-owamp.es.net ↔ llnl-owamp.es.net	2014-03-04 00:00:43 - 2014-03-04 23:58:54	Intermittent Bursts, Intermittent Dips, Persistent Increase
10	atla-owamp.es.net ↔ nash-owamp.es.net	2014-03-04 00:00:38 - 2014-03-04 23:58:31	Intermittent Bursts, Intermittent Dips
11	atla-owamp.es.net ↔ nersc-owamp.es.net	2014-03-04 00:00:25 - 2014-03-04 23:58:47	Intermittent Bursts, Intermittent Dips, Persistent Increase
12	atla-owamp.es.net ↔ newy-owamp.es.net	2014-03-04 00:00:13 - 2014-03-04 23:58:50	Intermittent Bursts, Intermittent Dips
13	atla-owamp.es.net ↔ sdsc-owamp.es.net	2014-03-04 00:00:35 - 2014-03-04 23:58:42	Intermittent Bursts, Intermittent Dips, Persistent Increase
14	atla-owamp.es.net ↔ slac-owamp.es.net	2014-03-04 00:00:40 - 2014-03-04 23:58:48	Intermittent Bursts, Intermittent Dips
15	atla-owamp.es.net ↔ snll-owamp.es.net	2014-03-04 00:00:07 - 2014-03-04 23:58:40	Intermittent Bursts, Intermittent Dips, Persistent Increase
16	atla-owamp.es.net ↔ star-owamp.es.net	2014-03-04 00:00:07 - 2014-03-04 23:58:40	Intermittent Bursts, Intermittent Dips
17	atla-owamp.es.net ↔ sunn-owamp.es.net	2014-03-04 00:01:09 - 2014-03-04 23:58:31	Intermittent Bursts, Intermittent Dips, Persistent Increase
18	atla-owamp.es.net ↔ wash-owamp.es.net	2014-03-04 00:00:15 - 2014-03-04 23:59:01	Intermittent Bursts, Intermittent Dips

TABLE II: Long-term perfSONAR traces description

Trace ID	Source ↔ Destination	Time Range (Start - End)	Time Series Characteristics
1	fnal-owamp.es.net ↔ bois-owamp.es.net	2014-10-01 00:00:31 - 2014-10-30 23:59:47	Intermittent Bursts, Intermittent Dips, Persistent Increase
2	fnal-owamp.es.net ↔ hous-owamp.es.net	2014-10-01 00:00:04 - 2014-10-30 23:59:58	Intermittent Bursts, Intermittent Dips, Persistent Increase
3	fnal-owamp.es.net ↔ nersc-owamp.es.net	2014-10-01 00:00:25 - 2014-10-30 23:59:00	Intermittent Bursts, Intermittent Dips, Persistent Increase
4	fnal-owamp.es.net ↔ pppl-owamp.es.net	2014-10-01 00:00:01 - 2014-10-30 23:59:40	Intermittent Bursts, Intermittent Dips, Persistent Increase
5	fnal-owamp.es.net ↔ sacr-owamp.es.net	2014-10-01 00:00:01 - 2014-10-30 23:59:31	Intermittent Bursts, Intermittent Dips, Persistent Increase
6	fnal-owamp.es.net ↔ sdsc-owamp.es.net	2014-10-01 00:00:21 - 2014-10-30 23:59:37	Intermittent Bursts, Intermittent Dips, Persistent Increase
7	fnal-owamp.es.net ↔ slac-owamp.es.net	2014-10-01 00:00:57 - 2014-10-30 23:59:51	Intermittent Bursts, Intermittent Dips, Persistent Increase
8	fnal-owamp.es.net ↔ snll-owamp.es.net	2014-10-01 00:00:12 - 2014-10-30 23:59:05	Intermittent Bursts, Intermittent Dips, Persistent Increase
9	fnal-owamp.es.net ↔ sunn-owamp.es.net	2014-10-01 00:00:05 - 2014-10-30 23:59:34	Intermittent Bursts, Intermittent Dips, Persistent Increase
10	fnal-owamp.es.net ↔ wash-owamp.es.net	2014-10-01 00:00:08 - 2014-10-30 23:59:49	Intermittent Bursts, Intermittent Dips, Persistent Increase

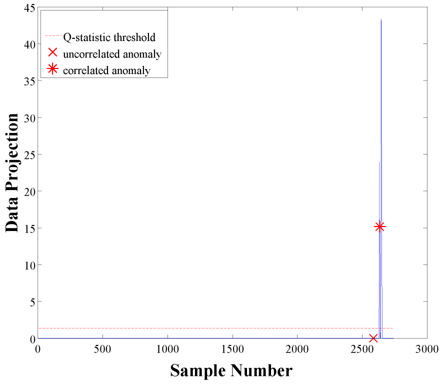

Fig. 11: Short-term traces: Detection of anomalies by APD on first principal component data projection with the help of Q-statistic

as shown in Fig. 11, our scheme identifies an anomaly above the Q-statistic as a correlated anomaly event, and identifies an anomaly below the Q-statistic as an uncorrelated anomaly event for improved detection accuracy. Based on the analysis above, we can judge a correlated anomaly occurred in a local domain (i.e., within ATLA) at 22:29:11- 22:38:34 time period. In order to validate this detection, we checked each of the traces using just the APD scheme to detect anomaly events in each trace. We found six traces to have the anomaly events within the same time period window.

We also applied our PCA-APD scheme to another set of month-long perfSONAR traces shown in Table II. Fig. 12 shows three detected anomalies using the first principal component and without Q-statistic. Upon using Q-statistic, our scheme successfully distinguished between correlated and uncorrelated anomalies as shown in Fig. 13. From the long-term measurement analysis, we determine that during 2014-10-29 06:10:56 - 2014-10-29 07:11:10 time period, a correlated anomaly occurred in the local domain (i.e., within FNAL).

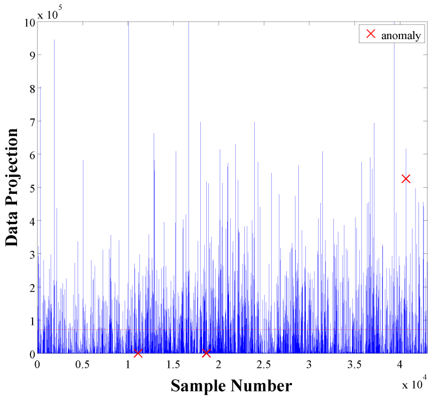

Fig. 12: Long-term traces: Detection of anomalies by APD on first principal component data projection

We next validate the accuracy of our 'black box' PCA-APD scheme in identifying correlated anomalies in the perfSONAR traces by leveraging a related ESnet topology map. Fig. 14 shows the ESnet topology map related to both our short-term and long-term measurement trace scenarios that gave us the unique opportunity to validate the findings of the PCA-APD scheme by looking for anomaly events in the network having similar time-signatures. For the short-term measurements, we found that the anomalies mostly occurred on the path

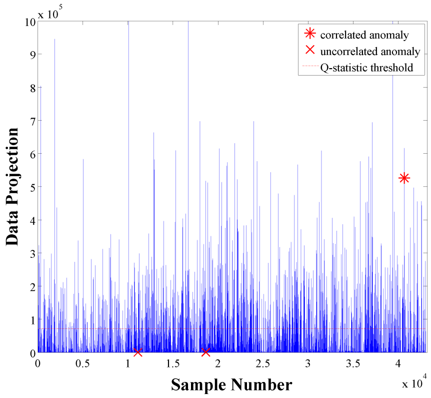

Fig. 13: Long-term traces: Detection of correlated anomalies by APD on first principal component data projection with Q-statistic

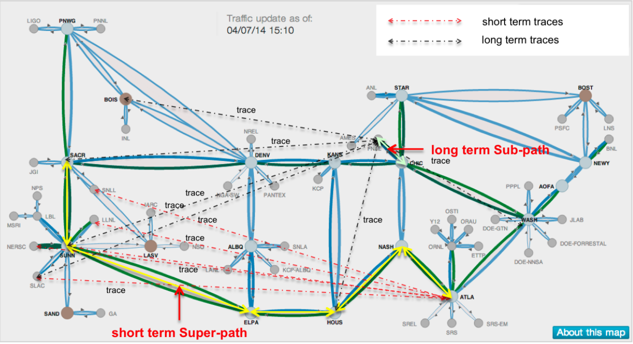

Fig. 14: Traces shown in the DOE lab sites topology map for source-side correlation analysis

SACR↔ATLA (as shown with text annotation and a yellow line in Fig. 14). Given that these are correlated anomalies according to the PCA-APD scheme, we guess that the other sites along this path are also experiencing anomaly events. Consequently, we collected other sites' data that constitute 'sub-paths' to the SACR↔ATLA 'super-path' corresponding to the same day and checked for anomalies with our PCA-APD scheme. We observed that all the anomalies occurred at the same time at about the time period of 22:29:28 - 22:37:41. We can judge from the above analysis that the root-cause of the correlated anomaly event exists in the path SACR↔ATLA.

A similar validation investigation using the ESnet topology map on long-term measurement traces revealed that the anomaly events mostly occurred on the path FNAL↔CHIC (as shown with text annotation and a cyan line in Fig. 14). However, in this case we searched for other 'super-paths' having FNAL↔CHIC as the common intermediate hop, i.e., 'sub-path' and checked for anomaly events with our PCA-APD scheme. We observed all such paths with anomaly events occurred within the same time window (2014-10-29 06:10:56-2014-10-29 07:11:10). In this case the evidence is seen on the path FNAL↔CHIC that is the root cause of correlated anomaly event. Such assertions of root-cause location cannot be made with PCA alone without clear network topology and measurement data. Although in the proposed scheme, we do not consider the availability of complete topology information, we can analyze the timestamp information which may provide indications as to when the anomaly event first occurred, and on which particular path. Thus, the PCA-APD scheme can effectively work in 'black box' scenarios, and its output information is helpful to determine the root-cause location of correlated network anomaly events that may impact data transfer performance of data-intensive applications.

VI. Conclusion

In this paper, we presented a novel PCA-based network-wide correlated anomaly detection scheme that: (i) uses principal component analysis to capture the maximum variance in a given multiple path measurement time series, (ii) applies adaptive plateau detector (APD) to detect anomaly events with fused data transformation by PCA, and (iii) leverages Q-statistic event correlation analysis in order to accurately filter out correlated and uncorrelated anomalies.

With the strength of our prior work in developing APD's accurate uncorrelated anomaly detection algorithm, our proposed PCA-APD scheme in this paper has the unique ability to detect both correlated and uncorrelated anomalies with high accuracy and low false alarms, in a timely manner. With event correlation analysis, our scheme is suitable for source-side anomaly localization to help network operators to diagnose the root-cause of bottlenecks, even when network topology information is not completely available.

We validated our scheme with both synthetic trace data and actual perfSONAR trace data collected from DOE Lab sites, and present case studies that validates the utility of our PCA-APD scheme. Our scheme's outcome can help a network operator to isolate and diagnose the root-cause of a correlated network-wide anomaly event as occurring within the local domain, or in an external domain.

References

[1] A. Hanemann, J. Boote, E. Boyd, J. Durand, et. al., "PerfSONAR: A Service Oriented Architecture for Multi-Domain Network Monitoring", *Proc. of Service Oriented Computing*, 2005. (http://www.perfsonar.net)

[2] perfSONAR Deployment Milestone - http://es.net/news-and-publications/esnet-news/2014/perfsonar-milestone

[3] J. Zurawski, M. Swany, D. Gunter, "Scalable Framework for Representation and Exchange of Network Measurements", *Proc. of IEEE TRIDENTCOM*, 2006.

[4] A. Hanemann, V. Jeliazkov, O. Kvittem, L. Marta, J. Metzger, I. Velimirovic, "Complementary Visualization of perfSONAR Network Performance Measurements", *Proc. of IEEE International Conference on Internet Surveillance and Protection*, 2006.

[5] P. Calyam, J. Pu, W. Mandrawa, A. Krishnamurthy, OnTimeDetect: Dynamic Network Anomaly Notification in perfSONAR Deployments, *Proc. of IEEE/ACM MASCOTS*, 2010.

[6] P. Calyam, M. Dhanapalan, M. Sridharan, A. Krishnamurthy, R. Ramnath, "Topology-Aware Correlated Network Anomaly Event Detection and Diagnosis", *Springer Journal of Network and Systems Management (JNSM)*, 2013.

[7] A. McGregor, H-W. Braoun, "Automated Event Detection for Active Measurement Systems", *Proc. of Passive and Active Measurement Workshop*, 2001.

[8] C. Logg, L. Cottrell, "Experiences in Traceroute and Available Bandwidth Change Analysis", *Proc. of ACM SIGCOMM Network Troubleshooting Workshop*, 2004.

[9] J. Jackson, G. Mudholkar, "Control Procedures for Residuals Associated with Principal Component Analysis", *Technometrics*, 1979.

[10] A. Lakhina, M. Crovella, C. Diot, "Diagnosing Network-Wide Traffic Anomalies", *Proc. of ACM SIGCOMM*, 2004.

[11] A. Soule, K. Salamtian, N. Taft, "Combining Filtering and Statistical Methods for Anomaly Detection", *Proc. of ACM IMC*, 2005.

[12] A. Mahimkar, J. Yates, Y. Zhang, A. Shaikh, J. Wang, Z. Ge, C. Ee., "Troubleshooting Chronic Conditions in large IP Networks", *Proc. of ACM CoNEXT*, 2008.

[13] L. Zonglin, H. Guangmin, Y. Xingmiao, Y. Dan, "Detecting Distributed Network Traffic Anomaly with Network-Wide Correlation Analysis", *Proc. of EURASIP J. Adv. Signal Process*, 2009.

[14] Y. Zhou, G. Hu, "Network-wide Anomaly Detection Based on Router Connection Relationships", *Proc. of IEICE Transactions*, 2011.

[15] P. Kanuparthy, D. Lee, W. Matthews, C. Dovrolis, S. Zarifzadeh, "Pythia: Detection, Localization, and Diagnosis of Performance Problems", *IEEE Communications Magazine*, 2013.

[16] M. Marvasti, A. Poghosyan, A. Harutyunyan, N. Grigoryan, "An Enterprise Dynamic Thresholding System", *Proc. of International Conference on Autonomic Computing*, 2014.