Distributed Spatial Anomaly Detection

PARMINDER CHHABRA[†] pchhabra@cs.bu.edu
CLAYTON SCOTT[§] cscott@eecs.umich.edu
ERIC D. KOLACZYK[‡] kolaczyk@math.bu.edu
MARK CROVELLA[†] crovella@cs.bu.edu

***Abstract*—Detection of traffic anomalies is an important problem that has been the focus of considerable research. Recent work has shown the utility of spatial detection of anomalies via cross-link traffic comparisons. In this paper we identify three advances that are needed to make such methods more useful and practical for network operators. First, anomaly detection methods should avoid global communication and centralized decision making. Second, nonparametric anomaly detection methods are needed to augment current parametric approaches. And finally, such methods should not just identify possible anomalies, but should also annotate each detection with some probabilistic qualifier of its importance.**

We propose a framework that simultaneously advances the current state of the art on all three fronts. We show that routers can effectively identify volume anomalies through cross-link comparison of traffic observed only on the router's own links. Second, we show that generalized quantile estimators are an effective way to identify high-dimensional sets of local traffic patterns that are potentially anomalous; such methods can be either parametric or nonparametric, and we evaluate both. Third, through the use of false discovery rate as a detection metric, we show that candidate anomalous patterns can be equipped with an estimate of a probability that they truly are anomalous. Overall, our framework provides network operators with an anomaly detection methodology that is distributed, effective, and easily interpretable. Part of the underlying statistical framework, which merges aspects of nonparametric set estimation and multiple hypothesis testing, is novel in itself, although the derivation of that framework is necessarily given elsewhere.

I. INTRODUCTION

Detecting unusual traffic patterns is a critical problem for network operators. An important specific instance is the problem of detecting *volume anomalies.* A volume anomaly is an unusually large or small volume of traffic occurring within some predefined time period. Such anomalies can be evidence of a wide range of operational problems, including malicious behavior (denial of service attacks, worm spreads), misconfiguration or equipment failure (traffic shifts), or changes in customer behavior.

Unfortunately, network traffic exhibits high variability in many different dimensions, and the large amount of noise in the system makes precise detection of volume anomalies a challenging problem. Approaches based on timeseries analysis, while well-studied, (e.g., [1], [3]) can be difficult to tune and use in many networks.

In contrast to timeseries methods, a more recent approach showing promise for volume anomaly detection takes a network-wide approach [10]. In the network-wide method, links with anomalous traffic volumes are not identified by comparing with past values, but rather by comparing with other values in the network at the same time. That is, network-wide traffic analysis makes use of typical relationships between traffic volumes on different links, and identifies points in time when individual links or sets of links violate those typical patterns.

However, as proposed to date, network-wide volume anomaly detection relies on comparing traffic on all or most network links simultaneously. This implies moving traffic measures to a central location for analysis. Unfortunately, such a centralized scheme suffers from at least two key problems. First, as networks increase in size, moving data to a central location is an expensive proposition and does not scale well. Second, a centralized scheme has a single point of failure making it susceptible to attack.

Thus, the first and primary goal of this paper is to develop a distributed scheme with an ability to identify anomalies that is on par with the global, network-wide approach. We assume that distributed detection will need to be implemented in routers (or devices attached to routers), and we note that a router with N links to other routers has $2N$ traffic measures that it can obtain by itself, without exchanging any information with other routers. The distributed method we introduce here then consists of two distinct stages, involving extremely low communication among routers. In the first stage, each router identifies a set of *candidate anomalies* by comparing traffic measures on only the small subset of $2N$ links local to it. More precisely, each router identifies outliers in the $2N$-dimensional space visible to it by exploiting local traffic correlations. Then, in the second stage, a set of *consensus anomalies* are produced through a simple adjacent-neighbor filtering algorithm that acts to submit locally declared candidate anomalies to a degree of verification. In this step, each router communicates its outlier information with its neighbors, and outliers are finally classified as anomalies by consensus if two connected routers both report an outlier for the same time bin. Note that this distributed scheme is no longer network-wide, but is still spatial rather than temporal. We therefore term it *spatial* volume anomaly detection.

To succeed in our proposed approach, we need to adopt a first-stage detection framework that is well-suited to the kind of variation that network traffic typically exhibits. *Generalized quantile sets* (GQSs) are natural to use here. Formally, given a probability distribution P, we define a β-level generalized quantile set to be a set of minimum Euclidean volume with probability mass at least β. In other words, GQSs encapsulate regions of space where the mass of P is most concentrated.

[†] Computer Science Deptartment, Boston University, Boston, MA, USA.
[‡] Department of Math and Statistics, Boston University, Boston, MA, USA.
[§] Department of EECS, University of Michigan, Ann Arbor, MI, USA.

As such, outliers are naturally characterized in terms of GQSs. The larger β is allowed to grow before a measurement falls in the corresponding β-level GQSs, the more extreme or 'outlying' that measurement may be considered. Note that in the case where P is a multivariate Gaussian distribution, the GQSs are just ellipsoids.

Unfortunately, while this approach is natural, applying it to network traffic can be difficult because for such data, the modeling and estimation of the underlying distribution P is quite a challenge. On the one hand, with sufficiently strong modeling assumptions, a parametric framework may be adopted for this purpose. For example, extensive studies of network traffic have shown that in some settings traffic can be modeled well as Gaussian. These settings are those in which traffic is highly aggregated, either in terms of number of concurrent flows or amount of time per sample [7]. However, the convergence of traffic volumes to Gaussian may not occur in practice if aggregation is insufficient and the traffic sources being aggregated show high variability in terms of their sending rates. Such high variability is not uncommon in network traffic [13], [16], and so in many cases the Gaussian model is not appropriate. The alternative is to adopt a nonparametric framework i.e., one in which minimal assumptions are made on the nature of the underlying distribution P. But generally larger amounts of data are required for estimation in nonparametric contexts.

As a result of these issues, the second goal of this paper is to explore the relative utility of *both* parametric and nonparametric GQS approaches for volume anomaly detection. In the parametric case, we assume the underlying distribution of the typical data (here, the traffic vector seen at a router) to be multivariate Gaussian. In the non-parametric case, we essentially infer the distribution from the data itself.

The third and final goal of this paper is to go beyond simple generalized-quantile-based *detection* and attempt to *annotate* the resulting outliers with a measure of significance. The reason that this is needed is that we are applying the generalized quantile test repeatedly over time. If we use a fixed mass constraint (β) we get a constant number of outliers (potential anomalies) at each router — a constant alarm rate. Thus, while one router may experience a large number of true volume anomalies, another router may generate the same number of detections while only experiencing a small number of true volume anomalies. It is therefore helpful to have a measure of how *severe* an anomaly is for comparison across routers. Another way to state the issue is that we need to vary the alarm rate threshold at each router since each can potentially include a differing number of false alarms.

Technically speaking, what we identify here within the volume anomaly detection problem is a version of the well-known *multiple testing* problem in statistics. This problem refers to the context where one subjects a number of separate tests to the same acceptance/rejection criterion that would be used when considering only a single test. As the number of separate applications of the test grows, it begins to outweigh the unlikeliness associated with any individual test declaring a detection based on nominal (i.e., non-anomalous) data. These errors are considered false positives or false discoveries. In principle, the number of such false discoveries can be controlled simply by expanding the acceptance region of the test (i.e., making the test more conservative). But such a gain comes at the cost of a corresponding decrease in ability to detect true anomalies. Alternatively, over the past decade methods have been developed for controlling the *rate* of false discoveries – rather than the number – which have proven to balance the detection of true and false positives much more successfully [2] [15]. We use so-called *false discovery rate (FDR)* methods here, in conjunction with our GQS-based detection strategy, to annotate our detections[1] .

Our overall contribution in this paper is a distributed, two-stage strategy for volume anomaly detection, that first utilizes a combination of generalized quantile sets and false discovery rate methods to identify a prioritized set of candidate anomalies. Then, it outputs a final set of anomalies verified by consensus through a simple exchange of candidates among neighbors. To assess the accuracy of this strategy, we apply it to traffic measurements from the Abilene network. Since we are interested in whether distributed spatial methods can perform as well as centralized whole-network methods, we compare our detections against those obtained using the subspace method as described in [10]. Comparing our distributed approach against the subspace approach gives us some indication of how well our approach does in the absence of any global information.

Our results show that, in general, distributed methods can perform nearly as well as the centralized method, despite the lack of global knowledge in the distributed case. We find that nearly all anomalies detected by the subspace method are also detected by our distributed spatial methods with low FDRs.

In comparing parametric and non-parametric approaches, we find that overall the non-parametric approach does nearly as well as the parametric approach with Gaussian-like data, and it does better where the data appear to be distinctly non-Gaussian. This is precisely what one would expect with our data, since these datasets are from a highly aggregated backbone-type network, and hence are usually (but not always) well modeled as multivariate Gaussian.

Finally, we show the utility of the FDR method in terms of identifying more precisely a set of the most significant anomalies in our data, and annotating those anomalies with a measure of their importance.

II. Background

Several techniques have been proposed to detect anomalies in network traffic. They can be largely classified into spatial and temporal techniques.

Early anomaly detection techniques used primarily temporal methods [3]–[5] Temporal techniques exploit patterns in a timeseries to expose network anomalies. Techniques proposed in [3], [4] use deviation from normal behavior in

[1] The joining of FDR methods and nonparametric GQS estimation is in fact new, and was developed by the authors in conjunction with the material in this paper, but is necessarily presented elsewhere [14].

a timeseries to identify network faults. Deviation from the norm is combined with the probabilistic framework of a Bayesian network to detect faults by [5]. In [1], the authors perform a signal analysis of network data that exploits its time frequency characteristics. In [9], a sketch-based change detection technique is used to detect significant changes in massive data streams. Time-series forecast models (ARIMA, Holt-Winters etc.) for anomaly detection detect significant changes by looking for flows with large forecast errors. In [18], the authors use temporal techniques (EWMA, Fourier / Wavelets) to first detect candidate anomalies from link traffic measurements. While each of these temporal methods has been shown to be effective in certain settings, the temporal approach requires careful timeseries modeling, which can be difficult to tune in practice due to the bursty and long-range dependent nature of network traffic.

Spatial methods [10] exploit correlation among links in the network to define normal traffic behavior. Unlike temporal techniques, these techniques do not require delicate parameter tuning to capture normal traffic behavior in data. However, these methods are inherently centralized, leading to communication burden across the network. The authors in [6] reduce the data movement needed in network-wide anomaly detection through judicious use of network-wide communication. Each network monitor continuously tracks principal components to within an error tolerance. A monitor sends its recent measurements to the central entity *only if* the local error tolerance is violated. However, that approach reduces but does not eliminate network-wide communication. The need to completely eliminate such communication is a prime motivation of our work; to our knowledge the current paper represents the only fully-distributed volume-based spatial approach to traffic anomaly detection.

Distributed signature generation techniques have also been proposed for anomaly detection. In [8], each anomaly detection monitor first identifies suspicious source IP addresses. Next, each monitor shares the source IP address among all monitors. The authors assume the availability of a multicast facility to all monitors. The authors in [17], define a framework within which systems from different administrative domains can participate in coordinated intrusion detection. Every overlay axis node maintains a global and a local view of intrusion and attack activity. Axis nodes receive summaries from their peers which are then used to create a view of global activity. Our notion of a distributed approach differs from the above in that only neighboring routers communicate. This allows for very low communication overhead.

III. Properties of Traffic Data

In this section we describe the data used, and point out properties of the data that motivate our approach.

As described in Section I, our approach seeks to minimize communication between routers. This implies that routers primarily make use of traffic measures on their own links. Routers are assumed to measure the volume of traffic passing over each link in fixed time intervals. Our methods are equally applicable to any traffic measures for which volume anomalies are important, such as bytes, packets or flows; in our examples we focus on byte traffic. A router having $2N$ unidirectional links thus obtains a timeseries of vectors $\{\mathbf{x}_i, i = 1, \ldots, T\}$ with $\mathbf{x}_i \in \mathbb{R}^{2N}$.

A. Data Used

To illustrate our approach we use data collected from Abilene, the Internet2 backbone network. The Abilene backbone carries traffic from universities in the US and is hence non-commercial. This network has a total of 11 PoPs across continental USA.

We use a total of four weeks of data, one week (Week I) from Apr 2003 (Apr 7-13) and three weeks (Week II-IV) from Dec 2003 (Dec 8-28), and process one week of data at a time. We refer to these data sets as Weeks I-IV. Traffic volume (in bytes) from the Abilene network is binned in 10 minute intervals. This gives a total of 1008 timepoints per week. For comparison, in [11] the authors study backscatter data over several weeks and find that typically, about 50% of the attacks are less than 10 minutes in duration and about 80% are less than 30 minutes. So, anomalies detected by our scheme are ones that are pronounced in traffic volume in a 10 minute interval, which account for a significant percentage of anomalies reported in [11].

Each router connects to between two and four other routers. So for our data, we have $\mathbf{x}_i \in \mathbb{R}^{2N}$ with $N = 2, 3,$ or 4. Hence, our measurements lie in a space of dimension between 4 and 8. To visualize such vectors we project to two dimensions in constructing the figures shown in this paper; but all of our methods are implemented in the full 4 to 8 dimensional space of actual measurements.

B. Correlation Properties

Spatial anomaly detection relies on the presence of correlations between traffic measurements on different links. Hence an important first question to ask is whether traffic observed at *an individual router's links* (as opposed to across all links in a network) exhibits correlations, and whether those correlations are strong enough to usefully inform the outlier detection process.

To answer this question we show examples of traffic measurements taken from the Abilene network. Figure 1 shows scatterplots of traffic on two links at each of the Atlanta, Chicago and Indianapolis routers; Figure 2 shows scatterplots of traffic on two links at each of the Chicago, Houston, and Sunnyvale routers.

The figures show that there are strong correlations in the traffic that effectively constrain the space of "normal" traffic and provide useful boundaries for outlier detection. For example, in Figure 1 large traffic volumes on one link correspond to large traffic volumes on the other link — so an outlier detection method that takes into account both values simultaneously will be more sensitive than one that examines each value separately. Likewise, in Figure 2 certain combinations of traffic values are

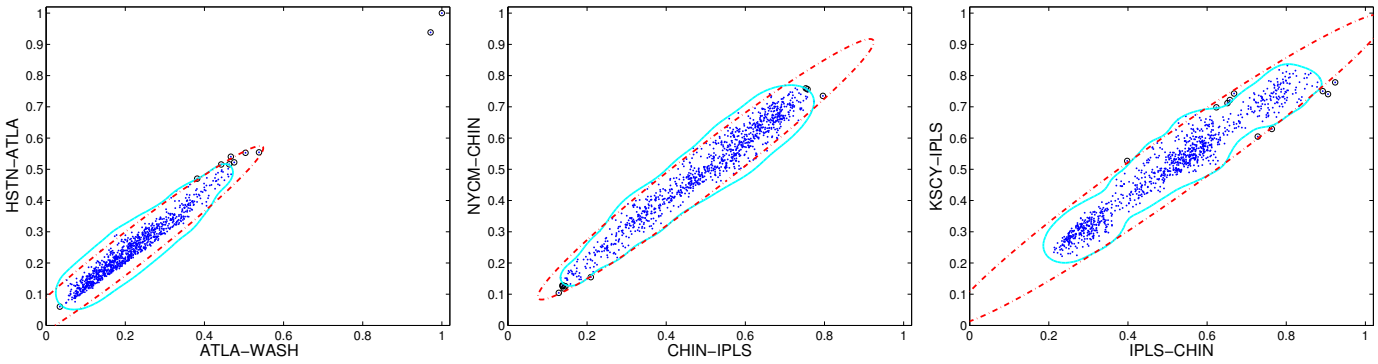


Fig. 1. Scatter plots of traffic volume at the links connecting to routers in Atlanta (Week III), Chicago (Week I) and Indianapolis (Week III).

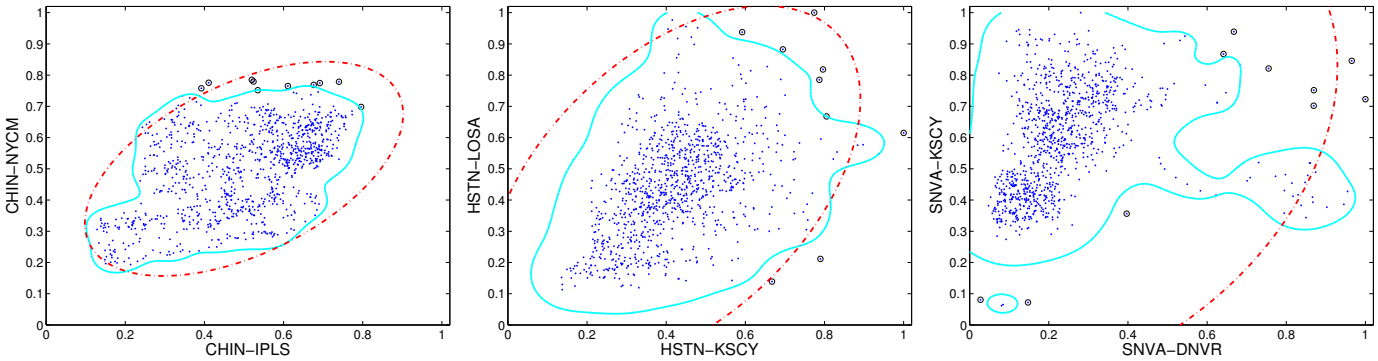


Fig. 2. Scatter plots of traffic volume at the links connecting to routers in Chicago (Week I), Houston (Week I) and Sunnyvale (Week II).

much more likely than other combinations, although in these cases the relationship is more complex.

Given that strong correlations exist between traffic measures on different links, the natural next question is: how should we describe the correlations in a manner useful for outlier identification? The most common approach in such problems is to use a parametric model to describe the data, and the most common parametric model used is the multivariate Gaussian. To illustrate the use of such a model for our data, we have plotted the ellipses corresponding to the $\beta = 0.99$ threshold of the maximum likelihood fit of a multivariate Gaussian to our data (details are in Section IV). If the data were well described as multivariate Gaussian, then the shape of the point cloud would roughly match the ellipse and 99% of the data points should lie inside the ellipse.

Visual inspection suggests that the multivariate Gaussian model is a good fit for data in Figure 1, but not for the data in Figure 2. Thus we conclude that while the multivariate Gaussian may be a good model for a router's link traffic in some cases, it is not so in all cases.

The irregular nature of the point clouds in Figure 2 suggest the need for a non-parametric description of the data useful for outlier detection. We present such a method in Section IV; here we simply show the results of applying the method, and again selecting a set containing 99% of the data. (Both the parametric and nonparametric methods used here are instances of GQS estimates). The nonparametric sets are shown in gray outline. In general, the nonparametric approach seems to yield a much better match to the distribution of the data shown in Figure 2.

To illustrate the outliers that would be identified by both methods, we show outliers for the nonparametric method with small circles, while outliers for the parametric method are simply the points that fall outside the ellipse. There are a number of observations that can be made. First, the two methods clearly identify different outliers, but the difference is much more pronounced when the data are irregularly distributed as in Figure 2. Second, when data are irregularly distributed, the multivariate Gaussian model is quite poor, establishing bounds that are far too wide, as seen in the case of the Sunnyvale router in Figure 2. Finally, the nonparametric method, although more general, does not necessarily do as well as the parametric method when data are approximately Gaussian. This can be seen in the case of plots in Figure 1, where the nonparametric method can select outliers than appear to be better described as normal data.

These observations suggest that it is important and worthwhile to consider both parametric and nonparametric approaches to outlier detection for this problem, which is how we proceed in the remainder of the paper.

IV. Detecting Anomalies

A. Generalized Quantile Sets and False Discovery Rate

In this section we describe the methodology whereby candidate anomalies are selected locally at each router. (The final declaration of whether one of these candidates is actually an

anomaly is made by the adjacent neighbor filtering described below.) Recall that at a router we observe traffic volumes $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ at T distinct time points, as described in section III.

We adopt the goals of (1) ranking the measurements $\mathbf{x}_1, \ldots, \mathbf{x}_T$ in order of their potential to be anomalous, and (2) assigning *annotations* $\gamma_1, \ldots, \gamma_T$, with $0 \leq \gamma_t \leq 1$, to each $\mathbf{x}_t \in D$. Our annotations respect the ranking established in the first goal, with larger annotations corresponding to more potentially anomalous timepoints. Furthermore, they provide an interpretable quantification of the degree to which each $\mathbf{x}_t$ appears to be anomalous. The annotations may be thresholded to obtain local decisions, or propagated to neighboring routers and aggregated in some way, as is done in the distributed detector described below.

To construct such an annotated ranking we adopt the framework developed by the second and third authors and developed in detail in [14]. To begin, we discuss the problem of ranking. Adopting a probabilistic framework, we view the typical (non-anomalous) local measurements $\mathbf{x}_t$ as independent realizations (ignoring temporal dependencies) from a common probability distribution P. A key assumption underlying our approach is that **anomalies are outliers** with respect to this distribution. Therefore, we rank the points $\mathbf{x}_t$ according to how extreme they are with respect to the nominal distribution P.

To make the notion of "extreme" precise in this multidimensional setting, the concept of generalized quantile sets [2] is adopted. Given $0 \leq \beta \leq 1$, the generalized quantile set at level β with respect to P is defined to be

$$G_\beta = \arg \min_{\text{all sets } G} \lambda(G) \quad \text{s.t. } P(G) \geq \beta.$$

Here λ denotes Euclidean volume (Lebesgue measure). The GQS is the smallest set (in terms of Euclidean volume) that encompasses at least $100\beta\%$ of the probability mass of P. If $f(\mathbf{x})$ denotes the density of P, then it can easily be seen that $G_\beta = \{\mathbf{x} : f(\mathbf{x}) \geq \eta\}$ for some $\eta > 0$, where η increases as β decreases. That is, GQSs are density level sets. Then, we rank $\mathbf{x}_t$ according to

$$\beta_t := \inf_{0 \leq \beta \leq 1} \{\beta : \mathbf{x}_t \in G_\beta\},$$

with larger values of β_t corresponding to more potentially anomalous volume patterns.

We now turn to the matter of assigning meaningful annotations. The values β_t are themselves an obvious, and indeed not unreasonable, candidate for such an annotation. Indeed, β_t is the probability that a nominal measurement is less extreme than $\mathbf{x}_t$. However, there is the need to interpret these values and, although the values β_t provide a meaningful interpretation of each $\mathbf{x}_t$ when viewed in isolation, they are not designed to be meaningfully interpreted *en masse*. This observation is a variation on the issue at the heart of the so-called 'multiple testing problem' in statistics. One version of the problem is that, even if all $\mathbf{x}_t$ are indeed nominal, for any β, we expect about $100(1-\beta)\%$ of the β_t to be greater than β. For example, if an isolated measurement $\mathbf{x}_t$ has $\beta_t > \beta = 0.95$, we would feel confident ascribing a large annotation to $\mathbf{x}_t$. However, if we gather $T = 1000$ measurements, then we expect about $1000 \times 0.05 = 50$ measurements to have $\beta_t > 0.95$ just by chance. Such an outcome is often unsatisfactory, particularly when nontrivial amounts of energy are expected to be used to follow up on discoveries, as is often the case in anomaly detection problems.

This problem has received a great deal of attention in the statistical literature over the past decade since the seminal paper of Benjamini and Hochberg [2], who developed an approach based on the *false discovery rate* (FDR). To describe this concept in our setting, let Q denote the distribution of the contaminated measurements. That is, Q is a mixture of P and the distribution on anomalies. Adopting Storey's framework [15], we may associate an FDR to $\mathbf{x}_t$ by

$$\text{FDR}(\mathbf{x}_t) := Q(\mathbf{x} \sim P \,|\, \mathbf{x} \notin G_{\beta_t}).$$

Thus, the FDR associated with $\mathbf{x}_t$ is the probability that, given that a "discovery" is made relative to $\mathbf{x}_t$ (i.e., a measurement is more extreme than $\mathbf{x}_t$), that discovery is in fact false. We therefore propose, as a more meaningful alternative to the values $\beta(\mathbf{x}_t)$, to annotate our ranked observations by the values

$$\gamma_t := 1 - \text{FDR}(\mathbf{x}_t). \tag{1}$$

These annotations, defined in terms of FDR, convey valuable information to a network operator, for example, who can afford to devote only limited time to the pursuit of false discoveries. Furthermore, it can be shown that these annotations preserve the ranking defined in terms of the β_t [14].

Since the annotations depend on the unknown probability distributions, these quantities must be estimated from the data $\mathbf{x}_1, \ldots, \mathbf{x}_T$, which are realizations of the contaminated mixture distribution Q. The details of this estimation procedure are given in [14], where the emphasis is on nonparametric setting but applies equally well to the parametric setting with minor modifications. A central ingredient in that estimation procedure is a method for GQS estimation. In particular, if $\tilde{f}$ denotes the density of the contaminated distribution Q, we must estimate the smallest level set of $\tilde{f}$ containing $\mathbf{x}_t$, say G_t, for each t.

One question addressed in the present work is whether the local Abilene data D is best modeled in a parametric or nonparametric fashion. To address this question, we compare outlier detectors based on parametric and nonparametric methods of estimating the level sets G_t. In both cases, we perform density estimation based on $\mathbf{x}_1, \ldots, \mathbf{x}_T$ to obtain an estimate of $\tilde{f}$. The set G_t is then estimated by thresholding the estimate of $\tilde{f}$ such that $\mathbf{x}_t$ is right on the boundary of the estimated level set.

[2] Generalized quantile sets are also known as minimum volume sets and minimum measure sets in the literature. While the term 'minimum volume sets' is perhaps most common, we prefer to avoid confusion with traffic volumes.

In the parametric approach, the data are assumed to be Gaussian, and the parameters of the Gaussian density, the mean and covariance, are estimated by maximum likelihood, yielding the sample mean and sample covariance of $\mathbf{x}_1, \ldots, \mathbf{x}_T$. A second parametric approach is also considered. Since the data D is by its very nature contaminated, it contains some outliers that might influence the maximum likelihood approach. Therefore, we also experimented with the method of robust Gaussian parameter estimation described in [12]. In the nonparametric approach, the density $\tilde{f}$ was estimated using a kernel density estimator having an isotropic Gaussian kernel. Selection of the kernel bandwidth is described in [14].

B. Adjacent Neighbor Filtering

As described so far, the distributed method involves no communication between routers. However we observe that the benefits of a distributed approach are maintained even if *strictly local* communication is employed. In particular, adjacent routers exchange traffic in normal operation, and so exchange of information for outlier filtering is reasonable.

Thus, having identified candidate anomalies locally using FDR-annotated generalized quantile sets, we use adjacent-router communication to add strength to the conclusion that an outlier is in fact an anomaly. The intuition behind this step is that when a true anomaly occurs on a link, the two routers on each end of the link should both report the timepoint as an outlier. Hence, only if two adjacent routers observe an outlier at the same timepoint do we classify it as an anomaly. We will see in Section V-A that this overall algorithm is effective in classifying outliers into anomalies and keeping the number of false positives low.

C. The Subspace Method: A benchmark

Our goal is to demonstrate that a distributed anomaly detection appraoch can perform as well as a centralized approach. A prominent example of the centralized approach is the subspace method, so we use the anomalies detected via the subspace method as our comparison case.

The subspace method uses Principal Component Analysis (PCA) to detect volume anomalies (which we will call p-anomalies) by starting with the traffic vector corresponding to measurements of all network links, and separating it into normal and anomalous components. The key idea in the subspace-based detection stage is that the subspace corresponding to maximal traffic variation can be identified with normal behavior; so normal and anomalous data components can be effectively separated by projecting traffic onto these two subspaces.

Residual traffic (the portion lying in the anomalous subspace) is thresholded using a $1 - \alpha$ confidence limit that corresponds to a false alarm rate [3] of α. The confidence limit is derived under the assumption that the traffic vector follows a multivariate Gaussian distribution. Space does not permit a detailed explanation of the subspace method; more can be found in [10].

[3] We note that in following [10], this use of false alarm rates does not correct for multiple testing, as we do in our proposed methods. To do so here in a comparable manner is beyond the scope of this paper.

V. Results

In this section, we assess the performance of our methods. To do so, we treat p-anomalies (those detected using PCA) as ground truth anomalies. This allows us to assess whether our distributed methods detect as well as the centralized (PCA) method. However, we note that our distributed methods may additionally detect true anomalies that were not detected by PCA. For the purposes of our evaluation, these will be treated as false alarms, even though they may not be. Thus the detection rates and false alarm rates we report are conservative with respect to true anomalies.

To represent ground truth we use the subspace method (PCA) at detection thresholds of $\alpha = 0.999$ and 0.995. Our distributed methods will use thresholds of $\gamma = 0.75$, 0.50, 0.25 and 0.10. (We will discuss the significance of these thresholds in Section V-C.) Events identified by PCA are called *p-anomalies,* and those identified by our distributed GQS approach, are called *q-anomalies.* A *timepoint* refers to a single time bin; an anomaly may span multiple timepoints. An event may occur on multiple links at the same timepoint or over consecutive timepoints on the same link or links. We call each such occurrence a *unique* event. In comparing the methods, we consider the distributed method to identify a unique PCA event if the q-anomaly includes some or all of the timepoint(s) of the p-anomaly. On the other hand, we consider the distributed method to have identified a PCA timepoint *only if* a p-anomaly and q-anomaly were reported at the same timepoint.

A. Evaluation of Proposed Methods

Table I compares the number of p-anomalies identified by the parametric and the non-parametric GQS methods for the four γ thresholds. The table shows that the distributed methods are effective at identifying p-anomalies. For PCA at $\alpha = 0.995$, the parametric method identified 34 and the non-parametric method identified 33 out of 38 p-anomalies. Similarly, for a PCA at $\alpha = 0.999$, both methods identified all the 19 p-anomalies.

Most p-anomalies are multi-timepoint events. An analysis of the data revealed that all of the unidentified p-anomalies were single timepoint events. Nonetheless, over 50% of the single timepoint p-anomalies are identified by the distributed method.

Detection rate increases with decreasing γ. However, it is important that the false alarm rate does not grow too large in order to obtain a high detection rate. To explore this question we show both detection and false alarm rates, in the form of Receiver Operating Characteristic (ROC) curves. These are shown, broken down by week, in Figure 3. The figure shows the ROC curves for the non-parametric method on the left, and for the parametric method on the right.

The figure shows that for most weeks and both method variants, high detection rates can be obtained at quite low false

γ threshold	Unique PCA Events $\alpha = 0.995$	PCA events identified by non-parametric	PCA events identified by parametric	Unique PCA PCA events $\alpha = 0.999$	PCA events identified by parametric	PCA events identified by non-parametric
γ = 0.75	38	23	24	19	15	15
γ = 0.50	38	28	27	19	16	16
γ = 0.25	38	31	32	19	17	19
γ = 0.10	38	33	34	19	19	19

TABLE I
COMPARING PCA EVENTS FOR α = 0.999 AND 0.995 AGAINST PARAMETRIC AND NON-PARAMETRIC GQS METHODS FOR A RANGE OF γ THRESHOLDS

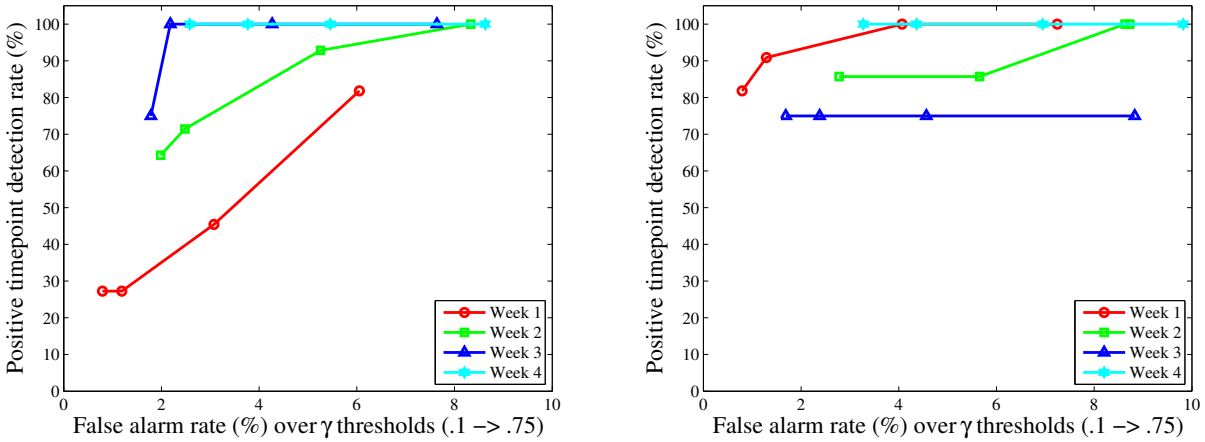


Fig. 3. Detection and false-alarm rates for the non-parametric (left) and the parametric (right) methods.

alarm rates. For example, for Week IV, detection rates of 100% can be obtained at false alarm rates as low as 2%, for either the parametric or non-parametric approach. Thus, not only can the distributed approach be an effective alternative to centralized detection methods, but it can do so with minimal numbers of false alarms.

Figure 3 also shows that while the parametric method outperforms the non-parametric method in Week I, the non-parametric method does well in Week III. Results from Weeks-II and IV are about the same. In summary, while neither one of the two distributed methods consistently outperforms the other, they are both successful in identifying the p-anomalies with relatively low false alarm rates. This motivates a closer comparison of the two approaches, which we present in the next section.

B. Parametric vs Non-Parametric

To explore the differences between parametric and non-parametric method, we compare results at individual routers.

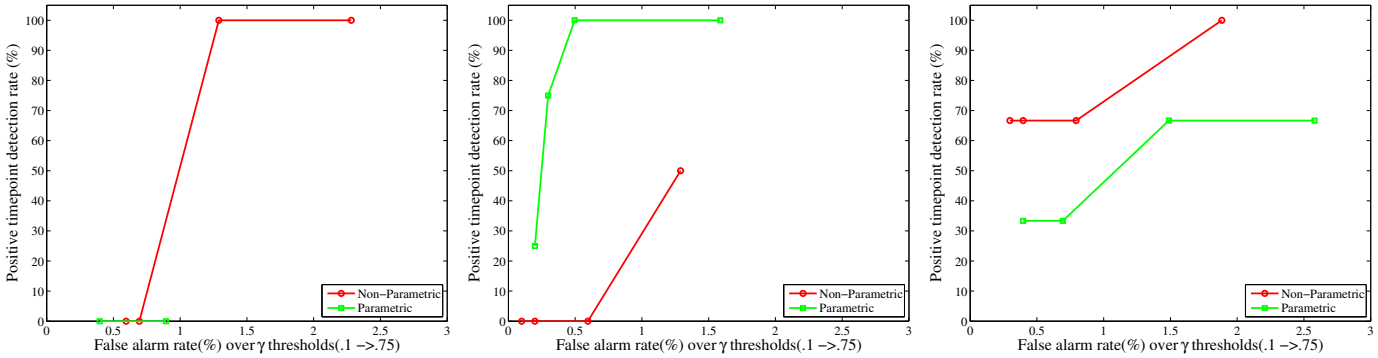
In Figure 4, we show ROC curves for three selected routers. The figure shows that the difference in performance between parametric and non-parametric methods varies strongly across routers. In the case of the Chicago and Denver routers, the non-parametric method outperforms the parametric method while for the Washington router, the parametric method does better.

To better understand why one method does better than the other, we look at traffic patterns at each router. In Figure 5 we show scatter plots of traffic volume at each of the three routers. Although there are multiple links at each router, we plot data for two links that illustrate the differences we found. It is clear that traffic at the Chicago and Denver routers are much more irregularly distributed than that of the Washington router, and that modeling them using a multivariate Gaussian distribution would be a poor choice. On the other hand, the data from the Washington router is well modeled as Gaussian.

As seen in Figure 4(a), the strongly non-gaussian nature of data at the Chicago router means that as we lower the γ threshold,the non-parametric method identifies all the p-anomalies at a γ threshold of 0.25. The parametric method on the other hand does not identify any p-anomalies even for the lowest γ threshold of 0.1. At the other extreme, the Gaussian nature of data at the Washington router (Figure 4(b)) implies the parametric method detects all the p-anomalies at a γ threshold of 0.25. The non-parametric method on the other hand is able to identify only 50% of the p-anomalies for a γ threshold of 0.1. For the Denver router, as expected, the non-parametric method identifies more p-anomalies than the parametric method for each γ threshold, as seen in Figure 4(c).

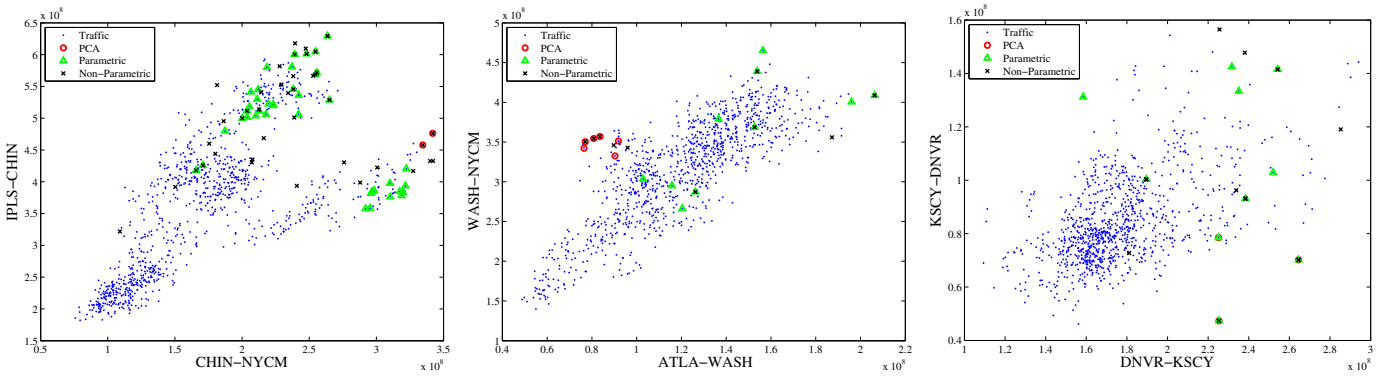
Finally, another property of the data (not shown) is that, as the number of links at a router increases, the parametric method tends to do relatively better. So for Sunnyvale and Kansas City routers (8 links in each router), the parametric method generally outperformed the non-parametric method.

We conclude that in most cases, the parametric method slightly outperforms the non-parametric method on our data. This is a consequence of the high aggregation level of traffic in the Abilene (backbone) network. However the differences in performance between parametric and non-parametric are not large, and the non-parametric method is additionally effective in cases where data is irregularly distributed — as would be expected in lighter-utilized networks or those with less-aggregated traffic, such as edge or enterprise networks.



(a) Chicago, Non-parametric (b) Washington, Parametric (c) Denver, Non-parametric

Fig. 4. ROC curves for Chicago (Week-III), Washington (Week-I) and Denver (Week-IV) routers for γ thresholds of 0.75, 0.5, 0.25 and 0.1



(a) Chicago, Non-parametric (b) Washington, Parametric (c) Denver, Non-parametric

Fig. 5. Scatterplots showing all anomalies for Chicago (Week-III), Washington (Week-I) and Denver (Week-IV) routers

C. Finding FDR using Gamma plots

Finally, we illustrate the benefits that accrue through the use of FDR based detection (γ thresholds). For comparison, we contrast FDR based annotation with the usual method, that is, the β value as described in Section IV.

Figure 6 shows both β and γ annotations for a particular router and week, using both the non-parametric (left) and the parametric (right) methods. Points are sorted in increase order of their β value, and the vertical stems denote the timepoints that are p-anomalies. As expected, the β annotation (which is essentially a statement of the fraction of observed data smaller than the given observation) increases linearly with the rank of the point considered. One implication of this fact is that setting a β threshold for detection (as is commonly done) is a somewhat arbitrary decision.

On the other hand, the range of γ values shows a sharp knee. This shows that points with similar β values can have very different likelihoods of being a true detection. Furthermore, setting a γ threshold is less arbitrary since there is a clear knee to guide the choice. Finally, we reiterate the underlying strength of the FDR approach to detection by γ thresholding: it allows the operator to directly express the severity of detections of interest, by specifying the likelihood that a detection is a false alarm (see eqn. (1)).

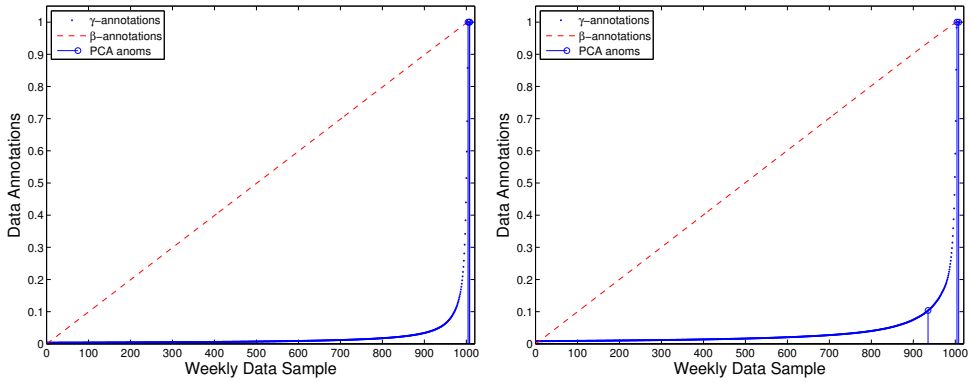
We also note that the γ annotations for the non-parametric method have a sharper knee than that of the parametric method. This implies that for a given γ threshold, the non-parametric method will report fewer datapoints than the parametric method. This is confirmed by the fact that the false positive rate of the non-parametric method is consistently lower, for any given γ than that of the parametric method in our datasets.

A clear illustration of the ability of the FDR method to control false alarms is shown in Figure 7. This figure compares the γ and β methods across each week. For each week, the first two bars show the number of timepoints detected by (1) setting a fixed β threshold (0.99) at each router and (2) setting a γ threshold yielding approximately the same number of detections (averaged over all four weeks). The third and fourth bars show the portions of the first two bars that are false alarms.

Figure 7 also shows a key property of the FDR method, namely, that a different number of unique timepoints are detected in each week. This is in contrast to the β threshold method which by definition detects the same number of timepoints in each week. The figure also shows that the FDR method results in significantly fewer false alarms than the β threshold method.

VI. Conclusions

In this paper we have shown that a distributed approach to spatial volume anomaly detection can be nearly as effective



(a)Non-parametric (b) Parametric

Fig. 6. Different annotations (β and γ) for all timepoints of Week-III Chicago router traffic.

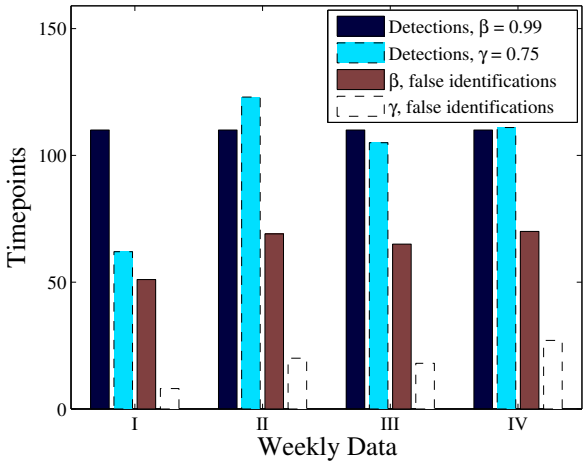


Fig. 7. Detection and false alarms for a given β and γ for Weeks I-IV

as centralized methods. This holds promise for more robust, fault-tolerant and attack-resistant methods for volume anomaly detection. We have also shown that both parametric and non-parametric variants of the generalized quantile set approach are effective for this problem, and provided insight into situations in which each variant is preferred over the other. Finally, we have illustrated the utility of using false discovery rate as a annotation and detection criterion, to make detections more informative to the operator.

A number of open questions remain. First, the scope of our evaluation did not allow detailed inspection of individual detections. This means that our methods may perform even better than we have reported here, but it also means that subtle differences in the nature of anomalies detected by centralized versus distributed methods are not known. We anticipate more detailed study in future work. Further, we are eager to extend these promising results to a larger class of traffic anomalies and explore the performance of our methods in different (edge, enterprise) networks.

VII. Acknowledgements

The first author would like to thank Anukool Lakhina for useful discussions on PCA. This work was supported by Intel and by NSF grants CCR-0325701, ANI-0322990 and CCR-0325701 and ONR grant N00014-06-1-0096.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, Nov. 2002.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc B*, 57(1):289–300, 1995.

[3] J. D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *LISA*, pages 139–146, 2000.

[4] F. Feather, D. Siewiorek, and R. Maxion. Fault detection in an ethernet network using anomaly signature matching. In *Proceedings of SIGCOMM '93*, pages 279–288, 1993.

[5] C. S. Hood and C. Ji. Proactive network fault detection. In *INFOCOM (3)*, pages 1147–1155, 1997.

[6] L. Huang, X. Nguyen, M. Garofalakis, M. I. Jordan, A. Joseph, and N. Taft. In-network pca and anomaly detection. In *Advances in Neural Information Processing Systems 19*, pages 617–624. MIT Press, Cambridge, MA, 2007.

[7] J. Kilpi and I. Norros. Testing the gaussian approximation of aggregate traffic. In *Proceedings Internet Measurement Workshop*, 2002.

[8] H. Kim and B. Karp. Autograph: toward automated, distributed worm signature detection. In *SSYM'04: Proceedings of the 13th conference on USENIX Security Symposium*, 2004.

[9] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch-based change detection: Methods, evaluation, and applications. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, 2003.

[10] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *Proceedings of ACM SIGCOMM 2004*, pages 219–230, Aug. 2004.

[11] D. Moore, G. Voelker, and S. Savage. Inferring internet denial-of-service activity. In *Proceedings of USENIX Security Symposium*, Aug. 2001.

[12] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1999.

[13] S. Sarvotham, R. H. Riedi, and R. G. Baraniuk. Connection-level analysis and modeling of network traffic. In *IMW '01: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 99–103, 2001.

[14] C. Scott and E. Kolaczyk. Nonparametric assessment of contamination in multivariate data using minimum-volume sets and FDR. Technical report, Univ. Michigan, 2007. Available at http://www.eecs.umich.edu/∼cscott/.

[15] J. Storey. The positive false discovery rate: A Bayesian interpretation of the q-value. *Annals of Statistics*, 31:6:2013–2035, 2003.

[16] X. Wang, S. Sarvotham, R. H. Riedi, and R. G. Baraniuk. Network traffic modeling using connection-level information. In *Proceedings SPIE ITCom*, August 2002.

[17] V. Yegneswaran, P. Barford, and S. Jha. Global intrusion detection in the domino overlay system. In *In Proceedings of NDSS*, 2004.

[18] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan. Network anomography. In *IMC'05: Proceedings of the Internet Measurement Conference 2005 on Internet Measurement Conference*, 2005.