Challenging the Supremacy of Traffic Matrices in Anomaly Detection

Augustin Soule
Thomson

Fernando Silveira
Thomson

Haakon Ringberg
Princeton University

Christophe Diot
Thomson

ABSTRACT

Multiple network-wide anomaly detection techniques proposed in the literature define an anomaly as a statistical outlier in aggregated network traffic. The most popular way to aggregate the traffic is as a *Traffic Matrix*, where the traffic is divided according to its ingress and egress points in the network. However, the reasons for choosing traffic matrices instead of any other formalism have not been studied yet. In this paper we compare three network-driven traffic aggregation formalisms: ingress routers, input links and origin-destination pairs (i.e. traffic matrices). Each formalism is computed on data collected from two research backbones. Then, a network-wide anomaly detection method is applied to each formalism. All anomalies are manually labeled, as a true or false positive. Our results show that the traffic aggregation level has a significant impact on the number of anomalies detected and on the false positive rate. We show that aggregating by OD pairs is indeed the most appropriate choice for the data sets and the detection method we consider. We correlate our observations with time series statistics in order to explain how aggregation impacts anomaly detection.

Categories and Subject Descriptors

C.2 [**Computer-communication networks**]: Network Operations—*Network monitoring, Network management*

General Terms

Performance

Keywords

Anomaly detection, Traffic aggregation

1. INTRODUCTION

Detecting unexpected changes in traffic patterns is a topic which has recently received much attention from the network measurement community. These anomalous changes do not only impact network performance, but sometimes represent security threats to users. Building anomaly detection systems is the first step towards securing the Internet, but recent research has proven this to be a very challenging problem.

We study a family of anomaly detection methods that operates in three steps, which correspond to three successive passes of data reduction. First, network statistics are collected on all network links in the form of flow descriptors. Second, flow descriptors are transformed into a set of time series. This transformation requires to pick (1) a network-wide aggregation format and (2) a set of features (in our case, entropy of IP addresses and port numbers). The final step consists in isolating abnormal events using one statistical outlier detection algorithm to the time series. This paper focuses on the second step, and more exactly on the analysis of the impact of the network-wide aggregation formalism on the anomaly detection step.

Since the anomaly detection is performed by extracting statistical outliers from normal traffic, we intuitively expect that if too many flows are aggregated, only the bigger anomalies will be visible. On the other hand if flows are unevenly dispersed in many time series, the statistical noise level is high, and thus too many meaningless events will be interpreted as anomalies.

Previous papers on network-wide anomaly detection [5, 10, 12] have demonstrated the efficiency of various statistical anomaly detection techniques. All these techniques detect anomalies in the *Traffic Matrix* time series, i.e. the time series of traffic between each pair of origin and destination routers in the network. In [5], the motivation for using the traffic matrix was to facilitate the detection of the entry and exit point of the anomaly into the network. The downside of traffic matrices is the complexity of their computation. But traffic matrix is not the only formalism. Other have successfully applied network-wide anomaly detection on input links [12] or random aggregation [6].

In this work we study the influence of three different aggregation formalisms on the anomalies discovered by an entropy based Kalman filter approach [10]. The formalisms in this work naturally emerge from network perspective, i.e., input links, ingress routers, and the traffic matrix formalisms. We validate the anomaly detection results on one week of measurements collected in November 2005 on GÉANT and Abilene, respectively the European and U.S. research networks. The data collected include routing and flow information for both networks. These data sets were measured under com-

IMC'07, October 24-26, 2007, San Diego, California, USA.

pletely different data-reduction parameters (i.e. sampling rate, temporal aggregation and IP anonymization level).

Over 3 500 anomalies were discovered in our data sets. We validate the accuracy of the detection on each formalism by inspecting all anomalies manually in order to extract the false positive rate. As a result of this work, we were able to: (1) compare different formalisms using the exact same method parameters, (2) better understand one key factor impacting anomaly detection, and (3) correlate this impact with metrics and characteristics from the data sets produced by each formalism.

The paper is organized as follows. Section 2 describes the complete methodology we used to measure the accuracy of different traffic formalisms. In Section 3 we present the results of our experiments and analyze them through different metrics. Finally, Section 4 concludes the paper and summarizes our findings.

2. EXPERIMENTAL METHODOLOGY

2.1 Data collection

Both Abilene [1] and GÉANT [2] are well known in the measurement community. Abilene has 11 Points of Presence (PoPs) that provide connectivity to research and academic networks in the United States. GÉANT is the European research network. It is composed of 22 PoPs. It interconnects national research networks and connects to the Internet.

Both networks collect routing information and sampled traffic statistics on input links. Abilene collects routing information through multiple Zebra BGP monitors connected to each router. GÉANT has one single Zebra BGP monitor which is part of the BGP mesh. In both cases, the BGP monitors record all BGP updates. Both networks use Juniper routers (flow statistics are recorded using *cflowd*). We collected and sanitized a week of full data for both networks in November of 2005.

As we mentioned in the introduction, the two networks use different values for measurement parameters, i.e. sampling rate, time aggregation and IP address anonymization. Table 1 summarizes these values. We have evidences that these parameters impact anomaly detection, as shown for example in [7, 4] who studied the influence of sampling rates on anomaly detection. The analysis of the impact of these parameters is outside the scope of this paper. However, for one of our experiments, and in order to compare the two networks in equivalent conditions, we have also re-sampled Abilene and then aggregated its information in 15 min time bins. Since NetFlow traces lose the information about individual packet arrivals, the best we can do is randomly sample each of the packets in the trace. We have anonymized the last 11 bits of GÉANT's IP addresses.

Another source of difference between anomalies detected in Abilene and GÉANT comes from the nature of traffic demands in each network. For instance, GÉANT provides Internet transit service to its customers while Abilene does not. Understanding the impact of traffic nature is complex, and outside the scope of this paper.

2.2 Traffic aggregation formalisms

Given a set of network-wide flow records, we can build different aggregation formalism. Even though one may choose any arbitrary aggregation scheme, we choose to focus on three specific mappings that have natural interpretations in the context of networks and correspond to three levels of aggregation: *ingress routers*, *input links*, and the *traffic matrix* (also known as Origin-Destination pairs in [8]). Next, we describe each of these traffic aggregation schemes in more detail.

	Abilene	GÉANT
IP anonymization	11 bits	None
Time aggregation	5 min	15 min
Packet sampling	1/100	1/1000

Table 1: Data-reduction parameters for each networks

	Abilene	GÉANT
Ingress Routers	11	22
Input links	187	77
Traffic Matrix	121	484

Table 2: Number of time series per formalism

Ingress routers. The flow statistics are collected within each router and periodically sent to a storage point. The simplest way to aggregate data consists of aggregating the flows by the router where they were collected. This scheme has the nice property of being very easy to implement. The amount of time series to be analyzed is the total number of ingress routers in the network. In large-scale networks, ingress routers aggregate large amounts of traffic. Given that entropy is a measure of the distribution of the values of a given feature, heavy aggregation can make it difficult to catch a small variation hidden by a normal variation on the remaining portion of the traffic [11].

Input links. Together with flow information, routers also record the SNMP index of the incoming interface, which allows us to aggregate the traffic per link, i.e. *(ingress router, input interface)* pair. In this formalism, the data is less aggregated than in ingress routers, but there is also an increased number of time series that must be processed by the anomaly detection algorithm. Another advantage of input link aggregation is that it allows the ISP to immediately identify the interface on which an anomaly is detected and to block the corresponding traffic.

Traffic Matrix. Flows are aggregated according to the routers they enter and leave the network. Finding the egress router though requires to perform a routing look up on each flow, which (1) requires to collect routing information and (2) adds a significant computation overhead. This formalism has first been successfully used to detect anomalies by Lakhina *et al.* [5].

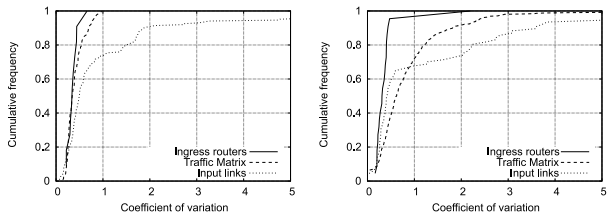
Table 2 summarizes the total number of time series that result from the three different traffic aggregation schemes. There are two noticeable differences between the two networks: in GÉANT the number of OD-flows exceed the number of input links while not in Abilene ; and despite GÉANT is much larger than Abilene (in traffic volume and PoP number), Abilene has a higher number of input links.

Figure 1 provides a general comparison of the flow variability in each data set. We compute the coefficient of variation, i.e. the ratio between the standard deviation and the mean of the number of flows for each of the time series in a formalism. The coefficient of variation is a standard metric for comparing variability in data sources with different

means. Intuitively, a large coefficient of variation is found on time series with high variance and small mean, such as in small and bursty links. Each plot shows the cumulative distribution functions for this metric over the time series in a given formalism. Ingress routers have the smallest flow variability among all formalisms. This is a direct consequence of the large amount of aggregation in those inputs, which makes the traffic fluctuations seem irrelevant near the total amount of traffic. What is most interesting to notice is that variability of input links seems greater than OD pairs. In Section 3 we relate that high variability with the incidence of false positives in anomaly detection.

2.3 Statistical anomaly detection

Given a traffic aggregation model, we are interested in extracting features that can ease the statistical outlier detection. Each aggregate is represented by time series of entropy values computed on four IP header fields, namely source and destination IP addresses and port numbers. Lakhina *et al.* established in [5] that a significant variation in entropy is an effective way to identify the presence of an anomaly in the data set.



(a) Abilene (b) GÉANT

Figure 1: Variability of the number of flows across multiple formalisms.

We are interested in anomaly detection methods which extract statistical outliers in entropy multivariate time series. We choose Kalman based approach described in [10] as a representative of this class of methods. Another very popular method in this class is the subspace method proposed in [5]. We have favored the Kalman filter over the subspace method because our previous experience has shown that there are still some open issues in the calibration of PCA parameters [9]. A similar study of aggregation formalisms with PCA and possibly other detection techniques is left for future work.

Intuitively, the Kalman filter works by modeling the traffic as a multivariate linear model, exploiting both the spatial and time correlation available in the data. At any point in time, one can use the model to predict the next values of the time series and compare those predictions against the actual measurements. If the prediction error is too high compared to the expected variance in the data, then a statistical anomaly is signaled at the space-time point where that condition is true. We refer to [10] for a more formal and detailed description of the method. All the results presented in this paper use a Kalman threshold of $6\,\sigma$ as suggested in [10]. This threshold detects a decent number of anomalies with a low false positive rate. The study of higher thresholds gives similar results and was not included to preserve space.

2.4 Manual labeling

We obtained a total of 3 541 statistical anomalies. *All* anomalies were manually inspected by one of the authors and labeled as true or false positive.

To make manual labeling easier and more reliable, we have designed a web based tool called *WebClass* [3]. WebClass receives the original traffic time series together with entropy values and the list of statistical anomalies computed by Kalman (WebClass is *not* a detection tool but a presentation tool). For each time bin (including those where an anomalous time bin has been identified), WebClass can display (1) the four features entropy time series, (2) packet, flow and byte count, and (3) top n flows counting for the largest amount of traffic. The labeler can pick any anomalous time bin, display detailed flow information for multiple consecutive time bins around the anomaly, zoom in/out, or scroll.

False positives are defined as statistical anomalies for which the labeler could not find a flow (or set of flows) that matches the observed entropy and volume variation in the four features displayed by WebClass. This label is then stored in the WebClass database with notes on why it was qualified as a true or false positive, and with the name of the labeler. WebClass then automatically selects the next anomaly to be labeled. With this tool an acquainted user can quickly match entropy changes with anomalous flows with a high level of confidence.

False negatives are real traffic anomalies for which Kalman did not detect a statistical anomaly; they are not studied in this work because the detection of false negative requires the systematic inspection of all time bins for every formalism and network, i.e., over one million time bins.

Thanks to WebClass, our data set, detection method and anomaly labeling can be shared and reused by other researchers. WebClass could help compare anomaly detection techniques, or verify previous results.

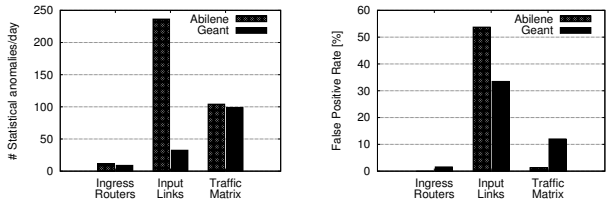
3. ANALYSIS OF RESULTS

In this section we analyze the results of our experiments on traffic data from Abilene and GÉANT. We should emphasize that even though we present results for two distinct networks, we do not intend to provide a thorough comparison of anomaly detection in different networks. As we mentioned in the previous sections, there are numerous differences between Abilene and GÉANT data, from measurement parameters to customer demands. Still, we verify that most of the observations are consistent across networks. Analyzing two networks should reduce the biases that would be expected from analyzing a single data set, and increase the confidence in our observations.

Performance of aggregation formalisms.

Figure 2 summarizes the results of the performance metrics we consider for each network and formalism. The plots show the total number of anomalies found in each data set, and the percentage of detections which correspond to false alarms, i.e., the false positive rate.

In Figure 2(b) the false positive rates for ingress routers are less than 2% in both networks. This would be a remarkable result if it were not for the fact that the total number of detections (and consequently the number of true positives) is much smaller than in all other formalisms. Intuitively, the



(a) Number of statistical anomalies per day (b) False positive rate

Figure 2: Performance of different formalisms in Abilene and GÉANT.

Formalism	Abilene	GÉANT
Ingress routers	0%	0%
Input links	68%	39%
Traffic Matrix	44%	76%

Table 3: Fraction of anomalies detected only in one formalism.

small number of detections in ingress routers comes from the fact that the time series of traffic are too much aggregated. More precisely, many anomalies are hidden within the bigger fluctuations of traffic.

Dividing the traffic by input interfaces is a straightforward way to reduce the level of aggregation. Indeed, it can be seen in Figure 2(a) that for Abilene data, there are over 20 times more statistical anomalies with input links than with ingress routers. This increase is not as impressive for GÉANT data, where it is only a factor of four. This difference between the two networks can be attributed to multiple reasons. First of all, according to Table 2, there are many more time series (i.e., input links) in Abilene than in GÉANT. Second, as we mentioned in Section 2.1 and Table 1, different measurement parameters across networks play an important role in anomaly detection metrics.

Despite having a higher number of detections than ingress routers, input links also display the highest false positive rates among all formalisms. Such false positive rates go as high as 53% with the original parameters of Abilene. That is clearly unacceptable for any detection scheme, since each false alarm would need to be verified by a human operator.

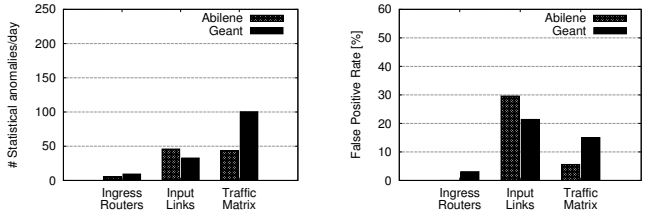
The traffic matrix formalism detects a large number of statistical anomalies, together with a small false positive rate. Nevertheless, the false positive rate for the traffic matrix in GÉANT seems somewhat higher than that of Abilene. Further in this section, we identify the main causes for the false positives in our data sets.

Table 3 displays the number of true positives that were detected in one formalism but not in the other two for each network. First, *all* the anomalies discovered in the ingress routers were also discovered using either input links or traffic matrix. This proves that the level of aggregation is too high in ingress routers. Second, the input links or traffic matrix formalism do detect different anomalies. The analysis of these differences has been kept for future work.

Impact of data-reduction parameters.

The plots in Figure 3 correspond to the same performance metrics evaluated in Figure 2 after having brought both networks to the same parameters in all formalisms. More precisely, we process the original data sets to have the same monitoring parameters as explained in Section 2.1, i.e. 1/1000 sampling, 15 minutes of time aggregation and 11 bits of IP anonymization. After this step, the total number of detections becomes much more consistent among the two networks. This suggests that the levels of packet sampling and time aggregation performed originally in GÉANT have a serious impact in the number of detections in input links. It also suggests that our comparison and methodology are meaningful.

While reduced from 50 % to 30 % the number of false positives in Abilene's input links is still at an unacceptable level for an operational network. Even though the data-reduction parameters impact the anomaly detection, it is not the only important factor.



(a) Statistical anomalies (b) False positive rate

Figure 3: Performance of formalisms after reducing the measurement parameters to the same level.

Impact of aggregation on anomaly size.

Figure 4 shows the results of an experiment to assess how the increased level of aggregation in ingress routers makes anomalies less noticeable. For each network, we consider the sets of anomalies found in either the input links or the traffic matrix. Let i represent a particular input in one formalism, i.e. an input link or an OD pair in the Traffic Matrix. Also, let t be a time bin where a statistical anomaly was detected. If p_t^i is the number of packets measured on time bin t and input i, then we compute the following value:

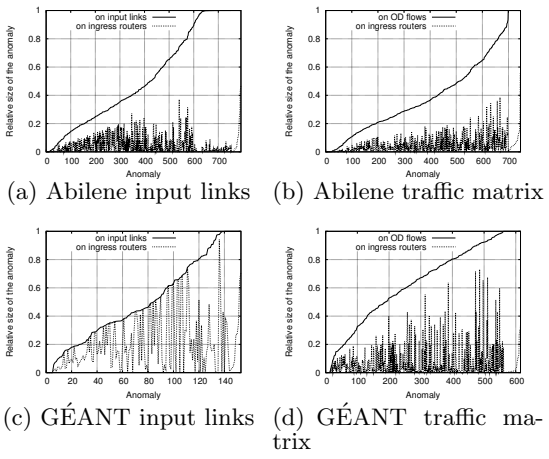
$$r_t^i = \frac{\left|p_t^i - p_{t-1}^i\right|}{\max\left\{p_t^i, p_{t-1}^i\right\}}. \quad (1)$$

This measures the relative deviation in the number of packets to the maximum between the current and previous time bins. Taking the absolute value on the numerator and the maximum value in the denominator enables us to analyze both increases and decreases with a single metric.

Given that any packet seen on input links or the traffic matrix formalisms can also be found with ingress routers, we can also estimate the relative size of the anomaly in the ingress router formalism. This can be done by replacing the denominator of (1) by the equivalent value measured over the time series in the corresponding ingress router.

We realize that the method we use, as described in Section 2.3, deals with time series of entropy values instead of packet counts. Nevertheless, the variations in entropy often correspond to events which increase or reduce the number of packets.

Figure 4 shows that in the Abilene data set there is a large decrease in the relative size of most anomalies with ingress routers. For the GÉANT plots, the difference is less evident, particularly with the input links formalism. It clearly explains why, on Figure 2, the increase in the number of detections from ingress routers to input links is not as pronounced in GÉANT as it is in Abilene.



(a) Abilene input links (b) Abilene traffic matrix

(c) GÉANT input links (d) GÉANT traffic matrix

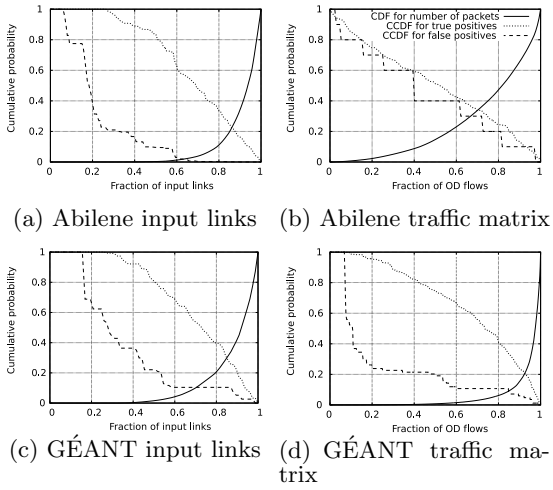
Figure 4: Ratio of the size of the anomalies vs the traffic on Abilene and GÉANT.

Impact of small inputs on the false positive rate.

Figure 5 compares the input links and traffic matrix formalisms in Abilene and GÉANT. For each data set, the corresponding plot has its inputs (i.e. input links or OD pairs) sorted on the x-axis according to size, measured as the total number of packets during the full week of the trace. One of the curves in each plot shows the cumulative distribution for the input size. The other two represent the complementary distribution functions for the number of true and false positives by input.

The plots 5(a) and 5(b) illustrate that at least 90% of the false positives are detected on input links which contributed to less than 2.5% of the total number of packets in Abilene and 3.5% of those in GÉANT. On the other hand, for the traffic matrix of Abilene, the very few false positives that are found, are relatively well spread across the different OD pairs. This spread matches the finding from figure 1 where the variability of the traffic matrix in Abilene is smaller than that of GÉANT leading to a smaller false positive rate. Finally, the traffic matrix from GÉANT on Figure 5(d) seems to concentrate most of the traffic on a few popular OD pairs. That can explain why the false positive rate is much higher in GÉANT than in Abilene (respectively 12% versus 1%).

Intuitively, the problem with small inputs (either links or OD pairs) is that even very small events can cause large deviations from the expected behavior. Recall that the general formulation of statistical anomaly detection (with which the Kalman filter method is compatible) is to look for time bins in which the local variance exceeds the global one by a certain number of times. Given that definition, small inputs with only occasional bursts of traffic may pose a serious problem by contaminating the method's perception of what is anomalous in the global setting of the network. Moreover, packet sampling will increase the amount of noise in the time series, which can be more harmful in small inputs than in traffic which is highly aggregated.



(a) Abilene input links (b) Abilene traffic matrix

(c) GÉANT input links (d) GÉANT traffic matrix

Figure 5: Distribution of anomalies and total traffic across different inputs.

4. CONCLUSIONS AND FUTURE WORK

In this paper we analyzed the impact of aggregation in the performance of a network-wide anomaly detection method. We analyzed three traffic aggregation formalisms that emerge naturally in the context of network measurements, namely ingress routers, input links, and traffic matrix. Our observations show that aggregation can be harmful to anomaly detection in two different ways. First, the relative size of an anomaly detected on input links or traffic matrices is considerably larger with these formalisms than with ingress routers, where there is too much aggregation. Second, a low level of aggregation increases the variability on some links. As a consequence the statistical noise increases, as well as the number of false positives.

We observe that the level of traffic aggregation in ingress routers is too large to allow effective anomaly detection in our data sets. On the other hand, with input link aggregation, the spread of traffic is extremely unbalanced, and this leads to false positive rates of up to 50%. The traffic matrix formalism seems to provide a good compromise between the number of true and false positives.

The work described in this paper is still preliminary. More formalisms, more detection methods and more data sets need to be studied. For example, we plan to study formalisms where the level of aggregation can be adjusted, such as in random aggregation [6]. We have already collected similar results with the PCA method described in [5]. However, more progress needs to be made on PCA understanding before we can have enough confidence in these results [9].

We have shown that the performances of the formalisms are strongly related to the variability of the data. An alternative way to reduce traffic variability would be to explore temporal aggregation and its influence on the anomaly detection. Using techniques such as the multi resolution anal-

ysis, we could identify timescales at which each formalism is optimal.

We also plan to study how different formalisms are biased towards detecting certain types of anomalies. As we mentioned, we have evidence that input links and the Traffic Matrix detect different subsets of anomalies. In order to do that we will need to further augment our data sets by identifying meaningful clusters of semantically related anomalies. With that in mind, we are working on improving our current classification tool.

5. REFERENCES

[1] `http://abilene.internet2.edu`

[2] `http://www.geant.net`

[3] `http://www.thlab.net/webclass/`

[4] Brauckhoff, D., Tellenbach, B., Wagner, A., May, M., and Lakhina, A. Impact of packet sampling on anomaly detection metrics. In *Proceedings of the ACM Internet Measurement Conference* (October 2006), pp. 159–164.

[5] Lakhina, A., Crovella, M., and Diot., C. Diagnosing network-wide traffic anomalies. In *ACM Sigcomm* (2004), ACM Press.

[6] Li, X., Bian, F., Crovella, M., Diot, C., Govindan, R., Iannaccone, G., and Lakhina, A. Detection and identification of network anomalies using sketch subspaces. In *Proceedings of the ACM Internet Measurement Conference* (October 2006), pp. 147–152.

[7] Mai, J., Chuah, C.-N., Sridharan, A., Ye, T., and Zang, H. Is sampled data sufficient for anomaly detection? In *Proceedings of the ACM Internet Measurement Conference* (October 2006), pp. 165–176.

[8] Medina, A., Taft, N., Salamatian, K., Bhattacharyya, S., and Diot, C. Traffic matrix estimation: Existing techniques and new directions. *ACM Sigcomm* (August 2002).

[9] Ringberg, H., Soule, A., Rexford, J., and Diot, C. Sensitivity of PCA for traffic anomaly detection. In *ACM Sigmetrics* (Jun. 2007).

[10] Soule, A., Salamatian, K., and Taft, N. Combining filtering and statistical methods for anomaly detection. In *ACM IMC* (Oct. 2005).

[11] Soule, A., Ringberg, H., Silveira, F., Rexford, J., and Diot, C. Detectability of Traffic Anomalies in Two Adjacent Networks. In *PAM* (Apr. 2007).

[12] Zhang, Y., Ge, Z., Greenberg, A., and Roughan, M. Network anomography. In *ACM IMC* (Oct. 2005).