Detectability of Traffic Anomalies in Two Adjacent Networks

Augustin Soule[1], Haakon Larsen[2], Fernando Silveira[3], Jennifer Rexford[2], and Christophe Diot[1]

[1] Thomson Research
[2] Princeton University
[3] Federal University of Rio de Janeiro

Abstract. Anomaly detection remains a poorly understood area where visual inspection and manual analysis play a significant role in the effectiveness of the detection technique. We observe traffic anomalies in two adjacent networks, namely GEANT and Abilene, in order to determine what parameters impact the detectability and the characteristics of anomalies. We correlate three weeks of traffic and routing data from both networks and apply Kalman filtering to detect anomalies that transit between the two networks. We show that differences in the monitoring infrastructure, network engineering practices, and anomaly-detection parameters have a large impact on which anomaly detectability. Through a case study of three specific anomalies, we illustrate the influence of the traffic mix, IP address anonymization, detection methodology, and packet sampling on the detectability of traffic anomalies.

1 Introduction

Identifying anomalous Internet traffic, such as malicious attacks, flash crowds, or traffic shifts, is a difficult and important challenge for Internet Service Providers (ISPs). In the past few years, researchers have introduced a promising new way to detect anomalies. Rather than scrutinizing the traffic on each link independently, the traffic is summarized in a link or traffic matrices and analyzed on all links simultaneously. Then, anomalies are detected by applying statistical analysis techniques to the matrices. This is known as "network-wide" anomaly-detection. This approach is very effective at detecting anomalies that are spread over multiple links, such as distributed attacks or traffic shifts caused by routing changes[3, 5].

Despite promising initial results, we still understand very little about network-wide anomaly detection methods. Relatively few papers have been published, and these studies unfortunately (1) do not describe the calibration of the methodology very accurately and (2) do not use the same measurement data sets. In addition, identifying and classifying *all* anomalies in a given traffic trace (in order to get a *ground truth* to which to compare the outcome of the detection methods) is extremely difficult, if not impossible on a large data set. The number of anomalies detected depends on many parameters that have not been studied systematically. In earlier papers, "manual tweaking" and "visual inspection" play an important role in the success of the anomaly-detection techniques. Therefore, network administrators cannot readily apply these network-wide anomaly-detection techniques "out of the box" or easily tune them for effective use in their networks. It will not be possible to use these methods in an operational network until we understand how anomaly “detectability” is influenced by network design, monitoring infrastructure, and anomaly-detection technique.

In this paper, we take a first step in this direction by studying traffic anomalies simultaneously in two backbone networks—GEANT and Abilene[1], focusing on anomalies that cross both networks. We analyze three weeks of time-synchronized traffic and routing traces for the two networks. Note that the goal of this work is not to answer all questions and explain every observation. We are far from being able to do so for reasons explained earlier. Our ambition is to identify problems and issues that need to be addressed before thinking of unsupervised and automatic anomaly detection in operational environment.

We summarize the traffic (for each time interval) in four entropy *Link Matrices*, each matrix corresponding to a given flow feature (i.e. source and destination IP address and source and destination port). To detect anomalies, we use the Kalman-filtering method introduced in [4]. The measurement data sets and detection methodology will be described in more details in the next sections. Note that for the purpose of this work, the detection method is not critical as long as it detects real anomalies with a low false positive ratio.

We use BGP routing information to identify the subset of the traffic that traverses both networks, and we perform anomaly detection on these reduced data sets, as well as on the full link matrices. Surprisingly, we find that many anomalies are detected in one network and not in the other. We show that it can be due to (i) the difference in monitored traffic sampling rate, (ii) the anonymization of the IP addresses, (iii) the calibration of the detection method (i.e., value of the detection threshold), or simply because of (iv) the traffic mix on the link where the anomaly is detected. To illustrate these claims, we analyze three specific anomalies where each of the potential causes listed above is involved in a missed anomaly detection in one of the two networks.

The paper is organized as follows. Section 2 describes our measurement data and formalism. Section 3 presents the anomaly detection methodology. We start the discussion of results (section 4) by general observations about the anomalies that are detected and missed in each network and the factors that impacts the detectability. Section 5 illustrates the general discussion with three specific anomalies where previously identified factors are indeed causing an anomaly to be missed. We discuss research challenges and future research directions in section 6.

2 Measurement Data

2.1 Collecting the Traffic and Routing Measurement

The data used in this paper has been collected in two academic networks, GEANT and Abilene. Abilene provides connectivity to research and academic networks in the US. It has 11 points of presence (PoPs) and 198 incoming links. One peculiarity of this network is that each Abilene customer must have a separate connection to the Internet since Abilene does not connect to the Internet. Abilene is very interesting from an anomaly-detection standpoint, as it is mostly used for experimental academic traffic. GEANT

[1] www.geant.net and abilene.internet2.edu

is the European Research Network. It interconnects national research networks, rather than directly connecting research institutions. GEANT is composed of 22 PoPs and 99 incoming links. It is connected to the Internet and provides transit service to its customers. During the time our data was collected, Abilene and GEANT were peering at two locations: between Washington DC and Frankfurt through an OC48 link and between New York and Amsterdam through a virtual LAN. Note that the traffic from multiple Autonomous Systems (ASes) is mixed in this VLAN. Therefore, we can not isolate the traffic going from Abilene to GEANT, and vice-versa, from the recorded data.

Both networks collect routing and sampled traffic statistics. Abilene collects routing information through Zebra BGP monitors connected to the routers. GEANT has one single Zebra BGP monitor which is part of the iBGP mesh. In both cases, the BGP monitors record all BGP updates. The flow statistics are recorded on each router using Juniper's *J-Flow* tool. GEANT routers' record one out of every 1000 packets and the flow information is exported to the Network Operation Center (NOC) every fifteen minutes. In Abilene packets are sampled at 1 out of 100, and flow information is exported every five minutes. In Abilene, the last 11 bits of each IP address are set to zero, preventing the identification of the source or destination host.

Merging all four datasets has been a serious challenge. We could only identify a period of 20 consecutive days between November 10 and November 30, 2005, where all datasets were complete in both networks. Routers and monitors are synchronized using NTP and each measurement record is labeled with a timestamp. The datasets are collected at different geographic locations. We can not guarantee that the clocks are perfectly synchronized. But the time granularity of the flow statistics is much larger than the NTP error and thus time synchronization is not an issue.

2.2 Aggregating the Traffic into a Link Matrix

We are primarily interested in anomalies for which the traffic transits from Abilene to GEANT and GEANT to Abilene. However, direct observation of anomalies on the peering links is not possible because of the presence of the VLAN. Therefore, we opted for network-wide detection in each network, as described in [3, 4, 7]. The traditional Traffic Matrix defined in these papers is sensitive to routing changes [6]. Given that we need to match traffic between two networks, routing errors could bias our observations. Therefore, we chose a traffic formalism which is insensitive to internal routing changes, i.e. incoming link matrix, all the incoming link traffic time series combined in a single matrix. Routing information is then used to identify the subset of flows that go from Abilene to GEANT, and vice-versa. The anonymization in Abilene did not bias this step as we did not identify any prefixes longer than 21 bits from GEANT to Abilene. Therefore, we use four different sources of data: link matrix from Abilene (A) and GEANT (G), and link matrices made of the flows that go from Abilene to GEANT and GEANT to Abilene. These two link matrices are noted respectively $A2G$ and $G2A$. Note that any anomaly detected in $A2G$ or $G2A$ should also be detected in G and A respectively.

2.3 Detecting Anomalies

The flow statistics are represented by time series of entropy values computed on four IP header fields, namely source and destination IP addresses and ports, as defined in [3]. The entropy measures how a distribution is spread over the range of values. We use the classic entropy equation for each feature f: $X_f(t) = -\sum_i p_f(i,t)\log(p_f(i,t))$ where $p_f(i,t)$ is the proportion of packets containing the feature value i during the time interval t. The entropy is low when the distribution is concentrated on a few values, and high when each value is equally probable. Lakhina established in [3] that a significant variation in entropy is an effective way to identify the presence of an anomaly in the data set. The four features entropy based detection also helps identifying the cause of the anomaly.

Network-wide anomaly detection is performed using the Kalman method applied independently on each of the four features. This method was introduced in [4] and in [5]. These papers can be read for details on the detection method. In short, the Kalman filter extracts the predictable part of the traffic time series according to a predefined model. The difference between the prediction of the model and the observed value is defined as the "innovation" of the time series. Anomalies are defined as a significantly large difference between the predicted value and the observation. This significant change is identify as abnormal whenever the absolute value of the innovation exceed T times the variance (σ) of the innovation.

However, network-wide anomaly detection returns a list of time bins were an anomaly should be present based on the interpretation of entropy feature variation (together with the links where the anomalous time bin has been detected). *Network-wide anomaly detection tools does not detect anomalous traffic.* To identify the anomalous traffic we perform a post-mortem analysis of all anomalous time bins observed in both networks. We compare the traffic in the anomalous time bin to the one in the time interval that just precedes it and look for some significant change in the traffic that could explain a certain combination of entropy variation. An entropy decreases correspond to a concentration of the feature distribution and an entropy increase denotes a dispersion of the feature distribution. It is easy to identify the traffic that caused the entropy to decrease as it is, most of the time, due to a flow that increases its traffic. On the other hand, it is very difficult to identify the traffic corresponding to an entropy increase as what we are now looking for is a dispersion of traffic. Therefore, anomaly classification is easier when one of the feature exhibits an entropy decrease. We also try to aggregate the anomalies that are detected in consecutive time bins. Aggregation is performed by matching feature entropy values, links where the anomaly has been detected, and the IP flows carrying the anomalous traffic. Once aggregation has been performed in each network, we match $A2G$ anomalies to G anomalies (and $G2A$ to A). This labeled data and the associated methods are available upon request.

3 General Observations

The goal of this section is to identify what factors impact the detectability of traffic anomalies. Table 1 summarize the number of anomalies detected in each data set. This table also gives the number of anomalies that are found in two data sets simultaneously (i.e. $A2G$ and G, $G2A$ and A). The threshold is similar in both networks and equal to 10 times σ with $\sigma = \{\sigma_1 \cdots \sigma_n\}$ the variance of the innovations. We chose this value because it resulted in zero false positive in [5].

Anomalies detected in					
A	G	$A2G$	$G2A$	$A2G \bigcap G$	$G2A \bigcap A$
78	14	58	10	5	3

Table 1. Number of anomalies observed between Abilene and GEANT for a threshold of 10σ. 2005/11/10 to 2005/11/30

78 anomalies are detected in Abilene and only 14 in GEANT. It is difficult to explain such a result, which only advocates for using a different threshold in both networks. We will come back on this issue below. GEANT is a larger network, but its sampling rate is lower. We conjecture that the very low sampling rate accounts for most differences in anomaly detection.

More anomalies are found in A going to G than in A coming from G. This phenomenon is more pronounced from A to G than from G to A, most probably because of sampling. This simply highlight the impact of the detection technique and the traffic data formalism on the detectability of anomalies. $A2G$ and $G2A$ are reduced data sets. They correspond to the subset of traffic captured in the origin network that is destined to the adjacent network. We conjecture that Kalman can extract anomalous behavior more easily in the reduced dataset.

This reduction of the data helps the Kalman method to detect anomalies. This is an interesting observation as it proves that both the method and the data set formalism impact the anomaly detectability.

We detected 58 anomalies in $A2G$ and only 5 (i.e. 9%) of these anomalies were detected in G. Similarly, 10 anomalies are detected in $G2A$ and only 3 of them are also detected in A (i.e., 33%). We suspect that the most probable explanation is the sampling rate in G, which is 10 times lower than in Abilene. The impact of sampling can also be observed on the number of $A2G$ anomalies detected in G, i.e. 12%. Moreover we expect that sampling affects differently the detectability of anomalies based on their nature. Thus the impact of the sampling rate on the anomaly detectability is not easy to evaluate. Recently a paper [2] studied the impact of traffic sampling on the detectability of the Blaster worm event using an entropy-based detection method. The paper shows that the worm is almost undetectable with a sampling rate of 1 out of 1000. In the same mindset, a theoretical result [1] shows that even a task as simple as ranking the flows according to the their size using sampled traffic requires a sampling rate greater or equal to 10%. We are far from these values in GEANT. However, not all anomalies in $G2A$ are also detected in A, which means that the sampling rate alone does not explain why we do not detect the same anomalies in both networks.

Anonymization could impact anomaly detection in Abilene. However, it is not clear whether the anonymization of the last eleven bits of the IP addresses reduces or increases the number of anomaly detected. However, we conjecture that anonymization will most probably change the way an anomaly is classified, by transforming the entropy dispersion of the IP address in an entropy concentration (showing one single IP address instead of multiple ones with the same prefix).

It is difficult to compare anomalies in the two networks with the same detection threshold. In the experiment below, we have chosen the threshold in G to be such that we obtain approximately the same number of anomalies in G and A. This threshold value is 5σ (when Abilene's threshold remains at 10σ). Table 2 show the number of anomalies in all data sets with this new threshold in G.

	anomalies detected in			
T_h in GEANT	G	$G2A$	$A2G \bigcap G$	$G2A \bigcap A$
Low	84	89	23	17

Table 2. Number of anomalies observed in GEANT for a 5σ detection thresholds. 2005/11/10 to 2005/11/30.

As expected, we now detect 84 anomalies in GEANT instead of 14 with a 10σ threshold which is five times more than with the 10σ threshold of table 1. Not surprisingly, the number of anomalies found in $A2G$ and G and in $G2A$ and A is also around 5 times more.

The 89 anomalies in $G2A$ are also easy to explain. Remember that $G2A$ is a different Link Matrix than G (in fact a subset the traffic contained in G). A higher number of anomalies in $G2A$ than in G confirms the impact of the detection method and of the data formalism on anomaly detectability, which has been discussed earlier in this section.

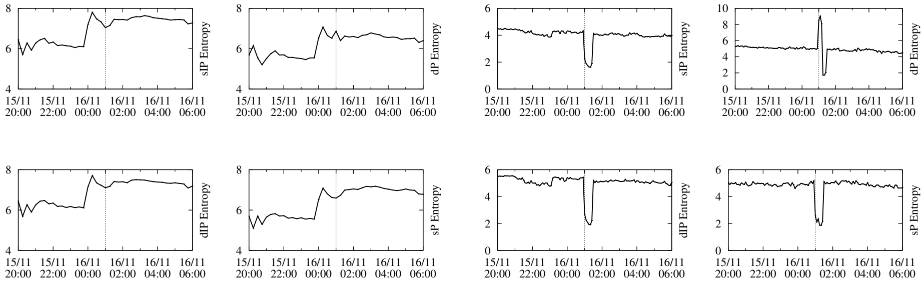
To summarize our observations, we have shown that the detection methodology, the data formalism, and the sampling rate do impact the number of anomalies that can be detected. We have seen that NOCs can play with the detection threshold to increase or decrease the number of anomalies detected. We suspect that IP address anonymization has a limited impact on anomaly detectability.

4 Case Study

The following three anomalies illustrate how the factors identified in the previous section can impact anomaly detection.

4.1 Impact of Sampling and Detection Threshold

This first anomaly is detected in the traffic flowing from Abilene to GEANT but is undetected in GEANT. It is an attacks against a SSH server that originated in Abilene on November 16^{th} between 01:00 and 01:30 GMT. A host in the university of Philadelphia starts scanning the network for vulnerable servers. It finds a reachable SSH server in Italy (at 01:15 GMT). Then the attacker tries to gain access to this server by flooding it with SSH packets.



(a) GEANT

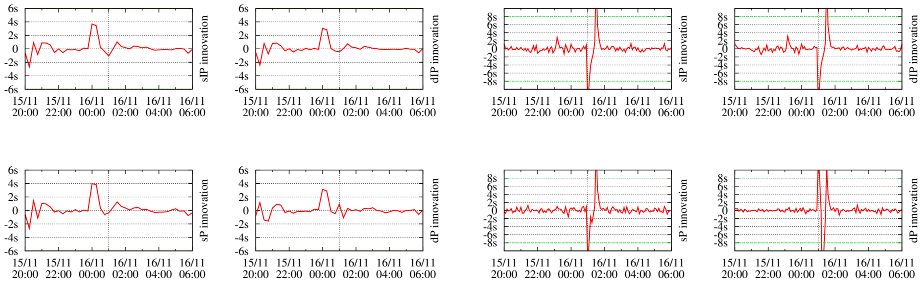
(b) Abilene

Fig. 1. Entropy of the four features during the SSH attack as seen in GEANT and Abilene.

In figure 1, we show the entropy of the four features (source IP, destination IP, source Port and destination Port) as seen by Abilene in its New York router 1(b) and in GEANT on its peering link to Abilene 1(a). The vertical line indicates the time at which the anomaly was detected in Abilene. The entropy plots on Abilene (fig. 1(b)) show that all the entropy values except the destination port decrease as expected in the case of port scans. The destination ports entropy increase shows that the attacker is doing a port scan on a few set of machines. At 01:15 GMT the attacker has found its victim and now targets its attack on a single port of the victim. As a consequence, the destination port entropy decreases.

This event can been seen on GEANT (fig. 1(a)) as a small increase followed by a small decrease on the destination port entropy. But the amplitude of the change is too small to be detected as an anomaly. Indeed, the total amount of anomalous traffic in Abilene is 84 000 packets. In GEANT, only 9 000 packets are sampled for the same traffic (i.e. approximately one tenth).

This anomaly being observed in GEANT, it is interesting to discuss whether a lower threshold in GEANT would have made this anomaly detectable. The figure 2 represents the time series of the Kalman innovation divided by its variance for the four features on



(a) GEANT

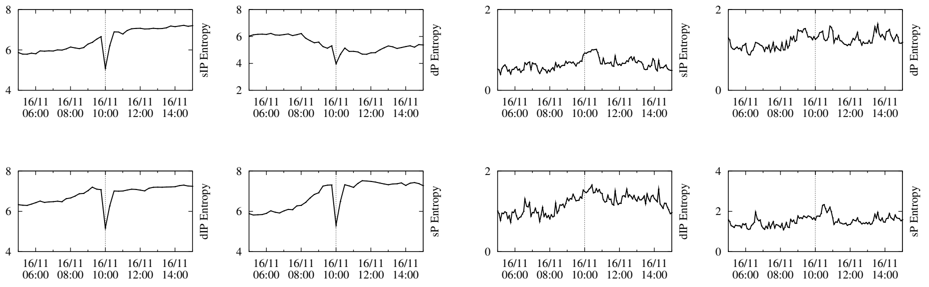
(b) Abilene

Fig. 2. Kalman innovation of the four features during the SSH attack as observed in GEANT and Abilene.

each network. As seen in the figure 2(b) any threshold lower than 12σ in Abilene will detect this anomaly. But inside GEANT the threshold need to be set to at most 1σ to be able to detect this event.

4.2 Impact of the Traffic Mix

This non malicious anomaly was detected in GEANT only on November 16^{th} at 10:00 GMT. It is characterized by a small number of SSH flows transferring a large amount of data between two hosts, one in the UK, and the other near New York. This transfer is performed over port 22, so we suspect these flows use SFTP. The four features observed in GEANT are shown figure 3(a). These features are the one we would expect in a such case. As in the case of a large file transfer between two hosts using a known application, the entropy of all four features should decrease as observed in figure 3(a). However, this anomaly was not detected in the Abilene despite a higher sampling rate and also despite that around 30 ,000 packets belonging to the anomaly were sampled on Abilene. The entropy of the features observed at this time on Abilene are shown figure 3(b). The reason why this anomaly goes undetected on Abilene is that at the same time, the entropy captures a concentration on port 80 due to on on-going massive HTTP transfer (220 000 packets in the anomalous time bin). Our anomalous file transfer is not big enough to significantly impact the entropy.



(a) GEANT

(b) Abilene

Fig. 3. Entropy of the four features during the large file transfer as seen in each network

This is a nice example of how an anomaly can hide in the network traffic. This anomaly illustrates that because of the traffic mix it might be impossible to detect an anomaly in a given network, whatever the value of the detection threshold is. Detecting an event hidden behind a predominant traffic requires either a different representation of the traffic, or detection in multiple networks. Summarizing the traffic as ingress-egress traffic matrix might separate the predominant traffic from the anomalous traffic and make it possible to detect the anomalous traffic.

4.3 Impact of Anonymization

Abilene anonymizes the IP addresses by inserting zeros on the last eleven bits. As explained in the previous section, it is difficult to evaluate the impact of such anonymization on the detection process. We did not find any anomaly that disappeared in Abilene because of anonymization. However, we found many instances of the following phenomenon, i.e. where anonymization impacts how the anomaly is classified.

On November 16^{th} at 05:00 GMT, we detected a port scan from a university connected to Abilene in Atlanta, to a sub-network connected to the Swedish router in GEANT. The maximum rate observed in Abilene was about 1 ,000 sampled packets every five minutes. The ingress link in Abilene is lightly loaded so even with this rate this anomaly was visible in the entropy of the features (fig. 4). The entropy of the port numbers increase as the distribution is spread. The distribution of the source IP is concentrated around the attacker's IP address as visible in the decreased entropy. But the entropy of the destination IP decreases indicating a concentration. In fact all the victims' IP addresses belong to the same sub-network and the side effect of anonymization is to make them look like a single IP address, creating an artificial concentration in the destination IP distribution.

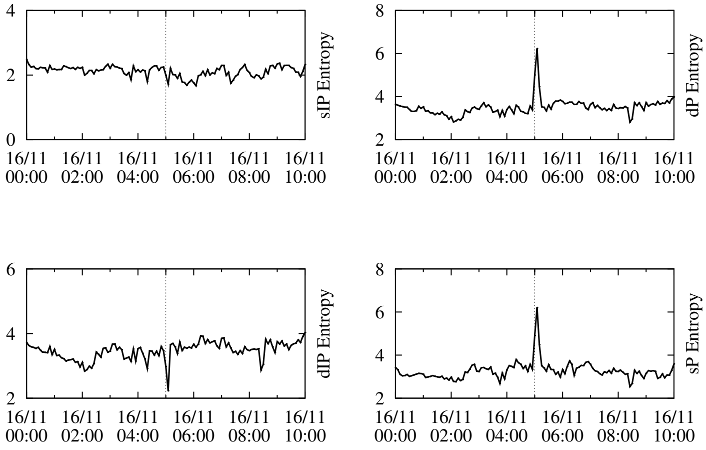


Fig. 4. Entropy of the four features during a Port Scan inside Abilene.

This observation has multiple consequences. We can imagine non anomalous traffic sent to multiple addresses in the same sub-network to be identified as an anomaly after anonymization. That should be easy to detect though. On the other hand, we could not imagine any scenario where anonymization could hide an anomaly. There is also a good side-effect of anonymization. As mentioned in section 2.3, identifying the traffic corresponding to an entropy increase (traffic dispersion) is usually very difficult. We could imagine to use anonymization as a way to transform a traffic dispersion in a traffic concentration. We keep the study of how to use anonymization techniques to help the anomaly classification for future work.

5 Conclusion and Future Work

We have shown that numerous factors impact the detectability of traffic anomalies in a given network. The major factors are detection methodology, data formalism, sampling and network traffic. IP address anonymization on the other hand might end being a feature that could make it easier to classify anomalies. However, its impact on anomaly detectability is still unclear. This work does not explain how each factor impacts the number and type of anomalies detected. However, it makes a clear case for (1) deeper analysis of anomaly detection techniques design and calibration and (2) Internet-Wide anomaly detection as a single method will not be capable to detect all anomalies in a network.

Using our two data sets, we are now starting a systematic analysis of two methods, Kalman and PCA, with different data formalisms, in order to understand how robust these techniques are and how to automatically choose the right operating parameters. Another important piece of work is to understand what is the minimum sampling rate in order not to miss anomalies. 1 for 1000 seems to be below that threshold.

A major concern is the lack of ground truth. We have started the annotation of the three weeks of traces used in this work. The annotated data set, including the anomalies we have detected, will be made available to the research community for comparison of observation and to facilitate the reproducibility of detection result, and the design of new detection techniques.

References

1. Barakat, C., Iannaccone, G., and Diot, C. Ranking flows from sampled traffic. In *ACM CoNEXT* (Dec. 2005).
2. Brauckhoff, D., Tellenbach, B., Wagner, A., Lakhina, A., and May, M. The effect of packet sampling on anomaly detection. In *ACM IMC* (Oct. 2006).
3. Lakhina, A., Crovella, M., and Diot., C. Diagnosing network-wide traffic anomalies. In *ACM Sigcomm* (2004), ACM Press.
4. Soule, A., Salamatian, K., and Taft, N. Combining filtering and statistical methods for anomaly detection. In *ACM IMC* (Oct. 2005).
5. Soule, A., Salamatian, K., Taft, N., and Nucci, A. Traffic matrix tracking using kalman filters. *ACM LSNI Workshop* (2005).
6. Teixeira, R., Duffield, N. G., Rexford, J., and Roughan, M. Traffic matrix reloaded: Impact of routing changes. In *PAM* (2005), pp. 251–264.
7. Zhang, Y., Ge, Z., Greenberg, A., and Roughan, M. Network anomography. In *ACM IMC* (Oct. 2005).