

Empirical Evaluation of Network-Wide Anomaly Detection

Fernando Silveira^{†*} Christophe Diot[†] Nina Taft^{*} Ramesh Govindan[‡]

[†]Thomson ^{*}University of Paris VI ^{*}Intel Research Berkeley

[‡]University of Southern California

Thomson Technical Report
Number: CR-PRL-2008-09-0004
Date: September 1, 2008

Abstract: While network-wide methods have become popular in the anomaly detection literature, there has been no quantitative evaluation of the advantage of such methods. In this paper we provide preliminary results of this analysis. Surprisingly, we observe that most of the anomalies found by a network-wide method are also found by a simpler single-link approach.



1. INTRODUCTION

Network operators face several challenges to keep backbone infrastructure working and clear from unwanted traffic. The main threats arise from malicious activity (e.g., botnets, worms), unexpected demands (e.g., flash crowds), and software or hardware problems (e.g., routing misconfigurations, link outages). Detecting, identifying and classifying these operational hazards is important but at the same time hard. One approach, known as anomaly detection, is to look for abnormal patterns in traffic measurements.

Within the current anomaly detection techniques, a class of methods exploit measurements from multiple vantage points. These are called *network-wide detection* methods [3, 6, 5], as opposed to others that focus on measurements from a single link. Network-wide methods are motivated by two arguments: (1) they can exploit the *spatial* correlation in the data to find anomalies more effectively; and (2) some anomalies are inherently distributed (e.g., botnet DoS attacks) which makes them harder to spot with data from a single link.

While it is intuitive that having more information generally leads to better detections, there has been no quantitative assessment of the advantage that network-wide methods have over single-link ones. Such an evaluation is important because network-wide methods require the initial step of bringing traffic measurements to a centralized collector. The willingness of operators to cope with this measurement overhead is likely to depend on whether network-wide methods can outperform simpler approaches.

In this paper, we empirically evaluate the performance of a network-wide anomaly detection technique, namely the Kalman filter [5]. We use a one month trace of 22 links from the Internet2 backbone. We run Kalman using the data from all links at once (i.e., network-wide), and also individually for each link. Surprisingly, we observe that the overlap between the two sets of detected anomalies is large (around 90%). Moreover, while the network-wide approach detects a few more anomalies, it misses others that are only found by the single-link strategy.

The rest of the paper is outlined as follows. We present our data set in Section 2 and the Kalman filter method for anomaly detection in Section 3. We show our experimental results in Section 4. Section 5 summarizes our conclusions and discusses future work.

2. TRAFFIC DATA

Network-wide anomaly detection methods require traffic measurements collected at multiple vantage points. We use data from the Internet2 backbone¹, which interconnects many research and education networks in

¹<http://www.internet2.edu/>

the USA. Several network-wide detection methods have been evaluated using data from that network [3, 4, 5].

Our data set is composed of flow records collected from 22 backbone links in Internet2 during August 2007. Some of the links in the Internet2 topology were not included in our study because they do not show up in our trace. This may have happened either because those links were not active in August 2007 or because the flow measurement software (J-Flow² in this case) was not enabled for the associated interfaces. Packets are randomly sampled at a 1/100 rate and binned in 5-minute intervals. All IP addresses in the traces have the last 11 bits set to zero to make hosts anonymous.



Figure 1: Internet2 links used in our analysis.

From the binned traffic, we compute two types of metrics. The first one is the total number of packets in each bin, which we call the traffic *volume*. Many popular anomaly detection methods are based simply on time series of traffic volume per bin [2, 1, 5, 3]. The second type of metric we study is the *entropy* of four header features: source and destination IP addresses and ports. Those metrics have been shown to catch more anomalies than simply the traffic volume [4]. Namely, if the number of packets with feature value i (say, destination port 80) in a time bin is given by x_i , then the entropy of the feature (i.e., destination ports) is given by:

$$H = \sum_{\forall i} x_i \log x_i \quad (1)$$

3. REFERENCE DETECTION METHOD

We use the Kalman filter detection method in our evaluations. While the full details of that method can be found in the work of Soule et al. [5], it has mainly two steps which we describe in a high level. First, the parameters of a linear model are calibrated on the time series of traffic (either volume or entropy) using an EM-based algorithm. Second, the model is used to predict

²<http://www.juniper.net/junos/>

each point in the time series given the previous ones. For the time bins containing only normal traffic, the prediction errors are supposed to be normally distributed with zero mean and unknown variance. We estimate their true variance by computing the sample variance of all errors in a trace. Thus the individual errors can be normalized by this estimate to produce a series of values with standard normal distribution. We call those numbers the assessment values of the time bins.

The higher the assessment value of a time bin, the more the traffic in that bin looks anomalous to the Kalman method. Clearly, we need to pick a threshold K for the assessment values in order to decide which time bins are anomalous and which are not. Given a target false positive probability p , we select the threshold value as the standard normal percentile at $1 - p/2$.

For our purposes of performing a comparative study, we choose the threshold value to minimize the number of false positives in the detection. Therefore, we use a threshold value of six, which corresponds to a target false positive rate lower than 10^{-9} .

We detect the anomalies in the network using two approaches: (1) finding anomalies in each link in turn and taking the union of those; and (2) detecting the network-wide anomalies simultaneously using the data from all the links. We refer to those approaches as our *input strategies*. Our study consists in comparing the sets of anomalies found by each strategy.

4. EXPERIMENTAL RESULTS

Figure 2 shows the total number number of anomalies found by Kalman in the whole network, when using either single-link or network-wide data. We show the results for volume anomalies as well as the augmented set of anomalies found with volume and entropy. The plot shows that the number is anomalies found in the network is roughly the same, regardless of using network-wide data. That holds for both volume anomalies and for volume and entropy ones. This means that having access to network-wide data, does not increase the number of detections in the network. It remains to measure the overlap between the two sets of anomalies to determine if the network-wide strategy finds the *same* events as the single-link one.

For each time bin, we compare the assessment values that Kalman produces under the two input strategies. Our rationale is that if the input strategy makes no difference, the assessment values should remain roughly the same. Figures 3 and 4 show scatter plots of the assessment values across input strategies. Each point corresponds to a time bin in one of the network links. We draw a diagonal line representing the case where assessment values are independent of input strategy.

In each plot, we show lines on each axis that correspond to a threshold value of six. Remember that the

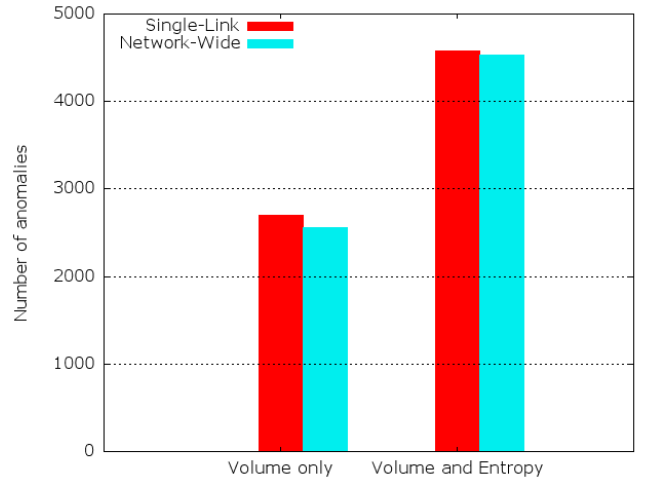


Figure 2: Number of anomalies found in each input strategy.

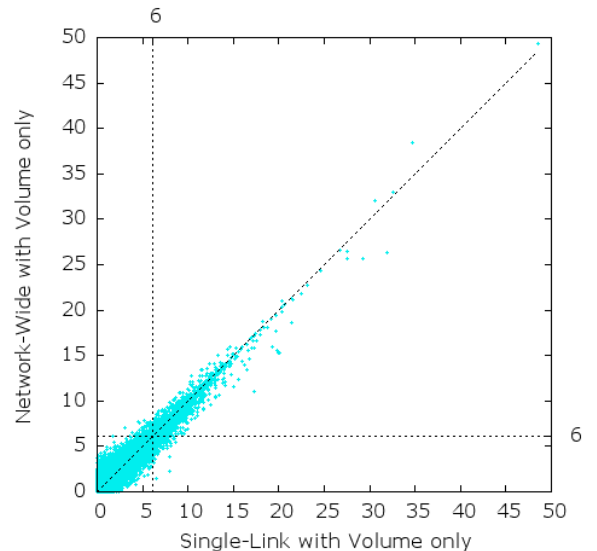


Figure 3: Scatter plot of assessment values with volume only.

assessment values are compared to the threshold to determine if a time bin is anomalous or not. Thus, points to the right of a vertical line are anomalies found by the single-link strategy, and points above the horizontal line are anomalies found by the network-wide strategy. Note that the two threshold lines divide each plot in four quadrants. The top-right quadrant corresponds to points in space and time considered anomalous by both input strategies. Conversely, the top-left and bottom-right quadrants are points respectively detected by the

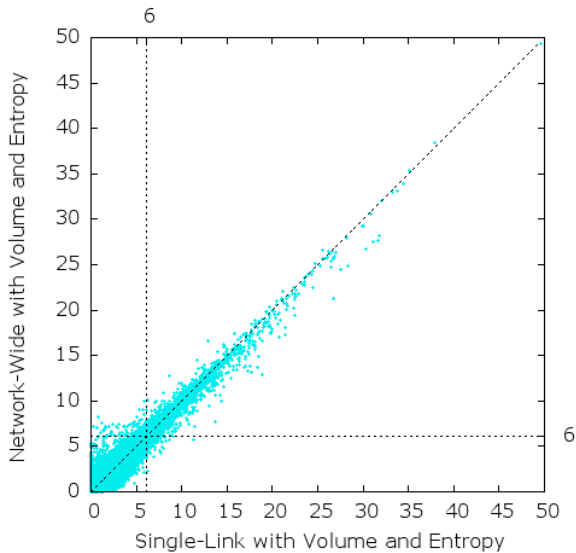


Figure 4: Scatter plot of assessment values with volume and entropy.

network-wide or the single-link strategies in isolation.

We measure the overlap between the input strategies as the fraction of network-wide anomalies which can also be found using only single-link data. The overlap is 90% among the volume anomalies, and 92% when we consider volume and entropy. This means that no more than 10% of the anomalies are missed if we constrain ourselves to a single-link method. Moreover, the single-link strategy also finds some anomalies which are ignored by the network-wide one. Namely, when only volume is used, 15% of the anomalies found by looking at single-links are not found by the network-wide approach. Thus figure lowers to 10% when entropy is also used.

In summary, our results show that the overlap between the network-wide and single-link input strategies is very high. Essentially, each strategy detects around 10% of “unique” anomalies, and overlaps with the other one on the remaining 90%. Moreover, this overlap is not very sensitive to the use of volume only or volume and entropy. We leave the investigation of the causes for these results to future work.

5. CONCLUSIONS AND FUTURE WORK

We have quantified the advantage of a network-wide anomaly detection method over a simpler single-link strategy in a one month trace of 22 links in the Internet2 backbone. We observed that the two approaches detect mostly the same anomalies, independently of whether these are detected on traffic volume or feature entropies.

This preliminary result leads to a number of ques-

tions to be addressed in future work. First, we need to determine if our results using data from the Internet2 backbone can be generalized to other data sets. Second, we would like to understand if our results are limited only to Kalman or if they extend to other methods as well, such as PCA. In summary, we need to determine which factors, related to either data or the methods, leads to detections that are unique to the network-wide strategy. We might leverage on such knowledge to design new methods that maximize the number of new detections. We intend to pursue this questions in our future work.

6. REFERENCES

- [1] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of IMW*, pages 71–82, 2002.
- [2] J. D. Brutlag. Aberrant behavior detection in time series for network monitoring. In *Proceedings of LISA*, pages 139–146, 2000.
- [3] A. Lakhina, M. Crovella, and C. Diot. Diagnosing network-wide traffic anomalies. In *Proceedings of SIGCOMM*, August 2004.
- [4] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *Proceedings of SIGCOMM*, August 2005.
- [5] A. Soule, K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In *Proceedings of IMC*, 2005.
- [6] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan. Network anomography. In *Proceedings of IMC*, pages 1–14, 2005.