Identifying statistically anomalous regions in time series of network traffic

Fernando Silveira
Thomson

Christophe Diot
Thomson

1. INTRODUCTION

Traffic anomalies are specific types of events and conditions which differ from the normal or expected network behavior, and therefore should be brought to the attention of network operators. Examples of such events include, but are not limited to link or router outages, Denial-of-Service (DoS) attacks, flash crowds, port scans and misconfigured devices.

Traffic anomalies usually cause visible changes in time series of traffic descriptors such as packet counts or entropy of IP header fields. These anomalies can be separated from daily fluctuations in traffic demands because the latter are well structured [1]. Anomalies are often a visible exception to the daily patterns.

Given the vague notion of a traffic anomaly as something unusual, the popular approach to the detection problem is to look for statistical anomalies, i.e., outliers detected by some statistical test. The main techniques proposed so far [1, 3, 4] work by filtering the predictable trends, such as periodicity, and then looking for outliers in the residual time series. Then, anomalies are the points in the time series whose distance to their baseline value is large when compared to some estimate of the standard deviation.

However, traffic anomalies last generally longer than a single time bin and current techniques fail to identify the period of time in which different yet related statistical anomalies occur. We call that an anomalous region, and it is an intermediate point between a statistical anomaly and its underlying traffic anomaly, since it provides additional information about the duration of an anomaly, but does not uncover its root cause. Intuitively, an anomalous region should contain the anomalous traffic or a condition which triggered a statistical anomaly. More formally, we seek a contiguous time interval which is anomalous with respect to its surrounding neighborhood, but whose individual time bins are consistent with each other.

Figure 1 shows a motivating example for anomalous regions. The plot focus on an 8-hour interval for a given OD-pair in the Abilene network. The four curves on top correspond to the entropy time series of source and destination IPs and ports, as originally proposed in the methodology of [3]. Three relatively long-lived traffic anomalies (each lasting for about one hour and a half) take place in this time series.

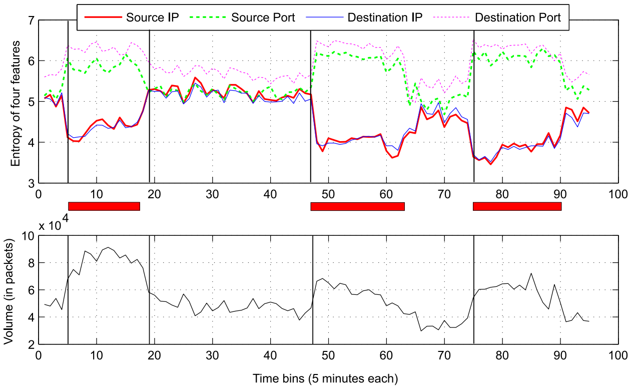

Figure 1: A recurring pattern in the time series of entropies and packets. Vertical lines denote the statistical anomalies detected by methods in the literature. The red horizontal bars highlight anomalous regions in time.

Methods like PCA [3] and the Kalman filter [4] detect basically subsets of the change points, illustrated by the vertical lines on Figure 1. Nevertheless, it is interesting to notice that the time bins in between and around those edges are also anomalous, i.e., significantly different, compared to the surrounding neighborhoods of normal behavior. In fact, given only the vertical lines in Figure 1, we cannot distinguish the start and end from each of the three visible anomalies.

Moreover, the three anomalies displayed in Figure 1 appear to have a distinguishable pattern in the way the

*CoNEXT'*07, December 10-13, 2007, New York, NY, U.S.A.
Copyright 2007 ACM 978-1-59593-770-4/ 07/ 0012 ...$5.00.

four entropies change (IP entropies decrease while port entropies increase to same levels in the three episodes). This suggests that these anomalies also share a common underlying root cause. Indeed by looking at the flow data, we have verified that a same pair of source and destination hosts adds roughly 20,000 sampled packets per time bin, using a number of TCP connections for bulk traffic.

A method that can output these regions instead of merely the change points will be useful as a preprocessing step for a more elaborate technique that seeks to pinpoint the root causes of a statistically significant change, i.e., to map the statistical anomalies back to the traffic anomalies.

Even though techniques based on multi-resolution analysis [1] can expose anomalies that occur in different time scales, in this abstract we propose an approach that can identify the length of an anomaly by looking only at the finest granularity of time series data.

2. PROPOSED METHOD

In order to make anomalous regions emerge from data automatically, we consider clustering the data points in the time series of traffic. In a general way, having n features for each time bin (among entropies, volumes, etc), we define a distance metric and partition the time bins so as to minimize the pairwise distances within a same set, and maximize them between different sets. Clustering is widely studied in pattern analysis and has applications in several fields [2].

In practical terms, consider a set of time bins, $\mathcal{I}$, where each $i \in \mathcal{I}$ is characterized by a 5-dimensional point, $x_i = \{H(\mathrm{sIP}_i), H(\mathrm{sP}_i), H(\mathrm{dIP}_i), H(\mathrm{dP}_i), N_i\}$. We denote by $H(X_i)$ the sample entropy of variable X_i measured over the time bin i, and by N_i the number of packets at time bin i.

We then apply hierarchical clustering [2] on the set of 5-dimensional points that represent the individual time bins. We use the *standardized euclidean metric* [2] as the distance metric for the clustering method. This is basically the regular euclidean distance with each squared coordinate previously normalized by its sample variance. This compensates for the differences in scale, e.g., comparing packets against entropy.

An interesting challenge comes from attempting to identify the anomalous regions simply from aggregate traffic descriptors, i.e., without looking deeper into the individual flows that are being measured. Given the fact that patterns in anomalous regions are visually noticeable, such as in Figure 1, employing a machine learning technique seems like a natural approach.

Figure 2 shows an interval containing a two-phase anomaly. Namely, a source host sends many small packets towards several ports of a target destination (the destination port entropy goes up) and later it focuses on a single port (all features concentrate). We apply hierarchical clustering to the set of points and get the regions in the bottom of Figure 2.

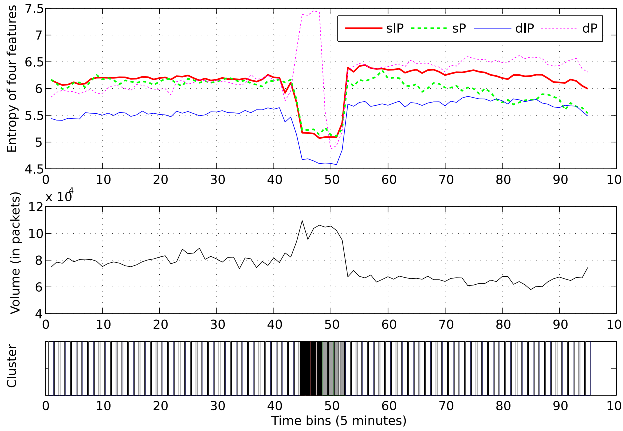

Figure 2: A two-phase anomaly and a segmentation produced by clustering.

A future improvement on such a technique will be to include some sort of temporal rule. Naturally, we expect the time bins of an anomalous region to be concentrated in time. In the examples we showed in Figures 1 and 2 we just regard time bins as loose points in a 5-dimensional space independent of the time variable.

We apply clustering as an off-line procedure, after detecting statistical anomalies to cluster time bins around the anomaly into semantically related points. One way could be starting with a small interval (say 2 hours), and applying clustering to subsequently larger intervals around the same point. At some point the anomalous region (the cluster where the original anomalous time bin fell) should stop growing, and the dominant cluster becomes that of normal traffic. In the long run, anomalies should be the exception and not the rule.

In this work we proposed automatically identifying anomalous regions, in place of statistical anomalies. This is a useful definition since it can aid in subsequent root cause analysis. We presented an approach which involves clustering points in the time series of measurements. We showed examples that encourage us to refine the method in a number of future directions.

3. REFERENCES

[1] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In *Proceedings of the ACM IMW*, pages 71–82, 2002.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[3] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *Proceedings of the ACM SIGCOMM*, pages 217–228, August 2005.

[4] A. Soule, K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In *Proceedings of the ACM IMC*, pages 331–344, 2005.