Multivariate SVD Analyses For Network Anomaly Detection

Jeff Terrell
Kevin Jeffay
F. Donelson Smith
Department of Computer Science
University of North Carolina at Chapel Hill

Lingsong Zhang
Haipeng Shen
Zhengyuan Zhu
Andrew Nobel
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

We are investigating the use of signal analysis methods for near real-time anomaly and intrusion detection. Recently, methods such as wavelet analysis [1] and principle component analysis [2-4] have been applied to network measurement data as a means for automatically detecting anomalies in networks. Anomalies have included both local events such as flash crowds as well as global events such as routing anomalies. We build on and extend these works in an attempt to automatically identify smaller scale, local anomalies such as denial-of-service attacks.

This research is work in progress; however, at present we believe there are three aspects to our work that are novel. First, we perform a principle components analysis on multivariate data rather than univariate data. A multivariate approach allows us to detect anomalies that do not have a strong signature in any of the time series of individual features. Second, we are looking at the utility of features based on entropy measures of measurement data such as packet size, source port, and destination port. Finally, we are attempting to perform these analyses with minimal delay by using analyses of past intervals (*e.g.*, the principle components) as a form of training data to tune the selection of anomalies in the current interval.

The following briefly outlines our approach and describes sample results we have obtained to date.

APPROACH

Our approach is a traditional principle components analysis of time series data using singular value decomposition (SVD). Unsampled network traffic is aggregated into bins of a certain size, such as 10*ms* or 1*s*. Then, a set of features is extracted for each bin. This results in a data matrix where columns correspond to traffic features and rows correspond to time bins. In the results presented here, six features were used: bytes per bin, packets per bin, number of active connections per bin, packet size entropy within a bin, source port entropy within a bin, and destination port entropy within a bin. For example, using a 1*s* bin size, a data matrix for an hour of measurement data contains 3,600 rows (bins) and 6 columns (features).

SVD applied to a data matrix X produces a decomposition of X into a product of matrices $U\,S\,V^T$, where U is a matrix of singular columns (or left singular vectors) u_i, S is a diagonal matrix of singular values, and V is a matrix whose transpose gives the singular rows (or right singular vectors) of X. The principal components matrix of X is simply the product of X and V (which is also equal to $U \times S$.) Each principal component PC_i is a column vector of length 3,600 representing a time-varying pattern common to several features. Also, the principal components are arranged in order, such that PC_1 is the most significant, explaining the most variation of the original data matrix, and the lower order principal components are driven more by the multivariate outliers, being more representative of anomalous behavior. We can define the significance of a principal component i as

$$W_i = s_i^2 / (\sum_{j=1}^{6} s_j^2).$$

Note that $W_1 \geq \ldots \geq W_6$ and $\sum_{i=1}^{6} W_i = 1$.

We can then select k lower-order principal components such that their total significance is maximally less than 5%. A weighted sum of squares of these selected components results in a statistical testing time series that quantifies anomalous behavior. With an underlying assumption of normality, this time series will have a marginal gamma distribution, and we can use this distribution to define an automatic threshold of a given significance level, such that points above this threshold are considered anomalous. This is just one example of a useful statistical test; others are being considered.

RESULTS

To date we have applied our methods to the analysis of traffic on the main link connecting the UNC campus to the Internet. This is a gigabit Ethernet link whose utilization ranges from 30% to 80% over the course of a day (*i.e.*, 300-800 Mbps of traffic). All of the features we have analyzed are derivable from TCP/IP packet header traces. We are currently working with consecutive 1 hour-long traces of headers from this link. At peak times of the day such packet header traces represent approximately 30 gigabytes of data. The SVD analysis of a data matrix with 100*ms* bins (36,000 rows by 6 columns) requires approximately 10 seconds of real-time on a commodity PC. The processing required for 10*ms* bins (360,000 rows) requires approximately 20 seconds. Furthermore,

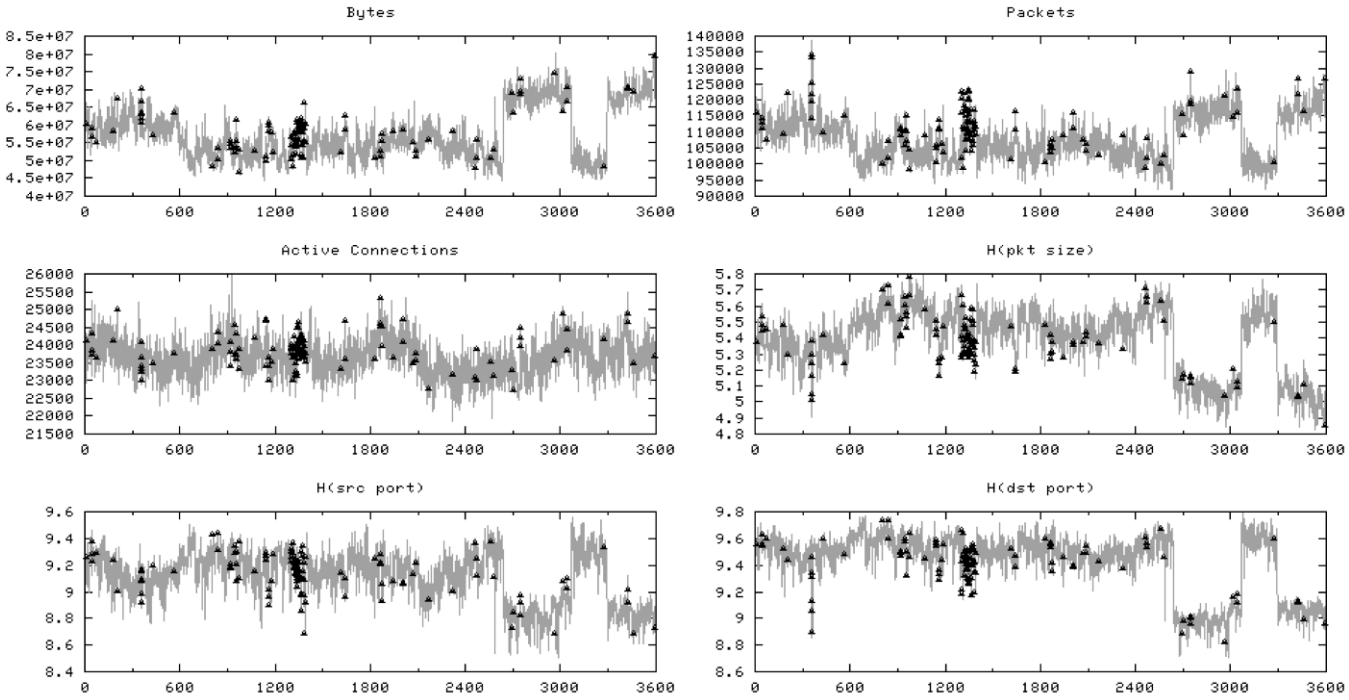

Figure 1: Measured time series of bytes, packets, active connections, and packet size, source port, and destination port entropies for a 1 hour trace and a 1 second bin size. Overlayed on this plot are the times (dots) at which an SVD analysis of the combined set of six features detected anomalies.

feature collection on a dual-core 3GHz PC can keep up with the rate of traffic. This indicates that it should be possible to perform this analysis frequently enough to discover anomalous network events as they are happening, in order to benefit network operators.

Figure 1 shows the original time series for each of the six features we currently use as well as the times at which statistical anomalies were automatically detected via an SVD analysis. A manual investigation of the trace revealed anomalies caused by single machines such as flooding mail servers as well as anomalies caused by multiple machines such as distributed port scans. Moreover, our analysis has shown that the method detects anomalies that are not apparent in any of the features individually. This implies existing methods that only analyze single features in isolation would not have detected these anomalies. For example, note the density of detected anomalies around times 1,300-1,400 seconds. In this interval none of the individual features give any obvious indication of an anomaly. However, a manual analysis of the trace uncovered a concentration of network port scans occurring.

While we believe these and similar results are encouraging, much work remains. We are still exploring the discriminating power of the features we employ and are working on assessing and characterizing both the false positive and false negative rates of this method.

ACKNOWLEDGEMENTS

This work was supported in parts by the National Science Foundation (grants ANI 03-23648, EIA 03-03590, CCR 02-08924), and the IBM Corporation.

REFERENCES

[1] P. Barford, J. Kline, D. Plonka, A. Ron, *A Signal Analysis of Network Traffic Anomalies*, SIGCOMM Internet Measurement Workshop, Marseille, France, Nov. 2002, pp. 71-82.

[2] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. Kolaczyk, N. Taft, *Structural Analysis of Network Traffic Flows*, ACM SIGMETRICS Performance Evaluation Review, Volume 32, Number 1 (June 2004), pp. 61-72.

[3] A. Lakhina, M. Crovella, C. Diot, *Characterization of Network-Wide Anomalies in Traffic Flows*, SIGCOMM Internet Measurement Conference, October 2004, pp. 201-206.

[4] A. Lakhina, M. Crovella, C. Diot, *Diagnosing Network-Wide Traffic Anomalies*, ACM SIGCOMM Computer Communication Review, Vol. 34, No. 4 (October 2004), pp. 219-230.