Using Adaptive Lossless Compression to Characterize Network Traffic

Karyn Benson and Lisa M. Marvel
U.S. Army Research Laboratory, Adelphi, Maryland 20783, Email: {karyn.benson,marvel}@arl.army.mil

Abstract—Detecting anomalies in network traffic is a challenging task, not only because of the inherent difficulty of identifying anomalies such as intrusions [1] but also because of the sheer volume of data. In this paper, we attempt to extend existing work in the field of steganalysis to the problem of detecting anomalies in network traffic. By losslessly compressing network traffic using an adaptive compression algorithm, we postulate that it is possible to characterize normal network traffic. Once typical traffic has been defined, it is possible to identify anomalous traffic as the traffic that does not compress well.

I. SUMMARY

In the area of steganalysis, lossless compression was used to model images and detect the presence of steganography [2], [3]. The compression was designed to compress images without steganography. Images that contain steganography could be winnowed because they did not compress well using such compressors. Eiland, et. al. [4] used a compression to calculate an estimate of information distance. Using this distance they were able to reliably detect systems scans in network traffic. Here we propose to combine these ideas to use compression to characterize network traffic. Deviations from the characterization are anomalies, which can assist in intrusion detection or identifying changes in the allocation of network resources (useful from a quality of service standpoint).

The basic approach is to construct an adaptive lossless compressor for certain network traffic. The compressor must have the potential to use and improve a pre-loaded dictionary. When a particular string of characters appears frequently, it is added to the dictionary. Items that occur less frequently will eventually be replaced. These additions and deletions are called dictionary updates. Once a steady state dictionary has been defined, drastic changes in the number of updates will signify a deviation in network traffic characteristics. The cause of the anomalous behavior can be determined by investigating the data that triggered the dictionary updates.

In this paper, we present our initial results demonstrating performance and correlation with actual network operations.

II. PRELIMINARY RESULTS

For our lossless compressor, we used an implementation of the Lempel-Ziv-Welch (LZW) with a least-recently used (LRU) replacement strategy. A dictionary size of $2^{16}$ entries was employed so that items in the dictionary could be encoded using two bytes. We collected and compressed netflow data [5] from a single collection site for a 48 hour period. The average network traffic rate for the period sampled was 200 megabytes per hour. A graph of compressor dictionary updates over a 48 hour period is shown in Figure 1. Data collection started at 9pm on Sunday and ended 9pm Tuesday.

During the first hour the percentage of new dictionary entries is 100% indicating the initial dictionary is being built. (This data point was removed from the graph for readability.) The number of dictionary updates starts at a fairly constant rate, until seven hours into our traffic (2 am). Presumably the increase corresponds to when batch processing and updates are performed. Eleven hours into our traffic (6 am) there is another increase; this can be attributed to the beginning of the workday. In both cases, a change in network operations has occurred and dictionary updates during that time could be used to investigate the source of the traffic anomaly. Twenty-four hours later the behavior is repeated to a lesser extent. The lower percentage of updates may be attributed to this day also being a weekday and some of the network flows that occurred on Monday morning remained in the dictionary during the compression of Tuesday’s traffic.

Since, it is plausible that changes that are function of operation scenarios - such as traffic that corresponds with business hours versus non-business hours. This characteristic could be quantified by building and pre-loading dictionaries.
specific to the collection site, day of the week and period of day. We can use the compression ratio and frequency of updates as a gauge to determine the boundaries of operation scenarios for dictionary construction. Once these dictionaries are defined, we expect the number of overall updates to decrease, leaving the non-characteristic anomalous network traffic to remain as dictionary updates.

III. CONCLUSIONS AND FUTURE WORK

In the paper, we have presented preliminary results for a method that will use lossless compression for network traffic characterization. We believe an increase in dictionary updates can be correlated with a change in network traffic behavior.

For future work, we intend to explore the following research questions:

- What are the best boundaries for the dictionary definitions?
- What is the true and false positive anomaly detection rate?
- What side information could be used to diminish the false positive rate of anomaly detection?

REFERENCES