Identifying Anomalous Network Traffic Through the Use of Client Port Distribution

By Josh Goldfarb
US-CERT

December 2005

Executive Summary

This particular approach to IP flow analysis examines server ports (0 to 1023) and the client ports that exchange flows with those server ports. This analysis operates under the assumption that for each server port, the number of flows from each port chosen by client machines should be relatively uniform. In other words, similar numbers of flows from each of the chosen client ports to a given server port are expected. If a large deviation from the norm is observed, that traffic is considered to be of interest and is flagged for further analysis. US-CERT has tested this analysis technique on a large, enterprise network with a large amount of network flow data. Details of this method of analysis are discussed in the next section of this paper.

1 INTRODUCTION

1.1 Motivation

A tremendous amount of Internet traffic passes through an enterprise network in a given day. With this large amount of data, it is difficult to separate significant security events from routine noise. Analysis methods to filter noise and bring attention to anomalous traffic are valuable when analyzing IP flow data. By examining client port – server port pairs, US-CERT has developed an effective approach to analyzing IP flow data.

2 METHOD DETAILS

The analysis method follows three steps. These steps are described here and have been scripted to conduct this analysis in an automated fashion.

Step 1: Group and Count Client Ports

In step one of the analysis, IP flow data is searched. Data from the desired time period is pulled from all IP flow data and set aside for further analysis. The data is then grouped by server port. For each server port, the client ports that exchanged flows with that server port are grouped, and the number of flows is summed. Traditionally, client ports range from 1024 to 65535, however all ports, 0 to 65535, are considered when conducting analysis. This is because this analysis method allows for discovering certain malicious activity taking place between the lower ports, such as Operating System (OS) fingerprinting, which is usually accomplished from port 0 to port 0.

Step 2: Calculate Mean and Standard Deviation for Each Server Port

In the next step of the analysis, statistics for each server port are calculated. The sample mean of client port flows is calculated, along with the standard deviation of the sample population. These statistics are calculated for those server ports for which there are flows from more than C client ports, where C represents the number of client ports for which there are flows to a given server port. $C = 5$, which was chosen through experimentation, seems to work well. $C < 5$ tends to give a large number of false positives, while $C > 5$ does not significantly reduce the number of false positives and may in fact eliminate some traffic of interest.

Step 3: Determine Traffic of Interest

In the final step of the analysis, traffic of interest is identified and further analysis is conducted on that traffic. Client ports with flow counts greater than S standard deviations above the sample mean are considered traffic of interest. $S = 3$, which was chosen through experimentation, seems to work well. $S < 3$ tends to produce a large number of false positives due to its tendency to include data that is considered to be within the bounds of expected variation. $S > 3$ tends to eliminate much of the traffic that should probably be flagged as traffic of interest.

A sample output of client port – server port pairs is provided in Appendix A. The analyst can use this output of client port – server port pairs to query the data set aside earlier. The data

returned by this query contains only flows exchanged between a given server port and the client ports identified as suspect by the analysis method. This results in a compact dataset from which the analyst can draw conclusions.

3 DISCUSSION

3.1 Results

This method of analysis has shown promise. The method has allowed analysts to pull out suspicious traffic from a very large dataset in a matter of minutes. The entire process described above is automated, freeing the analyst from the details of the method and allowing the analyst to focus on analyzing the traffic of interest. The method does, however, return some false positives. Parameters such as C and S can be modified as necessary to reduce the number of false positives or reduce the volume of traffic of interest that is discarded because it did not meet the given threshold.

3.2 Further Exploration

Further exploration into this method is currently being pursued. One way to improve this method is by introducing statistical rigor to the analysis. A more sophisticated statistical analysis would provide insight into how to reduce the number of false positives and ensure that a higher percentage of traffic of interest is identified. Future enhancements to this analysis might also include automating the follow-on queries that are run when traffic of interest is identified. This could free the analyst even further, allowing the analyst to concentrate on analysis rather than remembering syntax. The ultimate goal of this analysis method is to improve its accuracy and sophistication in order to improve the discovery of suspicious IP flow data, while reducing the number of false positives generated.

APPENDIX A: SAMPLE OUTPUT OF CLIENT PORT – SERVER PORT PAIRS

Port 37

sPort | count

60414 | 249

3897 | 428

60417 | 459

Port 161

sPort | count

1075 | 154

1026 | 235

Port 445

sPort | count

2272 | 37

4965 | 37

3123 | 38

2265 | 40

3122 | 44