An Analysis Of Internet Scanning

Thomas Dooner, Department of EECS; **Brian Stack**, Department of EECS; **Mark Allman**, Adjunct Faculty, Department of EECS

Abstract

On the Internet, automated hosts continuously scan wide swaths of networked computers. These scanners oftentimes are searching for vulnerable hosts, creating maps of Internet address space, or attacking hosts. When these scanners make connections, logs are made at the entry point of the scanned host's network. Using historical Internet traffic data from the Lawrence Berkeley National Lab, we analyze connection logs to determine scanning trends.

By analyzing the distribution of scanned hosts, scans of a singular host's ports, scan activity over time, and other properties, we hope to gain a deeper understanding of the motivation of these scanning hosts. This will allow us to then analyze current detection techniques and determine if potential for improvement exists. More accurate filters will provide substantive network security and speed benefits.

Background

With the growth of the number of hosts on the Internet, a corresponding growth in automated scanning for vulnerabilities has occurred. Since 1994, the beginning of our dataset, scanning volume has increased exponentially.

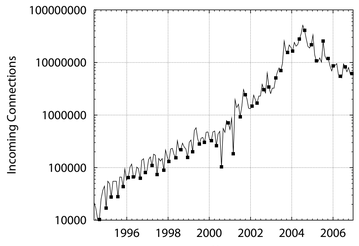
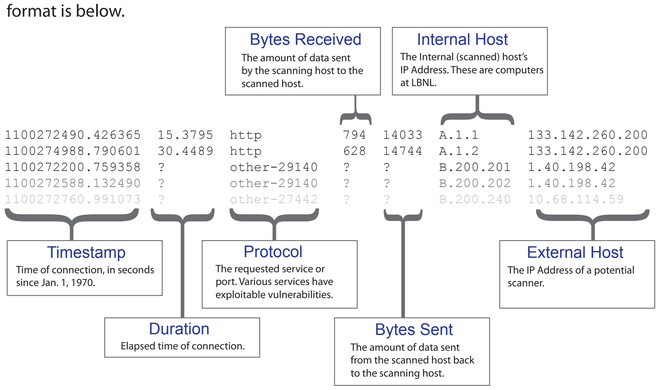


Figure 1: Scanning Rate As A Function Of Time[1]

A server today might have many web-accessible servers running: HTTP servers, E-mail servers, FTP servers, and many other services provide attack vectors for remote hosts. Scanners, seeking to exploit vulnerabilities in these services, scan wide swaths of the Internet's roughly 2^{32} (4.2 billion) total addresses. As connection speeds increase and botnets become more prevalent, sophisticated distributed scanners could potentially survey the entire address space looking for vulnerable hosts.

Procedure

Since 1996, all connections to Lawrence Berkeley National Labs' approximately 130,000 IP addresses have been recorded. Log data is stored in one file per day in the BRO log format developed by the International Computer Science Institute (ICSI). These files, heavily compressed, reach into the hundreds of megabytes. To analyze this data, we generated a series of Python tests to run on a small sample of these log files -- the first day of every month. Prior work[1] has established a threshold for determining if an external host is a scanning device. With this insight, we can filter all irrelevant results from the dataset and run complicated analyses. A brief overview of the BRO log format is below.



Results

First, we analyzed data from individual scanners over the entire dataset in an attempt to categorize scanners based on behavior. A cumulative distribution function plot of total Unique Internal Hosts scanned for each scanner yields a natural break at 250 internal hosts scanned per scanner. Approximately 92% of scanners fall below this threshold and 8% of scanners fall above this threshold.

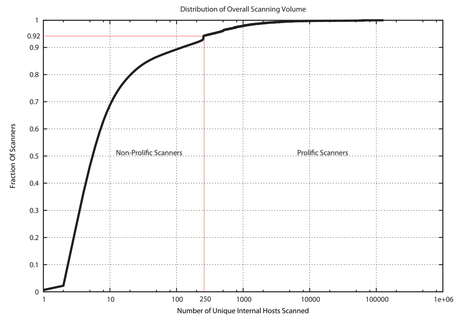


Figure 2: Characteristic-Based Division Of Scanners

Having defined 250 unique internal hosts as the threshold, we will refer to scanners which scan more internal hosts to be "Prolific" scanners and scanners which scan fewer to be "Non-Prolific" scanners.

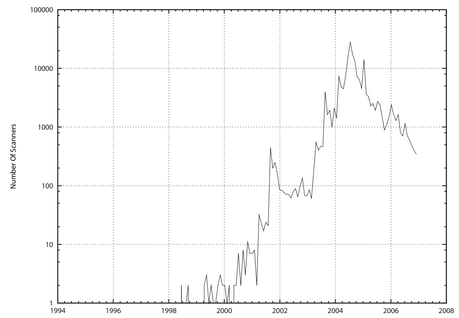


Figure 3: Prolific Scanners Over Time

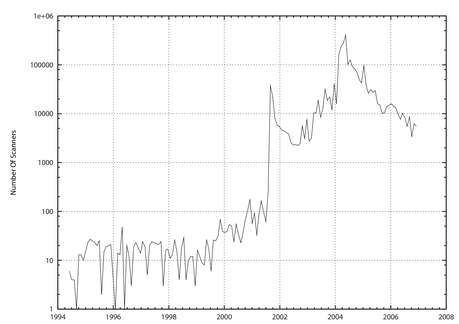


Figure 4: Non-Prolific Scanners Over Time

It can be seen that at any given time, there are approximately ten times as many non-prolific scanners as prolific scanners. The peak in 2004 is consistent across all our data for scan trends.

In addition to categorizing the scanners by scanning intensity, we can also categorize them by the fingerprint of their scanning activity. For now, we have identified two categories for scanners.

A **linear scanner** is one which follows a well defined pattern in selecting the IP addresses to scan. The change from one address to the next is a constant factor. An example of a linear scanner can be found in the sample scanning section. By using a linear regression, we can fairly easily detect any such scanning behavior. We classified a scanner as linear if the coefficient of determination (R^2) is greater than 0.99. This allows us great breadth in catching scanners into this category -- as scanners with both positive and negative slopes can be determined, as well as any scanners which scan every other internal host, or every third internal host.

A **random scanner** is a scanner that we determine to be following no set pattern for selecting the IP address to scan. The change from one address to the next can not be predicted, and the distribution fits a uniform model. To determine uniformity of the connections made by a scanner, we used the Anderon-Darling test[2]. Generally the Anderson-Darling test is used to find normal distributions, but we have implemented it to search for a uniform distribution of internal hosts.

An **unclassified scanner** is a scanner that has not yet been determined to be linear or random. Over time we will attempt to find a classification for all scanners that have a great enough number of targeted IP addresses.

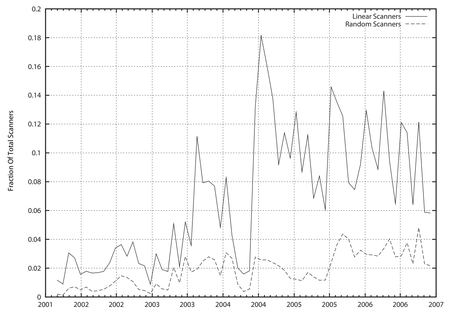


Figure 5: Categorization of Scanners Over Time

Since 2001, our thresholds determine there to be more linear scanners than random scanners. However, it is important to note that the y-axis of this graph peaks at 19%, hinting that there are a vast quantity of unclassified scanners that need further analysis.

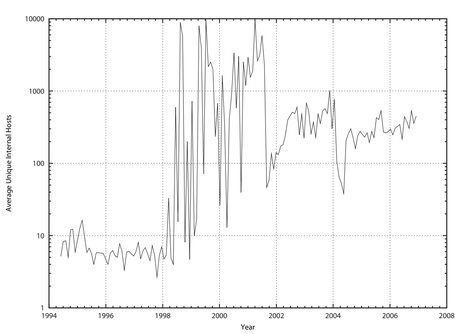
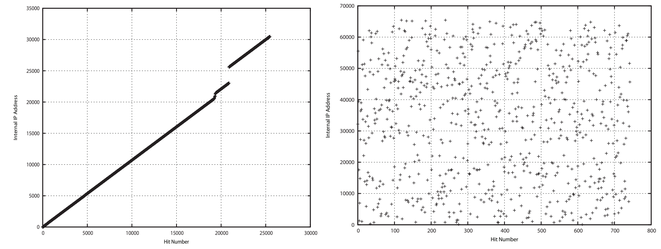


Figure 6: Average Internal Hosts Per Scanner

Sample Scanners



Figures 7 and 8: Two Example Scanners (one linear and one random)

Above are two visual representations of a scanner's activity. The x-axis is the chronological ordering of hits and the y-axis represents the Internal IP Address scanned (for convenience, these are converted to integer form).

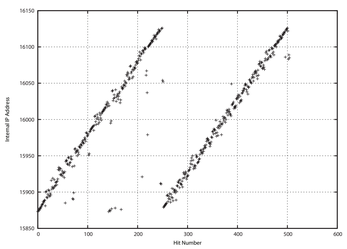
The left graph represents a linear scanner -- as the hits continue chronologically, the Internal IP Address continues to increase by a constant amount.

The right graph, however, represents a random scanner -- one where the hits follow no predictable pattern.

Conclusions

We feel that we have made good progress in finding our results. With a certain degree of certainty, we can give classifications to two groups of scanners, linear and random, each of which consistently make up a considerable fraction of scanning over time. The tests we used are a good starting point for further exploration of the data sets we are analyzing, and will provide a solid foundation for future work.

Future Work



The long term goal of our research is to find new and interesting patterns in network scanning (such as the pattern depicted above). The next step towards this goal will be to look into the unclassified category of our scans and, through visual inspection, find new patterns to search for in addition to the linear and random scanning. Furthermore, our techniques for classifying scanners can be improved with better statistical methods. Finally, work can be done to explore more of the data available in the logs to give a more comprehensive analysis of the scanners.

References

[1] Mark Allman, Vern Paxson, Jeff Terrell. A Brief History of Scanning. ACM SIGCOMM/USENIX Internet Measurement Conference, October 2007.

[2] "Framework for IP Performance Metrics". p. 33. RFC 2330.