A Tale of Two CTs: IP Packets Rejected by a Firewall

George Corser
Oakland University
168 Dodge Hall
Rochester, MI 48309
989-780-3168
gpcorser@oakland.edu

ABSTRACT
Two distinct curve tendencies (CTs) characterize the flow of IP packets rejected by a firewall from specific source IP addresses. One flow model appears relatively flat and steady over time. The other manifests as a single sharp spike. This study examines a recent real-world firewall log which exhibits these two patterns.

Categories and Subject Descriptors
C.2.0 [**Computer-Communication Networks**]: General – *Security and Protection*

K.3.2 [**Computers and Education**]: Computer and Information Science Education – *Curriculum*

General Terms
Security

Keywords
Firewall log analysis, visualization, spreadsheets

1. INTRODUCTION
They come in less at times. They come in bursts at times. IP packets rejected by firewalls may conform to one of two curve tendencies (CTs): the steady flow, or the one-time gush.

The distinction is critical to firewall performance analysis. For example, suppose a firewall rejects a large number of packets from a specific IP address during a given 24-hour period. Examining the average number of packets per source IP address may not improve understanding of the firewall activity if all of the packets happened to arrive in a single 5-minute burst. A brief explosion of rejected packets from a single IP address may indicate a denial of service (DOS) attack, where a steady stream of rejected packets may suggest that an attacker is systematically scanning for available open ports.

This paper reports the extent to which a recent real-world firewall log exhibits these two disparate patterns. Data are available at http://secs.oakland.edu/~gpcorser/firewall_data. IP addresses have been altered or anonymized for security reasons in some cases.

Rather than using specialized commercial software, this study employs mathematical and visualization techniques commonly available in spreadsheet application programs. Because of their broader availability and lower cost, spreadsheet analysis methods may better serve educators, students and informal network analysts.

Information Security Curriculum Development Conference 2012
October 12-13, 2012, Kennesaw, GA, USA.

2. BACKGROUND
This study examines firewall logs exclusively. Computer system log analysis exists in many forms, however. Oliner, Ganapathi and Xu provide a broad overview of the general topic [1].

2.1 Firewall Logs
A firewall is a piece of computer hardware or software. It controls the flow of network traffic, specifically internet protocol (IP) packets, from the un-trusted internet to the trusted local area network (LAN). A firewall resides between the internet and a LAN, and allows acceptable IP packets to pass through to the LAN, while discarding unacceptable IP packets.

A firewall log is a computer file. A firewall can be configured to store a message in a firewall log for every packet processed, or for a subset of the packets processed. In this study, the firewall log contains only information regarding rejected IP packets. However, sometimes it is useful to record approved packets.

Firewalls may process hundreds of thousands of IP packets per second. Logging every processed packet may sometimes be cost prohibitive in terms of processing time and storage space for the log. Consequently, system administrators often limit the amount of information recorded in logs.

Table 1. Firewall log (excerpt).

Date and Time	Message[1]
5/8/2012 11:27am	%ASA-4-106023: Deny tcp src dmz:173.244.163.189/2383 dst dmz:0.0.0.0/445 by access-group "acl_out" [0x0, 0x0]
5/8/2012 11:27am	%ASA-4-106023: Deny tcp src dmz:195.96.235.54/3861 dst dmz:0.0.0.0/445 by access-group "acl_out" [0x0, 0x0]
5/8/2012 11:27am	%ASA-4-106023: Deny icmp src dmz:128.68.102.107 dst dmz:0.0.0.0 (type 3, code 3) by access-group "acl_out" [0xe92546aa, 0x23864e5d]

Messages produced by the firewall and posted to the firewall log require interpretation. They may or may not provide insight regarding the particular problem being researched. (See Table 1.)

[1] Certain IP addresses (labeled: 0.0.0.0) have been anonymized at the providers request.

IP packets provide to the firewall only a limited amount of information, such as the nature of the data contained in the packet and the source and destination IP addresses. The firewall maintains an internal clock which allows it to know with relative precision the time and date a given IP packet was processed, but this clock must be synchronized periodically to verify that the firewall's time is accurate. Further, any or all of the information in an IP packet may be "spoofed," faked or forged, which means information from the firewall log cannot be accepted as fact without verification.

Attackers, well-aware that log analysis may be used by defenders, may modify attacks to avoid rejection by a firewall or to avoid detection by log analysis. Attackers may use spoofing, or they may spread attacks over time, to prevent the firewall from rejecting packets that arrive too closely together from the same source IP address.

Not all packets rejected by a firewall represent prevented attacks. Programming errors and even proper and normal network activity may sometimes be rejected by a firewall from time to time for a wide range of reasons.

2.2 Source Data

This report presents data from a firewall log of a publicly-traded "diversified international merchandising and marketing services company." [2] The company operates in the United States, Mexico, Canada, Romania, Turkey, Greece, South Africa, India, China, Japan, Australia. The corporation provided a log file for a firewall which rejected 165,308 IP packets during an approximately 24-hour period, from Tuesday, May 8, 2012 at 10:46am through Wednesday, May 9, 2012 at 11:12am.

2.3 Visualization Using Spreadsheets

Each organization faces different security concerns at different times. Consequently, firewall protection must be tailored to the particular institution and the specific situation. Log analysis cannot always be automated, so both technical and non-technical people may examine firewall logs. "Since humans will likely remain a part of the process of interpreting and acting on logs for the foreseeable future, advances in visualization techniques should prove worthwhile" [1].

The organization which kindly provided the data that is the basis of this study utilizes sophisticated tools to examine its network performance—much more sophisticated than what is available in typical spreadsheet programs. Commercial software packages offer especially advanced visualization of log data. However, these packages can be costly and may require special expertise to install and use. Spreadsheets, on the other hand, are more commonly available to educators and students alike and offer relatively advanced tools for mathematical and graphical analysis. Spreadsheets also offer tools for customized programming, which allow examiners to present the data in a wide variety of views.

3. METHOD

The analysis followed this procedure: First, parse the log file to obtain source IP addresses. Second, create two histograms of the data—one showing rejection activity in one-hour time intervals, the other showing it in five-minute time intervals. Third, plot the time intervals between packets. Fourth, identify the source IP addresses most frequently rejected by the firewall. Finally, plot the patterns of rejection over time.

Table 2. Firewall log (excerpt, parsed).

Date and Time	Rejection Code	Type	Source IP Address
41037 .47769676	%ASA-4-106023	tcp	173.244.163.189
41037 .47769676	%ASA-4-106023	tcp	195.96.235.54
41037 .47770833	%ASA-4-106023	icmp	128.68.102.107

3.1 Parsing Source Data

The firewall log was opened in Microsoft Excel, which was used to parse specific data columns from the log rows. Consider Tables 1 and 2. Table 2 shows the same information as in Table 1, with Date and Time formatted to display a more detailed reading for the rejected packet (fraction of a day). Each Message was parsed into Rejection Code, Type and Source IP Address.

3.2 Creating Histograms

Two histograms were created. Figure 1 shows IP packets rejected by the firewall, by time of day, in time increments of one hour. Figure 2 shows the same information in five-minute increments.

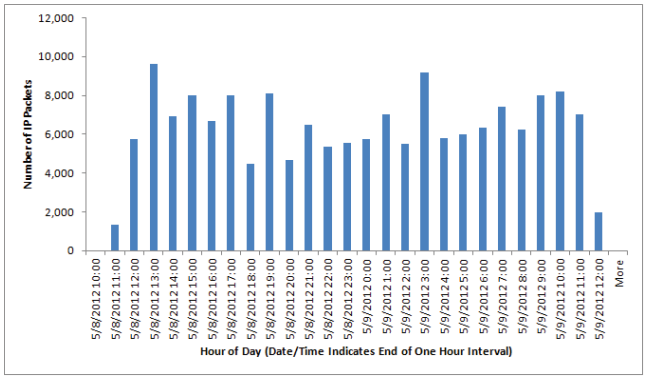

Figure 1. IP packets rejected by firewall, by time of day, in time increments of one hour.

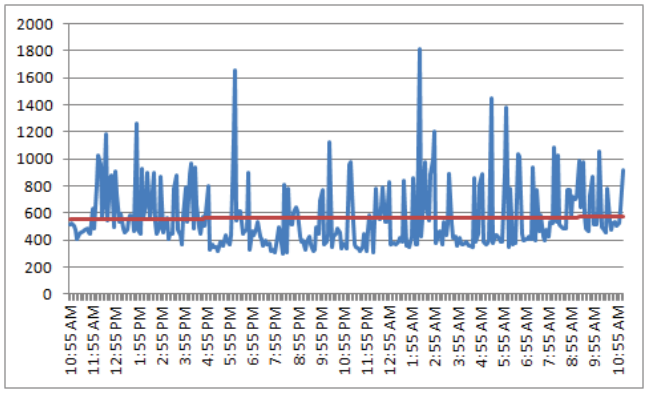

Figure 2. IP packets rejected by firewall, by time of day, in time increments of five minutes.

3.3 Plotting Time Intervals between Packets

Figure 3, a histogram, shows the number of packets rejected within a certain time interval after the prior rejected packet. Note Figure 3 uses a logarithmic scale on the y-axis.

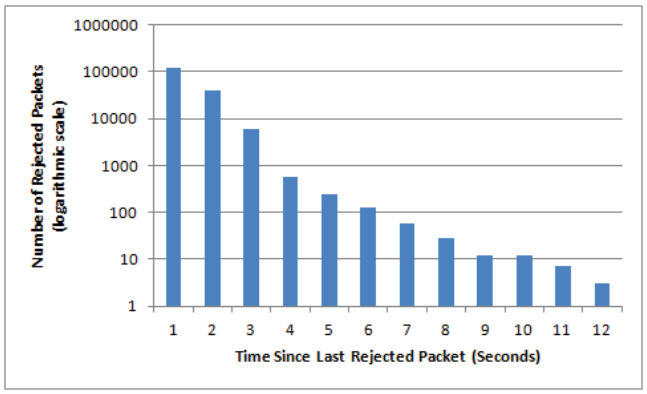

Figure 3: IP packets rejected in a 24-hour period, by time interval since prior rejected packet (logarithmic scale).

3.4 Identifying Source IP Addresses

A pivot table was prepared from the raw data, sorted by frequency, the number of times a packet from a given source address was rejected. The most frequently rejected source IP address is listed at the top, and the rest follow in descending order. See Table 3.

Table 3. IP packets rejected by firewall, by number of rejections and 5-minute time intervals in a 24-hour period.

Source IP Address	Intervals	Rejections
None	286	9745
0.0.0.0 (anonymized)	291	4769
205.171.93.37	238	1659
122.224.35.103	6	1569
203.140.9.1	146	1465
72.189.241.191	88	1323
189.203.203.1	111	1226
65.55.34.208	169	1009
65.55.90.21	169	1008
65.55.34.81	169	1005
All other (33253 other source IP addresses)	All intervals (Max: 291)	140530

3.5 Plotting Rejection Patterns

The most frequently rejected packets were reviewed to determine the pattern of rejection. The patterns for the top two rejected IPs identified in Table 3 are shown on the left hand side of Figure 4. These represent the "steady flow" pattern.

The data for the highest two peaks in Figure 2 revealed two IP addresses that accounted for the majority of the firewall rejection activity in those two particular five-minute time intervals. The rejection patterns for these addresses are displayed on the right hand side of Figure 4. These represent the one-time burst pattern.

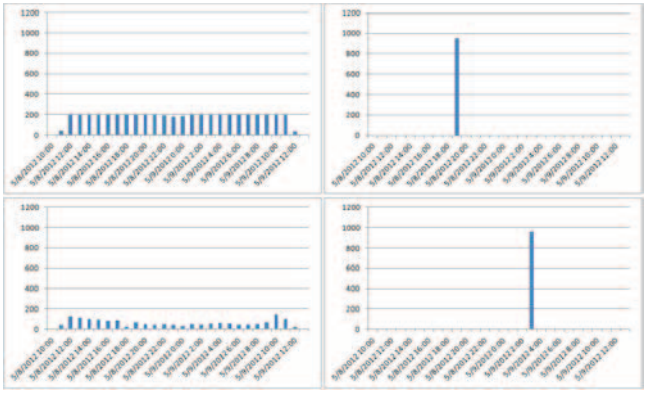

Figure 4. Rejection patterns for four source IP addresses, 0.0.0.0 (upper left), 205.171.93.37 (lower left), 78.188.170.222 (upper right), 78.131.58.184 (lower right).

3.6 Spreadsheet Limitations

Depending on the particular spreadsheet software, log analysis may be limited by maximum worksheet size and memory constraints of the computer on which the analysis is performed. If a firewall log exceeds (or approaches) these maximum sizes, spreadsheet analysis may be impossible (or impractical). In this study the size of the log file was well within the constraints established by spreadsheet applications.

4. RESULTS

Patterns of rejected source IP addresses often, but not always, fell into one of the two curve tendencies. There were 291 five-minute time intervals in the data. Excluding the nonexistent IP address and the anonymized IP address in Table 3, only 63 source IP addresses were rejected more than 291 times. In other words, packets from only 63 IP addresses could conceivably have been rejected in all 291 five-minute time intervals.

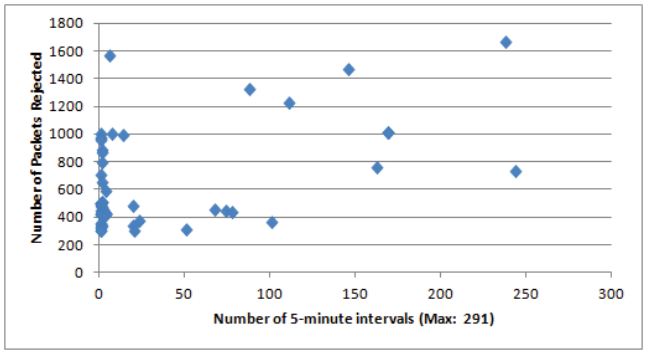

Figure 5. Number of packets rejected from specific source IP addresses, versus the number of 5-minute time intervals in a 24-hour period (max: 291).

Figure 5 plots the number of rejected packets for specific source IP addresses, versus the number of 5-minute intervals in which those packets were rejected, again, for the 63 most rejected source IP addresses, not including "none" and IP address 0.0.0.0. Note the visible gap: no source IPs were rejected in more than twenty-five (25) and less than fifty (50) 5-minute time intervals. Either IP packets were rejected in many time intervals (>50), or few (<25).

The gap was even wider when the number of packets rejected for a source IP exceeded 600, roughly double the number time intervals. Again, either IP packets were rejected in many time intervals (>87), or few (<15). See Figure 5.

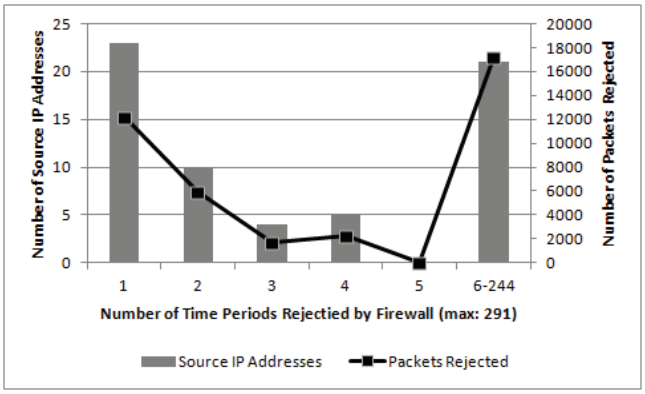

Figure 6. Number of source IP addresses and number of rejected packets versus number of rejections by firewall.

Figure 6 further reveals this phenomenon. The bars represent the number of source IP addresses (left vertical axis) with rejected packets. There were roughly the same number of source IPs that were rejected in only one time period as there were source IPs that were rejected in many (>5) time periods.

The line in Figure 6 represents the number of packets rejected (right vertical axis). Roughly the same number of packets were rejected over a long period of time, over (>5) time periods, as were rejected in single short bursts (in 1 time period).

How is this possible? The rightmost bar and line represent hundreds of time periods; the leftmost only one. This was done deliberately to show the disparate nature of the single-burst and the steady flow curve tendencies; that is, to show that precious few IPs experienced rejected packets in a medium number of time intervals.

4.1 Parsing Issues

All but 9,745 (5.9%) of 165,308 messages from the firewall log could be parsed to determine the source IP address. The other messages fell into a range of categories. Messages with no source IP address were excluded from our results so they have no bearing on the results or conclusions of this study, except for a slight effect on figures showing percentages of packets in relation to the total number of firewall log entries.

4.2 Histogram Issues

For the time interval histograms in this report, the choice of one-hour and five-minute time intervals was arbitrary. This research simply shortened the time intervals until a new pattern appeared, which was a fundamental component of the research method.

Figure 2 shows the same information in Figure 1, except using time increments of five minutes instead of one hour. The horizontal line in Figure 2 displays a linear regression. The Adjusted R Square was -0.00313236, indicating near-zero chance that the variance of the histogram is explained by the input variable. The Significance F was 0.762710387, indicating a 76.27% chance that the firewall output was a result of random chance. While the linear regression does not prove a pattern, the histogram suggests a baseline of rejected traffic underneath the regression line, and distinct spikes—not a purely chaotic fluctuation—above the regression line. This is the starting point for investigating the possibility of two types of rejections.

4.3 Time Intervals between Packets

The plot in Figure 3 shows the number of packets rejected by the firewall after a certain time, t, the time since the prior packet was rejected. For $t < 1$ *sec*, 117,934 packets (71.34%) were rejected, out of a total of 165,308. For *1 sec* $< t <$ *2 sec*, 40,301 packets (24.38%) were rejected. Only 125 packets (0.000756, 0.0756%) were rejected after time periods greater than 7 seconds since the prior rejected packet. This rapid drop suggests an exponential distribution of time intervals between rejected packets, as in a Poisson process. Figure 3 at first glance appears to display a negative exponential relationship between the number of rejections and the time intervals between rejections.

However, more thorough inspection reveals that the relationship is better described using a Weibull distribution. Rockwell Arena's Input Analyzer calculated the Kolmogorov-Smirnov p-value (<0.01), indicating that the null hypothesis (that the data conform to an exponential distribution) is extremely unlikely. But the p-value for the K-S Test for the Weibull distribution computed a much higher p-value (> 0.15), indicating that the null hypothesis cannot be rejected. The Weibull distribution, by the way, represents a special case of the generalized extreme value (GEV) distribution, which supports the main point of this paper: the data contain two divergent patterns.

One may ask, are the data random? Do the data represent an independent and identically distributed (IID) random variable? These questions can be debated. Network activity has dependencies. A three-way-handshake from a suspicious IP may not manifest in a single rejected packet, where an isolated ping might. The data do not exhibit perfect properties of IID random variables. Consider that there were 3,888 IPs with only a single rejected packet, and 20,720 with two rejected packets; 1,313 with three rejections and 3,987 with four. There appears to be a disproportionately high number of even numbers of rejections for the less-frequently rejected source IPs.

Recent studies such as [5] have reported success in analyzing network traffic patterns using multi-modal probability models under the IID assumption. This study does not address such fundamental questions as the randomness of the underlying data. The point, again, is simply that two curve tendencies manifested in this data from this particular firewall log.

4.4 Identifying Source IP Addresses

The number of total packets rejected and the number of time intervals chosen constrained the examination of source IP addresses. This study could only compare those source IP addresses with at least as many packet rejections as there were time intervals. Of 33,262 total IP addresses, only 63 satisfied this condition.

4.5 Plotting Curves

Figure 5 perhaps best shows the nature of the IP packets rejected by the firewall for the most-rejected source IPs. Steady-flow patterns involving many (>50) time intervals may represent one cluster. Short-burst patterns occurring over (<20) time intervals may represent another. The data show clear, though not necessarily perfect, adherence to the divergent patterns.

4.6 Other Limitations of the Results

The data do not suggest that firewalls always face dramatically different packet rejection patterns from source addresses over five-minute intervals. They merely show that the firewall under examination exhibited these curve tendencies on the particular day studied.

It is assumed that there was no spoofing activity that would have concealed less divergent patterns. This assumption is reasonable because the vast majority of rejected packets (93%) were flagged as attempted port scans. Port scans require conversations between clients and the server so even spoofed port scans would be expected to have the same conversational pattern, i.e. the same number of activity per time interval.

It is possible to spoof IP addresses while performing port scans, but this is usually done as cover. The spoofed port scan activity is used to conceal the real port scan activity, but the real port scan source IP address is in there somewhere. Anand, et. al. identified a possible solution to this problem by grouping source IP addresses by exhibiting similar patterns: "To avoid spoofing which may happen when an attacker IP address deliberately makes benign IP address to appear as scanner IP address to Scan Detection System, aggregate of destination IP addresses are also stored in the *Aggregate List*." [4] The point is that the spoofed pattern is similar to the genuine pattern, so the phenomenon of disparate curves is not affected.

Table 4. IP packets rejected by firewall, by rejection reason.

Rejection Code	Frequency
%ASA-4-106023 (port scan)	152956
%ASA-4-733100 (denial of service)	7188
All other (10 other rejection codes)	5164

The next-most-common rejection code was for "burst rate exceeded," indicating a possible denial of service (DOS) attack. This accounted for only 4% of the rejections. Since roughly half of the rejections, much more than 4%, conformed to the one-time burst flow model, clearly some source IPs with packets flagged as port scans had rejections in short bursts, while others had them in steady streams.

5. CONCLUSION

The firewall log examined in this report revealed very few source IP addresses with a high number of rejected packets in an intermediate number of time intervals. That is, packets from highly-rejected source IPs were either rejected in very few time intervals, or in very many.

Firewall log analysis remains a rich research area. Whether further study of these disparate patterns will lead us to enhanced security techniques remains to be seen. But we hope it will be a far, far better thing we do, than we have ever done; leading to a far, far better understanding than we have ever known.

6. REFERENCES

[1] Oliner, A., Ganapathi, A., and Xu, W. 2012. Advances and Challenges in Log Analysis. Communications of the ACM. 55, 2 (Feb. 2012), 55-61. DOI= http://doi.acm.org/10.1145/2076450.2076466

[2] Google Finance. http://www.google.com. Accessed: July 7, 2012.

[3] Kim, S. H., Wang, Q. H., Ullrich, J. 2012. A Comparative Study of Cyber Attacks. Communications of the ACM. 55, 3 (Feb. 2012), 55-61. DOI= http://doi.acm.org/10.1145/2093548.2093568

[4] Anand, T.; Waghela, Y.; Varghese, K. "A scalable network port scan detection system on FPGA," Field-Programmable Technology (FPT), 2011 International Conference on. pp. 1-6, 12-14 Dec. 2011. DOI= 10.1109/FPT.2011.6132712

[5] Mao, X.; Cheng, H.; Long, F.; Yang, A. "The Statistics and Analysis of Packet Traffic Distribution Characters in Access Network Based NetMagic Platform". Consumer Electronics, Communications and Networks (CECNet), 2012 2nd International Conference on. pp. 1468 -1472. Apr. 2012.