DARPA-2009 Intrusion Detection Dataset Report

Manaf Gharaibeh* Christos Papadopoulos*
**Colorado State University*

1 Introduction

The DARPA 2009 intrusion detection dataset is created with synthesized traffic to emulate traffic between a /16 subent (172.28.0.0/16) and the Internet. The dataset spans a period of 10 days between the 3rd and the 12th of November of the year 2009. It contains synthetic HTTP, SMTP, and DNS background data. The dataset comprises a variety of security events and attack types. This includes denial of service attacks and worms that are parametrized to exhibit various propagation characteristics. Our goal in this work is to characterize the dataset traffic and topology, and to provide an overview of many interesting security events and their interactions throughout the dataset.

Real traffic is typically preferable over a synthesized one for networks research and tools evaluation. However it is difficult to convince ISPs and organizations to provide their traffic to researchers. Mostly because of customers privacy concerns and their reluctant to reveal private information about their own business. It is also easier to include specific attacking scenarios in synthesized traffic that might not be available in real traffic. This dataset aims to present a basis for researchers and developers in the area of network monitoring and intrusion detection to evaluate and compare their approaches.

The previous DARPA dataset of 1999 was intended to simulate an Air Force base connected to the Internet [4] [3]. The 1999 dataset was designed such that some of the days do not have any attacks in order to train a detection system on normal behaviour. On the other hand, in the 2009 dataset all days include variety of attacks. The 1999 dataset was criticized for a number of issues. McHugh criticized many aspects of the dataset, mainly, the lack of statistical evidence that it is similar to real Air Force base traffic [6]. A concern shared by Mahoney and Chan [5] who criticized the dataset for having simulation artifacts. These artifacts include small range of attributes values compared to real traffic. The attributes include remote IP addresses, TTL, TCP options, and TCP window size. Such artifacts have effects on the evaluation of IDSs, especially the anomaly based detectors. They claim that idiosyncrasies of the traffic simulation could help identifying some intrusions. We did not investigate if the 2009 dataset has similar issues, that remains as a future work.

The dataset comprises about 7000 pcap files with around 6.5TB of total size. The size of each of the pcap files is just less than 1000MB. Each file typically covers around one to two minutes time window depending on the traffic rate. Different tools were used to analyze the dataset. The main tools used were tcpdump [1] and Argus tools [2]. We found it convenient for our analysis to aggregate the data in the pcap files into per day flows[1]. We used the *argus* tool to do that[2]. We then used a number of the Argus tools to analyse and generate statistics about the dataset.

The remainder of this report is organized as follows. In Section 2 we describe the dataset traffic and topology. Section 3 presents an overview of many security events found throughout the dataset. We finally conclude in Section 4.

2 Dataset Topology and Traffic

In this section we provide a detailed description for the DARPA 2009 dataset topology and traffic. As we do not have information about how the dataset was generated and what settings were used in the process, we try to extract this information from the dataset itself.

The DARPA 2009 dataset emulates a /16 subnet connected to the Internet. Figure 1 shows a simplified testbed for the dataset observed based on information extracted from the dataset traffic. The network 172.28.0.0/16 serves as this local subnet. A relatively small portion of the traffic does not involve addresses from network 172.28.0.0/16 but includes traffic within the subnet 192.168.61.0/24. The MAC addresses 0:21:56:ef:bc:0 and 0:13:80:5c:32:c0 are the interfaces through which the traffic between local network 172.28.0.0/16 and the rest of the Internet goes through and is captured by the outside sniffer. The MAC address 0:21:56:ef:bc:0 appears as the source MAC for packets coming out of the local network. It also appears as the source MAC address for the IP 192.168.61.5, which is sending BGP requests to IP 192.168.32.254. It also appears as the source MAC address for IP 192.168.61.66 packets, which are mostly ICMP unreachable messages sent out to non local IPs.

Table 1 lists observed MAC addresses in the 3rd of November (i.e. day 1) traffic. The table shows which IPs use these MACs as source addresses along with the total number of source and destination packets. The rest of the days follow a similar pattern. Note that the first two MAC addresses correspond to traffic related to subnet 172.28.0.0/16, while the others largely correspond to subnet 192.168.0.0/16 traffic. Except for the IP addresses 192.168.121.50 and 192.168.121.51 (both are DNS servers), the remaining 192.168.61.0/24 IPs are only exchanging packets internally.

Table 2 shows the number of observed subnets in IPv4 network classes A, B, and C.

[1] We chose days boundaries to be based on the timestamps in the pcap files names, which is UTC-5.

[2] We used the *argus* command with options *A* and *m* to preserve application bytes count and MAC addresses information

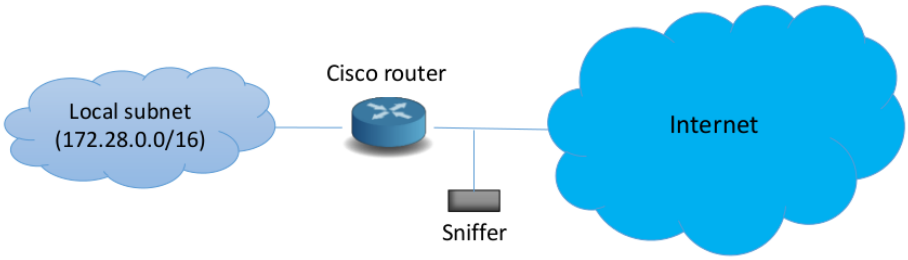

Figure 1: A simplified testbed setup that shows the local subnet 172.28.0.0/16 connected to the Internet. The local network internal traffic is not captured. Traffic between the local subnet and the Internet is captured via the outside sniffer.

Table 1: Day 1 MAC addresses grouped based on router manufacturer and IP addresses using them as sources.

MAC address	Src. addresses or prefixes using it	spkts	dpkts	Router manufacturer
00:21:56:ef:bc:00	172.28.0.0/16	283538089	413109181	Cisco Systems
00:13:80:5c:32:c0	Random Internet hosts	13879181	7823952	Cisco Systems
00:21:9b:fc:19:68	192.168.61.193	521670	427104	Dell Inc
00:1c:23:2b:17:84	192.168.61.197	179854	181534	Dell Inc
00:1e:4f:3e:45:1f	192.168.61.208	50998	11438	Dell Inc
00:1e:4f:3e:1a:ae	192.168.61.216	6026	4644	Dell Inc
00:1e:4f:3e:43:19	192.168.61.209	2640	6	Dell Inc
00:11:43:3d:95:97	192.168.61.200	1654	390	Dell Inc
00:1b:21:23:b9:90	192.168.61.194	1820501	1116748	Intel Inc
00:1b:21:3a:79:d5	192.168.61.196	85972	72842	Intel Inc
00:1f:f3:56:02:61	192.168.61.198	268966	182940	Apple Inc
00:1e:c2:f:a0:fc	192.168.61.201	54520	33894	Apple Inc
00:1a:70:14:d6:4a	192.168.61.214	5122	418	Cisco-Linksys, LLC
00:17:3f:9c:23:76	192.168.61.221	3832	3252	Belkin Corporation
00:1b:78:1f:49:78	192.168.61.202	2064	0	Hewlett Packard
00:50:4:64:27:e1	192.168.61.215	664	0	menicx international co., ltd.
00:10:60:33:9e:79	192.168.61.220	416	416	billionton systems, inc.

These are outside subnets that have at least one IP contacting our local subnet. The local subnet 172.28.0.0/16 comprises at least 138 class C prefixes. Each of them has at least one IP address that completed a 3-way handshake routine.

More than 21 thousands unique IPs were found to belong to the local subnet 172.28.0.0/16 in the 10 days of the dataset. On the other hand around 53 thousands unique IPs from

Table 2: Total number of outside subnets of network classes A, B, and C contacting the local subnet 172.28.0.0/16.

Subnet class	Number of network addresses
A	211
B	3544
C	3822

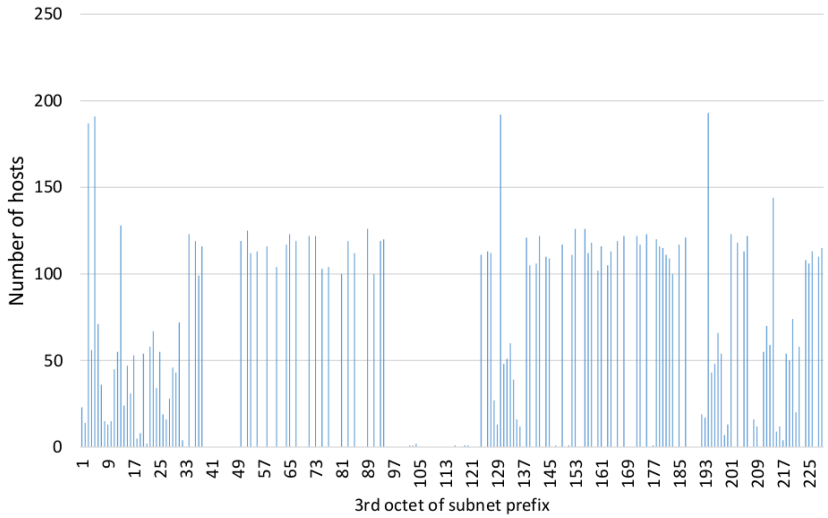

Figure 2: Hosts distribution under 172.28.x.0/24 subnets

outside the local subnet were observed for the same period. Just over 15 thousands local subnet IPs completed at least one 3-way handshake routine. Figure 2 shows the distribution of these hosts under the local slash-24 subnets. On the other hand, less than 18 thousands outside IPs were observed to complete at least one 3-way handshake routine. The hosts that completed 3-way handshake were responsible for more than 99.9% of the overall exchanged traffic. The total local subnet egress traffic volume observed over the 10 days is about 304GB, while the total ingress volume is about 5.89TB. Figure 3 shows the egress and ingress packet rate for the Local subnet 172.28.0.0/16 over the 10 days of the dataset. Figure 4 shows the corresponding bit rate over the same period. The packet and bit rates follow similar pattern from one day to another[3].

Figure 5 shows the aggregate volume in MB for ICMP, TCP, and UDP protocols per day. Other types like LLC and ARP are grouped together under *Others* category. As expected, TCP traffic volume is dominating other protocols followed by UDP. Table 3 shows the corresponding protocols' total number of packets per day, which again shows the dominance

[3]The time-line of the figures is UTC based, but recall that the dataset pcap files names correspond to UTC-5 timestamps

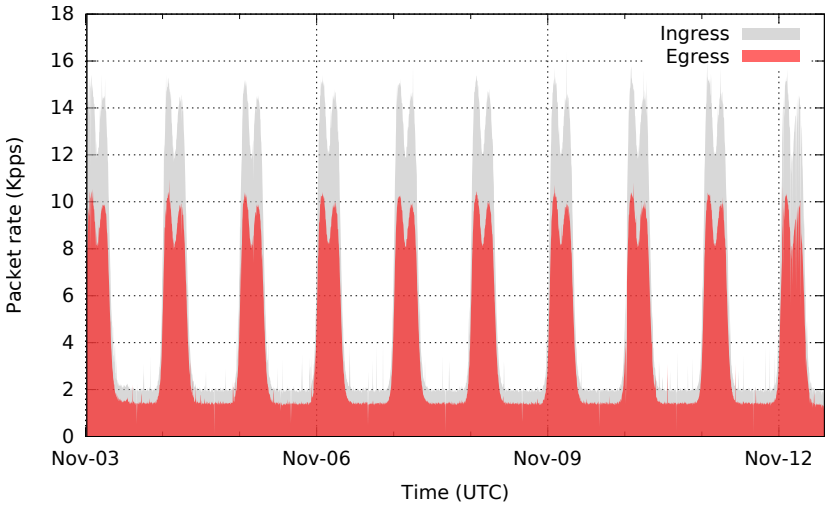

Figure 3: Local subnet 172.28.0.0/16 egress and ingress packet rate averaged over 5 minutes for all 10 days of the dataset.

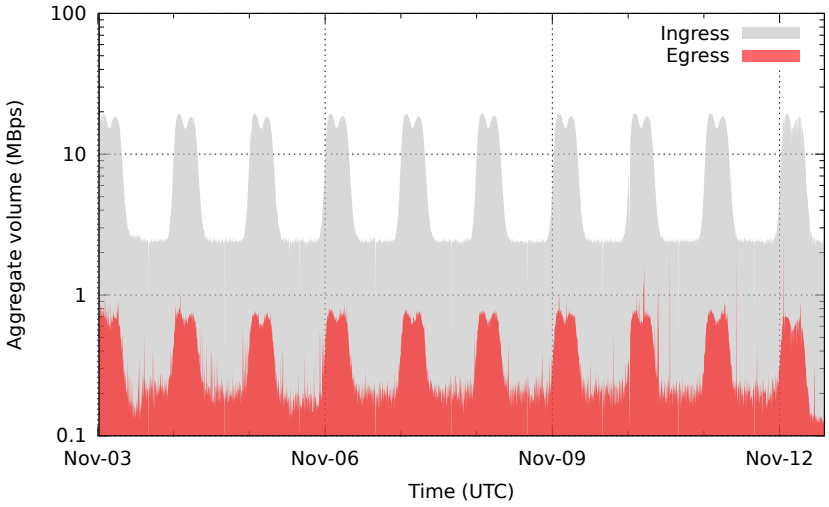

Figure 4: Local subnet 172.28.0.0/16 egress and ingress bit rate averaged over 5 minutes for all 10 days of the dataset.

Table 3: Protocols total number of packets per day.

Day	TCP	UDP	ICMP	Others
1	717911413	5229397	231980	314555
2	852676006	7593376	265857	481377
3	841793276	6996274	201786	474987
4	848319106	6419499	230163	475836
5	845163202	5559046	53246	477636
6	843108583	5544557	55763	473985
7	854003668	5951268	181063	476157
8	838681848	6517993	109259	470312
9	852333274	6258941	92599	477072
10	801158996	6425367	580217	439142

of TCP protocol.

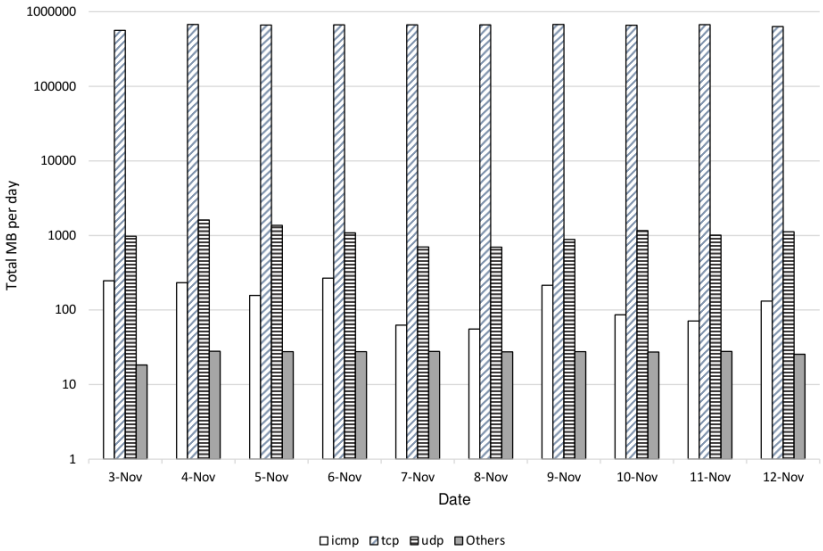

Figure 5: TCP, UDP, ICMP and other protocols aggregate volume (total MBytes per day).

We analysed UDP and TCP traffic to observe which services are more prominent. Table 4 shows a list of well known UDP ports referenced in the dataset. The list is sorted on the total number of packets sent to those destination ports throughout the 10 days of the dataset. The DNS on port 53 is the most dominant UDP application, a 3 order of magnitude ahead of the next one (NetBIOS Name Service). Table 5 does the same for TCP well known ports. HTTP on port 80 dominates all other applications followed by SMTP on port 25.

Table 4: UDP well known ports observed in the dataset and sorted on total number of packets sent to these ports throughout all days of the dataset.

Destination port	Service	Packets to port
53	Domain Name System (DNS)	35077369
137	NetBIOS Name Service	72518
123	Network Time Protocol (NTP)	31688
138	NetBIOS Datagram Service	16222
164	cmip-agent	12612
67	(BOOTP) Server, also (DHCP)	2166
547	DHCPv6 server	560

Table 5: TCP well known ports sorted on total number of packets sent to these ports throughout all days of the dataset.

Destination port	Service	Packets to port
80	Hypertext Transfer Protocol (HTTP)	3254848674
25	Simple Mail Transfer Protocol (SMTP)	46973422
22	Secure Shell (SSH)	17107432
21	FTPcontrol	10848271
445	Microsoft-DS SMB file sharing	3053938
113	identAuthentication Service/Identification Protocol	1500704
139	NetBIOS NetBIOS Session Service	618140
443	HTTPS (Hypertext Transfer Protocol over SSL/TLS)	12274
179	BGP (Border Gateway Protocol)	5293
23	Telnet protocolunencrypted text communications	2149

We next characterize the most common UDP and TCP applications. We identified prominent servers for Domain Name System (DNS), Simple Mail Transfer Protocol (SMTP), and HTTP. Only those that established connections were considered in the analysis. All subnet 172.28.0.0/16 DNS queries have the IP 172.28.10.6 as their source address, indicating that it is responsible for getting answers from DNS servers outside the local subnet[4]. A list of these DNS servers sorted on packets they received throughout the dataset period is shown in Table 6. Tables 7 and 8 show similar information about SMTP servers found outside and within the local subnet respectively. Finally, we found thousands of hosts within and outside the local subnet contacted on TCP destination port 80, Table 9 shows the corresponding statistics for local and outside HTTP servers that established at least one connection. Most of the HTTP traffic was between local hosts connecting to outside HTTP servers.

[4]The packets sniffer is set to capture traffic between the local subnet and the rest of the Internet. We do not see internal traffic within the local network

Table 6: Domain name servers outside local subnet found throughout the dataset 10 days and are sorted on packets received. All of these servers were only contacted by the local name server 172.28.10.6.

DNS server	Pkts received
192.168.121.50	3624558
192.168.121.51	3576562
192.228.79.201	1500119
192.33.4.12	1500118
198.41.0.4	1499287
128.8.10.90	1498043
192.112.36.4	1497895
192.5.5.241	1497737
192.203.230.10	1496938
128.63.2.53	1496271
192.58.128.30	1495567
192.36.148.17	1494545
193.0.14.129	1493453
199.7.83.42	1491977
202.12.27.33	1488975
128.61.199.84	81746
39.251.170.32	36179
94.191.225.83	36179
15.56.132.38	35538
152.81.216.191	30925
81.232.153.104	16606

3 Dataset Security Events

46 security events are documented throughout the the 10 days of the dataset. The ground-truth data for the security events is provided as spreadsheet file[5]. This ground-truth data contains only basic information about the events. This includes event type[6], source and destination IPs and ports, the start and end time of the event. In this section we give more details about many important security events and their interaction.

Distributed Denial of Service (DDoS) Attacks: Distributed denial of service (DDoS) attacks happen when multiple compromised systems are used to flood a target system with packets. As a result the victim target resources could be exhausted, rendering its services inaccessible. The data set comprises many DDoS attacks. All of which are targeting the destination IP address 172.28.4.7 on the HTTP port with SYN packets. A few to more than 100 different IPs are observed attacking the victim target at the same time throughout the dataset creating many SYN-flood attacks. Some IPs contribute significantly to the attacks

[5]The ground-truth data can be found using the following link: http://www.darpa2009.netsec.colostate.edu/

[6]We use the same events names from the ground-truth data in our discussion here for consistency.

SMTP server	Pkts received	Pkts sent
70.98.1.1	11933163	7389073
64.180.1.1	11728222	7204959
66.200.1.1	9293243	5710812
24.145.1.1	6405482	3939175

Table 7: SMTP servers outside the local subnet found throughout the dataset period.

SMTP server	Pkts received	Pkts sent
172.28.1.5	3249798	2380276
172.28.192.5	2816451	2071162
172.28.128.5	1545518	1050775

Table 8: SMTP servers found within the local subnet throughout the dataset period.

Table 9: HTTP servers observed within and outside the local subnet that established at least one connection. The numbers are aggregated over the dataset period.

	Instances	Src. pkts	Dst. pkts	Src. bytes(GB)	Dst. bytes (GB)
Local	1080	715172	1092977	0.1	0.22
External	5014	3214160090	4757755720	201	5985

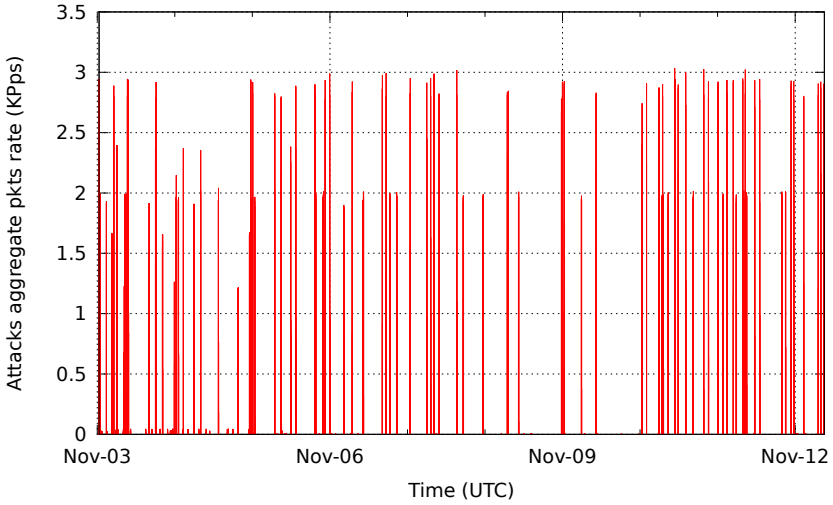

Figure 6: Syn-flood attacks aggregate packet rate averaged over 1 minute to address 172.28.4.7 throughout all the days of the dataset.

(e.g., IPs: 19.202.221.71, 19.202.221.72, and 19.202.221.73), while some other are stealthier and might be harder to detect. Each attack lasts for few minutes. Figure 6 shows the attacks' aggregate packet rate to the target victim throughout the 10 days of the dataset.

Malware DDoS Attacks: A number of local clients were compromised in the event *client compromise* via IP address 205.63.202.67. Table 10 shows a list of those clients. These clients were then used to launch *malware DDoS attacks*. Only IP address 152.162.178.254

Table 10: Local compromised hosts used in the malware DDoS attacks.

Compromised Hosts
172.28.133.12
172.28.198.166
172.28.195.194
172.28.1.203
172.28.20.99
172.28.129.124
172.28.133.86
172.28.1.134
172.28.133.158

was targeted on TCP destination port 499 (or iso-ill port) on the 4th, 5th, 6th, and 9th of November. One local source typically sends around 1MB of traffic volume per day to the victim. Multiple attacks are observed targeting the same destination in one day.

Spambots: 56 local hosts were compromised between the 3rd and 5th of November seemingly to send spam. Two outside IPs, 201.89.32.16 and 44.29.203.5, started the process. The event is labeled as *spambot client compromise* in the ground-truth data. These events were followed by the security event *spambot malicious download*, where three outside IPs, 68.91.226.37, 64.222.102.58, and 64.222.102.58 were used to download the malicious spambot program to the 56 compromised hosts, sending just over 14 KB of payload to each. After that, the compromised clients were observed in flows involving destinations, 66.200.1.1, 64.180.1.1, 24.145.1.1, and 70.98.1.1 on TCP destination port 25 (SMTP), and 77.91.104.22, 123.44.92.173 on TCP destination port 80 (www).

Scans: Under the event called *scan /usr/bin/nmap*, a number of outside IPs are scanning specific local IPs for specific ports or range of ports. IP 195.171.85.28 repeatedly scanned 13 hosts within subnet 172.28.18.0/24 for TCP ports 1257 and 3128 throughout the dataset. Port 1257 is scanned first then port 3128 within few seconds. IP 49.232.95.127 scan TCP ports 21 (ftp) and 4899 (radmin) of the host 172.28.97.169 few times every day except for the 4th and 9th of November. IPs 32.213.228.43 and 219.160.125.33 also scanned for TCP port 4899. on the other hand, the IP 211.144.70.97 scanned for TCP port 21. IP 122.201.178.72 scanned host 172.28.42.98 for ports from 64142 to 64353. IP 158.190.85.16 scanned host 172.28.116.90 for TCP port 80. IP 19.16.150.30 scanned for port 3129. Non of the scanned IPs were involved in other security events.

Under the event *failed attack or scan exploit/bin/iis_nsiislog.pl*, a number of outsiders are scanning for whole /24 subnets for TCP ports 4444 and 80. Table 11 shows each scanner and its target subnet under this attack.

The event *failed attack or scan exploit/bin/webstar_ftp_user* has two outside IPs scanning for TCP ports 4444 and ftp port 21. IP 15.174.249.80 scanned the whole 172.28.214.0/24,

Table 11: Scanners IPs and their target subnets under the event *failed attack or scan exploit/bin/iis_nsiislog.pl*. Same scanning activities are observed multiple times throughout the dataset.

Scanner IP	Scanned subnet
183.17.157.196	172.28.17.0/24
112.218.153.61	172.28.70.0/24
161.154.58.214	172.28.165.0/24
139.69.44.14	172.28.170.0/24
199.202.245.103	172.28.171.0/24

and IP 42.1.118.156 scanned the whole 172.28.5.0/24. We observe that a scanner first send few packets to a destination within the /24 subnet on the ftp port followed by packets to TCP port 4444. This is similar to the behaviour of the *worm W32.Reidana*. A worm that spreads using the MS DCOM RPC vulnerability. The worm attempts to download and execute a remote file via FTP and opens TCP port 4444.

Phishing Emails There are two separate sets of *phishing emails* events. Each of the two sets includes a sequence of events. The first set starts with the event *noisy phishing email exploit/malware/trawler*, which is only observed on the 10th of November. The outsider IP address 177.194.135.49 or (b1c28731.virtua.com.br) exchanged packets with the 3 SMTP local servers shown in Table 8 on destination port 25. Host 177.194.135.49 completes 3-way handshakes with the mail servers. Then in the event *noisy phishing email exploit/malware/trawler.pl*, 5 source IPs, 7.139.86.99, 98.127.202.173, 64.151.175.119, 205.148.165.66, and 177.194.135.49, communicate with the same mail servers on dst port 25 phishing for vulnerable clients. In the next event, *noisy client compromise + malicious download exfil*, 133 hosts from the local subnet 172.28.0.0/16, contact destination IPs 28.133.133.90, 7.139.86.99, 98.127.202.173, 64.151.175.119, and 205.148.165.66. These were the sources in the previous event except for the IP 28.133.133.90, which could be part of a stealthier attack. A malicious exfiltration code seem to be installed on the 133 local hosts in this step[7].

Two more independent security events seem to exploit the results of the above ones. In the first one, *noisy c2+ tcp control channel exfil nc*, the compromised local clients are observed sending tens of megabytes of payload to the same set of IPs from outside the local subnet. *noisy c2+ tcp control channel exfil fork* is the other event, which has only one source IP, 64.151.175.119, and 50 destinations, a subset from the compromised 133 local hosts.

The second set of *phishing emails* events starts with the event *phishing email exploit/-malware/trawler*. This event is vividly more frequent throughout the dataset compared to the noisy version in the first set of *phishing emails* events. It is observed throughout the days from the 3rd to the 10th of November. Table 12 shows a list of outside IPs for this

[7]The destination IP address 115.100.180.139 is one of the destinations of this event as indicated by the ground-truth data. However, we could not find it in the data itself. It might have been confused with the destination IP 28.133.133.90 as this one has the same sources and timestamps.

event. All of which contacted the mail servers in Table 8 on destination port 25. This is the first step to find and compromise clients and then use them in a number of attacks. The attacks involve downloading malicious code to the local hosts and exfilteration of their data. For example the following is notable sequence of events throughout the dataset:

- *phishing email exploit/malware/trawler*
- *post-phishing client compromise + malicious downlo*
- *post-phishing c2 echo*
- *post-phishing icmp exfil nc*

Another notable sequence of events:

- *phishing email exploit/malware/trawler*
- *post-phishing c2 heartbeat exploit/malware/malclie*
- *post-phishing c2 exploit/malware/malclient.pl*
- *post-phishing c2 + tcp control channel exfil explo*

There are many other related events. The following is list of them:

- *c2 + tcp control channel exfil exploit/malware/mal*
- *post-phishing tcp exfil nc*
- *post-phishing c2 + tcp control channel exfil nc*
- *c2+ tcp control channel exfil nc*
- *c2 exploit/malware/malclient.pl*
- *c2 heartbeat exploit/malware/malclient.pl*

Each of these events shares a subset of the IPs in Table 12, while the compromised clients are from the local subnet.

Table 12: Source IP addresses of the *phishing email exploit/malware/trawler* event.

Attacks outside IP addresses
24.252.33.237
23.3.252.153
130.236.192.227
138.106.196.169
82.144.22.171
195.95.157.19
43.79.200.176
130.149.81.31
212.18.56.231
62.197.83.31
34.252.80.110
212.45.250.61
16.225.81.115
44.239.223.135
115.240.12.124
116.70.60.210

4 Conclusion

We characterized the DARPA 2009 dataset background data and gave description for many of the security events within the dataset 10 days period of time. The dataset consists of synthetic HTTP, SMTP, and DNS background data. A variety of attacks spread through the dataset, including large scale DDoS attacks and worms with various propagation characteristics. While many of the attacks are targeting hosts within the local subnet 172.28.0.0/16, some of them utilizes this network local hosts to launch attacks against hosts from outside. Although the dataset is about 5 years old as of the time of writing this report, we expect the dataset would still be useful for evaluating intrusion detectors. Especially in the case of anomaly based approaches, which do not have knowledge of the attacks as in the case signature based detectors, which are more concerned about recent intrusions.

References

[1] tcpdump and libpcap. http://www.tcpdump.org/, February 2014.

[2] C. Bullard. Audit record generation and usage system (argus). http://www.qosient.com/argus/index.shtml, February 2014.

[3] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das. The 1999 darpa off-line intrusion detection evaluation. *Comput. Netw.*, 34(4):579–595, Oct. 2000.

[4] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das. Analysis and results of the 1999 darpa off-line intrusion detection evaluation. In *Recent Advances in Intrusion Detection*, pages 162–182, 2000.

[5] M. Mahoney and P. Chan. An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection. In *Proceeding of Recent Advances in Intrusion Detection (RAID)-2003*, volume 2820, pages 220–237, September 2003.

[6] J. McHUGH. Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Trans. Inf. Syst. Secur.*, 3(4):262–294, Nov. 2000.