ISOT Dataset Overview

The ISOT dataset is the combination of several existing publicly available malicious and non-malicious datasets.

We obtained and used two separate datasets containing malicious traffic from the French chapter of the honeynet project [1] involving the Storm and Waledac botnets, respectively. Waledac is currently one of the most prevalent P2P botnets and is widely considered as the successor of the Storm botnet with a more decentralized communication protocol. Unlike Storm using overnet as a communication channel, Waledac utilizes HTTP communication and a fast-flux based DNS network exclusively. To represent non-malicious, everyday usage traffic, we incorporated two different datasets, one from the Traffic Lab at Ericsson Research in Hungary [2] and the other from the Lawrence Berkeley National Lab (LBNL) [3]. The Ericsson Lab dataset contains a large number of general traffic from a variety of applications, including HTTP web browsing behavior, World of Warcraft gaming packets, and packets from popular bittorrent clients such as Azureus. We also incorporated all the datasets from the LBNL trace data to provide additional non-malicious background traffic. The LBNL is a research institute with a medium-sized enterprise network. The LBNL trace data consists of five datasets labeled D0…D4; Table 1 provides general information for each of the datasets. The recording of the network trace happened over three months period, from October 2004 to January 2005 covering 22 subnets. The dataset contains trace data for a variety of network activities spanning from web and email to backup and streaming media. This variety of traffic serves as a good example of day-to-day use of enterprise networks.

Table : LBNL datasets general information

	D_0	D_1	D_2	D_3	D_4
Date	Oct 4, 04	Dec 15, 04	Dec 16, 04	Jan 6, 05	Jan 7, 05
Duration	10 min	1 hour	1 hour	1 hour	1 hour
Number of Subnets	22	22	22	18	18
Number of Hosts	2,531	2,102	2,088	1,561	1,558
Number of Packets	18M	65M	28M	22M	28M

In order to produce an experimental dataset with both malicious and non-malicious traffic, we merged the above datasets into a single individual trace file via a specific process. First we mapped the IP addresses of the infected machines to two of the machines providing the background traffic. Second, we replayed all of the trace files using the TcpReplay tool on the same network interface card in order to homogenize the network behavior exhibited by all three datasets; this replayed data is then captured via wireshark for evaluation. Figure 1 depicts this merging process.

The final evaluation data produced by this process was further merged with all datasets from the LBNL trace data to provide one extra subnet to even simulate a real enterprise size network with thousands of hosts. The resulted evaluation dataset contains 22 subnets

from the LBNL with non-malicious traffic and one subnet (172.16.0.0/16) as illustrated in Figure 1 with both malicious and non-malicious traffic and this traffic appears to be originating from the same machines.

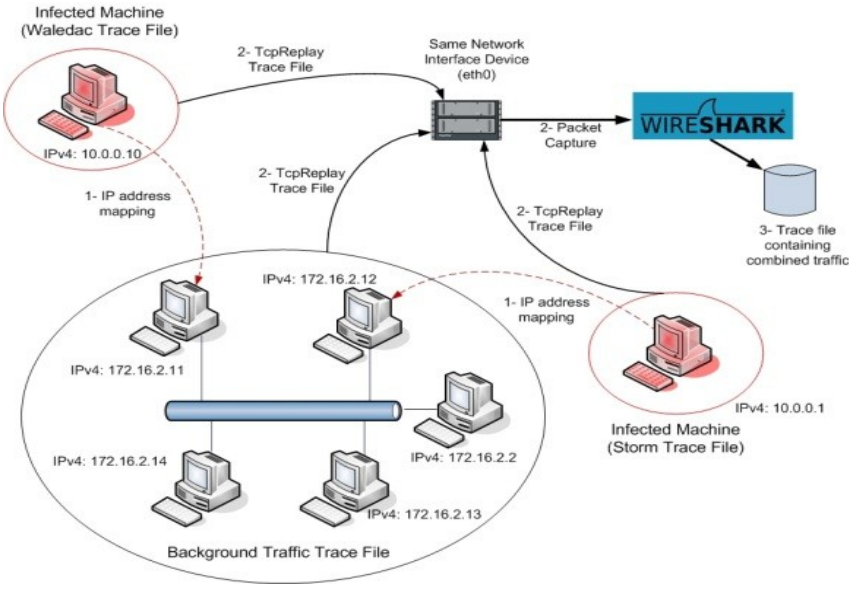

Fig . Dataset merging process

Table : List of machines that generate malicious/non-malicious traffic and corresponding labels.

IP Address	Type of Traffic Generated	Label of Malicious Traffic
172.16.2.11	Malicious/ UDP (Storm)	Src/Dst MAC BB:BB:BB:BB:BB:BB
172.16.0.2	Malicious/ SMTP Spam (Waledac)	Src/Dst MAC AA:AA:AA:AA:AA:AA
172.16.0.11	Malicious/ SMTP Spam (Waledac)	Src/Dst MAC AA:AA:AA:AA:AA:AA
172.16.0.12	Malicious/ SMTP Spam (Storm)	Src/Dst MAC AA:AA:AA:AA:AA:AA
172.16.2.2	Non-Malicious	Normal Src/Dst MAC
172.16.2.3	Non-Malicious	Normal Src/Dst MAC
172.16.2.11	Non-Malicious	Normal Src/Dst MAC
172.16.2.12	Non-Malicious	Normal Src/Dst MAC
172.16.2.12	Malicious/ Zeus	Src/Dst MAC CC:CC:CC:CC:CC:CC
172.16.2.12	Malicious/ Zeus (C & C)	Src/Dst MAC CC:CC:CC:DD:DD:DD
172.16.2.13	Non-Malicious	Normal Src/Dst MAC
172.16.2.14	Non-Malicious	Normal Src/Dst MAC
172.16.2.111	Non-Malicious	Normal Src/Dst MAC
172.16.2.112	Non-Malicious	Normal Src/Dst MAC
172.16.2.113	Non-Malicious	Normal Src/Dst MAC
172.16.2.114	Non-Malicious	Normal Src/Dst MAC

It is assumed that all the traffic from the LBNL is non-malicious. Table 2 lists the IPs of the machines in the subnet 172.16.0.0/16 that generate malicious and non-malicious traffic and Table 3 provides some statistics about the unique flows in the dataset. In addition to our labeling, the traffic from Traffic Lab at Ericsson Research in Hungary is

labeled to the level of the flow type, such as HTTP, SMTP, FTP, and etc., which does not provide any malicious traffic. Using the combination of these traffic sets, we simulate the behavior of a real world bot infected subnet to the best of our ability while at the same time taking advantage of existing, well labeled data which we use for training and evaluation purposes.

Table : Total number of unique malicious and non-malicious flows.

	Unique Flows
Malicious	55,904 (3.33%)
Non-malicious	1,619,520 (96.66%)
Total	1,675,424 (100%)

References

[1] French Chapter of Honenynet http://www.honeynet.org/chapters/france

[2] G. Szab´o, D. Orincsay, S. Malomsoky, and I. Szab´o, “On the validation of traffic classification algorithms,” in *Proceedings of the 9th international conference on Passive and active network measurement*, PAM’08, (Berlin, Heidelberg), pp. 72–81, Springer-Verlag, 2008.

[3] *LBNL Enterprise Trace Repository.* [Online] 2005. http://www.icir.org/enterprise-tracing.

To Reference this dataset use:

“Sherif Saad, Issa Traore, Ali A. Ghorbani, Bassam Sayed, David Zhao, Wei Lu, John Felix, Payman Hakimian, "Detecting P2P botnets through network behavior analysis and machine learning", Proceedings of 9th Annual Conference on Privacy, Security and Trust (PST2011), July 19-21, 2011, Montreal, Quebec, Canada”