CTU-13 Dataset Preprocessing

Source : New dataset, CTU-13-Extended, now includes pcap files of normal traffic — Stratosphere IP

Id	Duration(hrs)	# Packets	#NetFlows	Size	Bot	#Bots
1	6.15	71,971,482	2,824,637	52GB	Neris	1
2	4.21	71,851,300	1,808,123	60GB	Neris	1
3	66.85	167,730,395	4,710,639	121GB	Rbot	1
4	4.21	62,089,135	1,121,077	53GB	Rbot	1
5	11.63	4,481,167	129,833	37.6GB	Virut	1
6	2.18	38,764,357	558,920	30GB	Menti	1
7	0.38	7,467,139	114,078	5.8GB	Sogou	1
8	19.5	155,207,799	2,954,231	123GB	Murlo	1
9	5.18	115,415,321	2,753,885	94GB	Neris	10
10	4.75	90,389,782	1,309,792	73GB	Rbot	10
11	0.26	6,337,202	107,252	5.2GB	Rbot	3
12	1.21	13,212,268	325,472	8.3GB	NSIS.ay	3
13	16.36	50,888,256	1,925,150	34GB	Virut	1

Contents of

Dataset File :

- CTU-Malware-Capture-Botnet-42
- CTU-Malware-Capture-Botnet-43
- CTU-Malware-Capture-Botnet-44
- CTU-Malware-Capture-Botnet-45
- CTU-Malware-Capture-Botnet-46
- CTU-Malware-Capture-Botnet-47
- CTU-Malware-Capture-Botnet-48
- CTU-Malware-Capture-Botnet-49
- CTU-Malware-Capture-Botnet-50
- CTU-Malware-Capture-Botnet-51
- CTU-Malware-Capture-Botnet-52
- CTU-Malware-Capture-Botnet-53
- CTU-Malware-Capture-Botnet-54

Preprocessing

(Truncated) PCAP files in the extended data set extracted using geek [The Zeek Network Security Monitor]().

To prepare the data for training the files will be converted :

PCAP > ZEEK LOGS > CSV > Structured CSV > ML TRAINING

Extracted Files :

- analyzer.log
- capture_loss.log
- **conn.log**
- loaded_scripts.log
- notice.log
- packet_filter.log
- stats.log
- telemetry.log
- weird.log

conn.log fields

ts	Timestamp of first packet seen
uid	Unique connection ID
id.orig_h	Originator’s IP address
id.orig_p	Originator’s port
id.resp_h	Responder’s IP address
id.resp_p	Responder’s port
proto	Transport protocol (TCP/UDP/ICMP)
service*	Application service (http, dns, ssl) if identified
duration	Connection’s total duration in seconds
orig_bytes	Payload bytes from originator
resp_bytes	Payload bytes from responder

conn_state	Overall state of connection (e.g., ESTABLISHED, REJ)
local_orig	Whether the originator is a local host
local_resp	Whether the responder is a local host
missed_bytes	Bytes not captured due to packet loss
history	Packet-level flags indicating handshake/data flow
orig_pkts	Number of packets sent by the originator
orig_ip_bytes	Total IP-layer bytes from originator (including headers)
resp_pkts	Number of packets sent by the responder
resp_ip_bytes	Total IP-layer bytes from responder (including headers)
tunnel_parents	Reference to any parent tunnel connection (UID)

```
#truncate \\t for comma delimitted data
cat conn.log | zeek-cut ts uid id.orig_h id.orig_p id.resp_h id.resp_p proto service duration
orig_bytes resp_bytes conn_state local_orig local_resp missed_bytes history orig_pkts
orig_ip_bytes resp_pkts resp_ip_bytes tunnel_parents | tr "\\t" "," > conn.csv
```

1. Random Forest
    1. Bot IP to normal IP ratio : 7% of total

2. When training

class inbalance problem

feature selection algorithm - cfs chisquared

Merged all scenarios in to one :

Feature selection algorithm showed that even with different viruses, the most viable features were the same. Since the botnets aim to achieve infection and similar malignant behavior, they should have no prolbme merged toegerth

very good features

Revision #11
Created 2025-02-20 02:09:02 UTC by Admin
Updated 2025-04-13 17:30:14 UTC by Admin