

Malicious Encrypted Traffic Detection

HITCON CMT 2018



白貓肥宅
Aragorn
aragorn51882@gmail.com

About Me

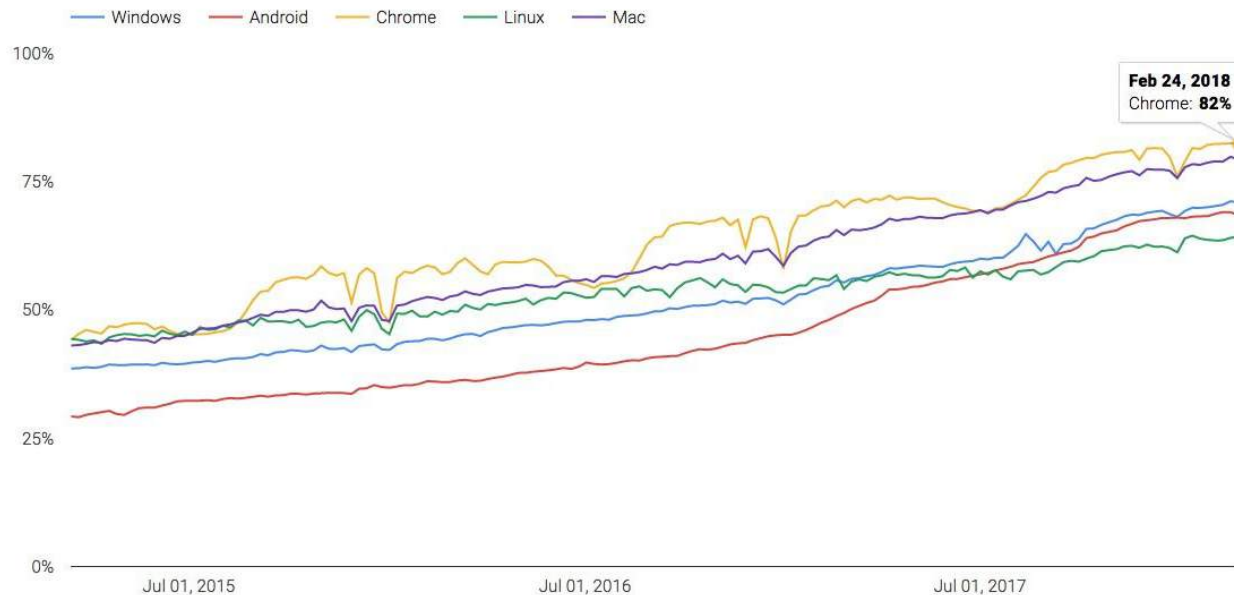
- Aragorn / 白貓肥宅
- Master of National Taiwan university
 - Security consultant in Somewhere
- NTUCSA (台大網路安全局)
- Malware Analysis、~~Operating Facebook fan page~~、
Packet Forensic、Penetration Test
- Speaker
 - 2016 TANET Network Technology Promotion Seminar -
Hacker Attack Techniques: APT Attack & Ransomware
Introduction



HTTPS Encrypted Traffic

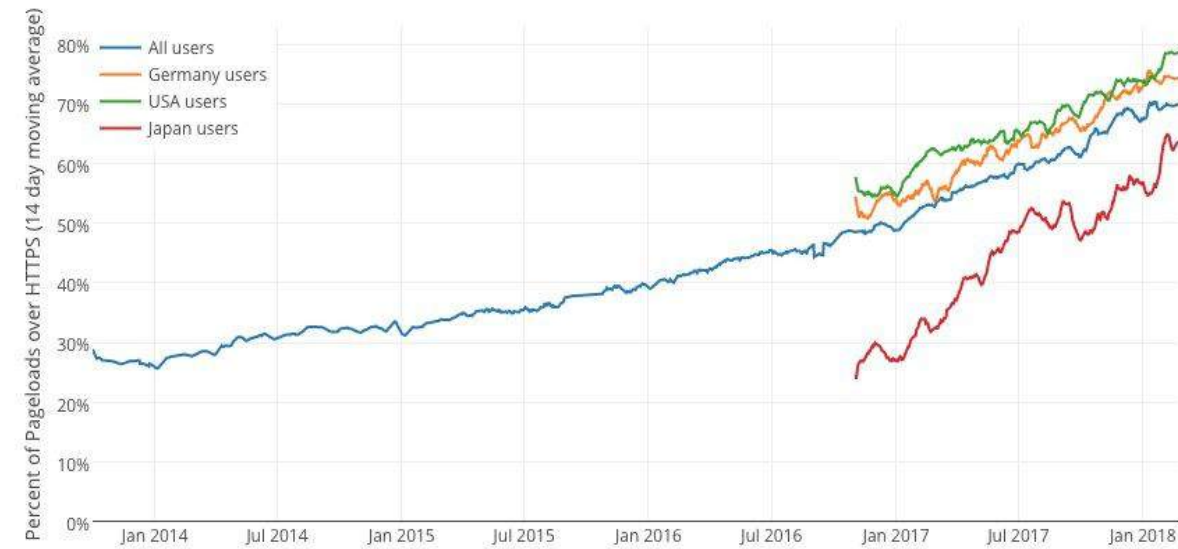
- Since the end of 2016, Google and Mozilla have released statistics, and more than **half** of their browser users have **used HTTPS** protocol encryption

Percentage of pages loaded over HTTPS in Chrome by platform



Percentage of Web Pages Loaded by Firefox Using HTTPS

(14-day moving average, source: Firefox Telemetry)



HTTPS Encrypted Traffic(cont)

- In March of this year, Cisco's latest survey found that HTTPS traffic **reached 50% in October 2017**, compared with only 38% of the overall in November 2016, the usage rate can be said to increase significantly.
- NSS Labs predict that there will be **3/4 of the network traffic** in 2019, and **encryption** will be used.

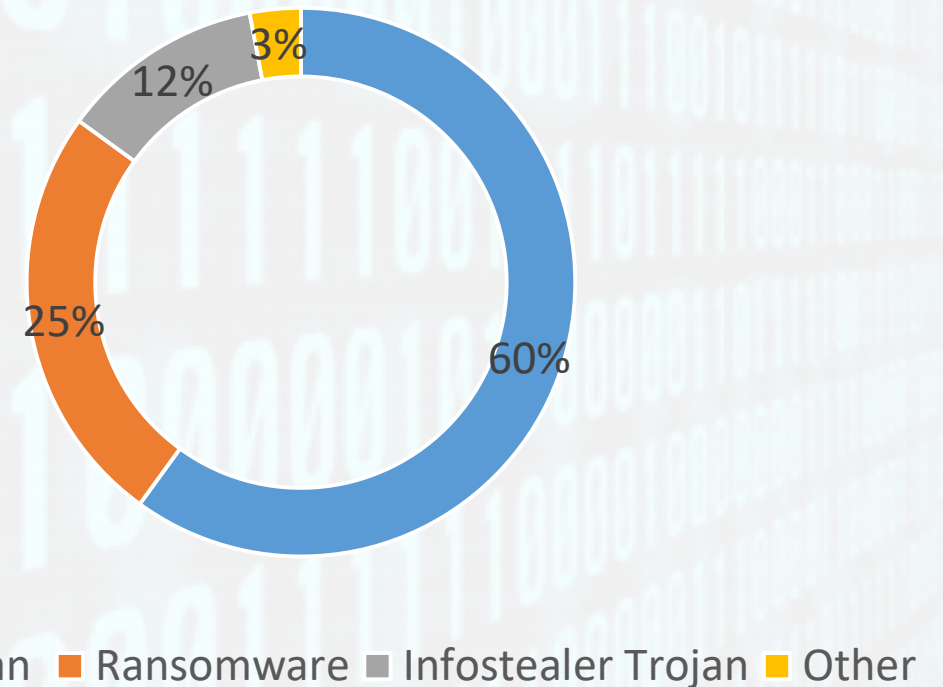


Malicious Encrypted Traffic

- According to Cisco's sampling, the proportion of malware that communicated via TLS encrypted connections **was 2.21% in 2015**, and increased to **21.44% in May 2017**.
- 10-12% of all Malware uses HTTPS
 - <https://blogs.cisco.com/security/malwares-use-of-tls-and-encryption> (Jan 2016)
- 37% of all Malware uses HTTPS
 - <https://blog.cyren.com/articles/over-one-third-of-malware-uses-https> (June 2017)
- From all HTTPS malware, 97% uses port 443, and 87% uses TLS
- In addition to TLS, SSL encryption, and technologies such as **VPN, I2P, and Tor encryption**, network security is facing great challenges.

Malicious Encrypted Traffic

- Exploit kits
 - using SSL/TLS-enabled advertising networks injects malicious scripts into legitimate websites
- Malware
- Adware
- Malware callbacks



Source : ZSCALER

Malware with Encrypted Traffic

Name	Type
Gamarue/Andromeda	Modular botnet
Sality	File infector, modular botnet
Necurs	Information stealer, backdoor, botnet
Rerdom	Click-fraud, botnet

["Dridex", "KINS", "Shylock", "URLzone", "TorrentLocker", "CryptoWall", "Upatre", "Spambot", "Retefe", "TeslaCrypt", "CryptoLocker", "Bebloh", "Gootkit", "Geodo", "Tinba", "Gozi", "VMZeus", "Redyms", "Qadars", "Vawtrack", "Emotet", "Trickbot"]

SSL Blacklist

- <https://sslbl.abuse.ch/>



SSL Blacklist

[Home](#) | [SSL Blacklist](#) | [Contact](#)

SSL Blacklist :: Home

SSL Blacklist (SSLBL) is a project maintained by abuse.ch. The goal is to provide a list of "bad" SSL certificates identified by abuse.ch to be associated with malware or botnet activities. SSLBL relies on **SHA1 fingerprints** of malicious SSL certificates and offers various blacklists that can be found in the [SSL Blacklist section](#).

If you are interested in SSL in general or you are looking for a way to implement SSL securely, you might want to have a look at the following links:

- [Qualys - SSL Server Tester](#)
- [Qualys - SSL Client Tester](#)
- [Qualys - SSL/TLS Deployment Best Practices](#)
- [BetterCrypto.org - Applied Crypto Hardening](#)
- [mbed TLS - An alternative open source and commercial SSL library \(formerly known as PolarSSL\)](#)
- [Hiawatha Webserver - An advanced and secure webserver for Unix that implements mbed TLS](#)

Below is an overview over all blacklisted SSL certificates. You can sort the list by clicking on any column title (please note that JavaScript must be enabled in your web browser in order to use this function). In addition, you can click on a SSL Fingerprint (SHA1) to receive more information about a specific entry in the SSL Blacklist.

If you are looking for a parsable format of the list below, you should take a look at [SSLBL Extended](#) (or for Dyre: [Dyre SSLBL Extended](#)).

[RSS](#) [SSBL RSS feed](#)
[RSS](#) [SSBL RSS feed \(Dyre only\)](#)

Overview of blacklisted SSL certificates (malicious Dyre C&C SSL certificates excluded):

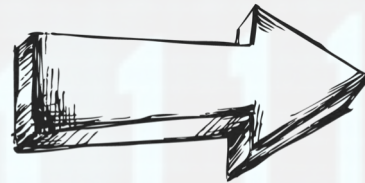
Listing date (UTC)	SHA1 fingerprint	Common Name	Listing reason
2018-07-24 13:38:13	911bfb76d3c056b8f61aece042fefccd2be0741	domain.com/O=My Company Name LTD./C=US	Godzilla C&C
2018-07-20 10:02:22	7be59f8f0811aabcb73c9f1c7df3b3e66f964ca0	C=XX, L=Default City, O=Default Company Ltd	PandaZeus C&C
2018-07-19 13:19:06	879c445c7a5b319ee04e3a1d1e3424f46b15064e	C=XX, L=Default City, O=Default Company Ltd	Malware C&C
2018-07-19 06:13:37	e9761aa8442c5a77d2d367cb6b4c5b0db97cda64	domain.com/O=My Company Name LTD./C=US	PandaZeus C&C
2018-07-17 10:57:39	4f3e38b897f1ac2dcc0e3834e3ef1d74c288f257	CN=domain.com, O=My Company Name LTD., C=US	PandaZeus C&C
2018-07-16 13:30:28	6cfebb47098abd1b3e1ecdcc14e294a3368488fa	qpdepkevla.mobi	Quakbot C&C
2018-07-16 13:30:00	e0d903bbddc642e5f7820b22d86eae9e15a7b2f8	lkbyae.org	Quakbot C&C

APT attack

- CVE-2017-0199 with abuses Powerpoint slide
 - **Remcos RAT** - REMCOS uses encrypted communication, including a hardcoded password for its authentication and network traffic encryption
- PLEAD 、 Shrouded Crossbow 、 Waterbear
- Keyboys - HP-Socket
- 遠銀 - splwew32.exe

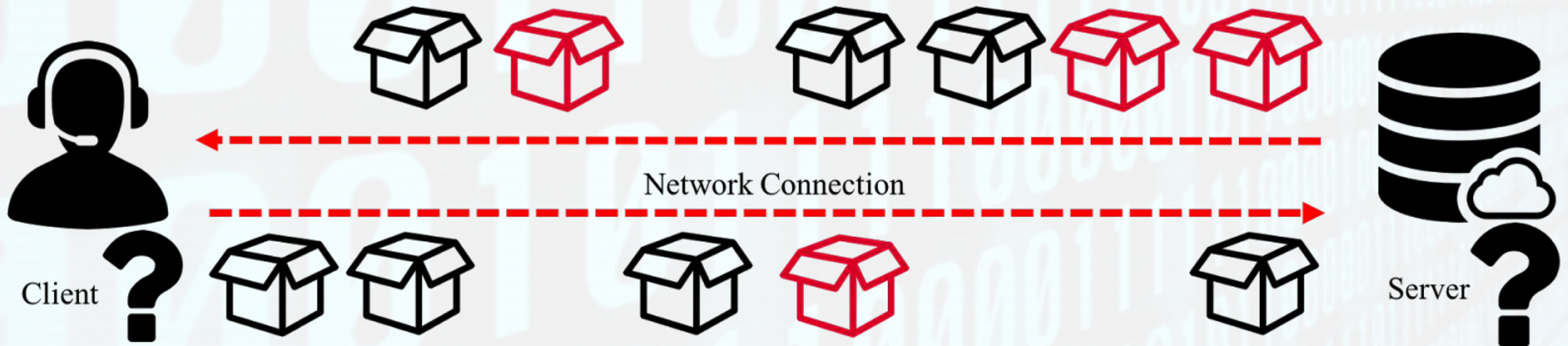
How to solve the problem?

- Change the **signature** based to **machine learning** based!

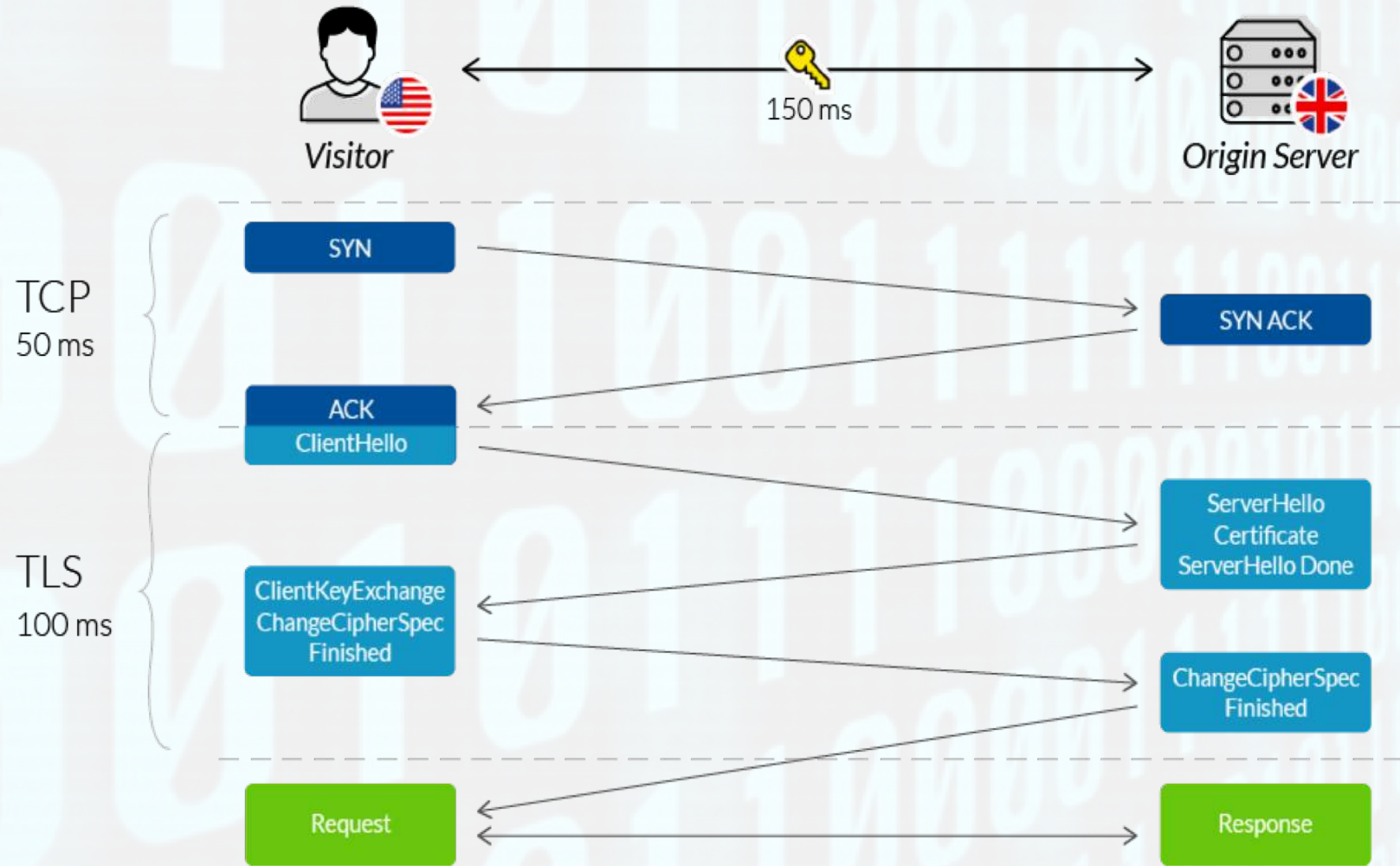


Our project: Deep Learning for Malicious Flow Detection

- To recognize the potential malicious behavior based on the net flow aspect especially for the encrypted net flow

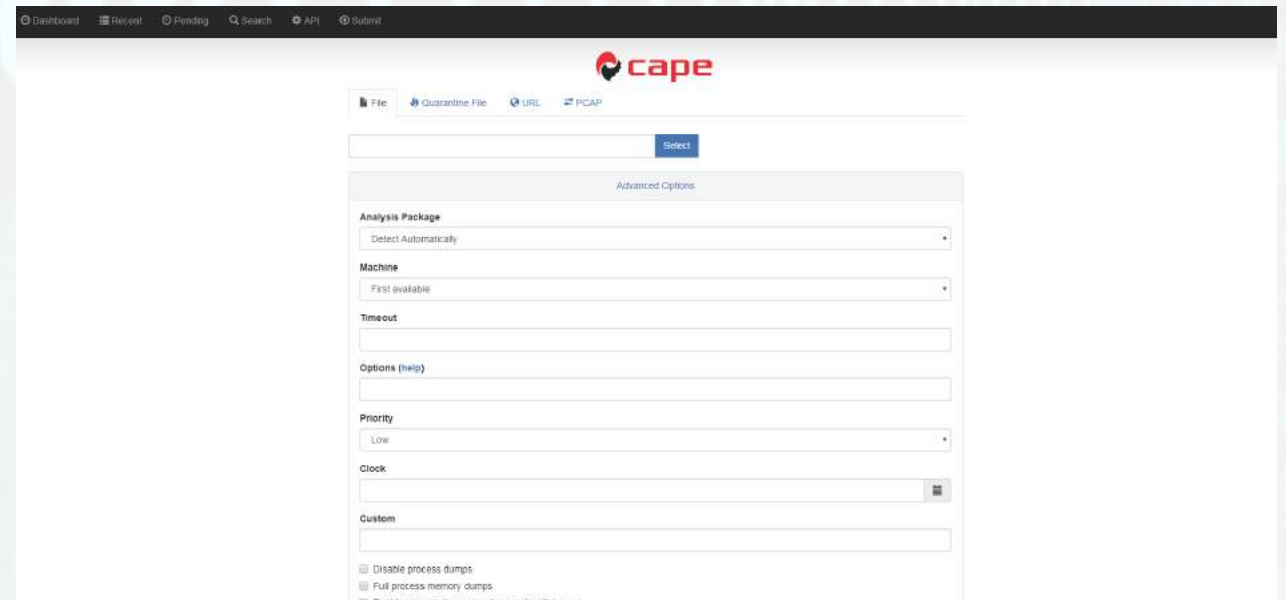


Encrypted Net Flow example: TLS



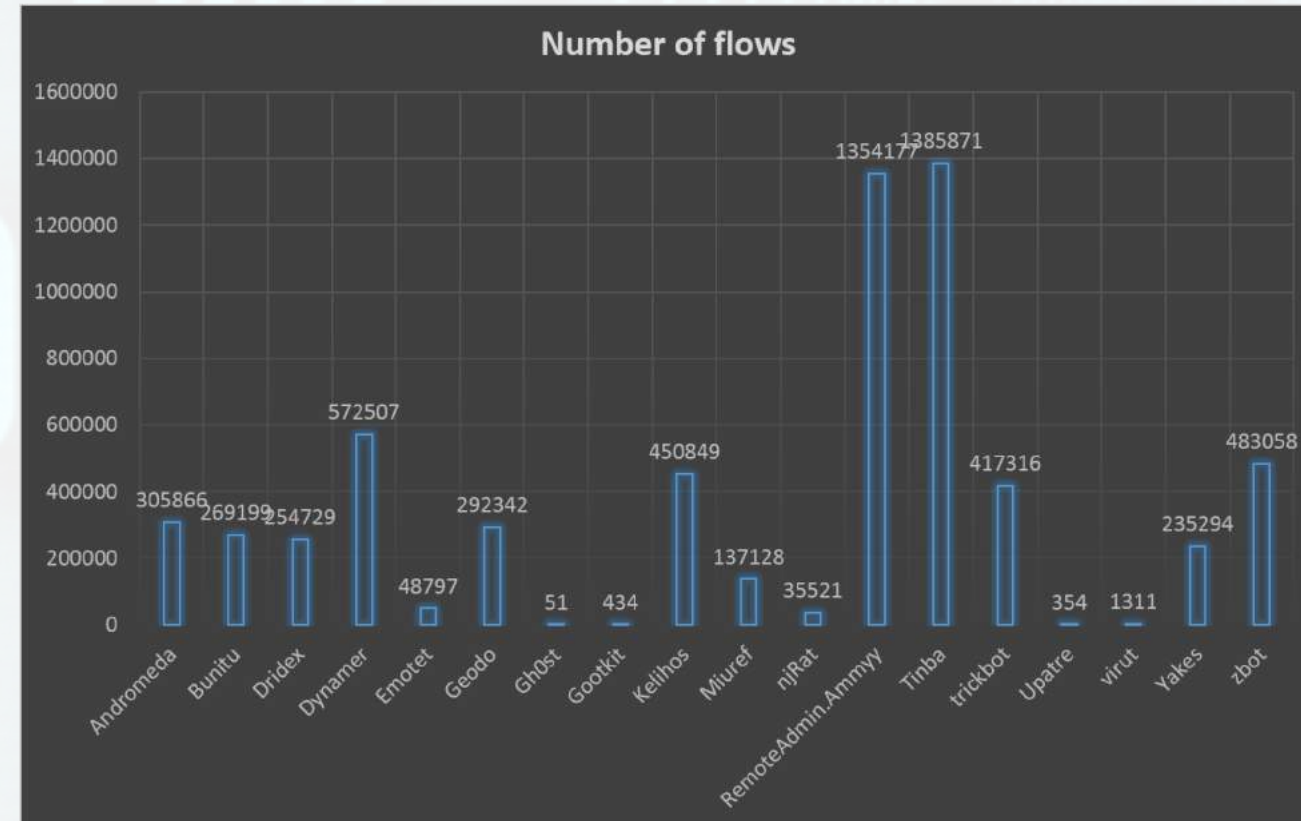
Dataset

- Pcaps/flows with HTTPS/VPN/Tor traffic
- Malware/VPN/Tor/Benign
- Capture with CAPE sandbox



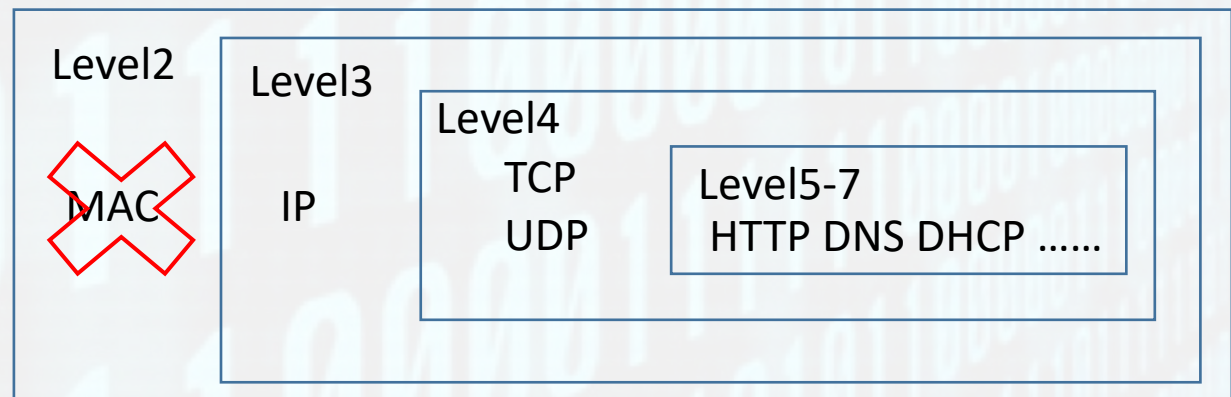
Dataset

- Malware traffic analysis
 - <https://www.malware-traffic-analysis.net/>
- CTU-13 dataset – public
 - Malware and Normal captures
 - 13 Scenarios. 600GB pcap
 - <https://www.stratosphereips.org/datasets-ctu13/>
- MCFP dataset – public
 - Malware Capture Facility Project
 - 340 malware pcap captures
 - <https://stratosphereips.org/category/dataset.html>
- Trend Micro Tbrain dataset
- UNB dataset – public
 - Tor-NonTor
 - VPN-NonVPN
 - <http://www.unb.ca/cic/datasets/index.html>
- Own malware/Tor dataset



Feature Engineering

- Cisco – joy
 - <https://github.com/cisco/joy>
- UNB – Flowmeter
 - <https://github.com/ISCX/CICFlowMeter>
- Bro logs
- Dpkt



Joy feature Intro



Packet Metadata

Feature	Type
Input/output IP	xxx. xxx. xxx. xxx
Input/output port number	Integer
Inbound/outbound bytes	Integer
Inbound/outbound packets	Integer
Total duration of the flow (ms)	Integer

HTTP:

- Request
 - `http_user_agent`
 - `http_accept_language`
- Response
 - `http_server`
 - `http_content_type`
 - `http_code`



DNS

- dns_domain_name
- dns_ttl(time to live)
- dns_num_ip
- dns_domain_rank

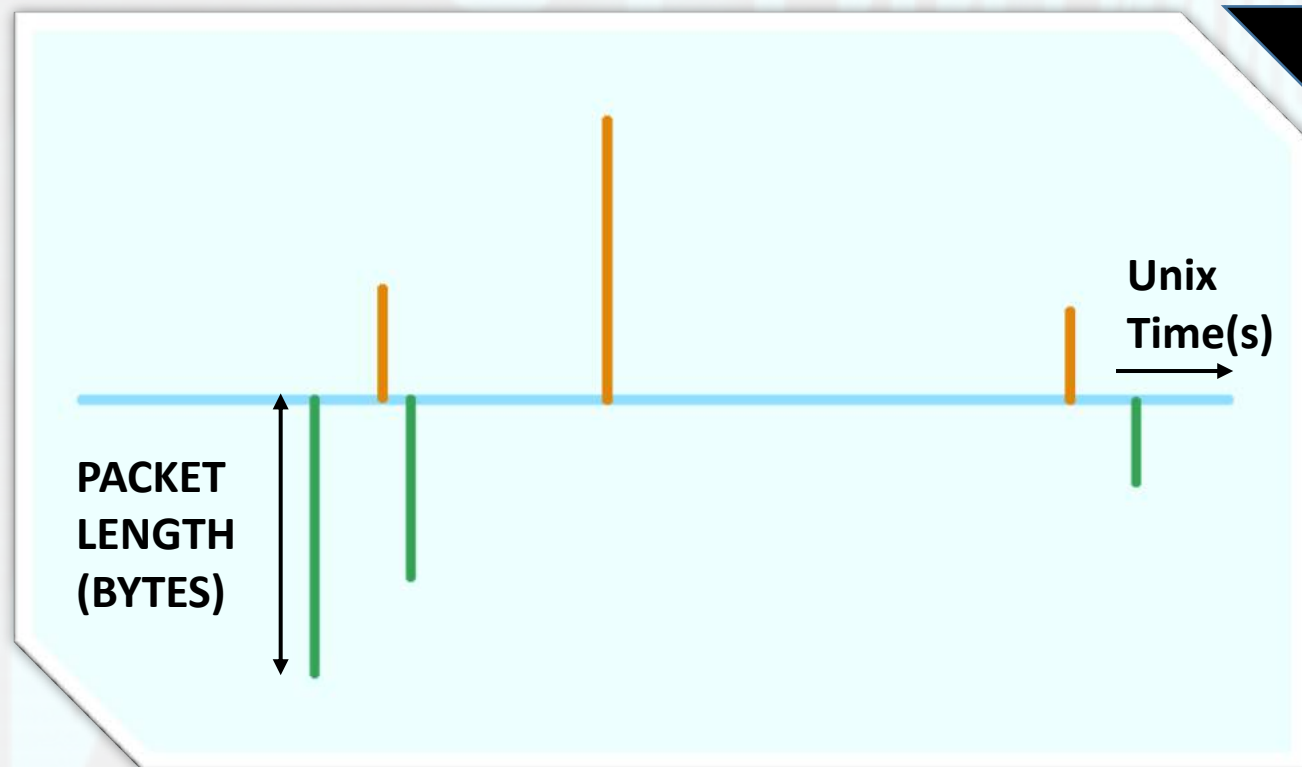


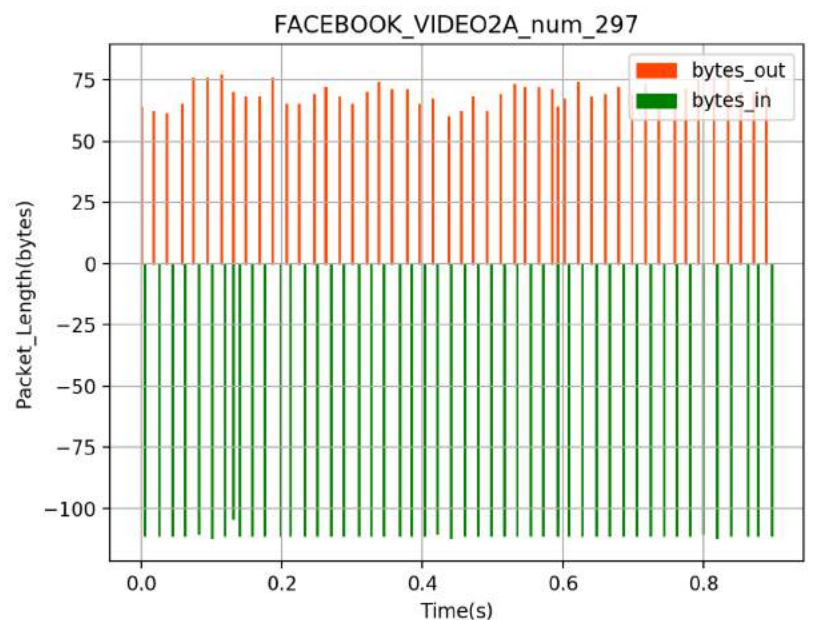
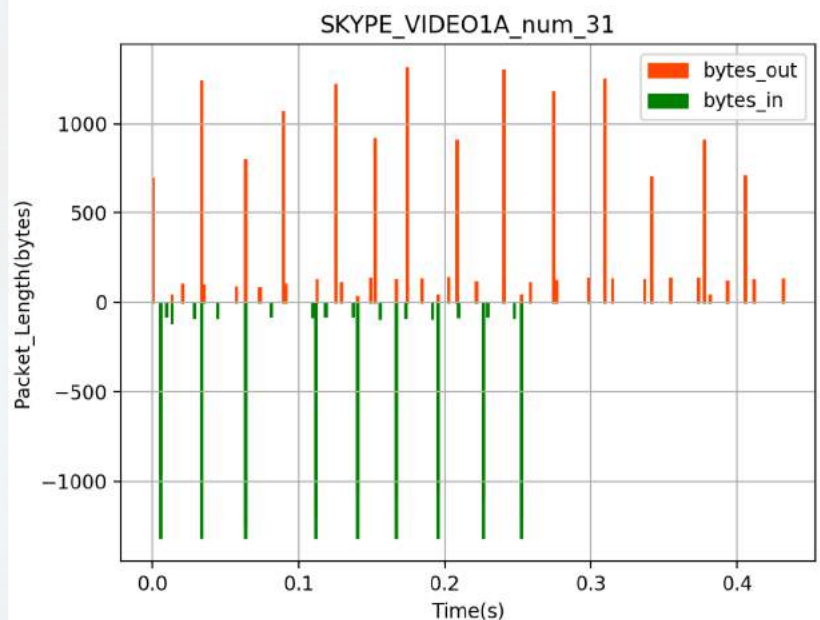
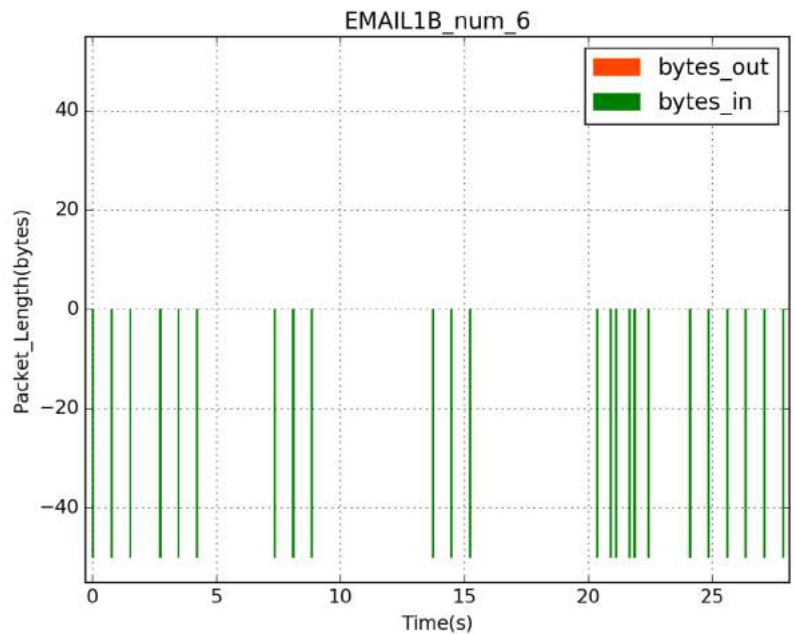
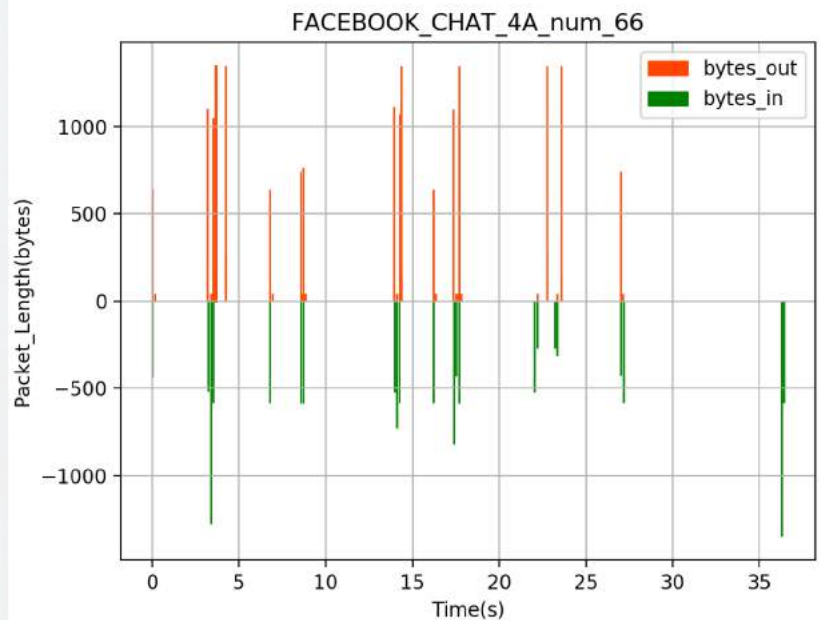
Sequence of Packet Lengths and Times (SPLT)

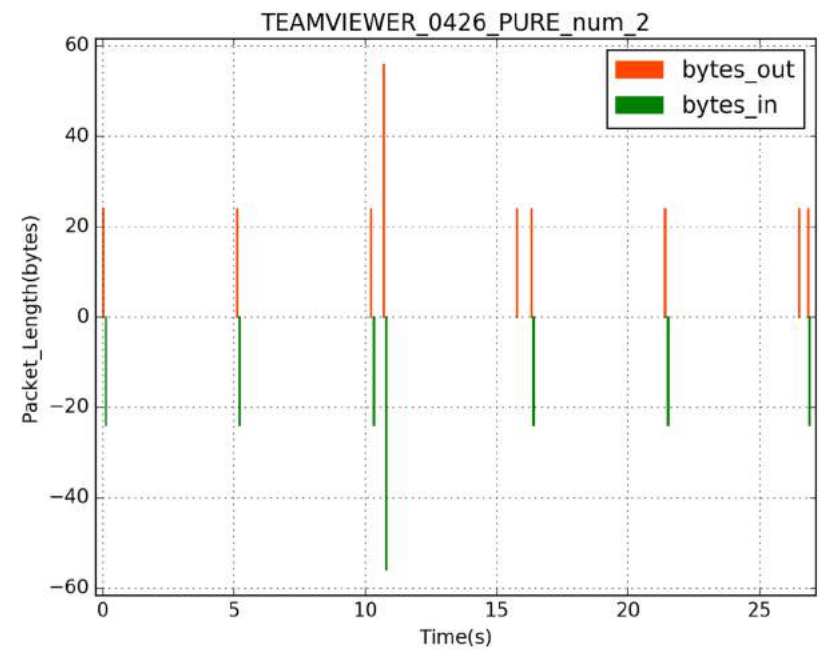
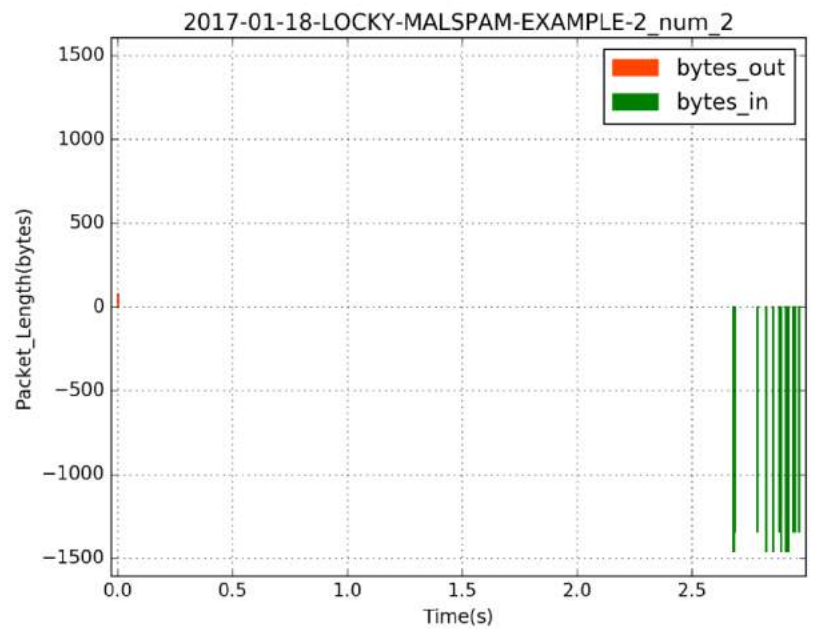
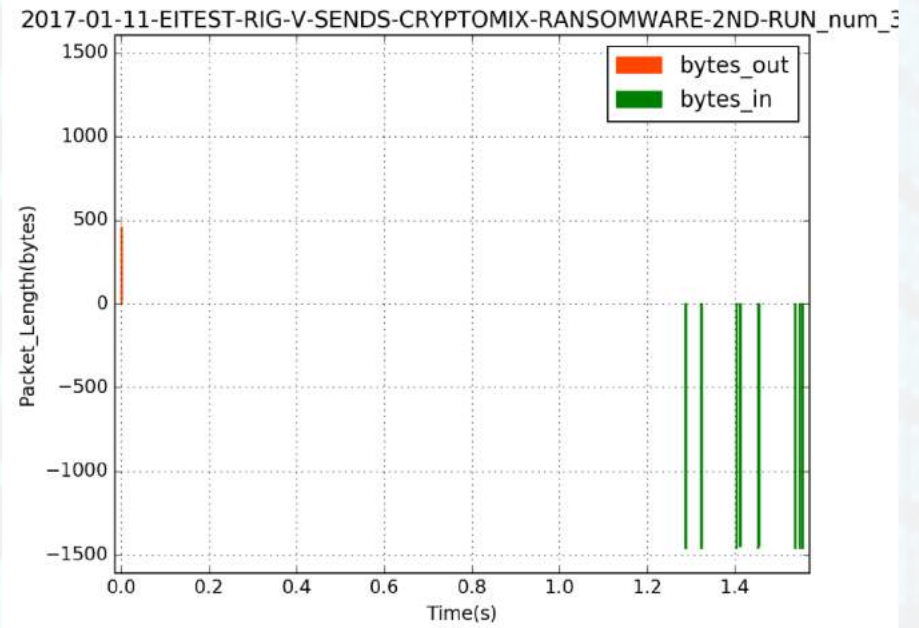
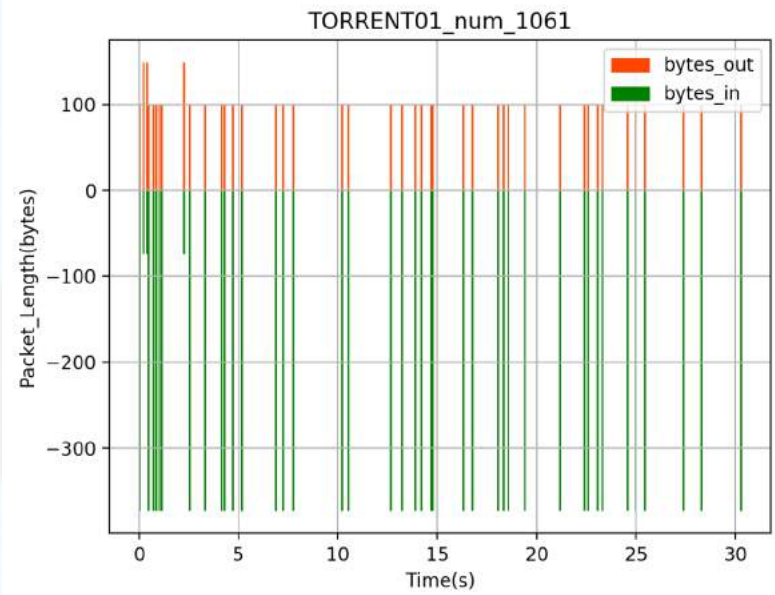
Malware Behavior	Network Behavior
Communication with command control server	Sequence of packet lengths
Write to the disk	Time interval between packet

- Size and Timing of the first few packets allow us to estimate the type of the data inside the encrypted channel

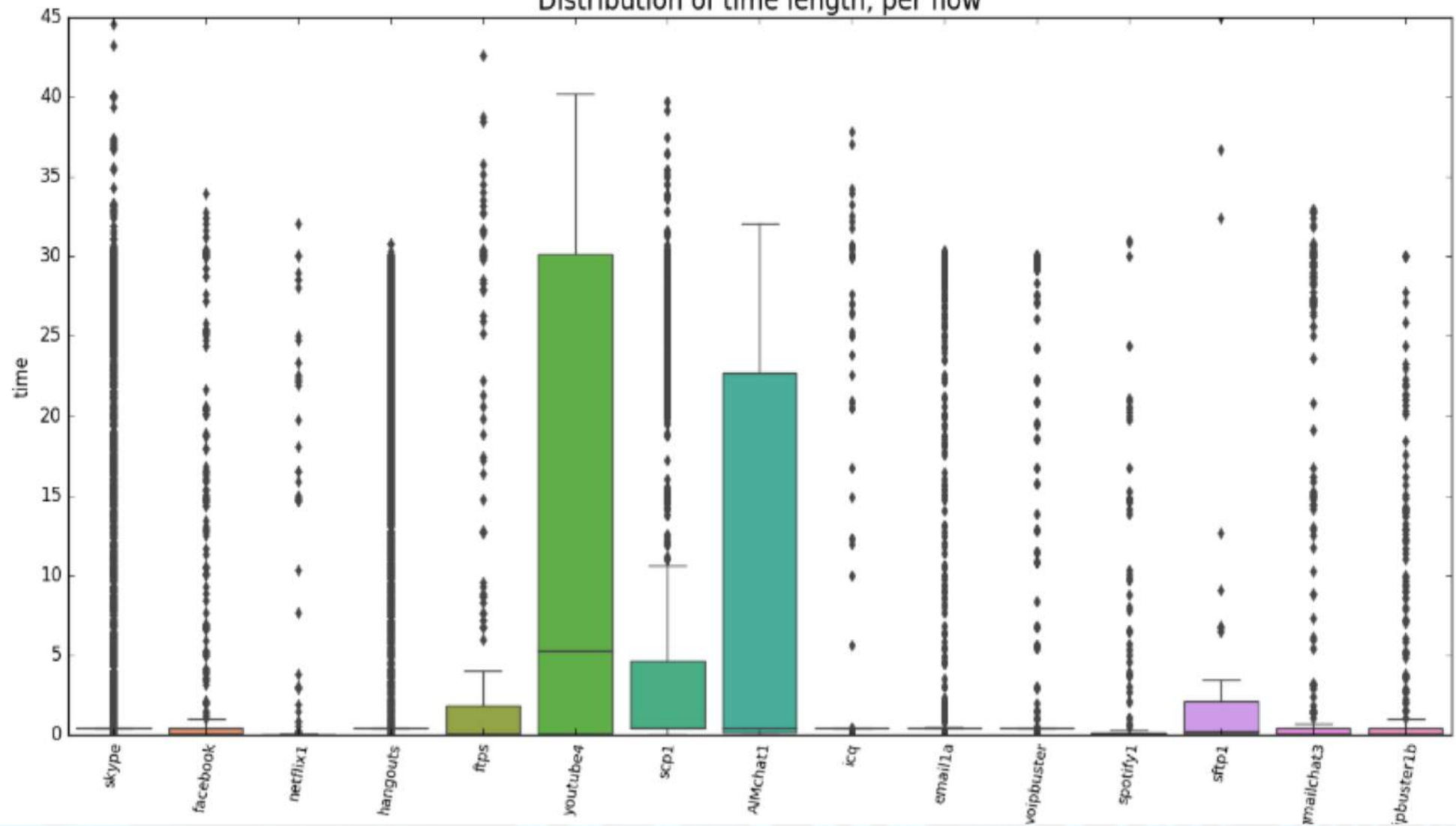
Visualize with SPLT







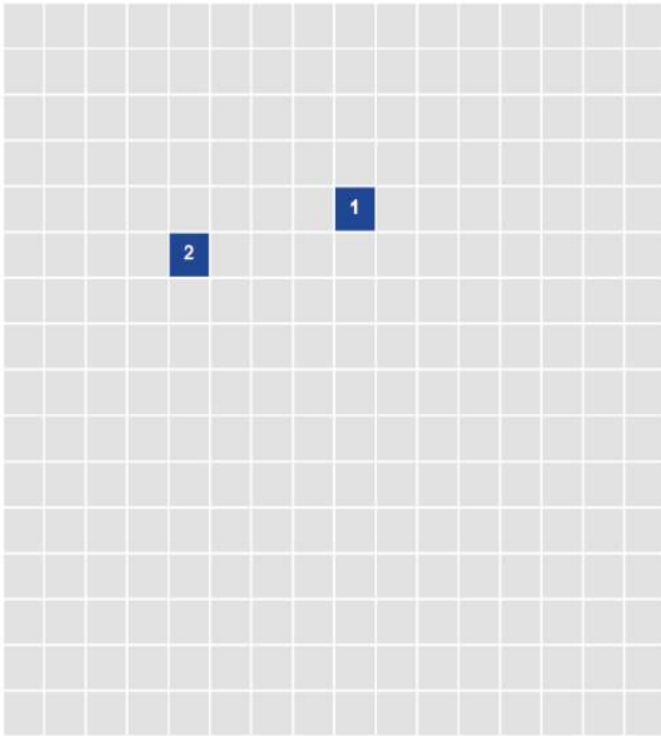
Distribution of time length, per flow



Byte Distribution

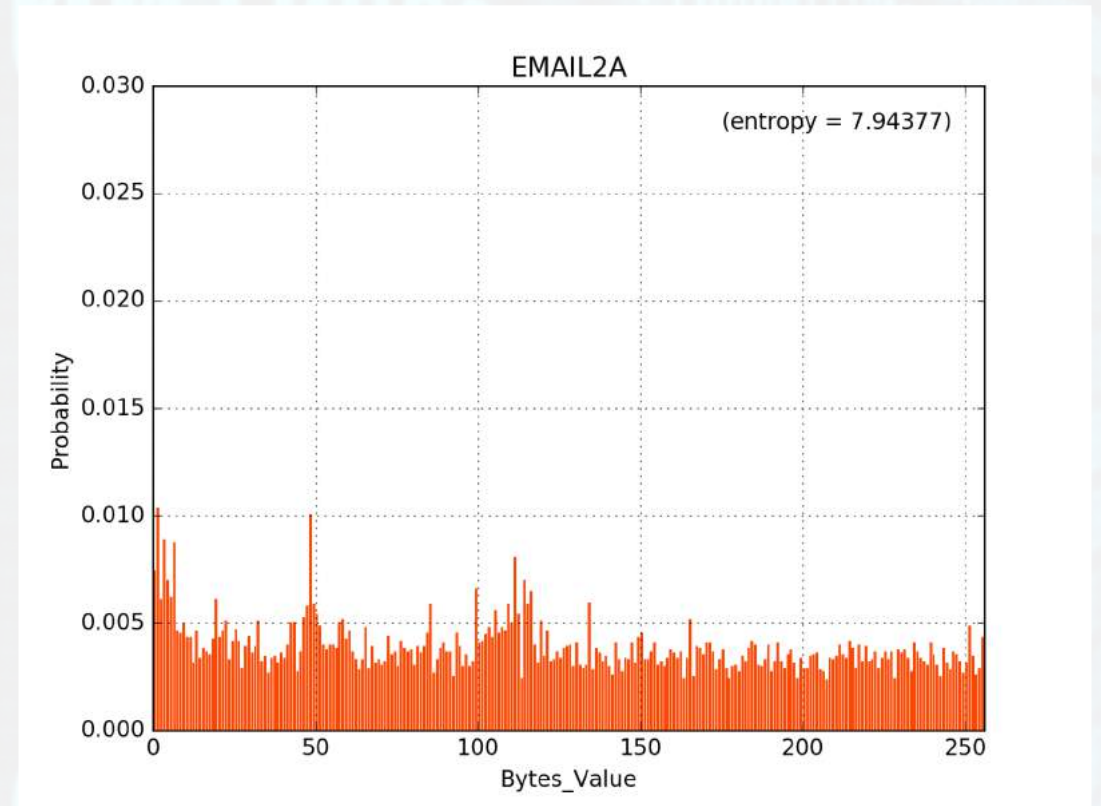
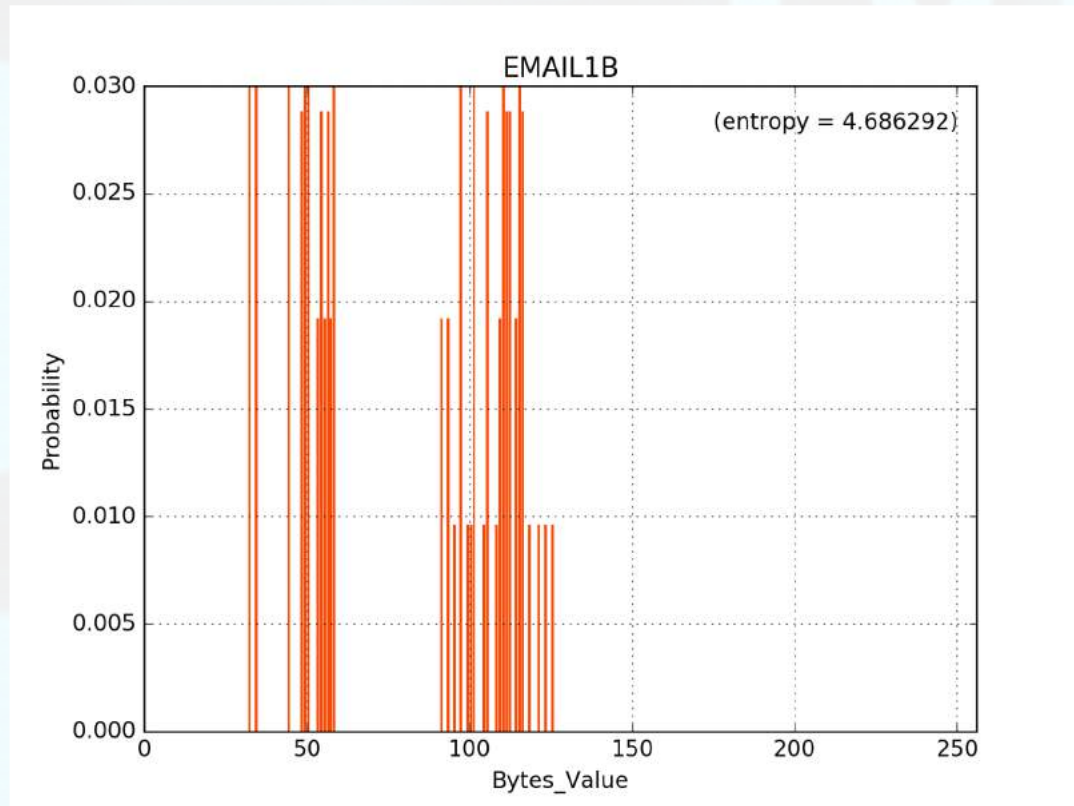
H T **T** P / 1 . 1 2 0 0 O K

□48 54 54 50 2f 31 2e 31 20 32 30 30 20 4f 4b



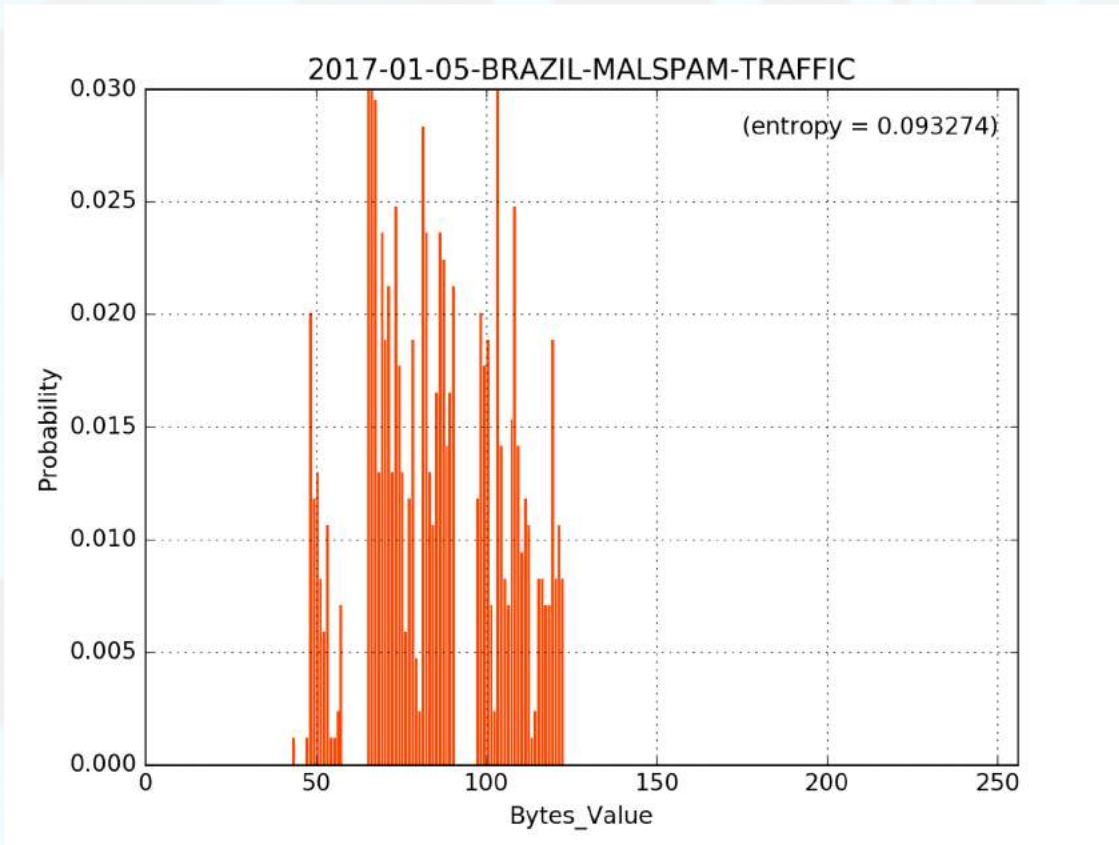
© 2015 Cisco and/or its affiliates. All rights reserved. Cisco Conf

Visualization with Byte Distribution

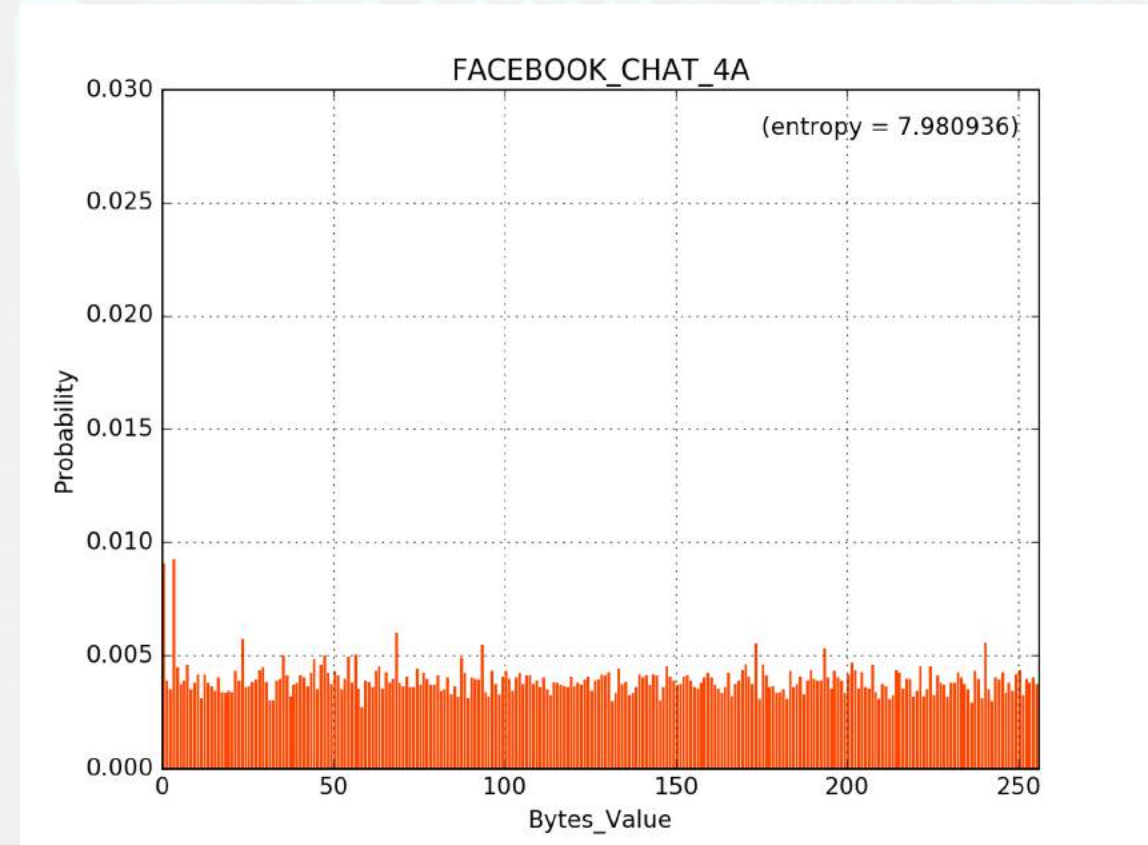


Email with **TLSv1.2**

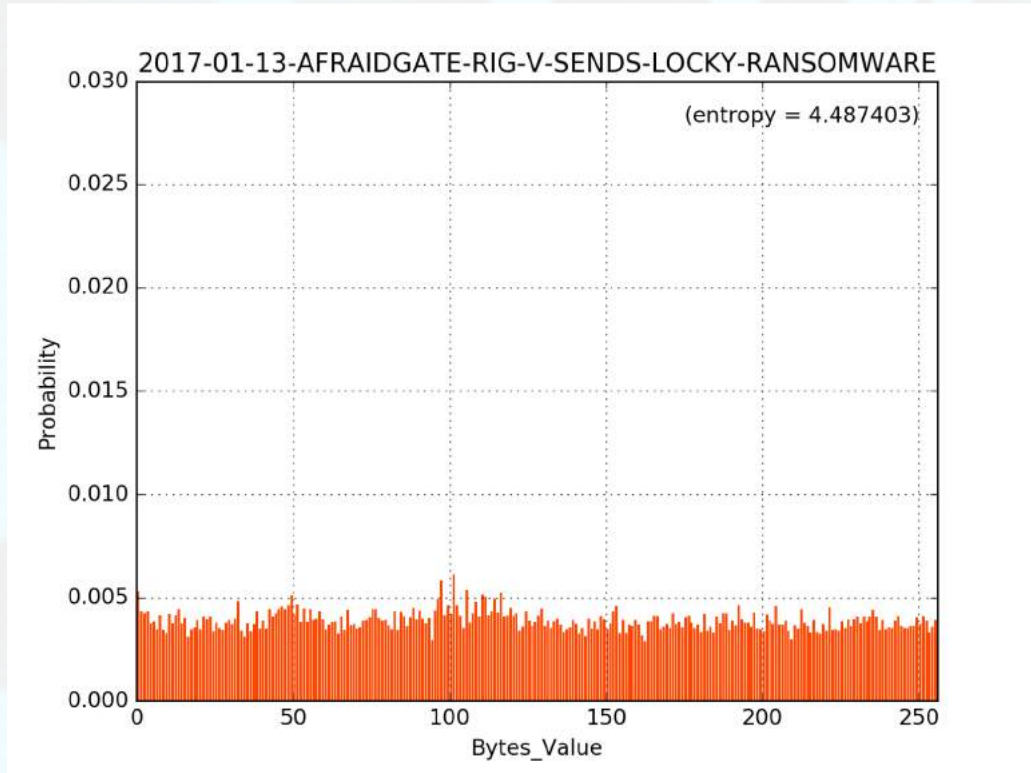
Malspam



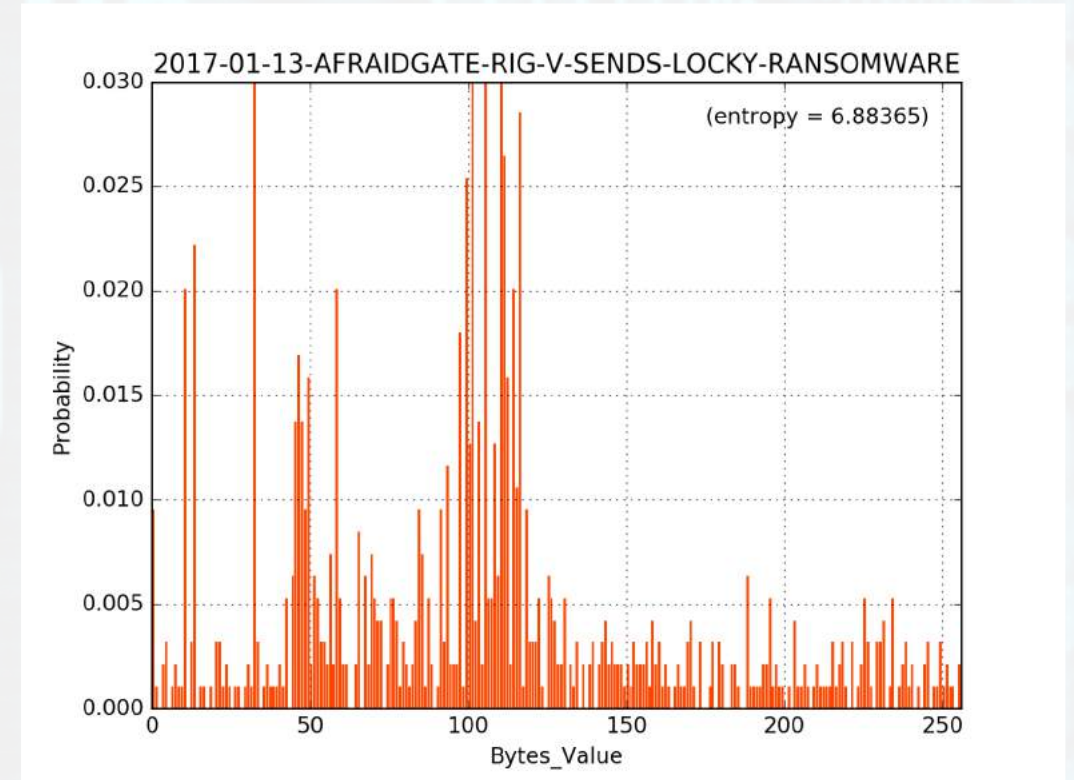
Facebook chat



Locky Ransomware

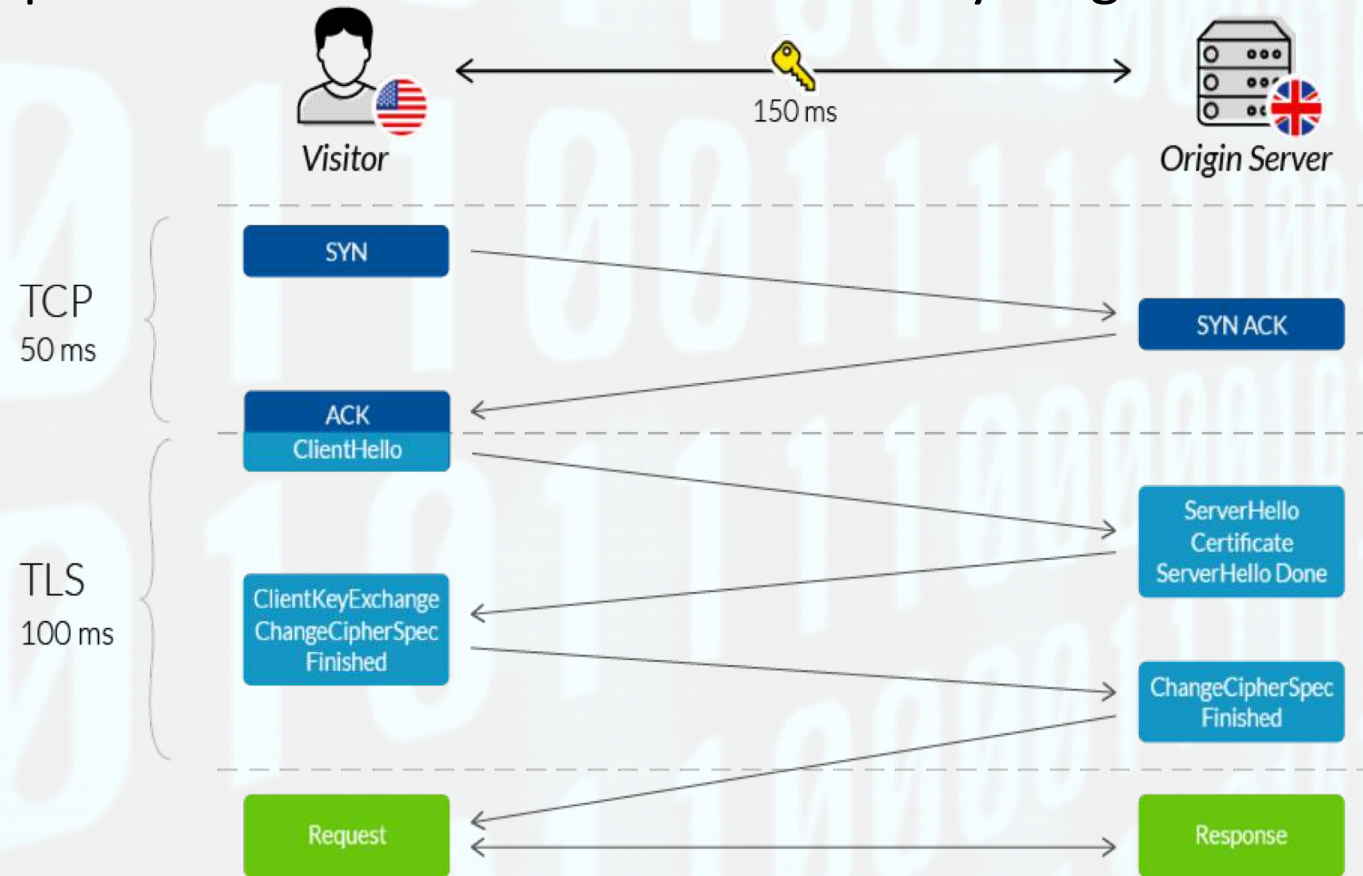


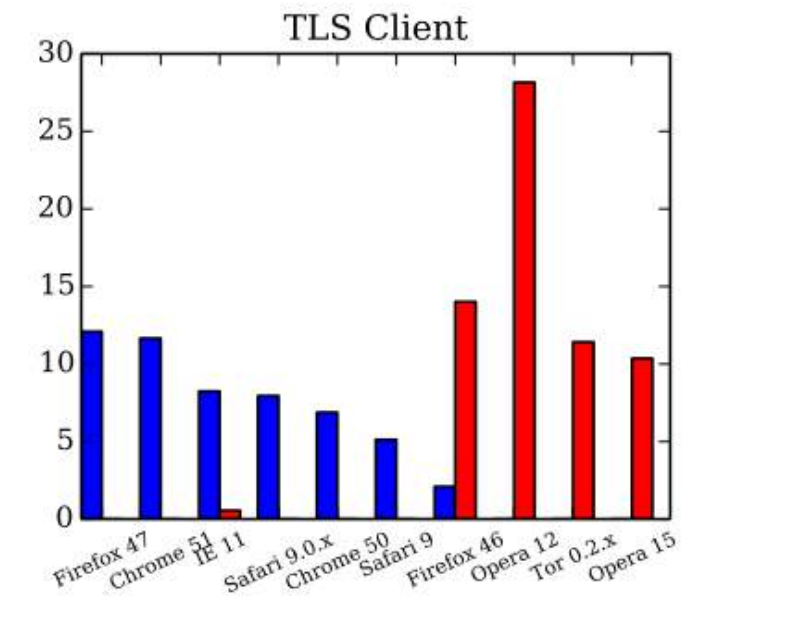
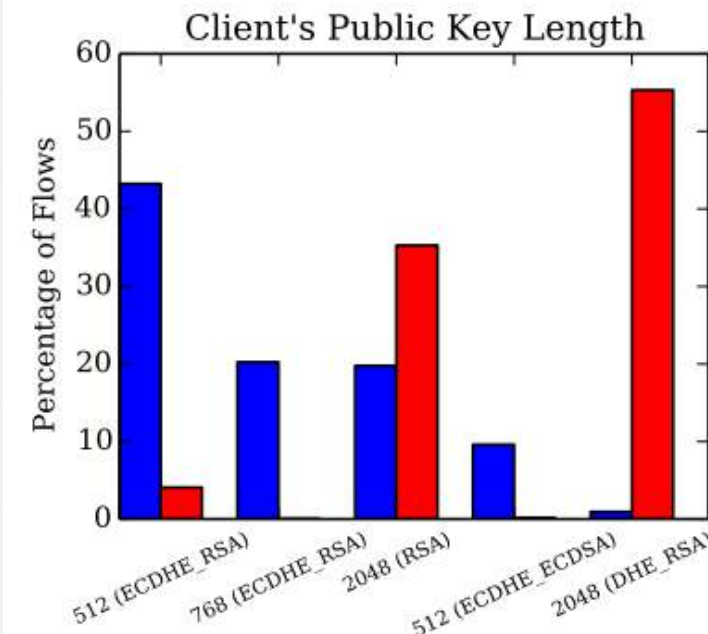
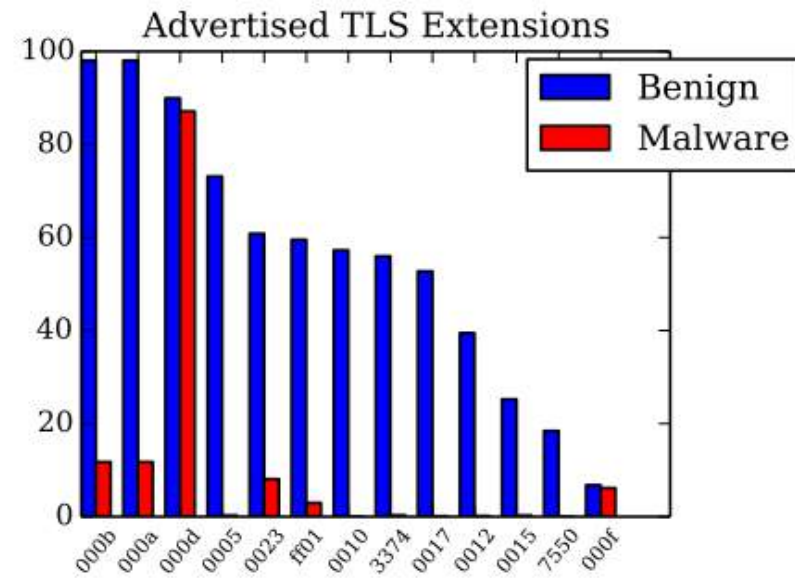
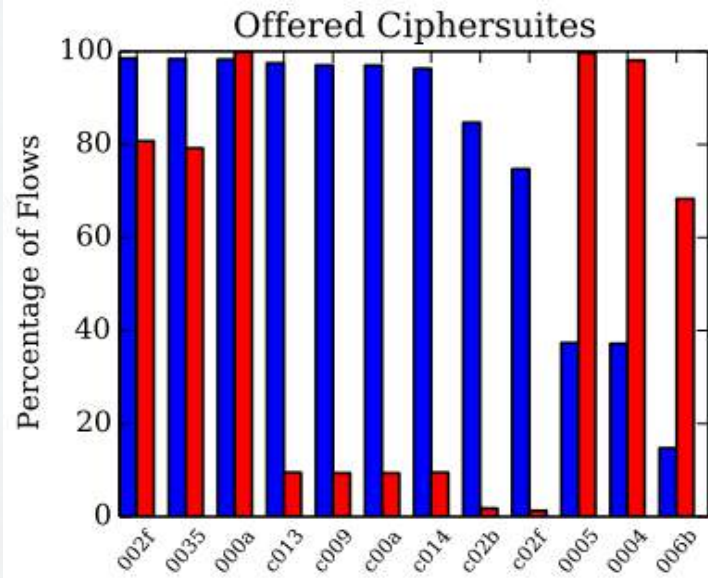
Locky Ransomware

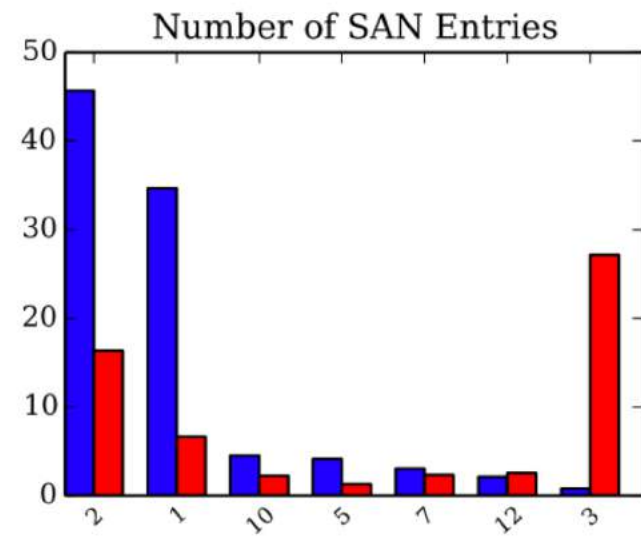
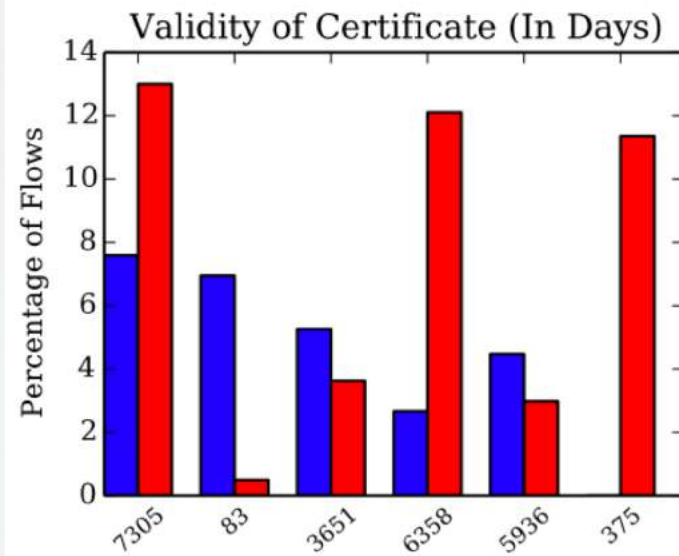
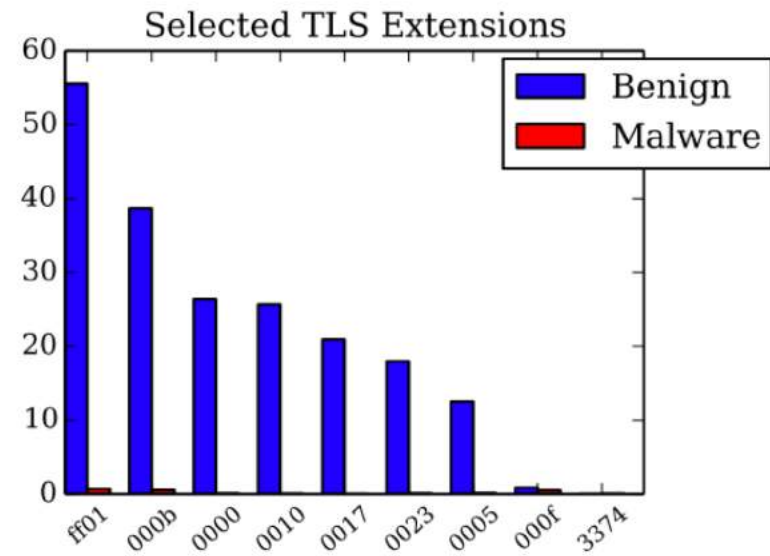
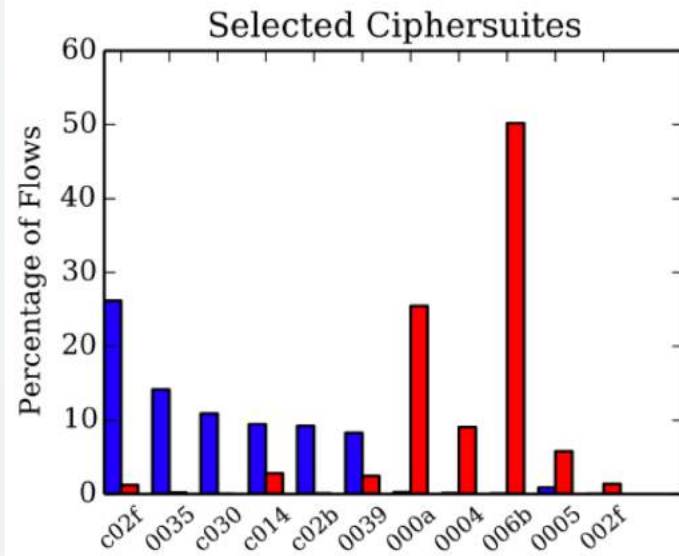


TLS Information

- TLS handshake info:
TLS Ciphersuite、TLS extension、Publickey length







Malware Family	Number of Flows	Unique Server IPs	Number of SS Certs	Selected Ciphersuite	Certificate Subject
Bergat	332	12	0	TLS_RSA_WITH_3DES_EDE_CBC_SHA	www.dropbox.com
Deshacop	129	38	0	TLS_RSA_WITH_3DES_EDE_CBC_SHA	*.onion.to
Dridex	103	10	89	TLS_RSA_WITH_AES_128_CBC_SHA	amthonoup.cy
Dynamer	372	155	3	TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256	www.dropbox.com
Kazy	1152	225	52	TLS_RSA_WITH_3DES_EDE_CBC_SHA	*.onestore.ms
Parite	275	128	0	TLS_RSA_WITH_3DES_EDE_CBC_SHA	*.google.com
Razy	564	118	16	TLS_RSA_WITH_RC4_128_SHA	baidu.com
Sality	1,200	323	4	TLS_RSA_WITH_3DES_EDE_CBC_SHA	vastusdomains.com
Skeeyah	218	90	0	TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256	www.dropbox.com
Symmi	2,618	700	22	TLS_ECDHE_RSA_WITH_AES_256_CBC_SHA	*.criteo.com
Tescrypt	205	26	0	TLS_RSA_WITH_3DES_EDE_CBC_SHA	*.onion.to
Toga	404	138	8	TLS_RSA_WITH_3DES_EDE_CBC_SHA	www.dropbox.com
Upatre	891	37	155	TLS_RSA_WITH_RC4_128_MD5	*.b7websites.net
Virlock	12,847	1	0	TLS_DHE_RSA_WITH_AES_256_CBC_SHA256	block.io
Virtob	511	120	0	TLS_RSA_WITH_3DES_EDE_CBC_SHA	*.g.doubleclick.net
Yakes	337	51	0	TLS_RSA_WITH_RC4_128_SHA	baidu.com
Zbot	2,902	269	507	TLS_RSA_WITH_RC4_128_MD5	tridayacipta.com
Zusy	733	145	14	TLS_RSA_WITH_3DES_EDE_CBC_SHA	*.criteo.com

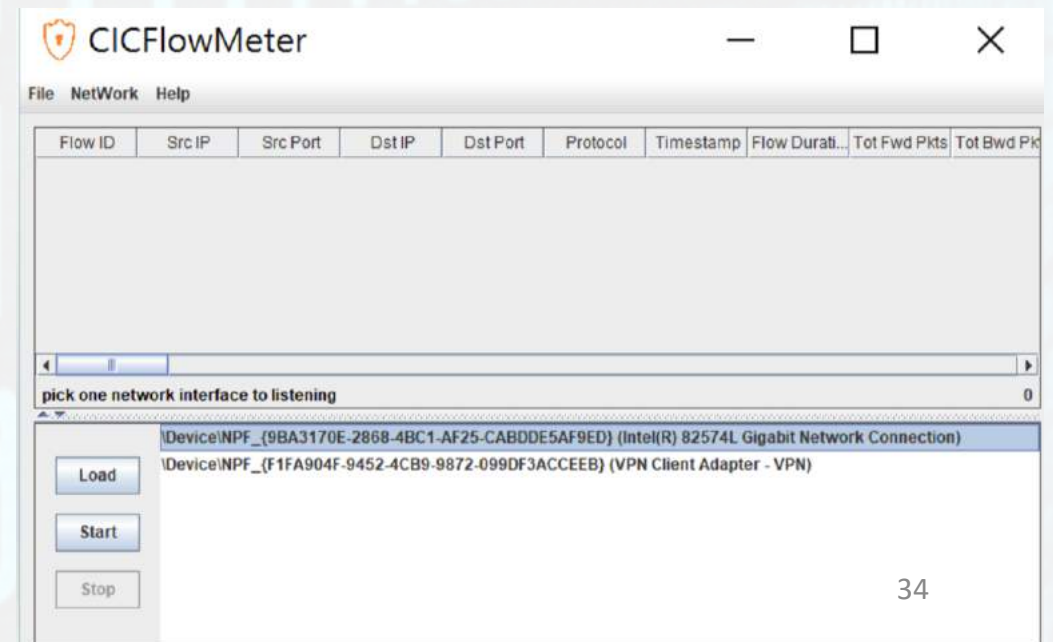
Source : <https://arxiv.org/abs/1607.01639>



CICFlowmeter Feature Intro

CICFlowMeter

- An open source tool
- Generate bidirectional flows from pcap files
- Extracts features from these flows
- Supports realtime generate bidirectional flows




Network basic Metadata

- Flow ID
- IP
- Port
- Protocol
- Timestamp



Time-based feature

- Flow Duration
- bytes/s
- packets/s
- packet length
- IAT(inter-arrival time)
- Flag
- Active time
- Idle time

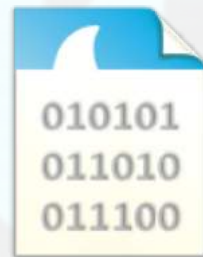
- 
- BWD 、 FWD(direction) 、 Total
 - Max 、 Min 、 Mean 、 Std

Bro logs

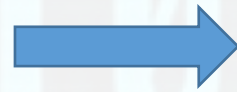
Idea from [Czech technical university in Prague](#)



Bro

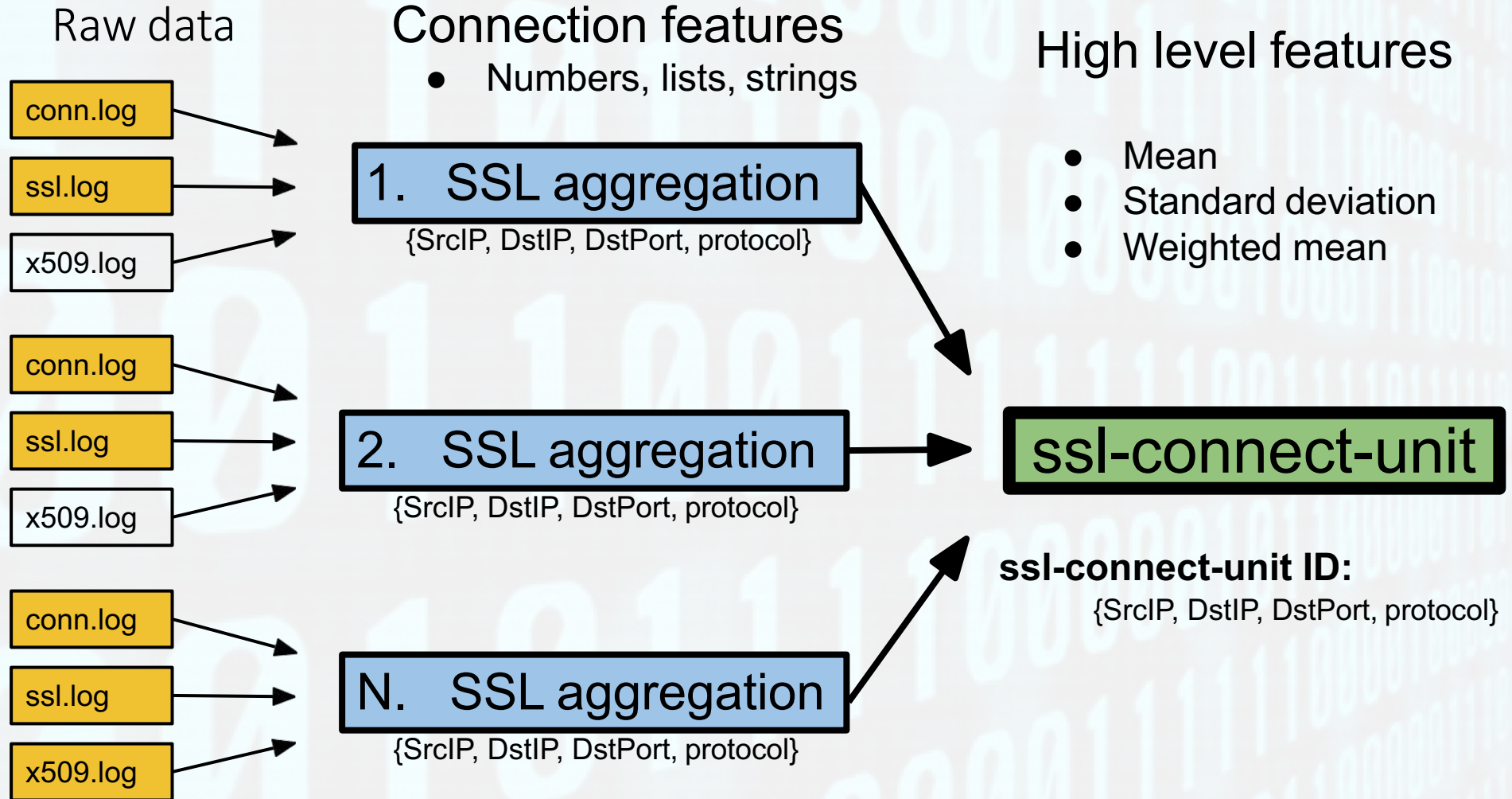


Bro IDS



Bro logs

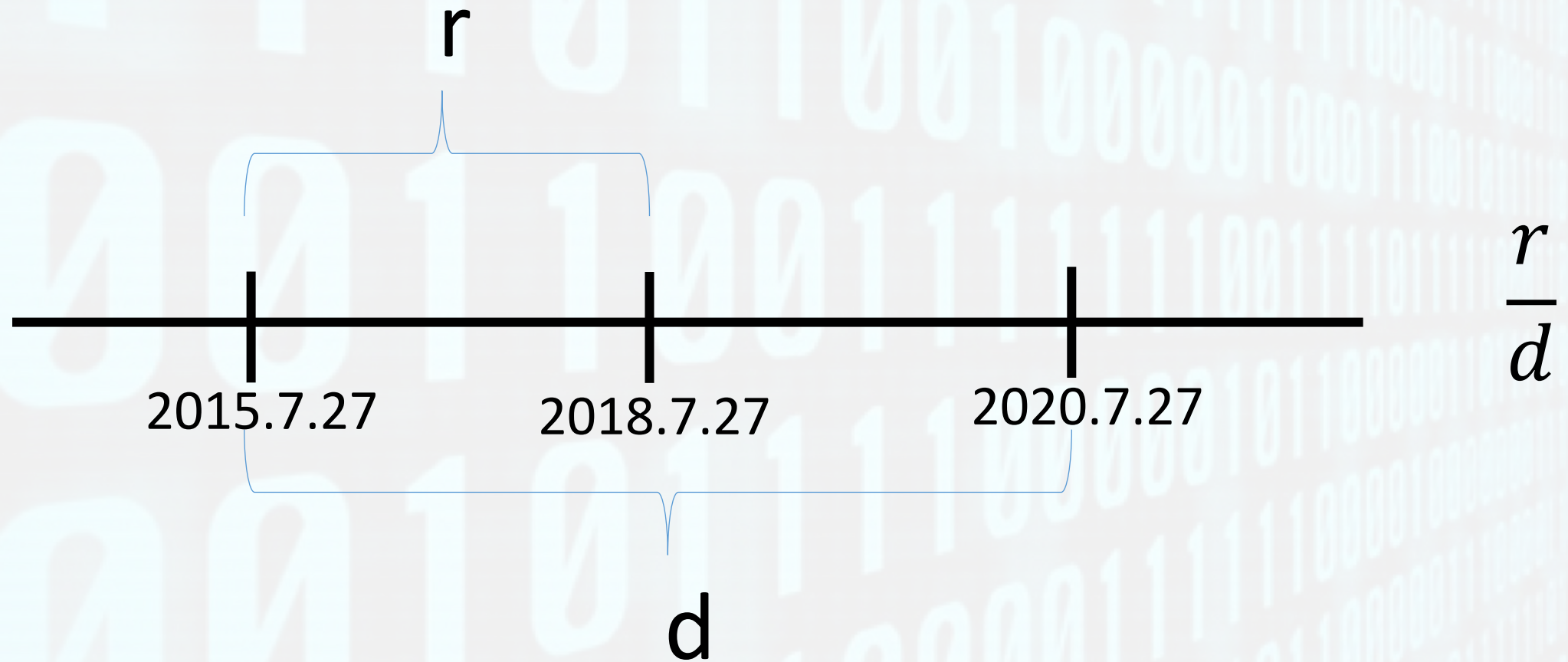
- Conn.log
- ssl.log
- X509.log
- dns.log
- http.log
- Files.log
-



40 Features of ssl-connect-unit

- Number of SSL aggregations
- Mean and standard deviation of duration
- Mean and standard deviation of number of packets
- Mean and standard deviation of number of bytes
- Ratio of TLS and SSL version
- Number of different certificates

Ratio of validity during the capture



Top 7 most discriminant features

- Certificate length of **validity**
- Inbound and outbound packets
- **Validity of certificate during the capture**
- Duration
- Number of domains in certificate (SAN DNS)
- SSL/TLS version
- Periodicity



Machine Learning methods

Quantity Dependent Backpropagation(QDBP)

- We introduce a vector F into backpropagation (eq (1)) and propose a QDBP algorithm which takes the disparity between classes into consideration and shows different sensitivities toward different classes.

- $\theta_i^{l^+} = \theta_i^l - \eta \times \frac{\partial Loss}{\partial \theta_i^l} \quad - (1)$

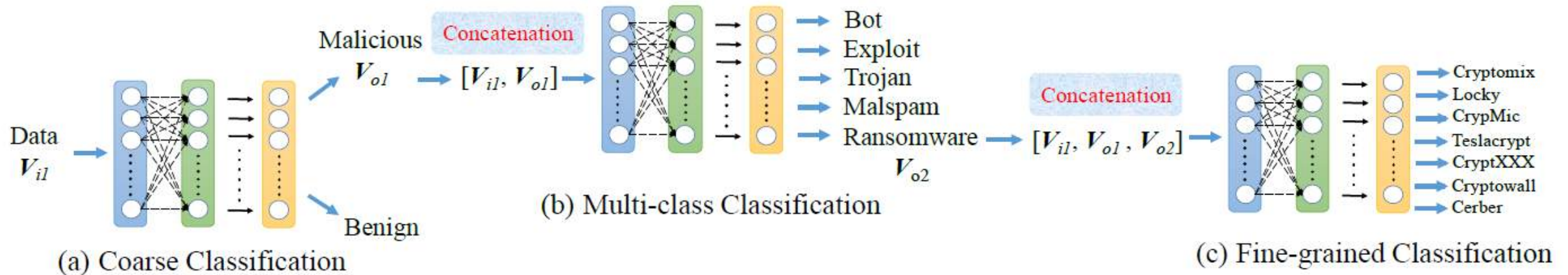
- $\theta_i^{l^+} = \theta_i^l - \eta \cdot F \cdot \nabla Loss \quad -(2)$

- $F = \left[\frac{c_1}{n_1}, \frac{c_2}{n_2}, \dots, \frac{c_N}{n_N} \right]$

- $\nabla Loss = \left[\frac{\partial Loss_1}{\partial \theta_i}, \frac{\partial Loss_2}{\partial \theta_i}, \dots, \frac{\partial Loss_N}{\partial \theta_i} \right]^T$

Tree-Shaped Deep Neural Network (TSDNN)

- To mitigate the imbalanced data issue, we propose an end-to-end trainable TSDNN model which classifies the data layer by layer.

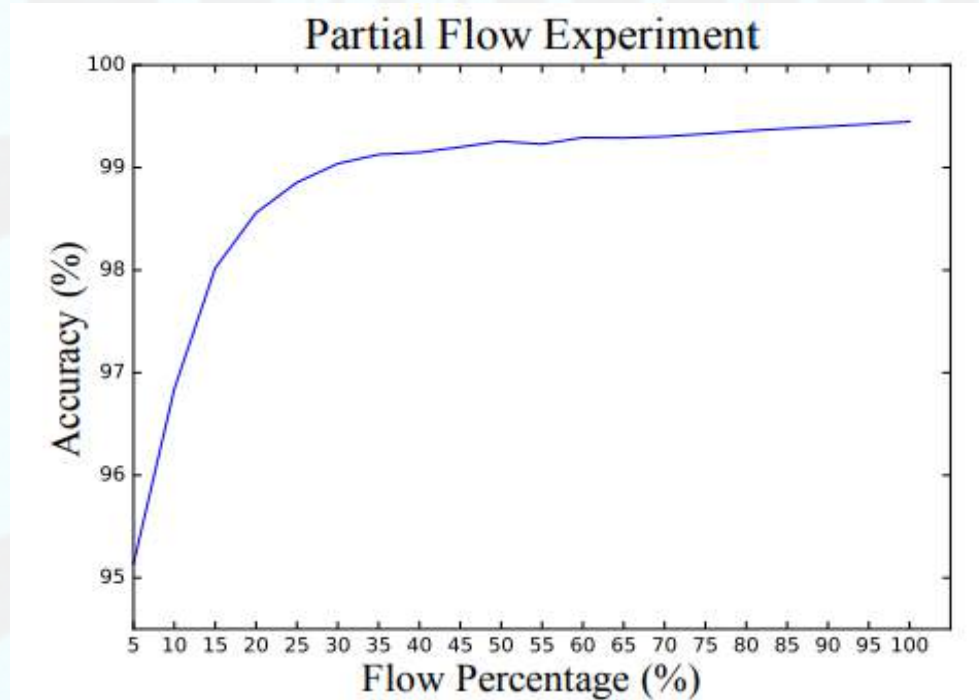


ACCURACY AND PRECISION OF DIFFERENT APPROACHES

Method	Accuracy	Precision
DNN + Backpropagation	59.08%	8.33%
DNN + Oversampling (10000 samples/class) [7]	85.18%	65.9%
DNN + Undersampling (45 samples/class) [8]	68.89%	49.45%
DNN + Incremental Learning [9]	78.84%	71.23%
DNN + QDBP	84.56%	62.3%
SVM (RBF)	83.87%	38.8%
Random Forest	98.9%	68.25%
TSDNN + QDBP	99.63%	85.4%

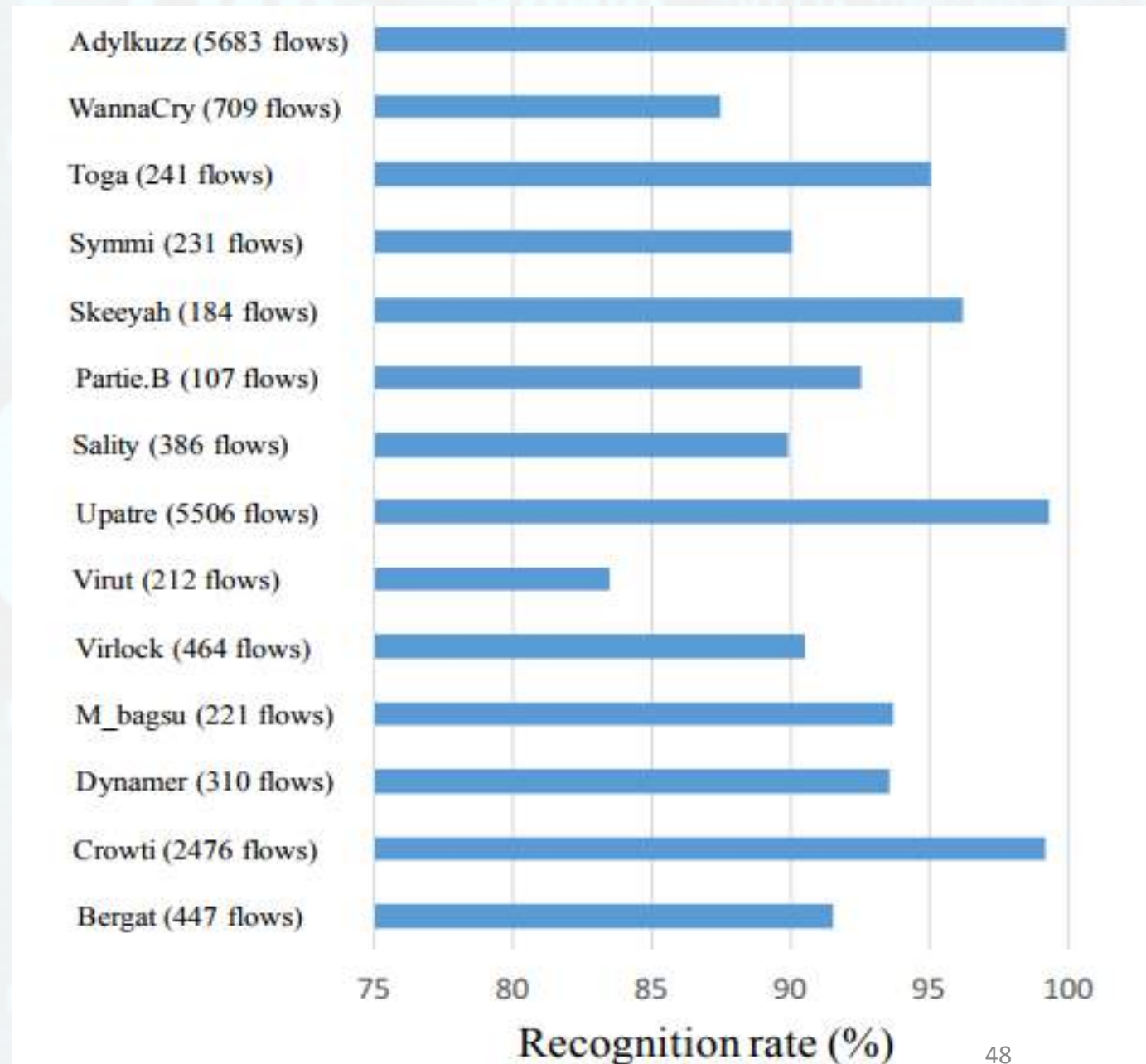
Partial flow Detection

- Our model is able to distinguish the malicious flow by only considering the first 5 % of the entire flow which shows the possibility of a realtime detection since the model can perceive the potential threats in the very beginning of the process without analyzing the entire flow.



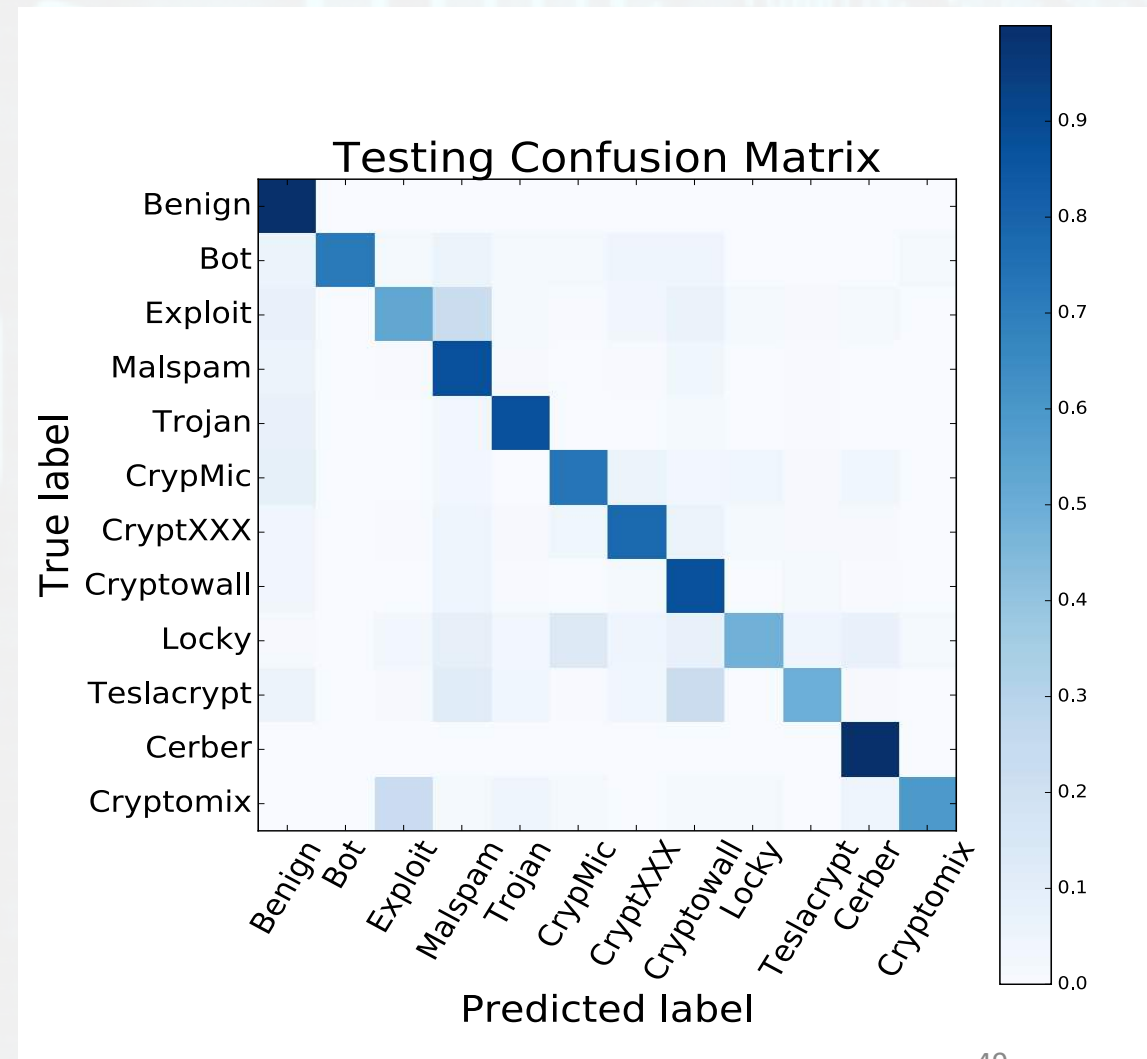
Zero-shot Learning

- We collect 14 different kinds of malware not in training data to evaluate the ability of our model to perceive potential threats.



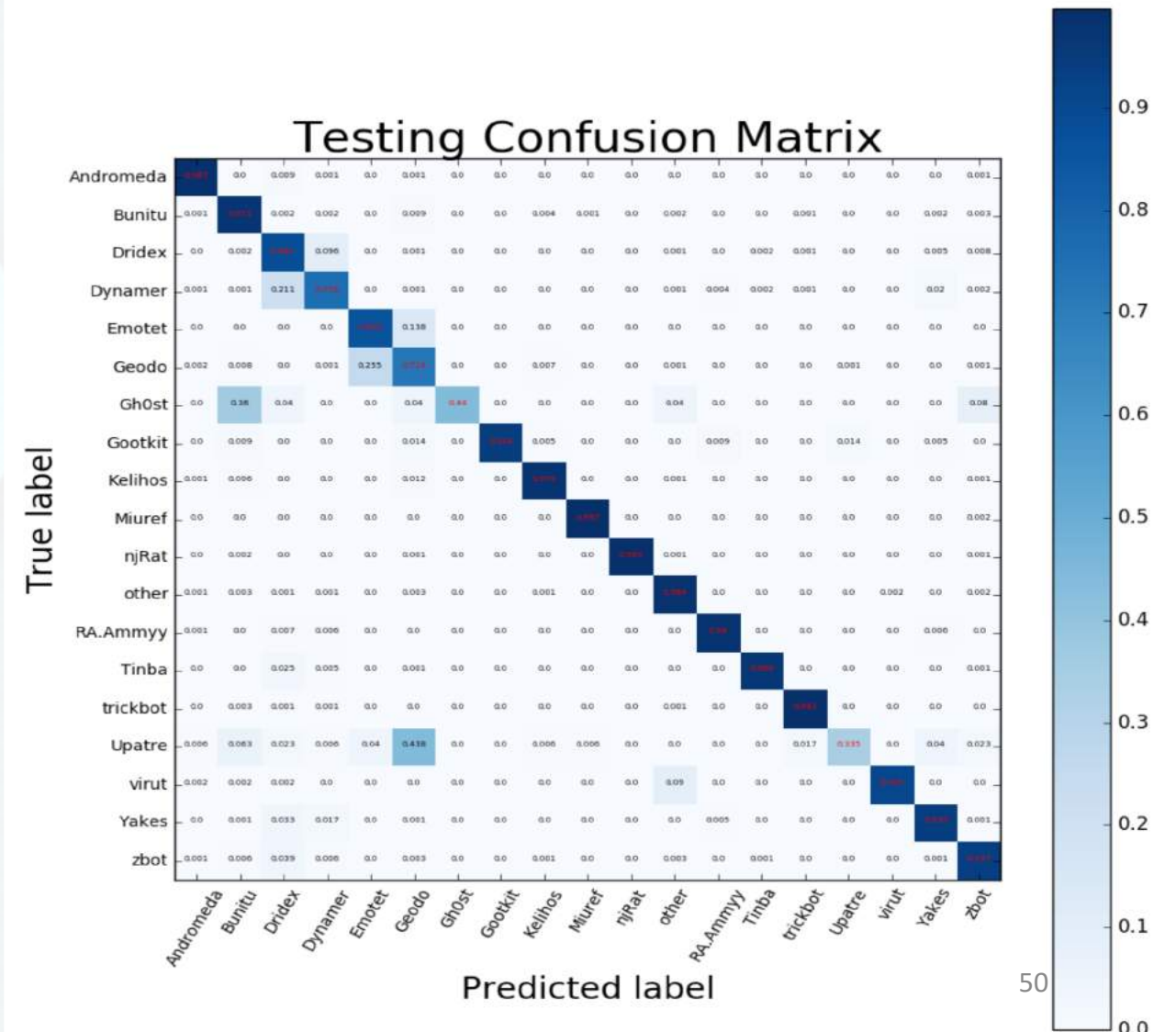
Multiclass Classification

- 12 classes
- Accuracy = 99.63%
- Precision = 85.4%



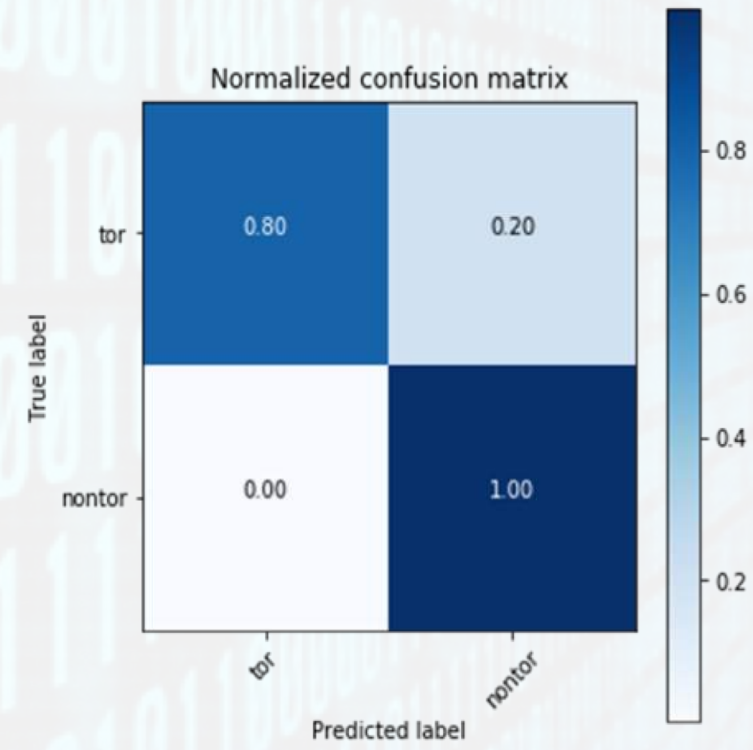
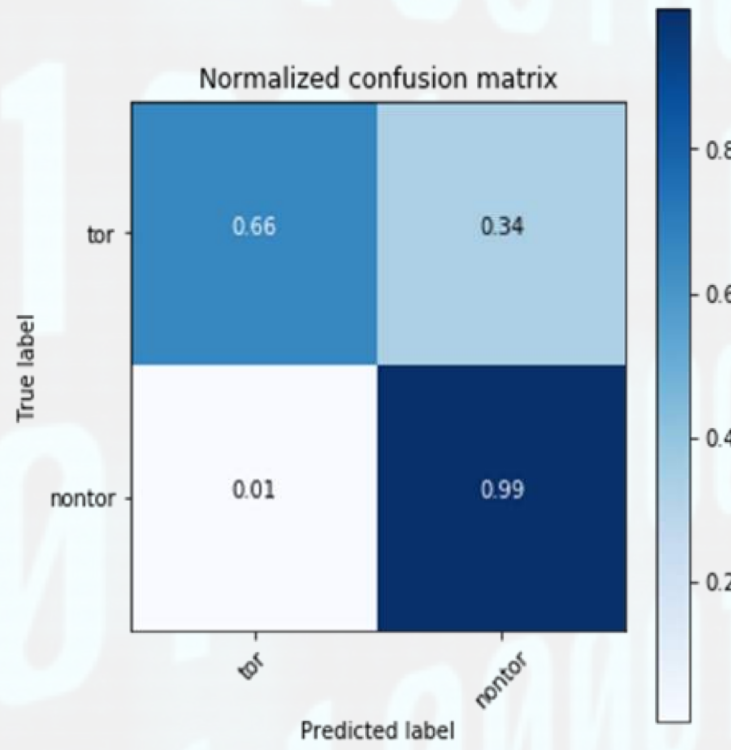
Multiclass Classification

- 19 classes
- Accuracy = 92.84%
- Precision = 87.32%



Tor-NonTor Classification

- Xgboost
- Accuracy = 98.7%
- Precision = 91.9%

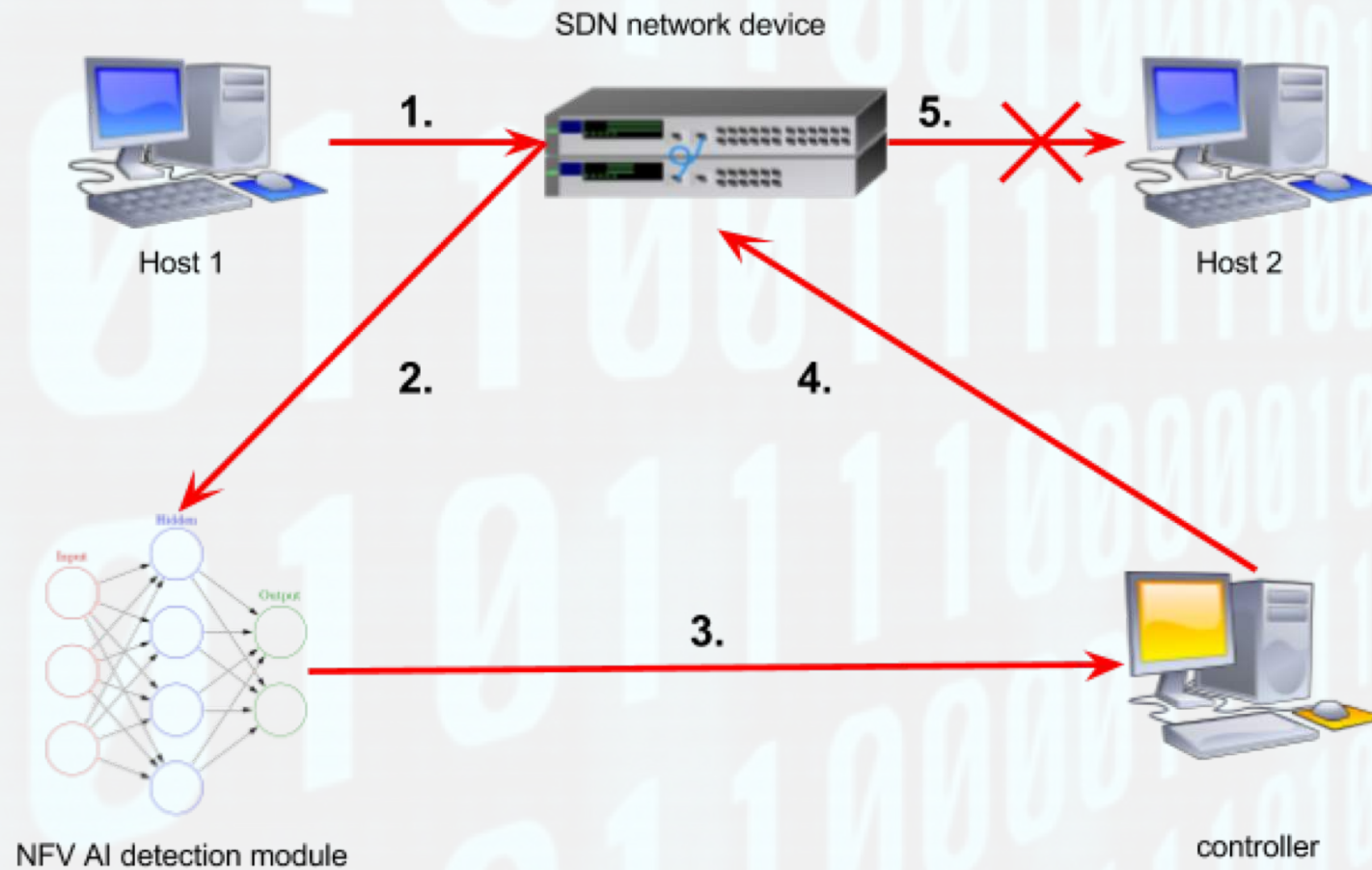


Application Classification among Tor

Algorithm	Accuracy	Precision	Recall	F-measure	
XGBoost	79.3	68.9	53.7	60.4	audio
		74.4	79.1	76.7	chat
		88.9	86.5	87.6	file
		66.8	56.0	61.0	email
		79.2	81.2	80.2	video
		84.2	86.5	85.3	voip
		96.6	92.7	94.6	p2p



Implementation on SDN



Demo



Special Thanks

- 林宗男教授
- **Project 成員:**
張育維、陳昀君、李宇哲、黃廉弼、蔡仲閔、劉錫臻、施柏諺、
盧冠蓉、蘇柏燁
- 謝謝小蘇邀稿
- 謝謝HICON議程組的肯定^_^

Reference

- **Deciphering Malware's use of TLS (without Decryption)**
<https://arxiv.org/pdf/1607.01639.pdf>
- **Characterization of Tor Traffic using Time based Features**
https://www.researchgate.net/publication/314521450_Characterization_of_Tor_Traffic_using_Time_based_Features
- **Detecting malware even when it is encrypted**
https://2018.bsidesbud.com/wp-content/uploads/2018/03/seba_garcia_frantisek_strasak.pdf
- **Deep Learning for Malicious Flow Detection**
<https://arxiv.org/pdf/1802.03358.pdf>



Thanks!



Email : aragorn51882@gmail.com

Facebook : 白貓肥宅