Towards Generalizable Network Threat Detection

Blake Anderson, PhD
September 2017
Data Science for Cyber-Security
Takeaways

• Network data is complicated:
  • Endpoint, network, and collection-point configurations impact data features
  • Applications/Protocols are constantly evolving

• Solving these challenges requires:
  • A detailed understanding of all aspects of the data generation process
  • More advanced data features to identify the underlying data generation process
TLS Overview

Client:

- ClientHello
- ServerHello/Certificate/ServerHelloDone
- ClientKeyExchange/ChangeCipherSpec/Finished
- ChangeCipherSpec/Finished
- Application Data

Server:

- Client: I support crypto!
- Server: I support that crypto, and I’m me!
- Client: Take this secret and let’s encrypt!
- Server: Your secret looks good; let’s encrypt!
- Client/Server: encrypted data!
Threats

Premises

Internet

Interception
Threats
Threats

- Premises
- Internet
- Infection
- Interception
Malware Trends

Malware's Exclusive Use of TLS

Malware's Exclusive Use of HTTP
Bestafera Malware - Uses TLS Exclusively; Key Logging and Exfil of Financial Data
Packet Lengths and Arrival Times

Google Search

Bestafera
TLS Client Fingerprint

Bestafera

Secure Sockets Layer
- TLSv1 Record Layer: Handshake Protocol: Client Hello
  - Content Type: Handshake (22)
  - Version: TLS 1.0 (0x0301)
  - Length: 214
- Handshake Protocol: Client Hello
  - Handshake Type: Client Hello (1)
  - Length: 210
  - Version: TLS 1.0 (0x0301)
- Random
  - Session ID Length: 0
  - Cipher Suites Length: 120
- Cipher Suites (60 suites)
  - Compression Methods Length: 1
  - Compression Methods (1 method)
  - Extensions Length: 48
- Extension: ec_point_formats
- Extension: elliptic_curves
- Extension: SessionTicket TLS
- Extension: Heartbeat

TLS Clients

(v:1.0.1r) OpenSSL
Tor
Outlook
Skype
Thunderbird
Data Collection and Training

- Metadata
- Packet lengths
- TLS
- DNS
- HTTP
Deploying Classifiers/Rules
Why is this Approach is Successful?

• We leverage unencrypted features about the application *and* the transmitted data:
Why is this Approach *not* Successful?

- We are assuming the benign training data generalizes.
  - The endpoints are configured similarly
  - The network is configured similarly
  - The data is being collected in the same way
- We are assuming the malicious training data generalizes.
  - The derived features aren’t sandbox artifacts
- We are assuming our labels correct.
Endpoint Artifacts
Variability in HTTP Request Sizes

Variability in Packet Sizes (The Same Host/URI in all Requests)
Variability in HTTP Request Sizes, cont.
TLS Dependence on Environment

- 1,305 unique malware samples were run under both Win7 and Win10:
  - 207 samples used the exact same TLS client parameters in both environments
  - 1,098 samples used the library provided by the underlying OS

- This affects the distribution of TLS parameters.
  - It also has secondary effects w.r.t. packets/bytes
HTTP Dependence on Environment

• 2,662 unique malware samples were run under both Win7 and Win10:
  • 1,035 samples used the exact same HTTP User-Agent in both environments
  • 1,627 samples varied their HTTP User-Agent in the separate runs

• This affects the distribution of HTTP parameters.
  • It also has secondary effects w.r.t. packets/bytes
Source Port Allocation

<table>
<thead>
<tr>
<th>System</th>
<th>Registered</th>
<th>Ephemeral</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-1023</td>
<td>1024-49151</td>
<td>49152-65535</td>
</tr>
</tbody>
</table>

WinXP: 1025-5000

Win7: 49152-65535

Linux: 32768-61000
Client Source Ports in the Wild
Network Artifacts
Interacting TCP Stacks

- Delayed ACKs, Nagle’s algorithm, slow-start, MSS, server load, network latency
RTT

(a) Amsterdam

(b) Bangalore
Collection Point Artifacts
MiTM/Proxies

certificate/etc. record altered

client_hello/etc. records altered
Multi-Packet Messages – GRO

- Collection points significantly affect packet sizes
Dataset-Level Obstacles
Browser Release Schedules

- Firefox:
  - v48.0
  - v49.0
  - v50.0
  - v51.0
  - v52.0
  - v53.0
  - v54.0

- Chrome:
  - v52.0
  - v53.0
  - v54.0
  - v55.0
  - v56.0
  - v57.0
  - v58.0
  - v59.0

- Safari:
  - v9.1.2
  - v9.1.3
  - v10.0
  - v10.0.1
  - v10.0.2
  - v10.0.3
  - v10.1

Release dates:
- Jul-2016
- Aug-2016
- Sep-2016
- Oct-2016
- Nov-2016
- Dec-2016
- Jan-2017
- Feb-2017
- Mar-2017
- Apr-2017
- May-2017
- Jun-2017
New Protocols - TLS 1.3

- Most records types will be encrypted
- New record padding options
- Zero-RTT Data
Poorly Labeled Datasets

Internet

- google.com
- facebook.com
- reddit.com
- wgosfcn.net

Network

Malware Sandbox

Internet

- googleapis.com
- cloudflare.com
- knownbad.com

- 203.153.xx.xx
- 104.131.xx.xx
- 85.214.xx.xx
- 50.116.xx.xx

- amazon-adsystem.com
- cloudflare.com
Moving Forward
Assumptions

• We cannot obtain perfectly labeled training data
• We cannot continuously collect training data from an environment
• We can obtain high-level statistics about a target network
• We can choose which network data features are collected
Solution Overview

Enhanced NetFlow

Network Inference
OS Inference
Application Inference
Malware Detection

Endpoint/Network Context
OS, Applications, PMTU, RTT, Infection, ...

classifier description
auxiliary data
fingerprint rules

Malware Family

Labels
Noisy Labels - Setup

• Randomly flip labels in the benign/malicious datasets
  • 0.5% to 5.0% of the samples

• Training set, August 2015-May 2016:
  • 620,072 samples from an enterprise network
  • 208,368 samples from a malware sandbox

• Testing set, September 2016:
  • 735,195 samples from a distinct enterprise network
  • 21,114 samples from a malware sandbox
Noisy Labels - Results

![Graph showing accuracy for Enterprise and Malware datasets with different models under varying percentage of corrupted labels in the training set.]
Synthetic Samples - Setup

Malware
- ThreatGrid
  - 50,000 Flows

Ent_1
- 150,000 Flows

Modify Samples

Feature Selection

Classifier

Ent_2
- 150,000 Flows
# Synthetic Samples - Results

<table>
<thead>
<tr>
<th>True Training Set</th>
<th>Synthetic Training Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,075 total alarms</td>
<td>813 total alarms</td>
</tr>
<tr>
<td>118 “bad” flows</td>
<td>685 “bad” flows</td>
</tr>
<tr>
<td>957 “good” flows</td>
<td>128 “good” flows</td>
</tr>
</tbody>
</table>
Conclusions

- Network data is complicated:
  - Endpoint, network, and collection-point configurations impact data features
  - Applications/Protocols are constantly evolving

- Solving these challenges requires:
  - A detailed understanding of all aspects of the data generation process
  - More advanced data features to identify the underlying data generation process
References

- Blake Anderson, David McGrew; Machine Learning for Encrypted Malware Traffic Classification: Accounting for Noisy Labels and Non-Stationarity; KDD, 2017
- Blake Anderson, David McGrew; Identifying Encrypted Malware Traffic with Contextual Flow Data; CCS AISec, 2016
- Open source package for network data capture and analysis: https://github.com/cisco/joy
Thank You