{ "name": "Android 4.0.4", "cs": ['c014', 'c00a', '0039', '0038', 'c00f', 'c005', '0035', 'c012', 'c008', '0016', '0013', 'c0
{ "name": "Android 4.1.1", "cs": ['c014', 'c00a', 'c022', 'c021', '0039', '0038', 'c00f', 'c005', '0035', 'c012', 'c008', 'c01c

Detecting Threats, Not Sandboxes

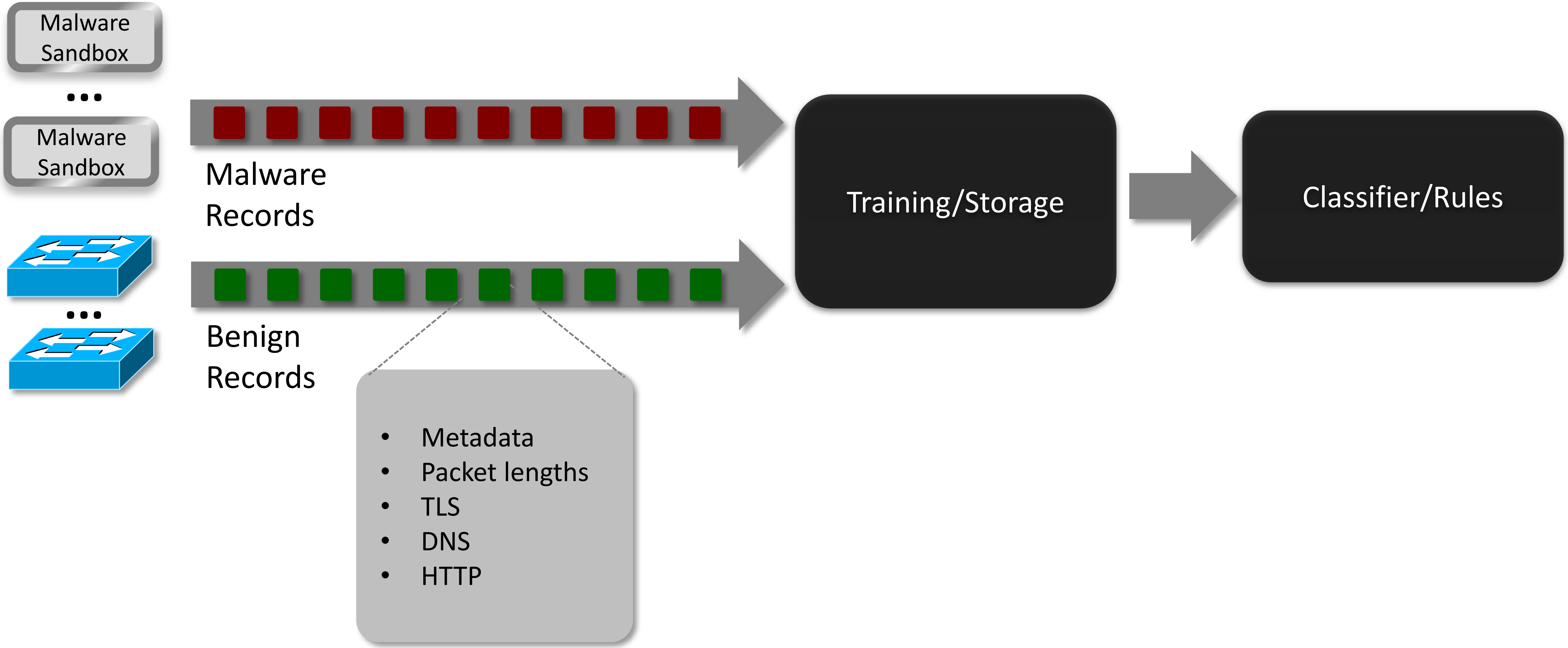
(Characterizing Network Environments to Improve Malware Classification)

Blake Anderson (blake.anderson@cisco.com), David McGrew (mcgrew@cisco.com)

FloCon 2017

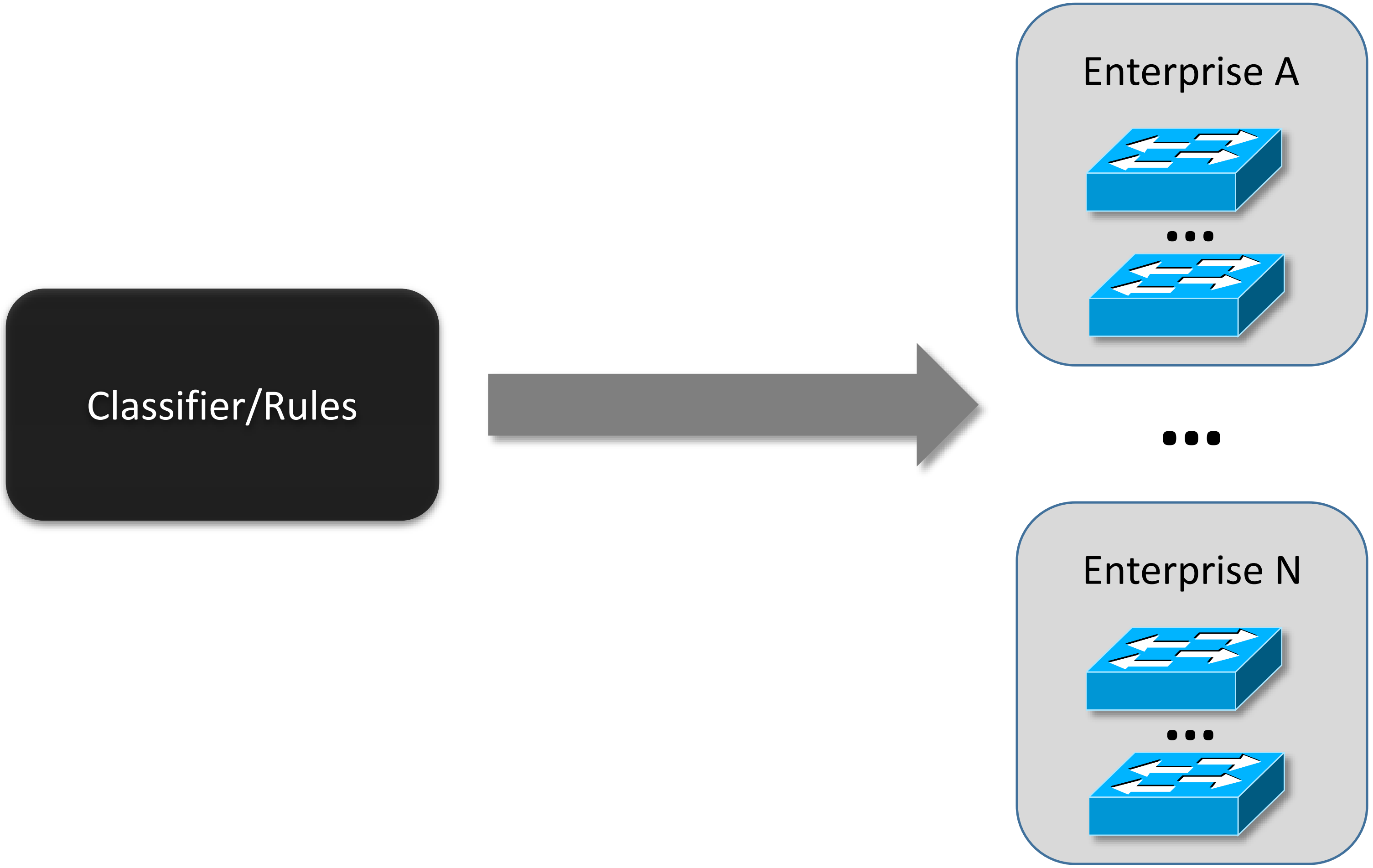
January, 2017

Data Collection and Training



- Metadata
- Packet lengths
- TLS
- DNS
- HTTP

Deploying Classifier/Rules



Problems with this Architecture

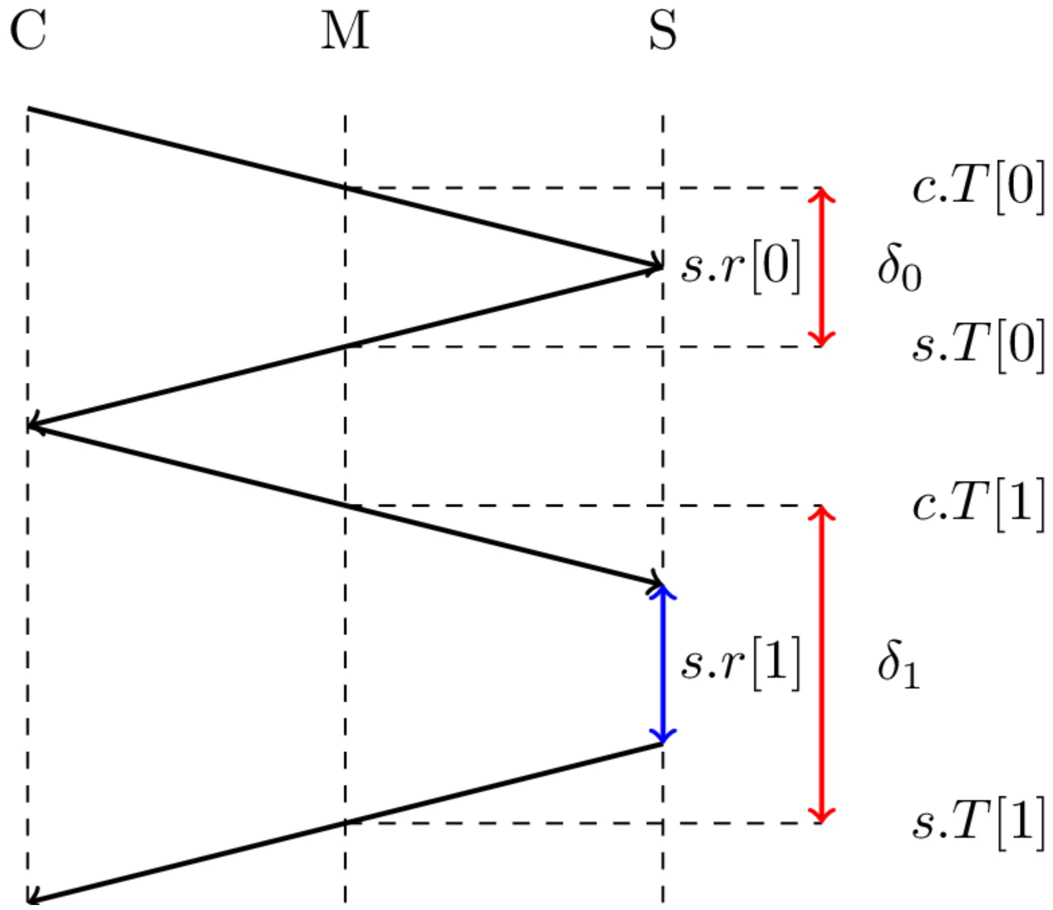
- Models will not necessarily translate to new environments
  - Will be biased towards the artifacts of the malicious / benign collection environments
- Collecting data from all possible end-point/network environments is not always possible

Network Features in Academic Literature

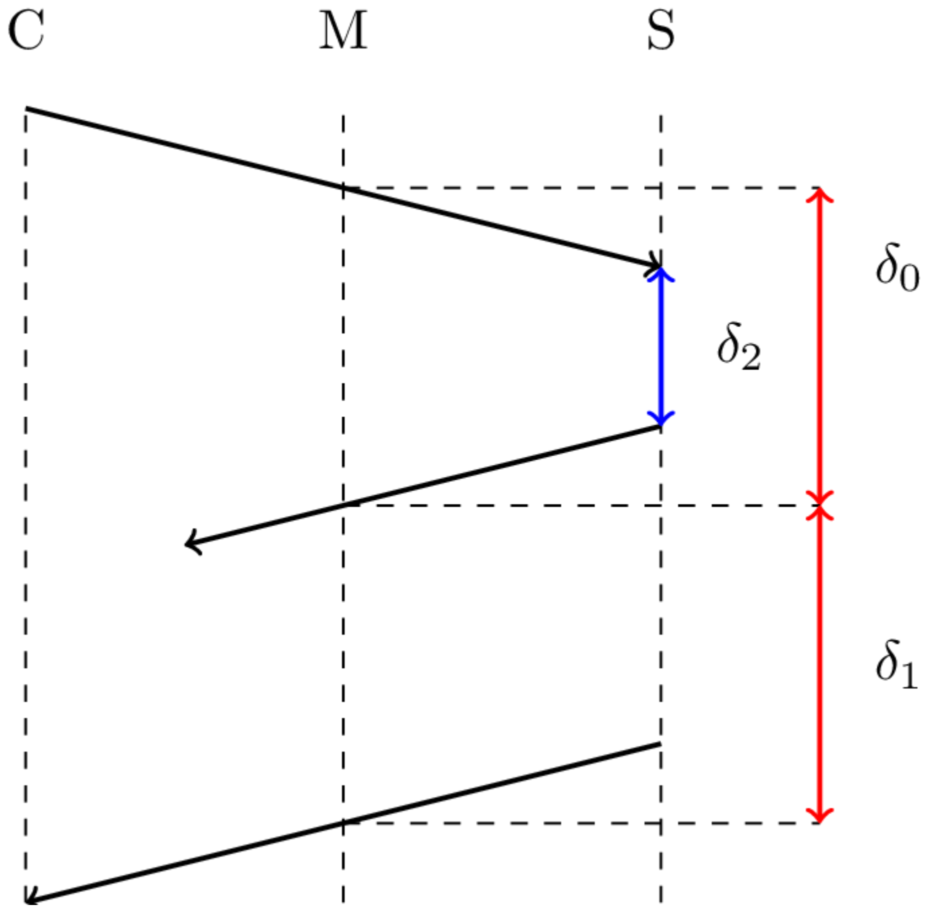
- 2016 – IMC / USENIX Security / NDSS
  - Packet sizes
  - Length of URLs
- 2012:2015 – CCS / SAC / ACSAC / USENIX Security
  - Time between ACKs
  - Packet sizes in each direction
  - Number of packets in each direction
  - Number of bytes in each direction

Network/Transport-Level Robustness

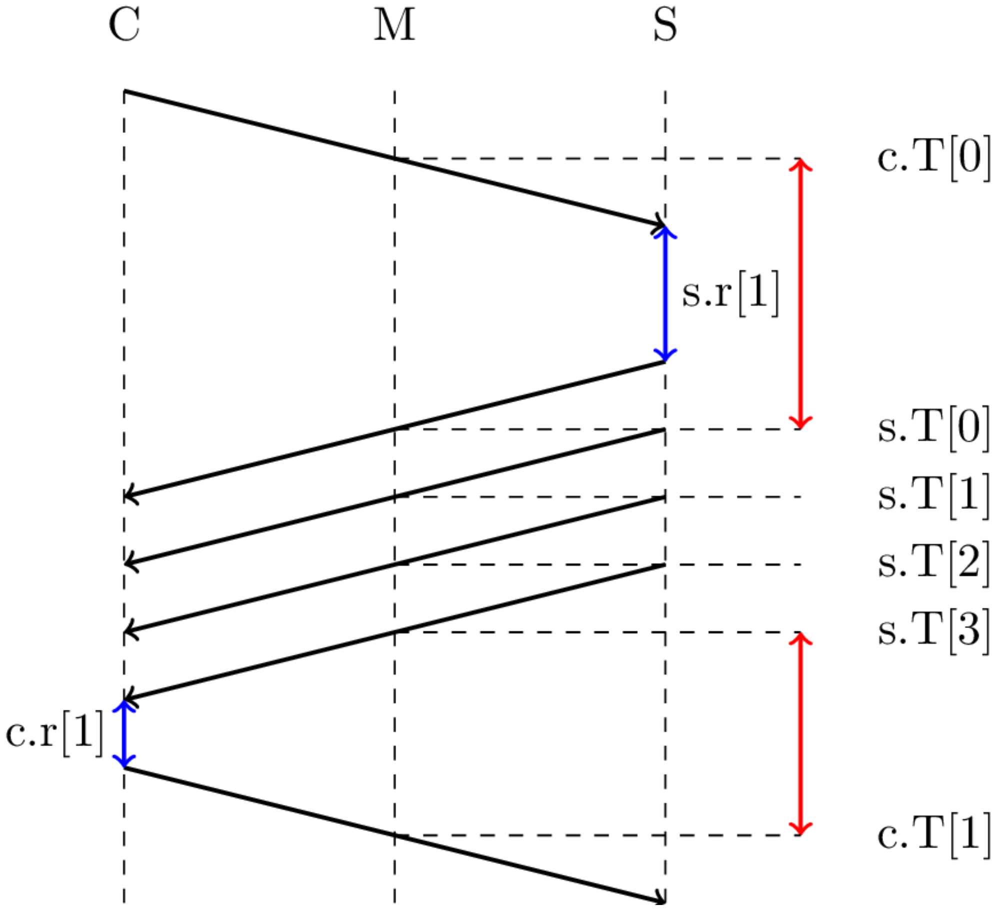
Ideal TCP Session



Inbound Packet Loss



Multi-Packet Messages

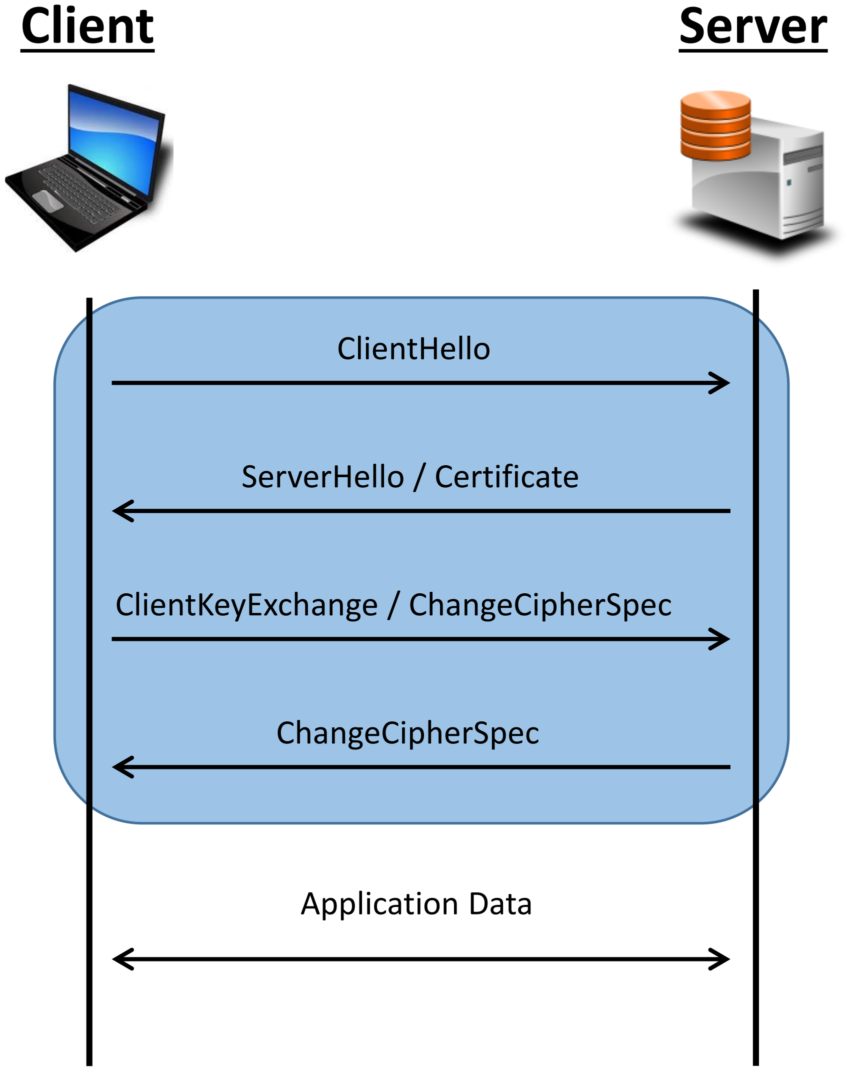


Collection Points / MTU / Source Ports

- Collection points significantly affect packet sizes
  - Same flow collected within a VM and on the host machine will look very different
- Path MTU can alter individual packet sizes
- Source ports are very dependent on underlying OS
  - WinXP: 1024-5000
  - NetBSD: 49152-65535

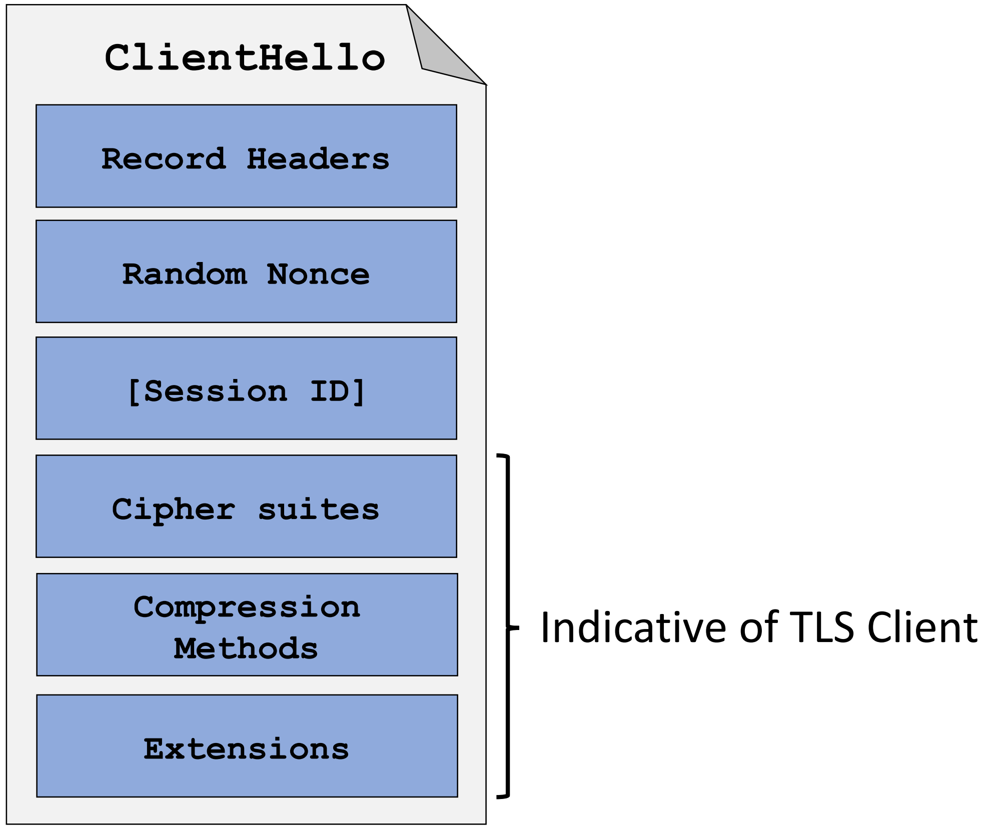
Application-Level Robustness

TLS Handshake Protocol



Client
Server
ClientHello
ServerHello / Certificate
ClientKeyExchange / ChangeCipherSpec
ChangeCipherSpec
Application Data

TLS Client Fingerprinting



ClientHello

Record Headers

Random Nonce

[Session ID]

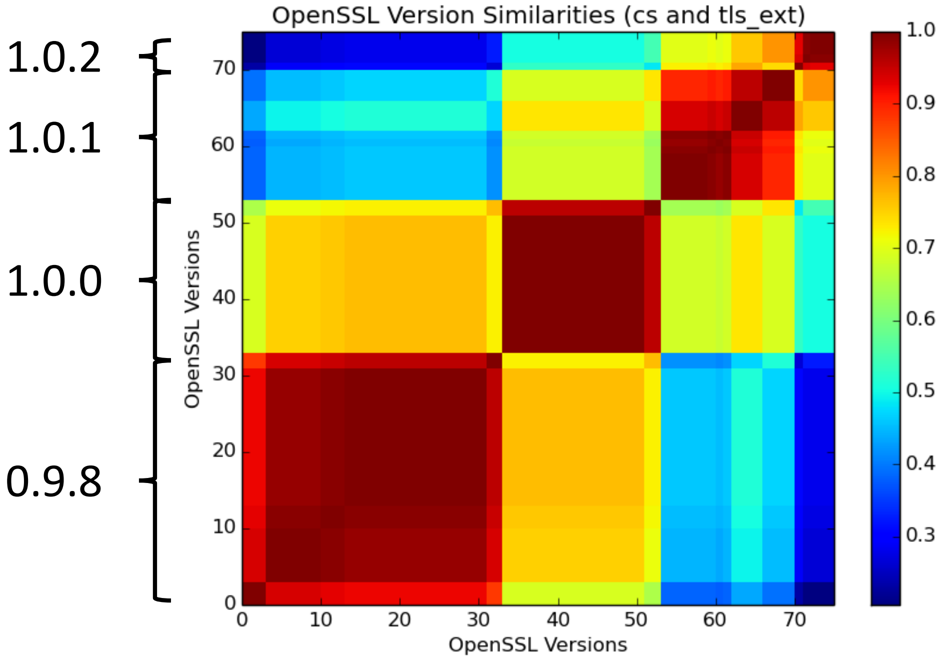
Cipher suites

Compression Methods

Extensions

Indicative of TLS Client

OpenSSL Versions



TLS Dependence on Environment

- 73 unique malware samples were run under both WinXP and Win7
  - 4 samples used the exact same TLS client parameters in both environments
  - 69 samples used the library provided by the underlying OS (some also had custom TLS clients)
- Effects the distribution of TLS parameters
  - Also has secondary effects w.r.t. packet lengths

HTTP Dependence on Environment

- 152 unique malware samples were run under both WinXP and Win7
  - 120 samples used the exact same set of HTTP fields in both environments
  - 132 samples used the HTTP fields provided by the underlying OS’s library
- Effects the distribution of HTTP parameters
  - Also has secondary effects w.r.t. packet lengths

Solutions

Potential Solutions

- Collect training data from target environment
  - Ground truth is difficult
  - Models do not translate
- Discard Biased Samples
  - Not always obvious which features are network/endpoint-independent

- Train models on network/endpoint-independent features
  - Not always obvious which features are network/endpoint-independent
  - This often ignores interesting behavior
- Modify existing training data to mimic target environment
  - Not always obvious which features are network/endpoint-independent
  - Can capture interesting network/endpoint-dependent behavior
  - Can leverage previous capture/curated datasets

Results

- L1-logistic regression
- Meta + SPLT + BD
  - 0.01% FDR: 1.3%
  - Total Accuracy: 98.9%

- L1-logistic regression
- Meta + SPLT + BD + TLS
  - 0.01% FDR: 92.8%
  - Total Accuracy: 99.6%

Results (without Schannel)

- L1-logistic regression
- Meta + SPLT + BD
  - 0.01 FDR: 0.9%
  - Total Accuracy: 98.5%

- L1-logistic regression
- Meta + SPLT + BD + TLS
  - 0.01 FDR: 87.2%
  - Total Accuracy: 99.6%

Conclusions

- It is necessary to understand and account for the biases present in different environments
  - Helps to create more robust models
  - Models can be effectively deployed in new environments
- We can reduce the number of false positives related to environment artifacts
- Data collection was performed with: Joy

Thank You