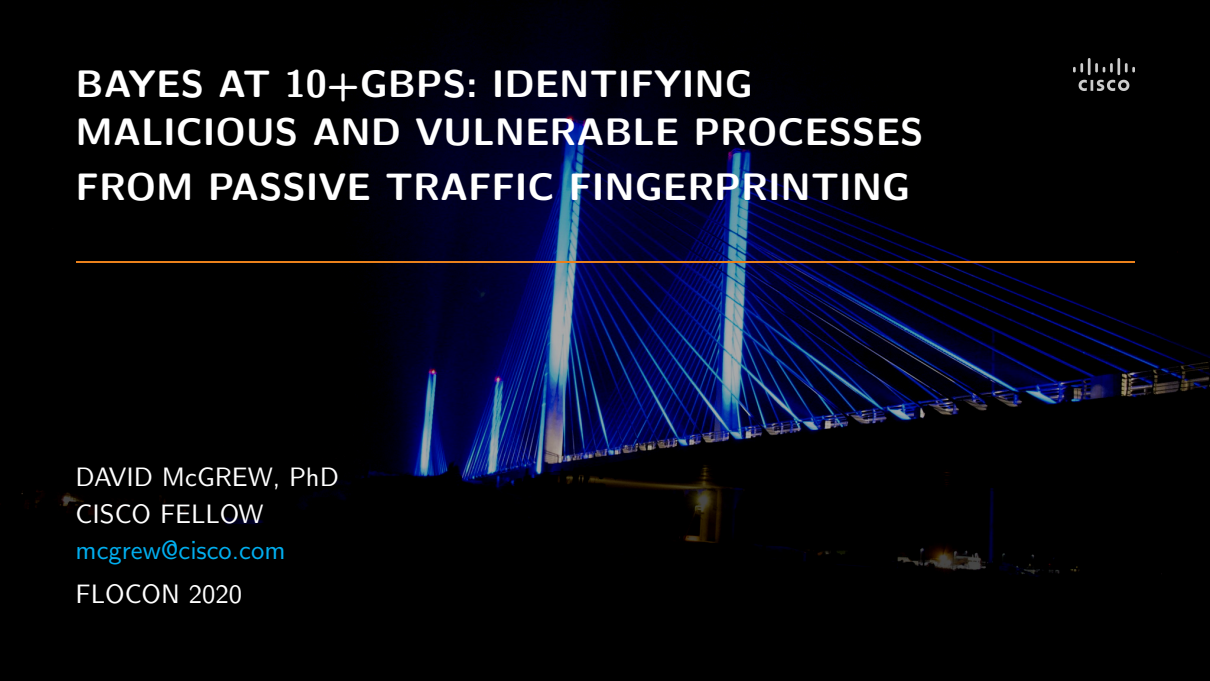
BAYES AT 10+GBPS: IDENTIFYING MALICIOUS AND VULNERABLE PROCESSES FROM PASSIVE TRAFFIC FINGERPRINTING



DAVID McGREW, PhD
CISCO FELLOW
mcgrew@cisco.com

FLOCON 2020

PEOPLE

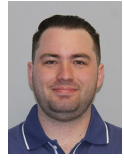
David McGrew
Security Research

Blake Anderson
Security Research

Brandon Enright
CSIRT

Adam Weller
CSIRT

Lucas Messenger
CSIRT



BACKGROUND: TECHNOLOGY TRENDS DISRUPTING VISIBILITY

End Host Monitoring

- **Consumerization (BYOD)** makes deployment hard
- **Virtualization** makes circumvention easy
- **Cloud computing** clones and relocates software
- **IoT** devices don't support agents

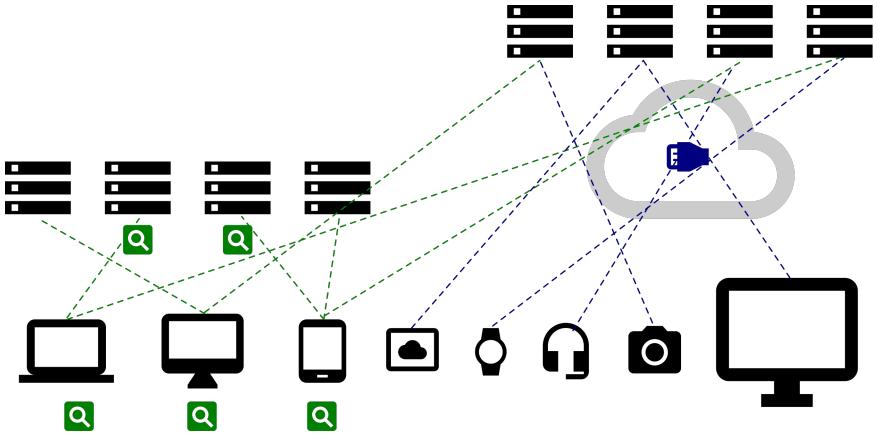
Network Monitoring

- **Cloud and distributed architectures** require network encryption
- **Network encryption** impedes session data inspection
  - Intrusion detection, data leakage detection, attack detection, ...
- **New protocols** like QUIC, Wireguard, ...

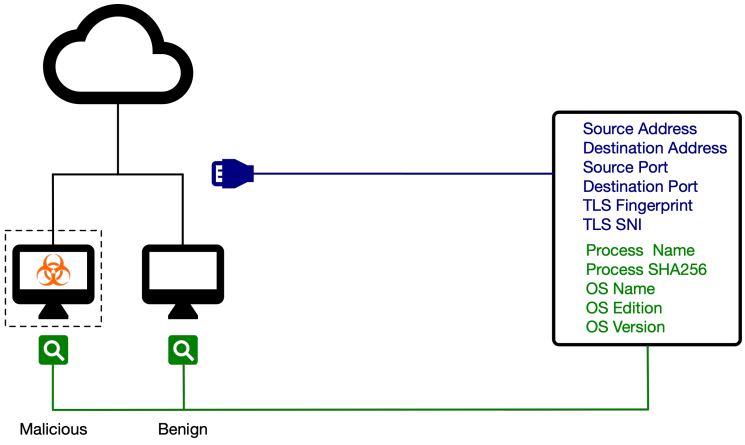
OUR GOALS

- Infer [malware] process from observations of TLS sessions
- High accuracy
  - Extensive ground truth data
  - Use all characteristic data features
  - Generalize to any Internet destination
- Support high data rates on server class hardware, with easy deployment
- Immediate inference from initial packet(s)
- Enrich CSIRT Splunk system
- Interpretability

VISIBILITY USING NETWORK AND END HOST MONITORING



TRAINING DATA FROM NETWORK/END-HOST FUSION



Produces ∼200M new labeled session records *per day*

Host data: AnyConnect NVM, network data: Mercury

PROCESS INFERENCE EXAMPLE

$$c: \begin{bmatrix} \texttt{Destination Port:} & \texttt{443} \\ \texttt{Destination Address:} & \texttt{192.168.60.1} \\ \texttt{TLS ProtocolVersion:} & \texttt{0301} \\ \texttt{TLS CipherSuites:} & \texttt{0035...0003} \\ \texttt{TLS Extensions:} & \texttt{None} \\ \texttt{TLS Server Name:} & \texttt{None} \end{bmatrix} \rightarrow \begin{bmatrix} \texttt{Process:} & \texttt{chrome.exe} \\ \texttt{Version:} & \texttt{76.0.3809.132} \\ \texttt{SHA-256:} & \texttt{5616...9acc} \\ \texttt{Category:} & \texttt{browser} \\ \texttt{OS:} & \texttt{WinNT} \\ \texttt{OSversion:} & \texttt{10.0.17134} \\ \texttt{OSedition:} & \texttt{Enterprise} \end{bmatrix}$$

TLS FINGERPRINT DATABASE


```
{
  "str_repr": "(0303)(0081c02cc02bc030c02f009f009ec024c023c028c027c00ac009c014c013009d009c003d003c0035002f000a)...",
  "total_count": 4187,
  "process_info": [
  {
      "process": "OneDrive.exe",
      "sha256": "53135CD348E8E80BEE5B156F2F95EE81F1176B818768A4421CA775A99F9D313C",
      "application_category": "storage",
      "count": 516,
      "classes_ip_as": {
        "8075": 373,
        "8068": 143
      },
      "classes_hostname_domains": {
        "windows.net": 214,
        "sharepoint.com": 176,
        "live.com": 95,
        "msn.com": 18,
        "windows.com": 9,
        "microsoft.com": 4
      },
      "os_info": {
        "(WinNT)(Windows 10 Enterprise)(10.0.17134)": 516
      }
  },
  ...
}
```

FINGERPRINT DATABASE STATISTICS

Sources

Source	Fingerprints	Sessions
Malware Sandbox	5,633	$3.61 \cdot 10^{7}$
End Host Agent	7,909	$5.43 \cdot 10^{9}$
Unlabeled	64,214	$4.10 \cdot 10^{10}$
Total	69,310	$4.65 \cdot 10^{10}$

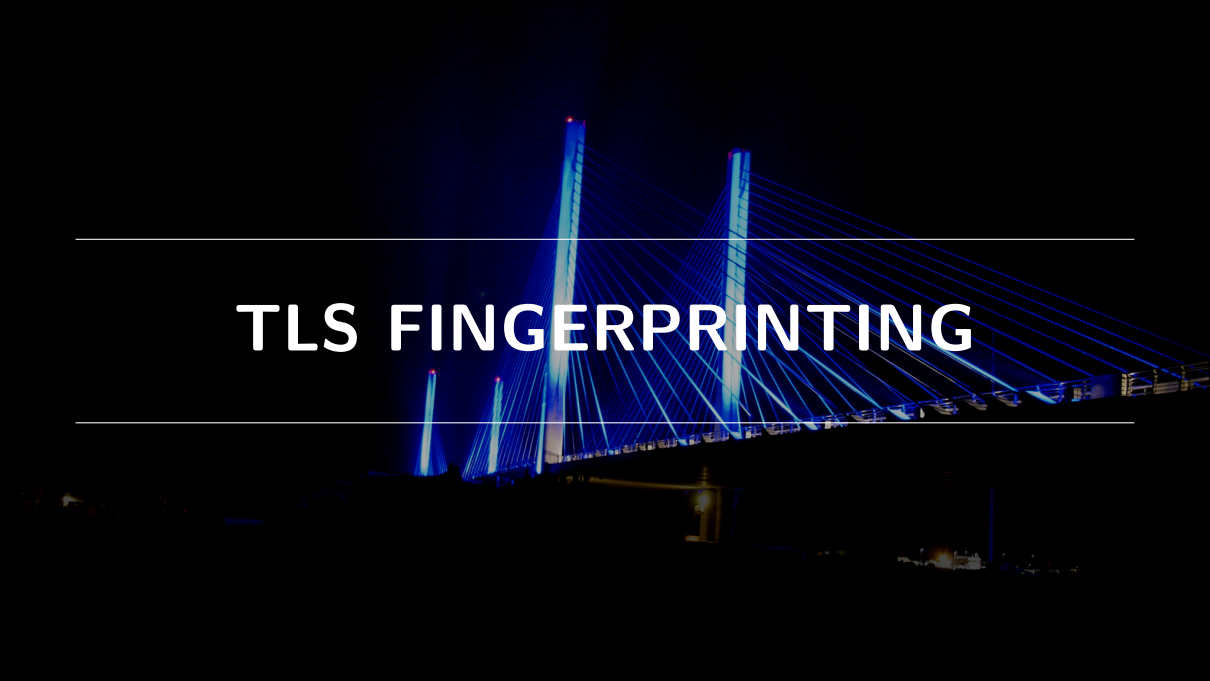
Application Categories

Category	Population
browser	6416
programming	1839
communication	1429
system	1046
email	725
productivity	627
storage	597
gaming	334
vpn	269
sysadmin	231
security	223
music	188
enterprise	166
photography	141
credential_manager	58
remote_desktop	57
misc	52
video	23
health	3
virtual_machine	2

Strings per Process

Number of Strings	Population
1	5559
2	1436
3-4	771
5-8	461
9-16	197
17-32	85
33-64	46
65-128	11
129-256	3
257-512	2

TLS Beyond the Browser: Combining End Host and Network Data to Understand Application Behavior, ACM IMC 2019



TLS FINGERPRINTING

DATA FEATURES AND ANALYSIS

String Analysis

Features are 'just bytes'

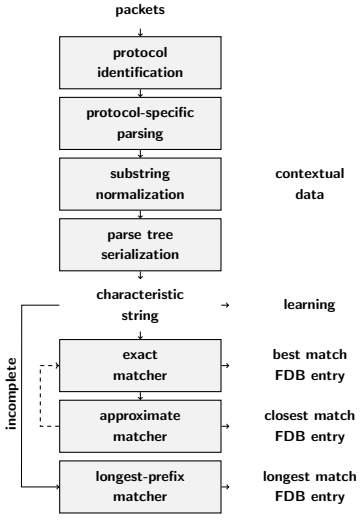
- TLS Version
- TLS Ciphersuite Offer List
- TLS Extension List

Context Analysis

Features have semantic meaning

- IP Destination Address (subnets)
- TCP Destination Port (ranges)
- TLS Server Name (domains)

CHARACTERISTIC STRING PROCESSING



SELECTIVE PACKET PARSING

```
16                      ContentType
03 01                   ProtocolVersion
02 00                   RecordLength
01                      HandshakeType
00 01 fc                HandshakeLength
03 03                   ProtocolVersion
e5 2c a9 01 ...fa 69 46 Random
20                      SessionIDLength
a1 f1 67 1b ...0a 17 69 SessionID
00 14                   CipherSuiteVectorLength
00 39 00 38 ...2f 00 07 CipherSuiteVector
.                       CompressionMethodsLength
.*                      CompressionMethodsVec
00 0a                   ExtensionsVectorLength
00 00                   ExtensionType
00 18                   ExtensionLength
00 16 00 00 ...63 6f 6d ExtensionData
00 0b                   ExtensionType
00 02                   ExtensionLength
01 00                   ExtensionData
```

Characteristic String

((0303)(00390038...2f0007)((0000)(000b00020100)))

- Bracket notation expresses parse tree
- Strings are self-typing
- General and flexible

SEMANTIC ANALYSIS OF DESTINATION CONTEXT

(NAÏVE) BAYESIAN INFERENCE

$$\text{process} = \underset{\text{all processes}}{\text{argmax}}\ \mathbf{P}(\text{process} \mid \text{fingerprint}, \text{da}, \text{dp}, \text{sni})$$

- Inference on fingerprint **and** destination context
- Interpretable
- **ML model captures knowledge of the Internet**

GENERALIZING THROUGH INTERNET CONTEXT

The fundamental goal of machine learning is to generalize beyond the examples in the training set - Pedro Domingos

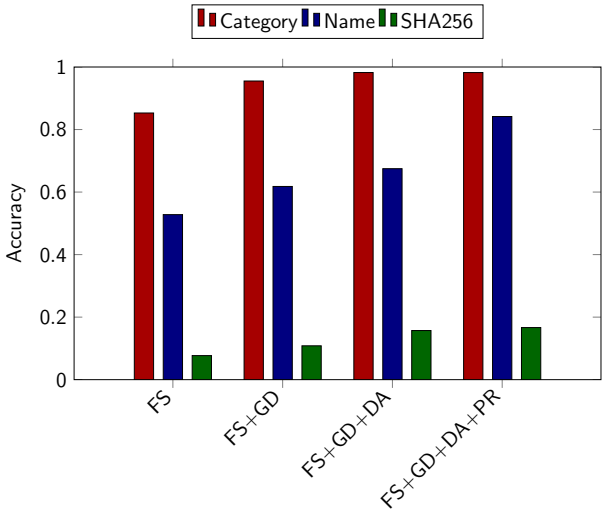
- Problem: what probabilities do we assign to addresses outside the training set?
- Solution: compute probabilities over *equivalence classes* of addresses
  - Addresses are equivalent if they are in the same BGP AS, or related via DNS, or owned by the same company, or related via PKIX

$$\mathbf{P}(f_i \mid z) = \prod_{j=1,p} \mathbf{P}(\gamma_i^j(f_i) \mid z).$$

INFERENCE EXAMPLE

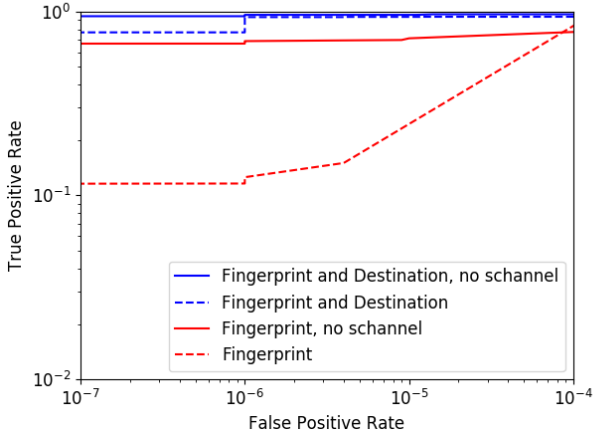

```
{
  "fingerprints": {
    "tls": "(0303)(c02bc02fc02cc030c00ac009c013c01400330039002f0035000a00ff)((0000)(000b000403000102)(000a001c00 ... 01))"
  },
  "tls": {
    "sni": "www.mku4kwjx7t.com"
  },
  "analysis": {
    "process": "tor.exe",
    "score": 0.999988,
    "malware": 1,
    "p_malware": 1
  },
  "sa": "64.100.12.6",
  "da": "62.210.5.178",
  "pr": 6,
  "sp": 4743,
  "dp": 443,
  "time_start": 1564612518.326139
}
```

PROCESS IDENTIFICATION ACCURACY



FS	Fingerprint String
DG	Generalized Destination Info
DA	Destination Address
PR	Prior Result

MALWARE TLS SESSION IDENTIFICATION



Single-session analysis of TLS features with[out] destination context, with[out] schannel


MERCURY

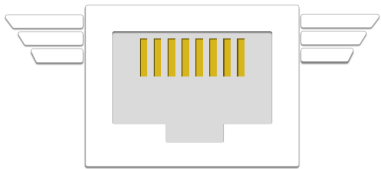
MERCURY: PACKET METADATA CAPTURE AND ANALYSIS

Goals

- 20+ Gbps on modern servers
- Minimal dependencies
- Linux AF_PACKET/TPACKETv3
- FPs: TLS, TCP, HTTP, DHCP
- Online NB inferencing
- FPDB updated weekly

Download

`https://github.com/cisco/mercury`



Disclaimer: accuracy requires using an FPDB appropriate for the network

FUTURE WORK

- Publish characteristic string analysis details
- Improve Naïve Bayes analysis with more context
- Collaborate to extend database and improve analysis
- Fingerprint more protocols
- Combined Operating System / Process inference
- Robustness across disparate networks


cisco
THANK YOU