Ground-Truth Driven Cyber Security Research: Some Examples

Mustaque Ahamad, Georgia Tech, NYU Abu Dhabi and Pindrop
Paul Royal, Georgia Tech
Terry Nelms, Georgia Tech & Damballa
Roberto Perdisci, University of Georgia
Bharat Srinivasan, Georgia Tech
Payas Gupta, NYU Abu Dhabi
Vijay Balasubramaniyan, Pindrop Security

Background

- Georgia Tech Information Security Center
  - Founded in 1998
  - About a dozen faculty, 30+ PhD students
  - MS degree program in cyber security
- Research philosophy
  - Data-driven and high impact research
- Research thrusts
  - Understanding emerging threats, mobile security, converged networks security & crypto

Data Driven Cyber Security Research

- Security is about assumptions and guarantees
- What assumptions can we make about the nature of threats?
  - Evolution from hackers and criminals to nation-states
- Ground-truth based approach
  - Observe, understand and defend
- Allows validation in a realistic setting

Agenda: Examples of Data-Driven Research

- GTISC MTrace System (Paul Royal)
  - Scalable malware analysis
- ExecScent
  - Malware family attribution via communication templates
- Phoneypot
  - Securing the emerging telephony ecosystem
- Data sharing and coordination challenges

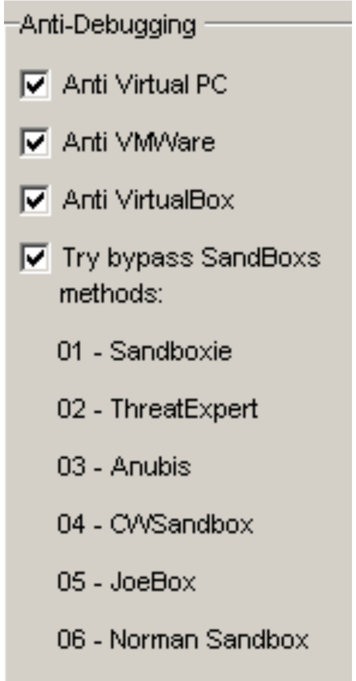
Example 1: Mtrace: Malware Analysis (Paul Royal)

- Malware is the centerpiece of current threats on the Internet
  - Botnets (spamming, DDOS, etc.)
  - Information Theft
  - Financial Fraud
- Used by Real Criminals
  - Criminal Infrastructure
  - Domain of Organized Crime

Malware Cont’ d

- There is a pronounced need to understand malicious software behavior
- *Malware analysis* is the basis for understanding the intentions of malicious programs
  - Threat Discovery and Analysis
  - Compromise Detection
  - Forensics and Asset Remediation

Malware Analysis - Transparency



Anti-Debugging

☑ Anti Virtual PC

☑ Anti VMWare

☑ Anti VirtualBox

☑ Try bypass SandBoxs methods:

01 - Sandboxie

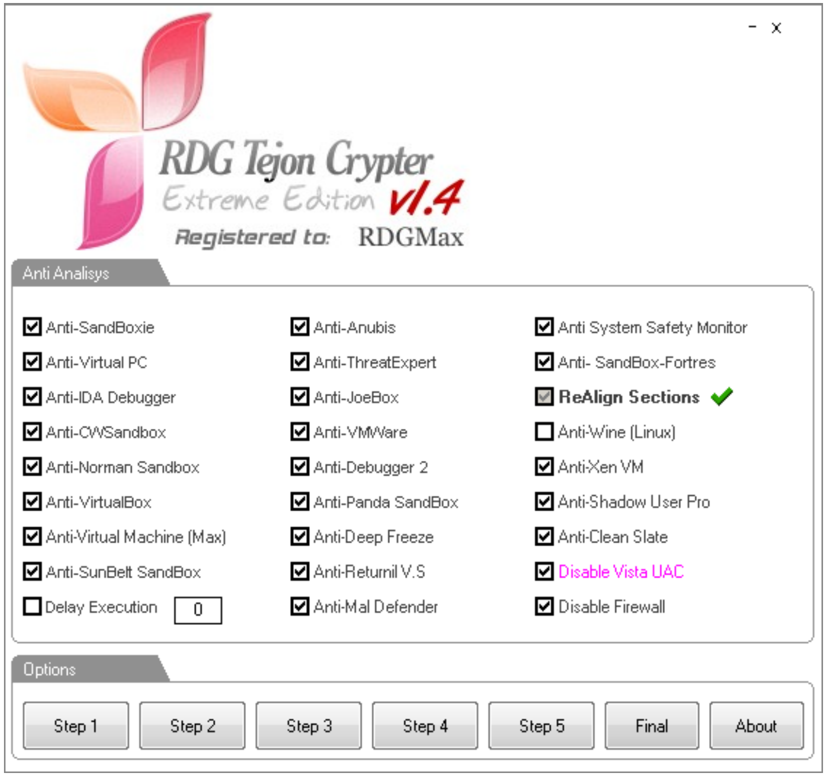
02 - ThreatExpert

03 - Anubis

04 - CWSandbox

05 - JoeBox

06 - Norman Sandbox



RDG Tejon Crypter
Extreme Edition v1.4
Registered to: RDGMax

Anti Analisys

☑ Anti-SandBoxie
☑ Anti-Virtual PC
☑ Anti-IDA Debugger
☑ Anti-CWSandbox
☑ Anti-Norman Sandbox
☑ Anti-VirtualBox
☑ Anti-Virtual Machine (Max)
☑ Anti-SunBelt SandBox
□ Delay Execution 0

☑ Anti-Anubis
☑ Anti-ThreatExpert
☑ Anti-JoeBox
☑ Anti-VMWare
☑ Anti-Debugger 2
☑ Anti-Panda SandBox
☑ Anti-Deep Freeze
☑ Anti-Returnil V.S
☑ Anti-Mal Defender

☑ Anti System Safety Monitor
☑ Anti- SandBox-Fortres
☑ ReAlign Sections ✓
□ Anti-Wine (Linux)
☑ Anti-Xen VM
☑ Anti-Shadow User Pro
☑ Anti-Clean Slate
☑ Disable Vista UAC
☑ Disable Firewall

Options

- Analysis tool/environment detection is a standard malware feature

Transparency Cont' d

- GTISC' s Idea: Use Intel VT as a malware analysis technology
  - External
    - No in-guest components to detect
  - Capable
    - Functionality sufficient to build analysis tools
  - "Equivalent"
    - Hardware-assisted nature offers same instruction-execution semantics
- Created tools supporting multiple tracing granularities
  - Coarse-grained tracing via SYSENTER_EIP_MSR displacement
    - e.g., System call tracing
  - Fine-grained tracing via TF injection
    - e.g., Precision automated unpacking

Malware Analysis - Automation

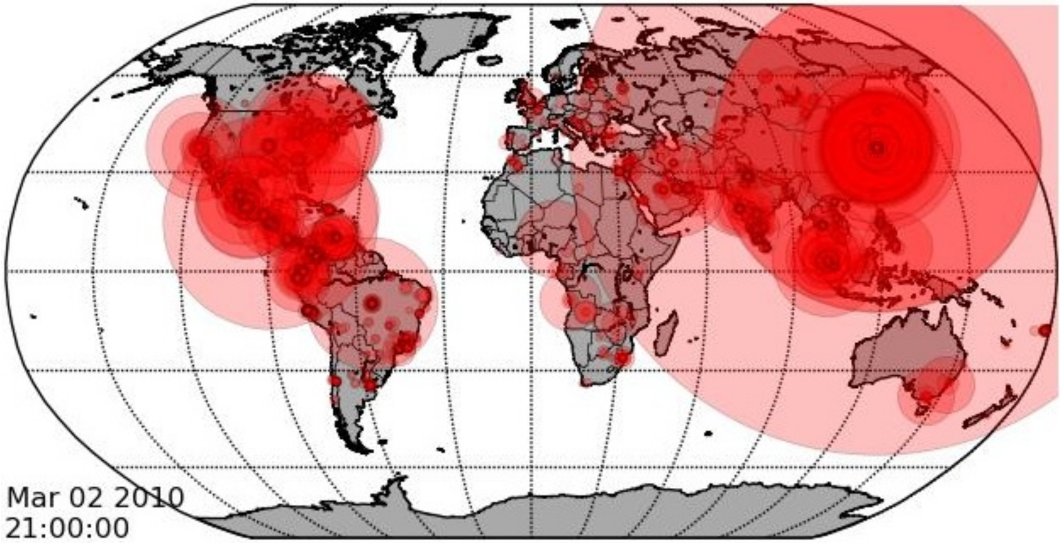
- DIY kits, packing tools, server-side polymorphism vastly increase volume of samples
- GTISC collects over 100,000 new samples each day
  - Collected from crawlers, mail filters, honeypots, user submissions, and malware exchanges
- Volume makes manual analysis untenable

Automation Cont’d

- GTISC has built a horizontally scalable, automated malware analysis framework
  - Each sample executed in a sterile, isolated environment
  - Intel VT used to ensure transparency
  - Structured representations of network actions placed inside intelligence database
    - C&C domains, anomalous outbound netflow, malicious download URLs, malware-generated email subjects, etc.
- Database used by corporate security groups, hosting providers, domain registrars, and law enforcement

Leveraging Intelligence - Mariposa

- Case Study: Mariposa
  - Large, data-stealing botnet
    - Used to steal credit card, banking information
    - Compromises in half of Fortune 1000
  - Before takedown, over 1M members



Mariposa Cont' d

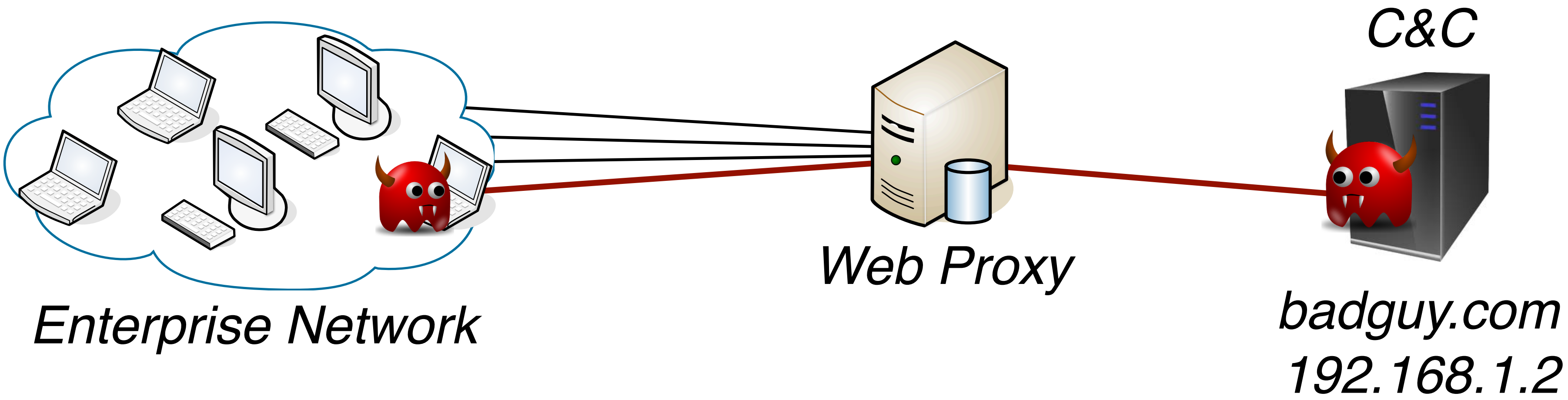
- Takedown Timeline
  - Spring 2009: Mariposa discovery
  - Fall 2009: International Mariposa Working Group (MWG) formed
    - Defence Intelligence, GTISC, Panda Antivirus, FBI, Guardia Civil (Spanish LEO)
  - December 2009: All C&C domains shutdown and sinkholed within hours of the first
    - Operators panic; log into domain management services from home systems
      - Warrants issued to operators' ISP
  - January 2010: Operators arrested
    - 800,000 financial credentials found on one operator' s home systems

Example 2: ExecScent: Mining for New C&C Domains in Live Networks with Adaptive Control Protocol Templates

Terry Nelms, Roberto Perdisci and Mustaque Ahamad

Appeared in Usenix Security Symposium, August 2013.

Modern Malware Networking



ExecScent Goals & Observations

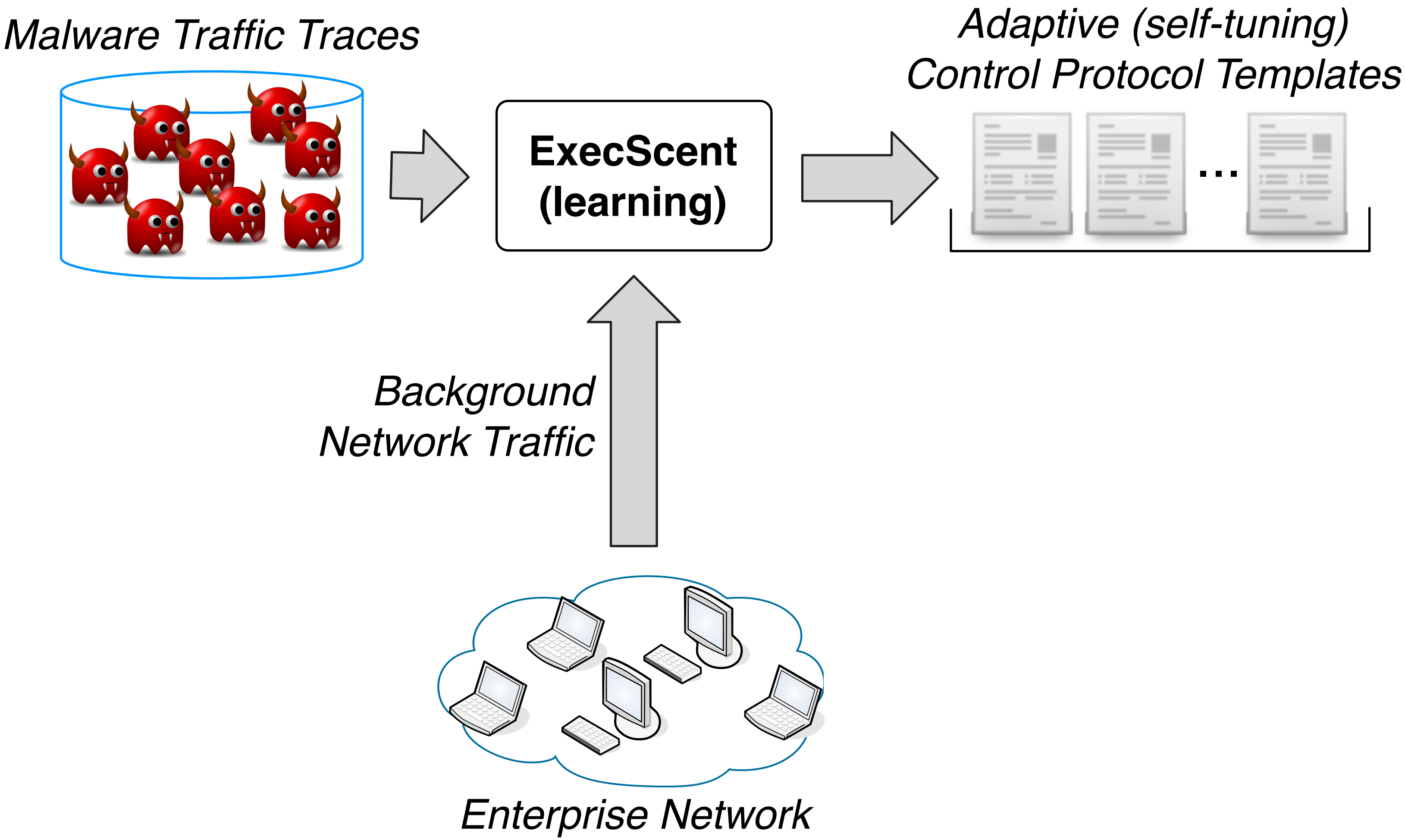
- Goals:
  - Network detection domains & hosts.
  - Malware family attribution.

- Observations:
  - C&C protocol changes infrequently.
  - HTTP C&C application layer protocol.

Adaptive Control Protocol Templates

- Structure of the protocol.
- Self-tuning.
- Entire HTTP request.

ExecScent Overview



Malware Traffic Traces

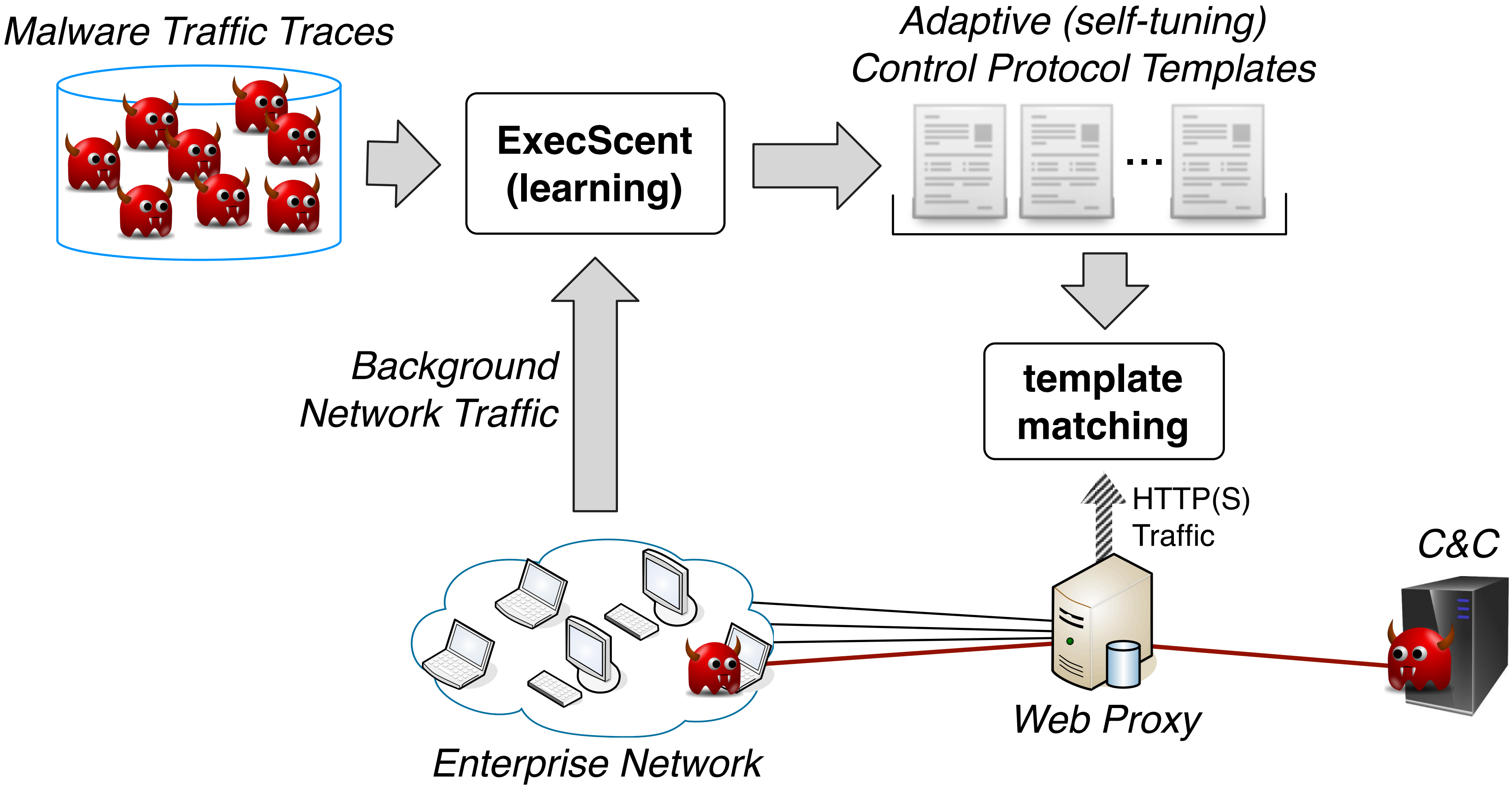
ExecScent (learning)

Adaptive (self-tuning) Control Protocol Templates

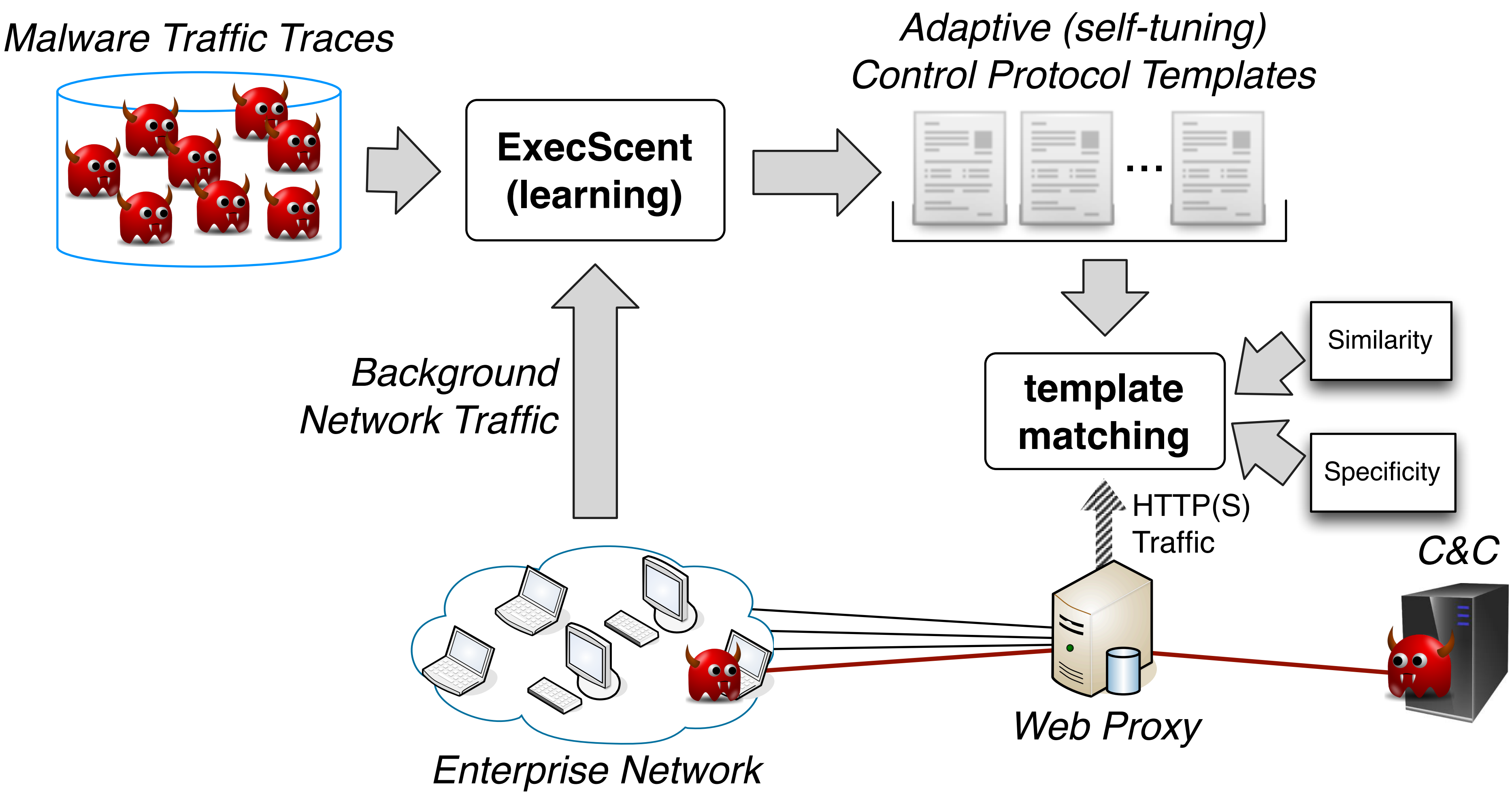
Background Network Traffic

Enterprise Network

ExecScent Overview



ExecScent Overview



Malware Traffic Traces

ExecScent (learning)

Adaptive (self-tuning) Control Protocol Templates

Background Network Traffic

template matching

Similarity

Specificity

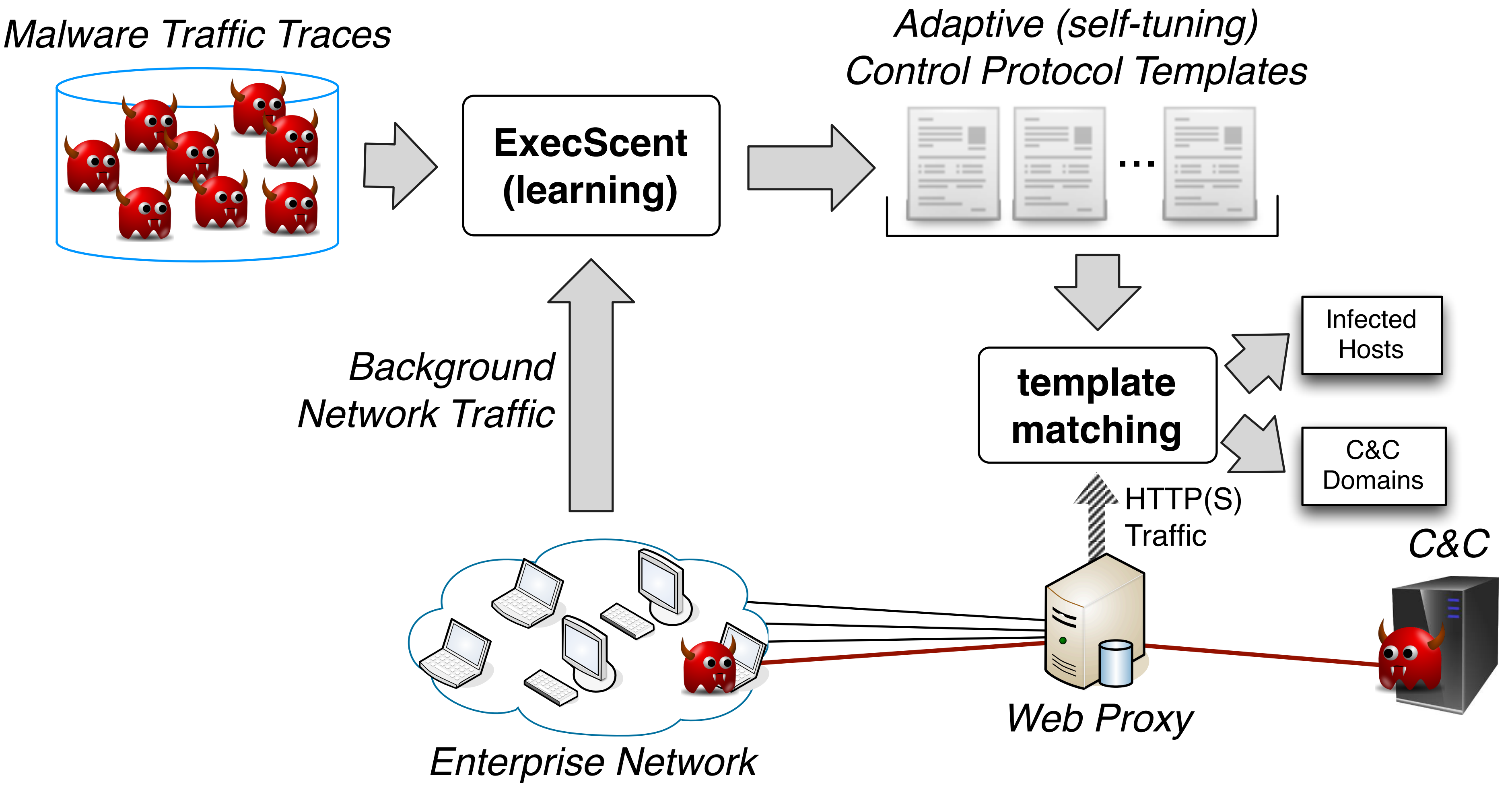
HTTP(S) Traffic

C&C

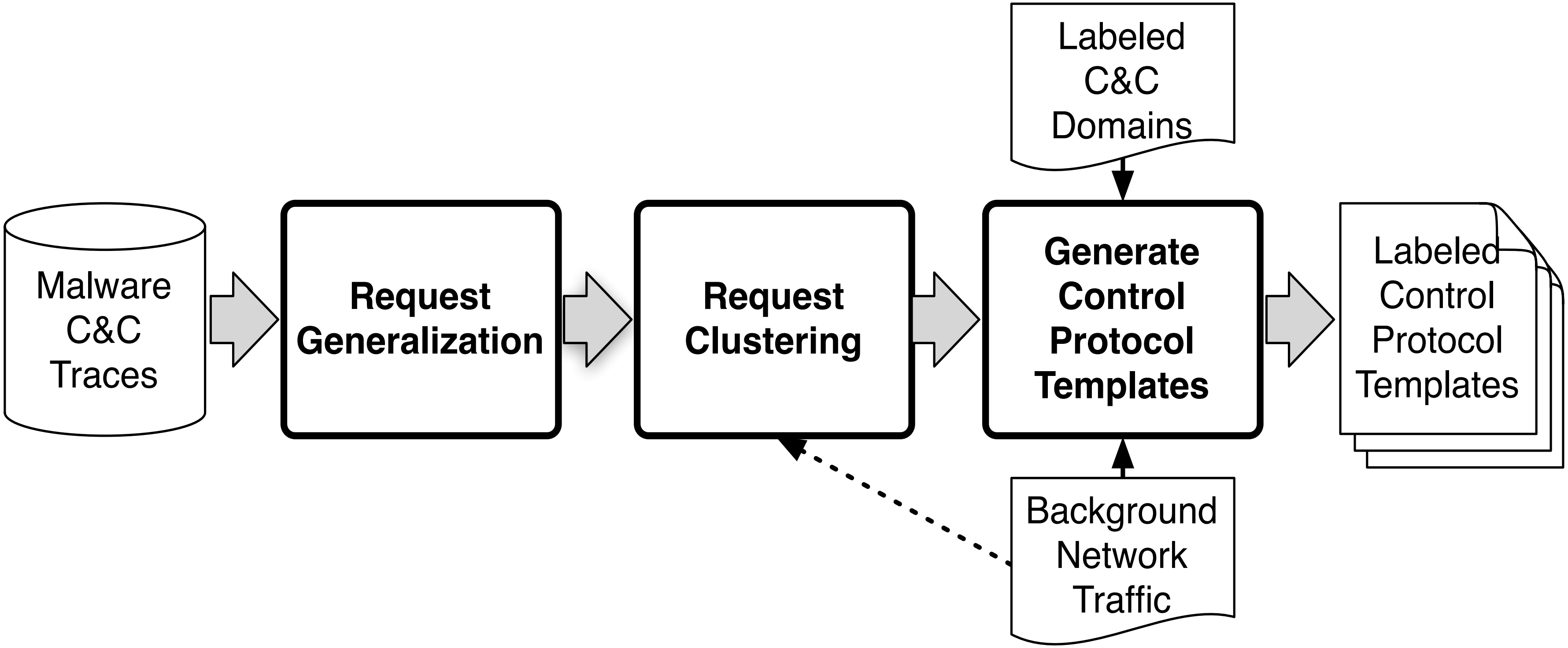
Web Proxy

Enterprise Network

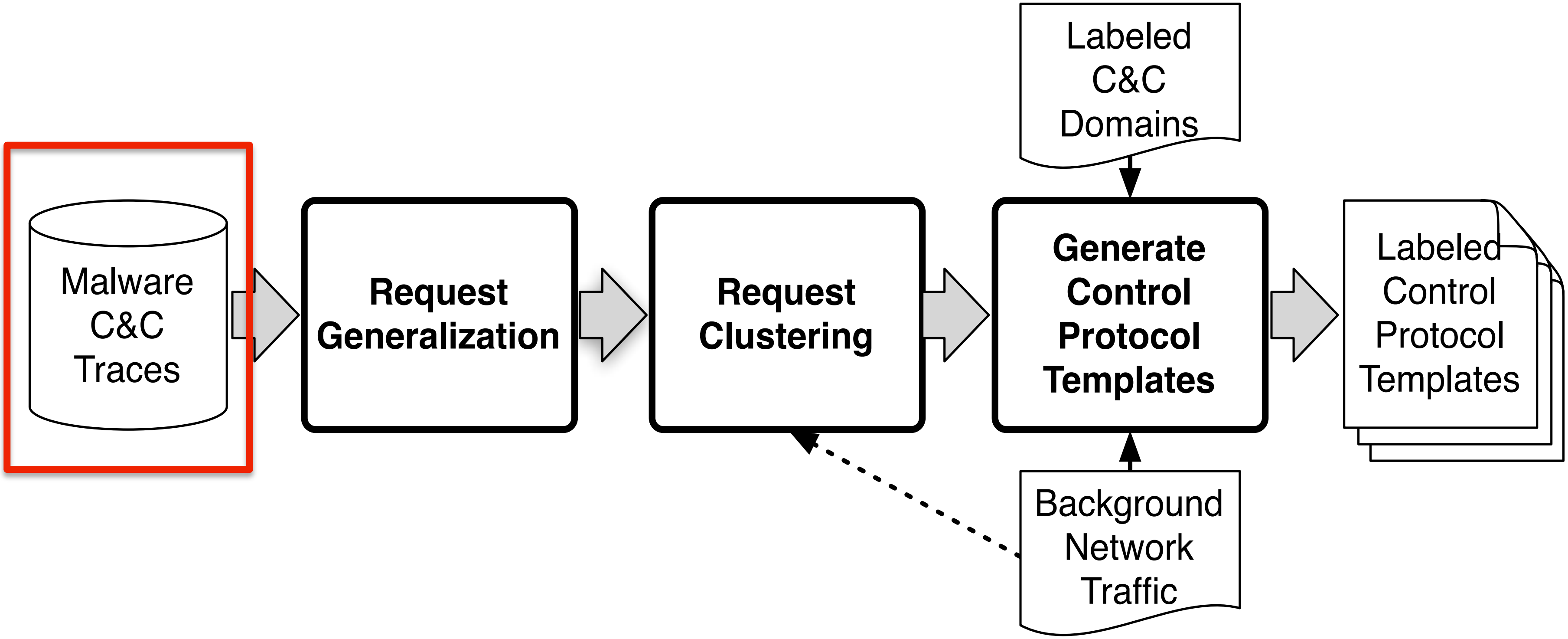
ExecScent Overview



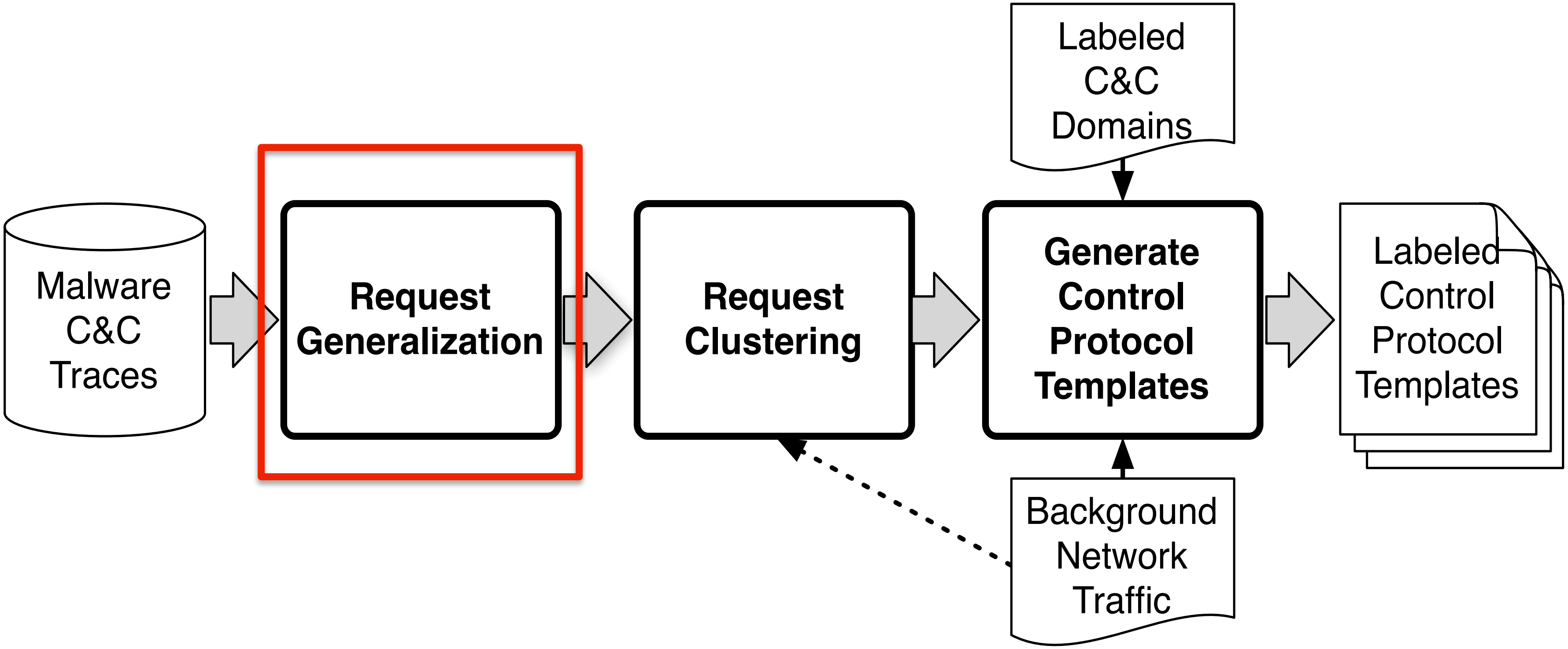
Template Learning Process



Malware C&C Traces


Labeled C&C Domains
Malware C&C Traces
Request Generalization
Request Clustering
Generate Control Protocol Templates
Labeled Control Protocol Templates
Background Network Traffic

Request Generalization


Malware C&C Traces
Request Generalization
Request Clustering
Labeled C&C Domains
Generate Control Protocol Templates
Background Network Traffic
Labeled Control Protocol Templates

Request Generalization

(a)

Request 1:

```
GET /Ym90bmV0DQo=/cnc.php?v=121&cc=IT
Host: www.bot.net
User-Agent: 680e4a9a7eb391bc48118baba2dc8e16
...
```

Request 2:

```
GET /bWFsd2FyZQ0KDQo=/cnc.php?v=425&cc=US
Host: www.malwa.re
User-Agent: dae4a66124940351a65639019b50bf5a
...
```

(b)

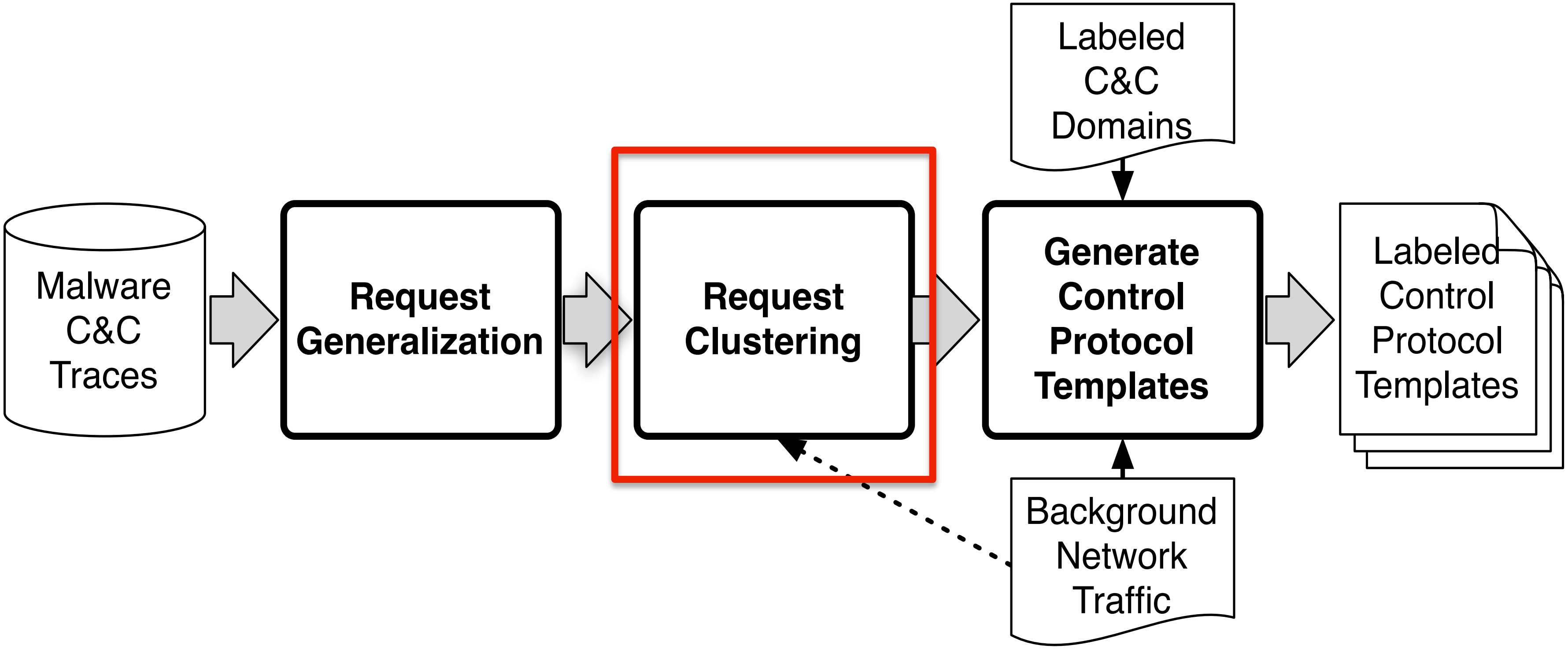
Request 1:

```
GET /<Base64;12>/cnc.php?v=<Int;3>&cc=<Str;2>
Host: www.bot.net
User-Agent: <Hex;32>
...
```

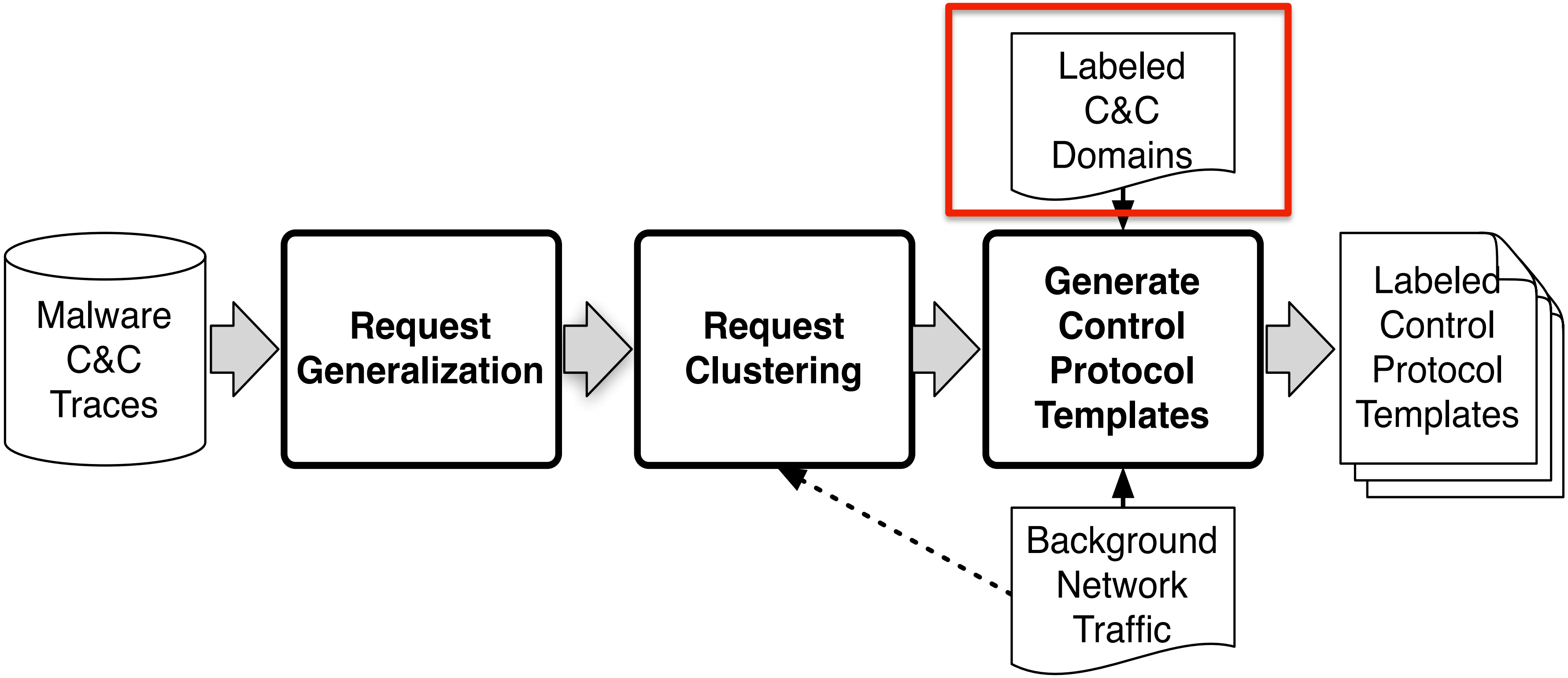
Request 2:

```
GET /<Base64;16>/cnc.php?v=<Int;3>&cc=<Str;2>
Host: www.malwa.re
User-Agent: <Hex;32>
...
```

Request Clustering

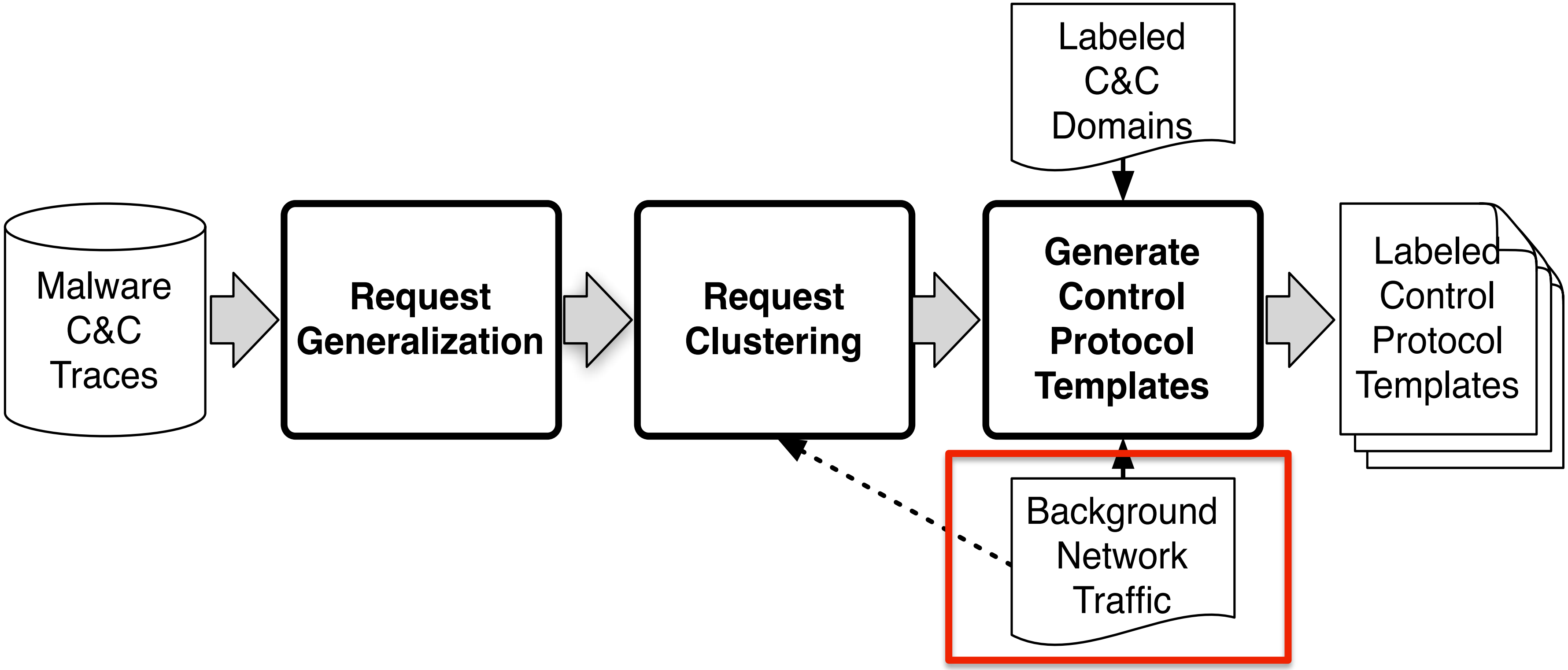

Malware C&C Traces
Request Generalization
Request Clustering
Labeled C&C Domains
Generate Control Protocol Templates
Background Network Traffic
Labeled Control Protocol Templates

Labeled C&C Domains

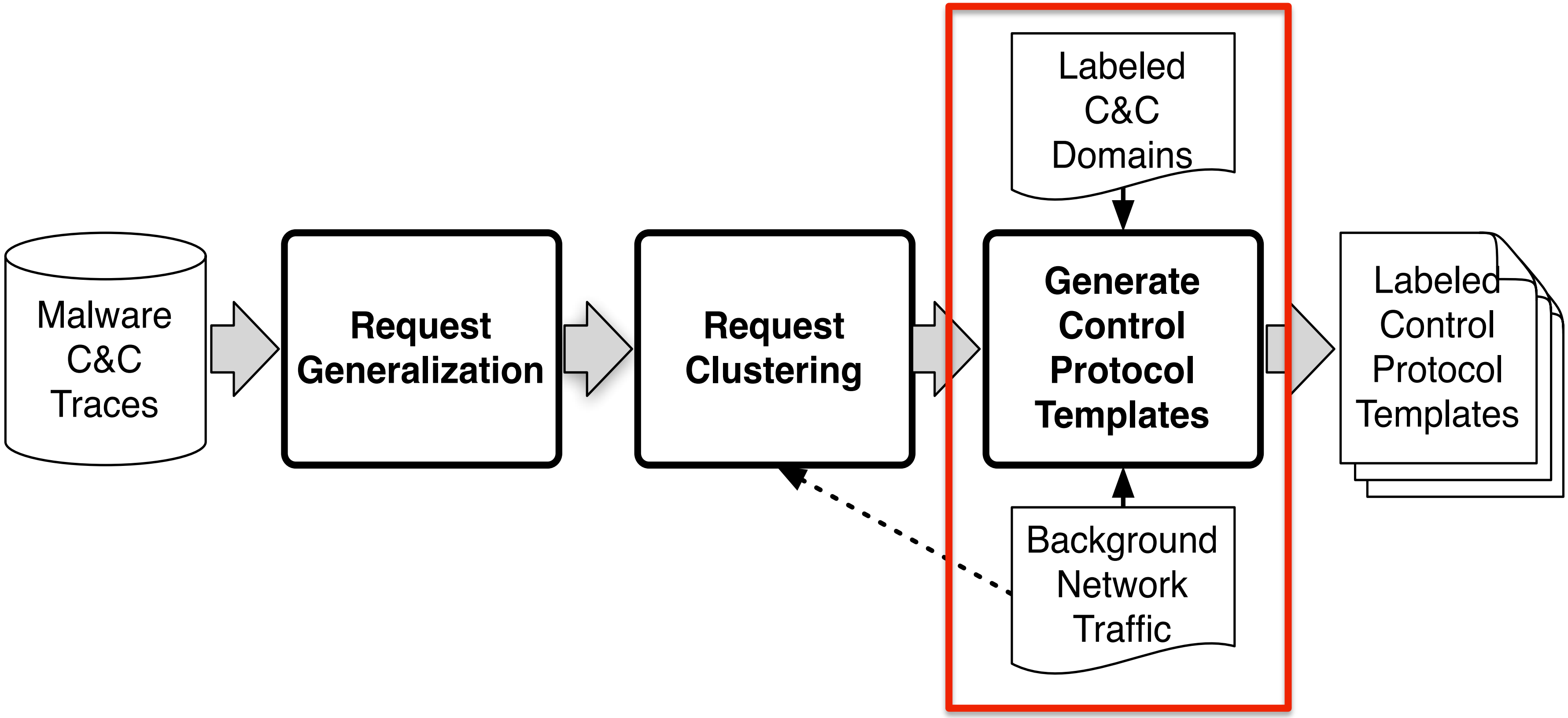

Labeled C&C Domains
Malware C&C Traces
Request Generalization
Request Clustering
Generate Control Protocol Templates
Labeled Control Protocol Templates
Background Network Traffic

2/15/14

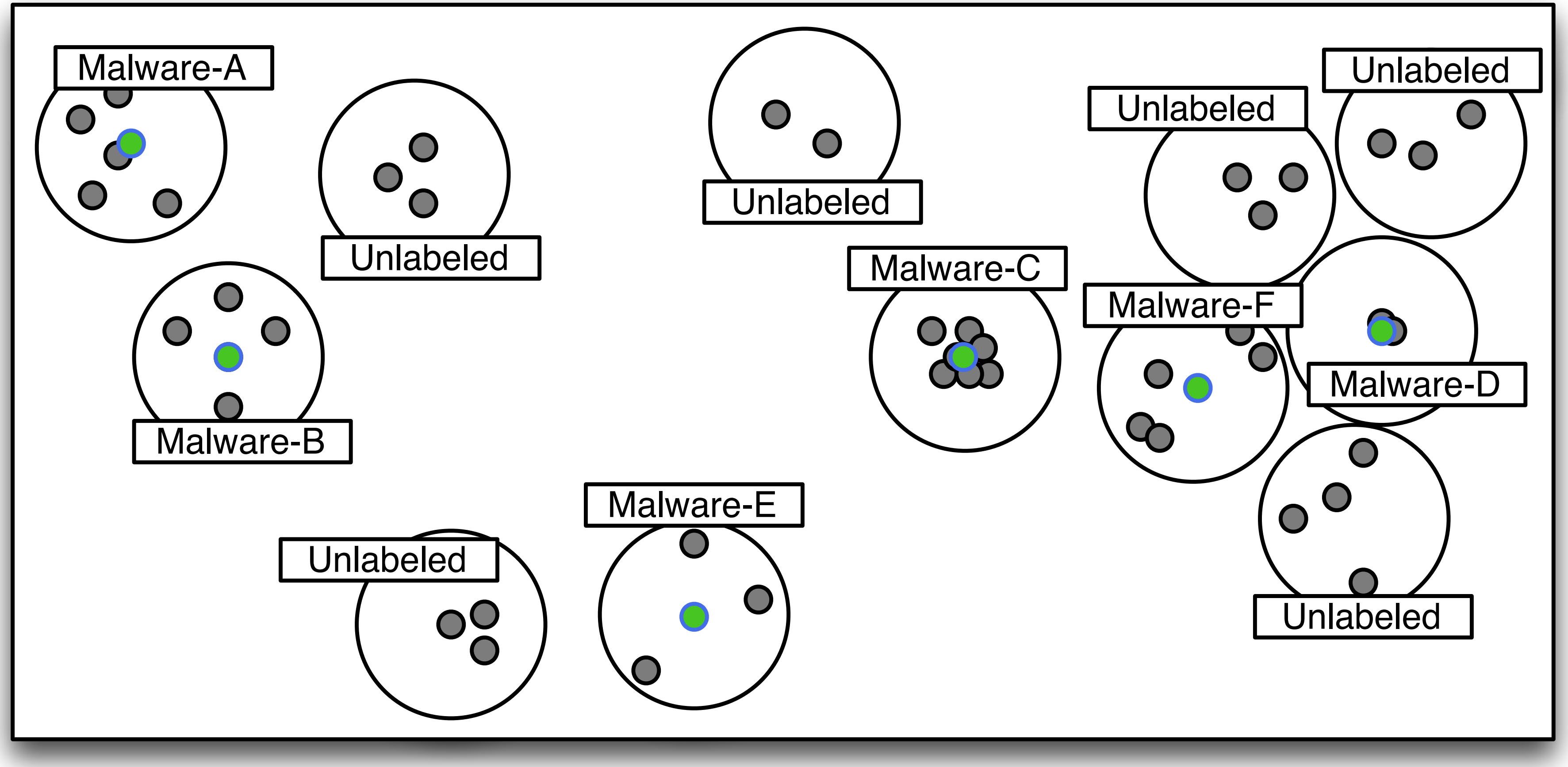
Labeled C&C Domains



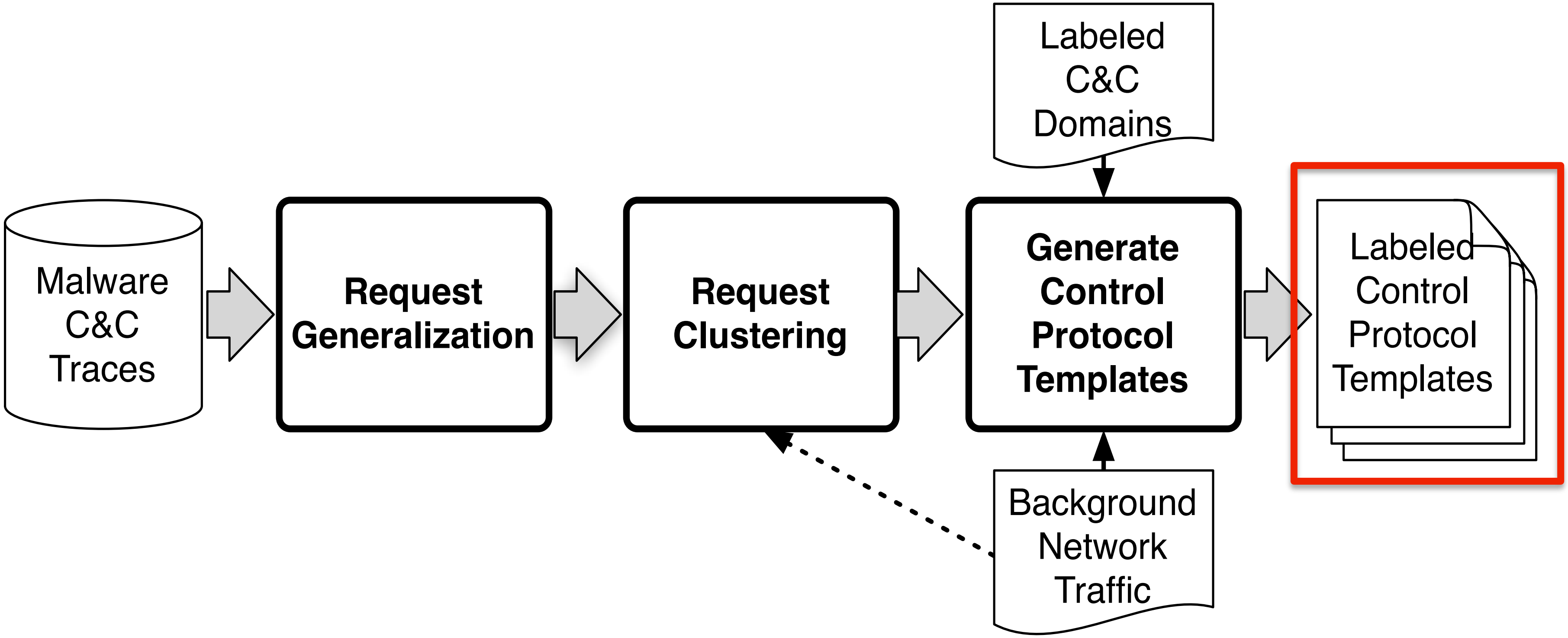
Generating CPTs



Generating CPTs



Labeled CPTs


Labeled C&C Domains
Malware C&C Traces
Request Generalization
Request Clustering
Generate Control Protocol Templates
Labeled Control Protocol Templates
Background Network Traffic

Labeled CPT

τ_1) **Median URL path**: /<Base64;14>/cnc.php

τ_2) **URL query component**: {v=<Int,3>, cc=<String;2>}

τ_3) **User Agent**: {<Hex;32>}

τ_4) **Other headers**: {(Host;13), (Accept-Encoding;8)}

τ_5) **Dst nets**: {172.16.8.0/24, 10.10.4.0/24, 192.168.1.0/24}

Malware family: {*Trojan-A*, *BotFamily-1*}

URL regex: GET /.*\?(cc|v)=

Background traffic profile:
specificity scores used to adapt the CPT
to the deployment environment

Template Matching

- Similarity
  - Measures likeness
  - Components
  - Weighted average
  - Match threshold
- Specificity
  - Measures uniqueness
  - Dynamic weights
  - Self-tuning

Input: req, CPT

Similarity: $s(\text{req}_i, \text{CPT}_i)$, for each component *i*

Specificity: $\delta(\text{req}_i, \text{CPT}_i)$, for each component *i*

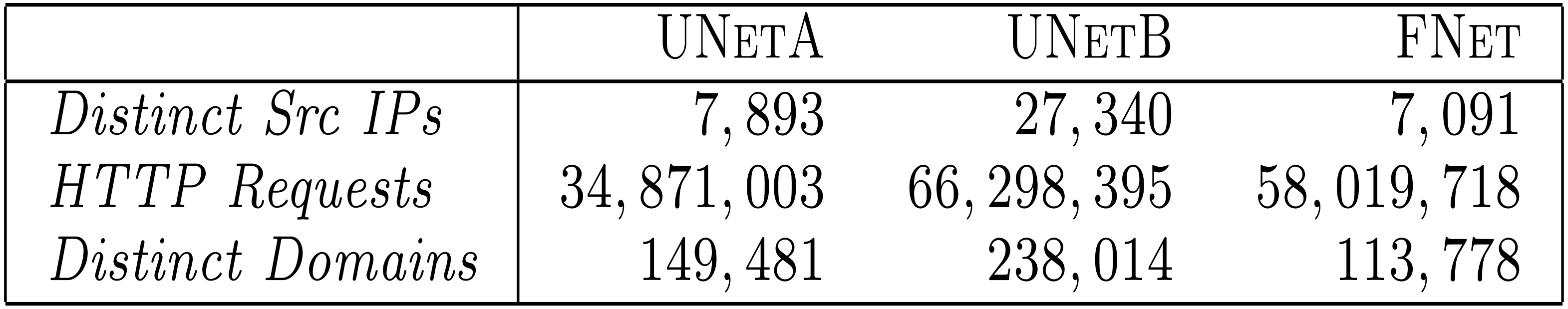
Match-Score: $f(\text{sim}, \text{spec})$

If Match-Score > Θ:
  return C&C Request

Similarity & Specificity Examples

- Example A (High Similarity, Low Specificity):
  - **/index.html** - Request
  - **/index.html** - CPT
- Example B (Low Similarity, High Specificity):
  - **/downloads/9908-7623-0098/images** - Request
  - **/VGVycnkgTmVsbXMK (<Base64, 16>)** - CPT
- Example C (High Similarity, High Specificity)
  - **/Ui4gUGVyZGIzY2kK (<Base64, 16>)**- Request
  - **/VGVycnkgTmVsbXMK (<Base64, 16>)**- CPT

Evaluation Deployment Networks



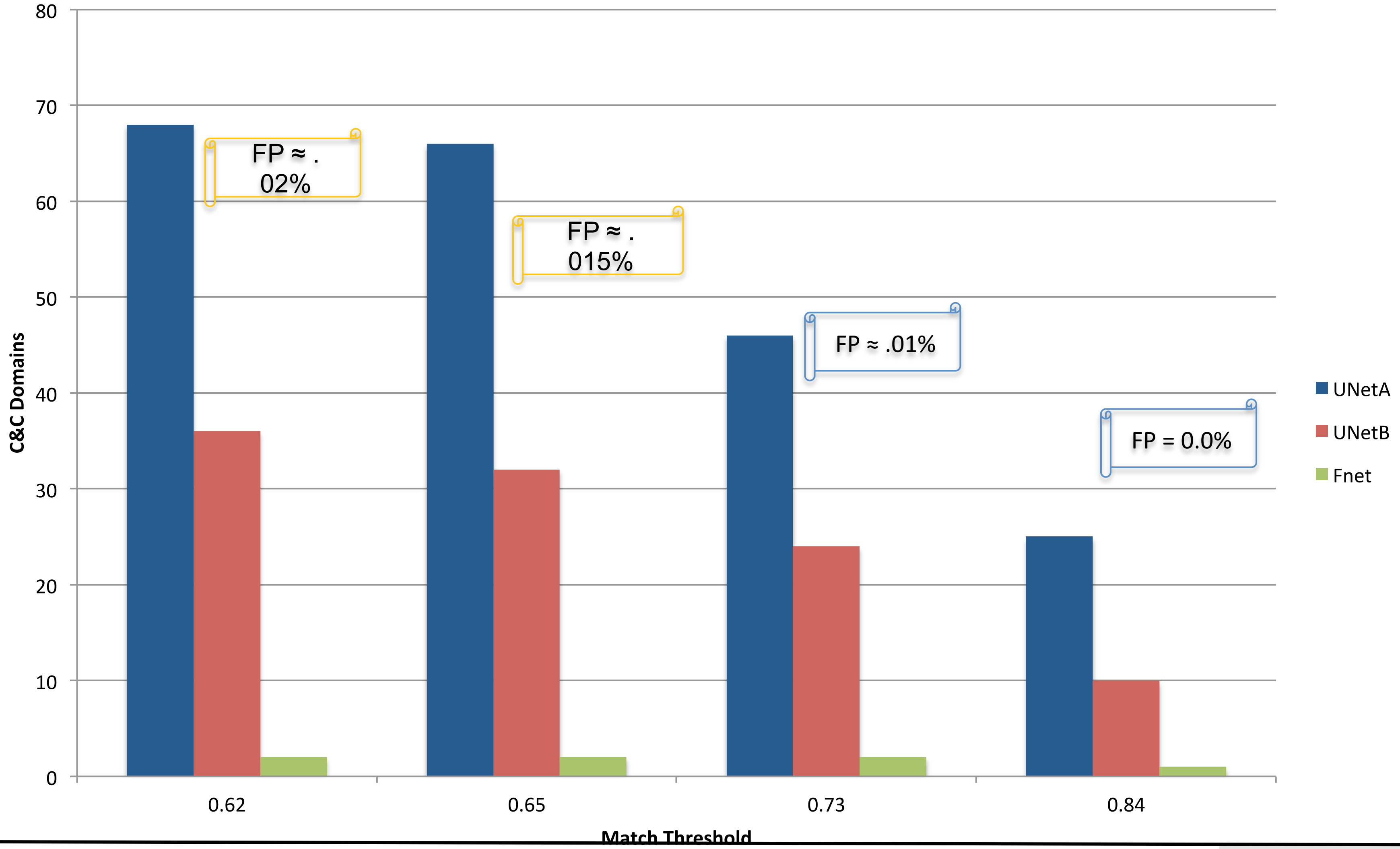
	UNETA	UNETB	FNET
Distinct Src IPs	7,893	27,340	7,091
HTTP Requests	34,871,003	66,298,395	58,019,718
Distinct Domains	149,481	238,014	113,778

- Evaluation ran for two weeks.
- CPTs updated daily beginning two weeks prior to evaluation.

Ground Truth

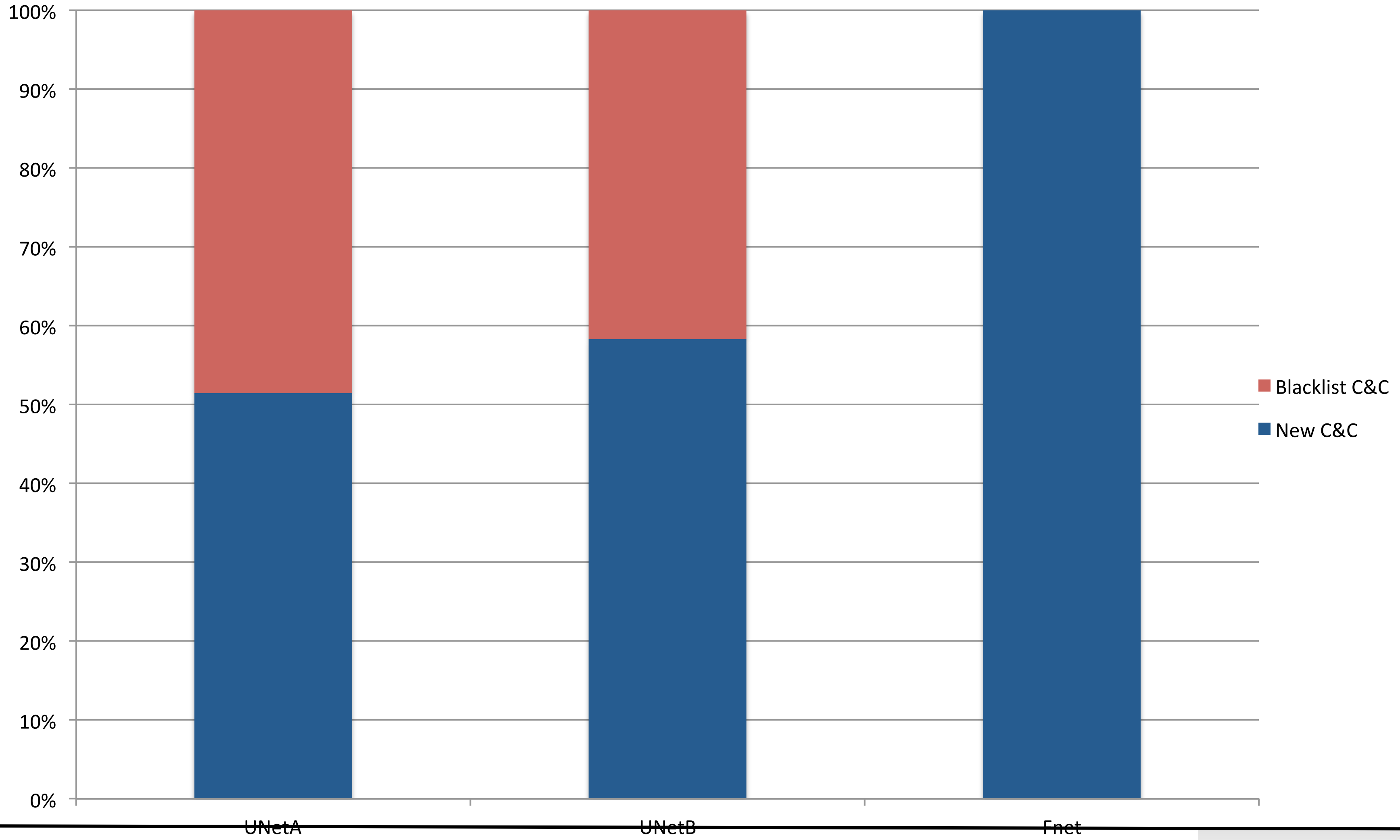
- Commercial C&C blacklist.
- Pruned Alexa top 1 million.
- Professional threat analysts.

Finding C&C Domains

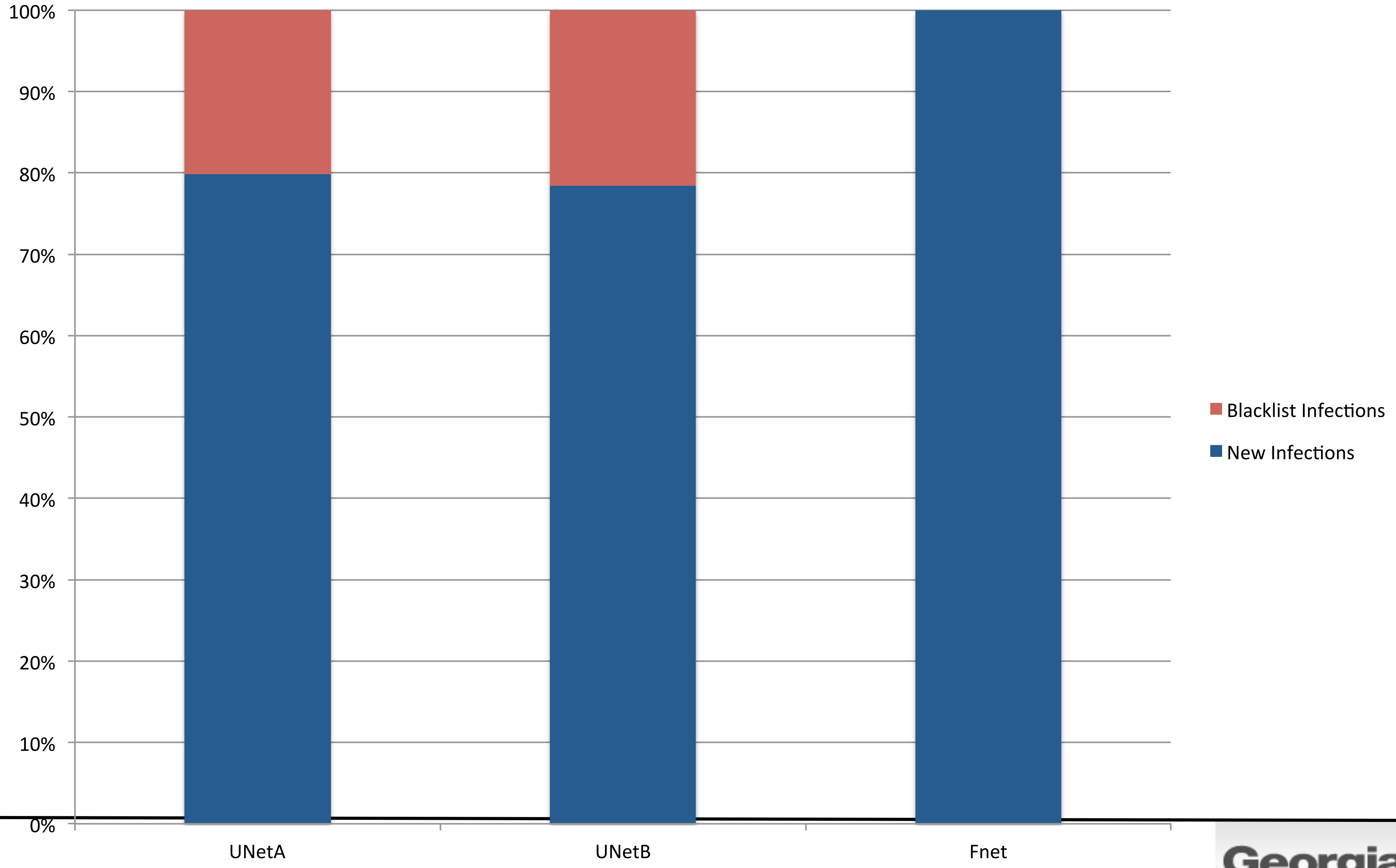


2/15/14

New vs. Blacklist Domains



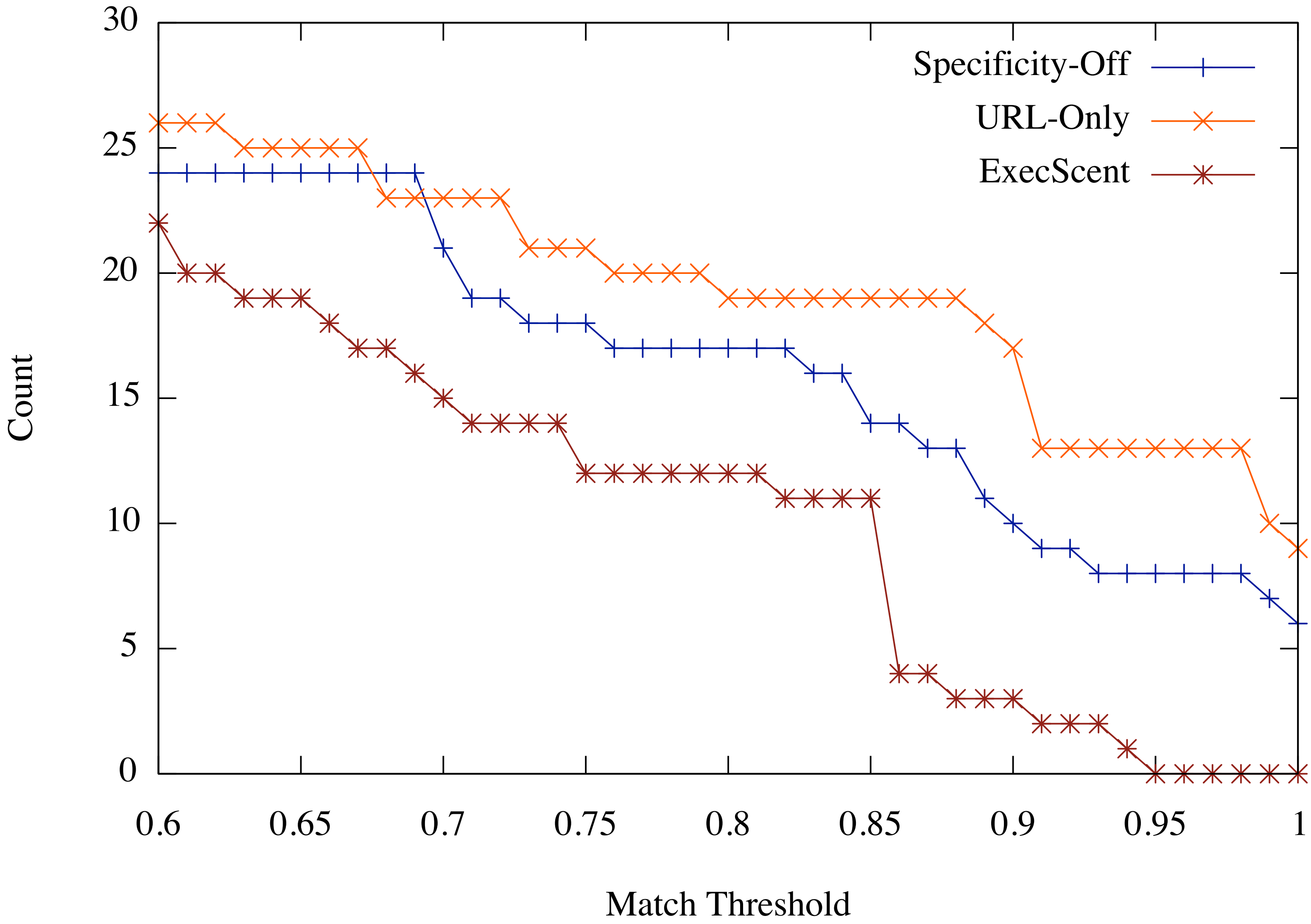
New vs. Blacklist Infected Hosts



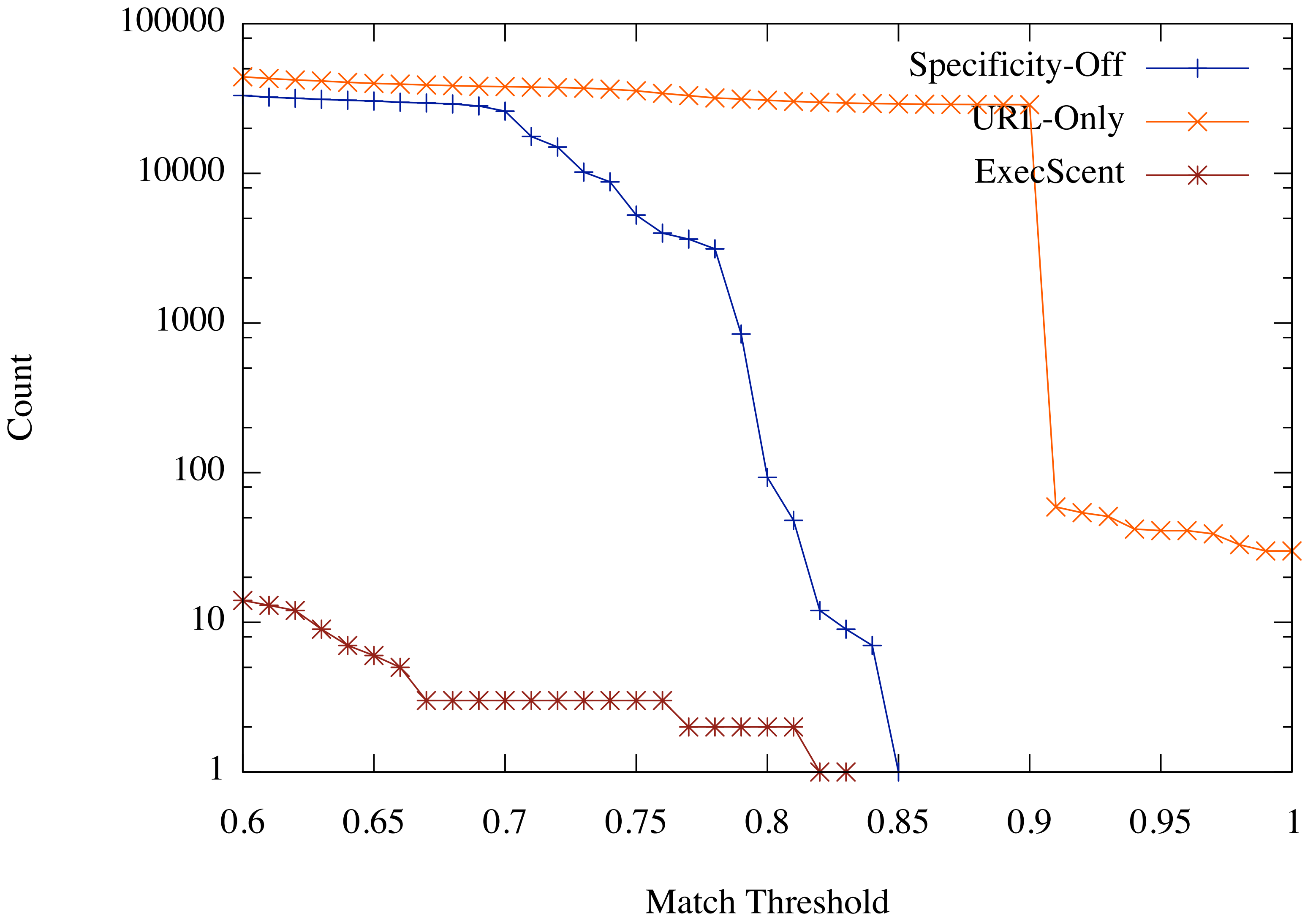
ISP Deployment

- Deployed the **65 newly discovered C&C domains** on **6 ISP networks** for one week.
- Counted the number of distinct source IP addresses contacting the domains daily.
- Identified **25,584** new potential malware infections.

Model Comparison - True Positives



Model Comparison – False Positives



Limitations

- Dependence on malware traces and labeled domains.
- Implement a new protocol when the C&C domain or IP address changes.
- Blend into background traffic.
- Inject noise into the protocol.

Conclusion

- Majority of C&C domains and infections discovered were not on a blacklist.
- C&C domains and IP addresses change more frequently than the protocol structure.
- Adaptive templates yield a better trade-off between true and false positives.
- ExecScent is currently deployed.

Example 3: Telephony Going the Internet Way

- Telephony used to be a trusted channel
  - We exactly knew the call path from source to its destination
- The new telephony landscape
  - Massive scale calling at little or no cost
  - Services like caller-id spoofing are widely available
  - Voice communication will increasingly become embedded into online applications
  - Hard to know “Who Calls me?”
- Have we seen something like this before?
  - Cyber criminals send email spam at massive scale, steal and monetize data, sell fake goods and even launch denial-of-service attacks.

Stealing Money with the Telephone

- Incoming Calls (Fraudster ➔ Victim)
  - Robocalling allows a fraudster to reach large number of targets
  - Telemarketers use it to to reach potential customers/victims
- Outgoing Calls (Victim ➔ Fraudster)
  - Driving traffic to premium numbers (IRSF)
  - Stealing data from victims who respond
- Fraud facilitating call centers (https://blogs.rsa.com/fraudster-operated-call-centers-emerge-in-the-underground-economy-to-facilitate-phone-fraud/)

Do We Have Data to Better Understand the Problem?

- FTC data has over five million records
  - Obtained a copy for research use, each complaint record has some information about the nature of the call, calling number (only 7 digits) and a timestamp
- Phoneypot: Georgia Tech/NYUAD/Pindrop/SUM/IIITD Telephony Honeynet
  - Using “seed” numbers that are carefully publicized at a variety of places
  - Using grey numbers
- Data from the web channel
  - Phone numbers in email spam, Youtube comments, Tweets?
- Crowd sourced data
  - 800notes.com, whocallsme.com etc.

Early Results of Data Analysis

- Are these the same guys we have seen elsewhere?
  - Nature of calls (e.g., what did a victim complain about)
- Is there evidence of caller-id spoofing?
  - How do we know for sure?
- How are victim numbers being harvested?
  - Web channel?

Nature of Services/Offers

Data Source	**Keywords**
FTC	• Credit Card, Bankcard, Lower interest, Cardholder services – over 60% • Home Alarm, Home protection, Emergency medical alert • Canadian pharmacy, Rx assistance
Twitter (analysis of about one million tweets that have phone numbers)	• Money, credit, bills, Rachael from card holder services • Drugs • Warranty • Education, degrees

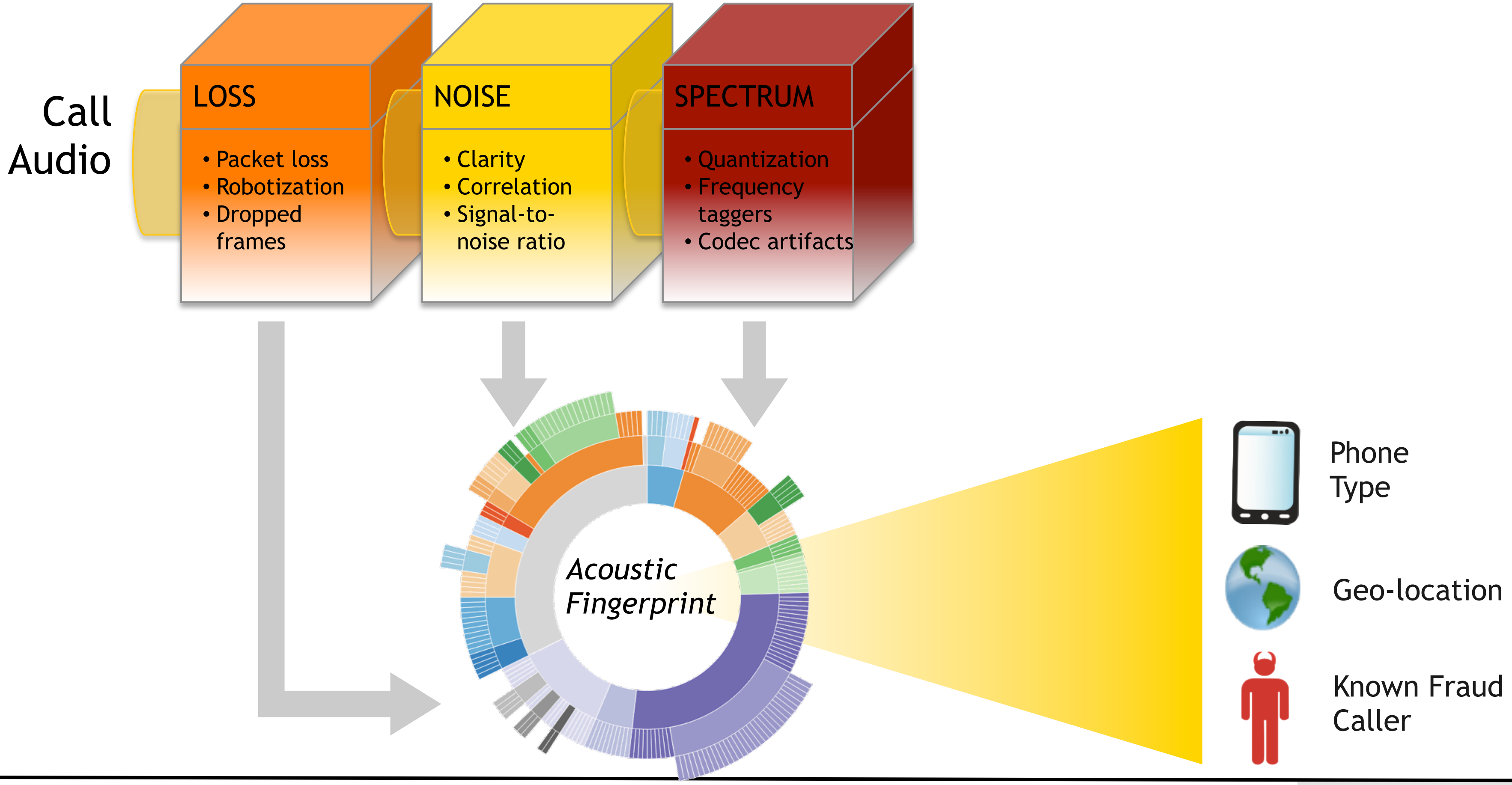
Phoneypot Story So Far (Pindrop/Georgia Tech only)

- Over 800 unsolicited calls over about two months
  - Received Rachel Calls, Pharmacy Calls, Free trip calls among other social engineering calls.
  - More VoIP calls but also good number of landline and cellular calls
  - About 1/3 calls were from outside of the United States

Other Observations from Phoneypot

- We are receiving dozens of calls, including on numbers that we added to do-not-call list
- Numbers are being scraped from the web channel
- Life history of a phone number seems to matter (qualifying process?)
- Calls from bored people who have nothing better to do with their time?

Detecting Caller-Id Spoofing via Acoustic Fingerprints [Pindrop Security]



Call Audio

LOSS

- Packet loss
- Robotization
- Dropped frames

NOISE

- Clarity
- Correlation
- Signal-to-noise ratio

SPECTRUM

- Quantization
- Frequency taggers
- Codec artifacts

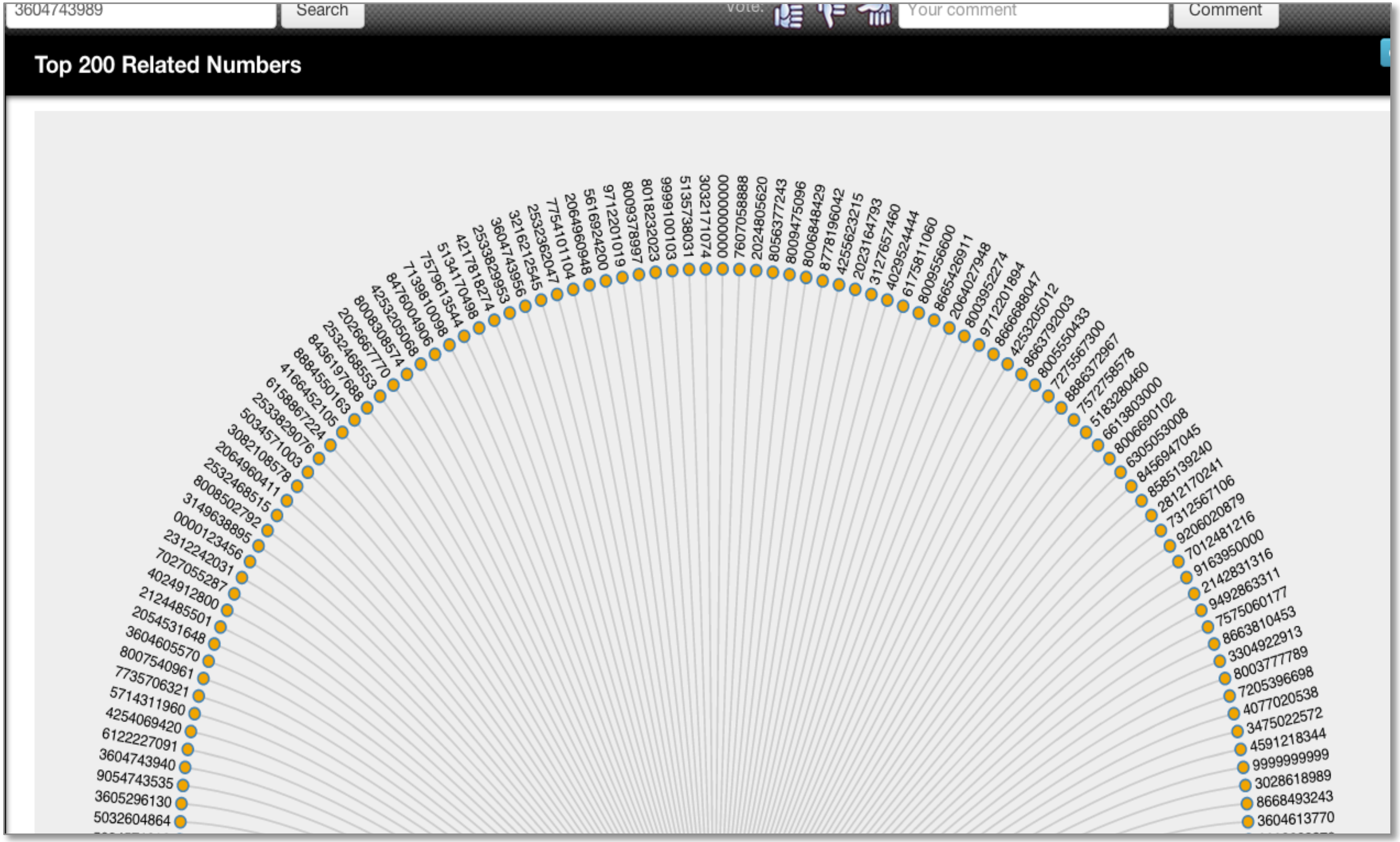
Acoustic Fingerprint

Phone Type

Geo-location

Known Fraud Caller

Is there Caller-id Spoofing?



Top 200 Related Numbers

Using Caller-id Spoofing to Craft Call Center Attacks

- Call centers have moved on to stronger authentication
  - Knowledge-based authentication
- Social engineering or weak KBA leads to password resets via the phone channel
- New password is used to attack the web channel
  - Funds transfer from online accounts

Next Steps

- How do we gain access to data to better understand the threat landscape?
- How do we “convict” or “blacklist” phone numbers like IP addresses or domains?
  - How do we stop calls coming from blacklisted phone numbers?
  - How do we stop people from going to bad numbers?
- How do we build stronger accountability (Know Your Caller)?
- How do we enhance trust in the telephony ecosystem?
  - Technology? Policy? Regulation? Awareness?

Getting Back to Data-Driven Research

- Data Sharing Challenges
  - Proprietary data and privacy issues
  - Going from data to actionable information
- Coordination
  - Building human trust networks
  - Proactive intelligence sharing
- Academic research centers are great places for facilitating data-driven research
  - Neutral, trusted places where industry, government and academia can come together

Conclusions

- Cyber threats are constantly evolving
- Getting ahead of the threats
  - Access to data from real networks
  - Effective analytics
  - Offering actionable intelligence
- Infrastructure for data collection, sharing and coordination
- Data is an excellent enabler for great research