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

Terry Nelms$^{1,2}$, Roberto Perdisci$^{3,2}$, Mustaque Ahamad$^{2,4}$

$^1$Damballa Inc.
$^2$Georgia Institute of Technology, College of Computing
$^3$University of Georgia – Dept. of Computer Science
$^4$New York University Abu Dhabi
Modern Malware Networking

Enterprise Network

Web Proxy

C&C

badguy.com
192.168.1.2
Malware Network Detection Methods

- Anomaly-Based
- Domain-Based
- URL-Regex
• 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

Malware Traffic Traces

ExecScent (learning)

Adaptive (self-tuning) Control Protocol Templates

template matching

HTTP(S) Traffic

Web Proxy

C&C

Background Network Traffic

Enterprise Network
ExecScent Overview

Malware Traffic Traces → ExecScent (learning) → Adaptive (self-tuning) Control Protocol Templates

Background Network Traffic

Enterprise Network → Web Proxy → HTTP(S) Traffic → C&C

template matching

Similarity

Specificity
ExecScent Overview

Malware Traffic Traces → ExecScent (learning) → Adaptive (self-tuning) Control Protocol Templates

Background Network Traffic → template matching → C&C Traffic

Infection Hosts → C&C Domains

Enterprise Network → Web Proxy → C&C

HTTP(S) Traffic
Template Learning Process

Malware
C&C
Traces

Request
Generalization

Request
Clustering

Generate
Control
Protocol
Templates

Labeled
C&C
Domains

Background
Network
Traffic

Labeled
Control
Protocol
Templates
Malware C&C Traces

Request Generalization

Request Clustering

Generate Control Protocol Templates

Labeled C&C Domains

Labeled Control Protocol Templates

Background Network Traffic
Request Generalization

- Malware C&C Traces
- Request Generalization
- Request Clustering
- Generate Control Protocol Templates
- Labeled C&C Domains
- 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 → Generate Control Protocol Templates

- Labeled C&C Domains
- Background Network Traffic

Labeled Control Protocol Templates
Labeled C&C Domains

1. Malware C&C Traces
2. Request Generalization
3. Request Clustering
4. Generate Control Protocol Templates
   - Labeled C&C Domains
   - Background Network Traffic

5. Labeled Control Protocol Templates
Labeled C&C Domains

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

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

Malware C&C Traces → Request Generalization → Request Clustering → Generate Control Protocol Templates

- Labeled C&C Domains
- Background Network Traffic

Labeled Control Protocol Templates
Labeled CPT

\( T_1 \) Median URL path: /<Base64;14>/cnc.php

\( T_2 \) URL query component: \{v=<Int,3>, cc=<String;2>\}

\( T_3 \) User Agent: \{<Hex;32>\}

\( T_4 \) Other headers: \{(Host;13), (Accept-Encoding;8)\}

\( T_5 \) 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: $s(req_i, CPT_i)$, for each component $i$

Specificity: $\delta(req_i, CPT_i)$, for each component $i$

Match-Score: $f(sim, spec)$

If Match-Score $> \Theta$: return C&C Request
• 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)
  – `/Ui4gUGVyZGlzY2kK (<Base64, 16>)` - Request
  – `/VGVycnkgTmVsbXMK (<Base64, 16>)` - CPT
Evaluation Deployment Networks

<table>
<thead>
<tr>
<th></th>
<th>UNetA</th>
<th>UNetB</th>
<th>FNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct Src IPs</td>
<td>7,893</td>
<td>27,340</td>
<td>7,091</td>
</tr>
<tr>
<td>HTTP Requests</td>
<td>34,871,003</td>
<td>66,298,395</td>
<td>58,019,718</td>
</tr>
<tr>
<td>Distinct Domains</td>
<td>149,481</td>
<td>238,014</td>
<td>113,778</td>
</tr>
</tbody>
</table>

• 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

C&C Domains Match Threshold

- UNetA
- UNetB
- Fnet

FP ≈ 0.02%
FP ≈ 0.015%
FP ≈ 0.01%
FP ≈ 0.0%

Match Threshold

0.62 0.65 0.73 0.84
New vs. Blacklist Domains

- UNetA
- UNetB
- Fnet

<table>
<thead>
<tr>
<th></th>
<th>UNetA</th>
<th>UNetB</th>
<th>Fnet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blacklist C&amp;C</td>
<td>50%</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>New C&amp;C</td>
<td>50%</td>
<td>60%</td>
<td>100%</td>
</tr>
</tbody>
</table>
New vs. Blacklist Infected Hosts

8/23/13

- UNetA
- UNetB
- Fnet

Blacklist Infections
New Infections
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 – False Positives

![Graph showing model comparison with false positives. The x-axis represents the match threshold, ranging from 0.6 to 1.0. The y-axis represents the count, ranging from 1 to 100,000. Three models are compared: Specificity-Off, URL-Only, and ExecScent. The graph illustrates the performance of these models at various match thresholds.]
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.
Questions?