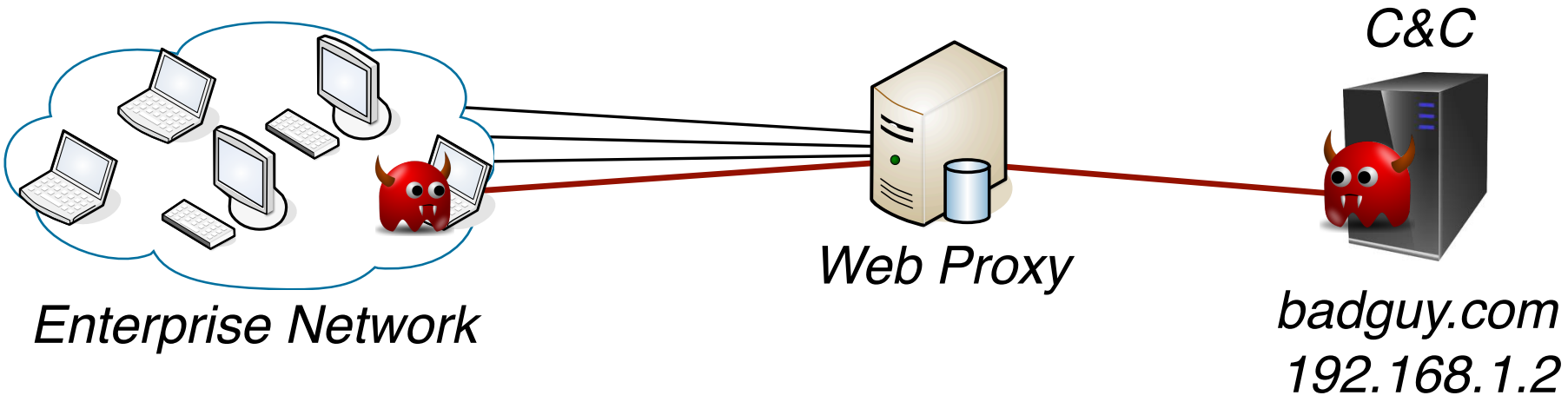
ExecScent: Mining for New C&C Domains in Live Networks with Adaptive Control Protocol Templates

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



Malware Network Detection Methods

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

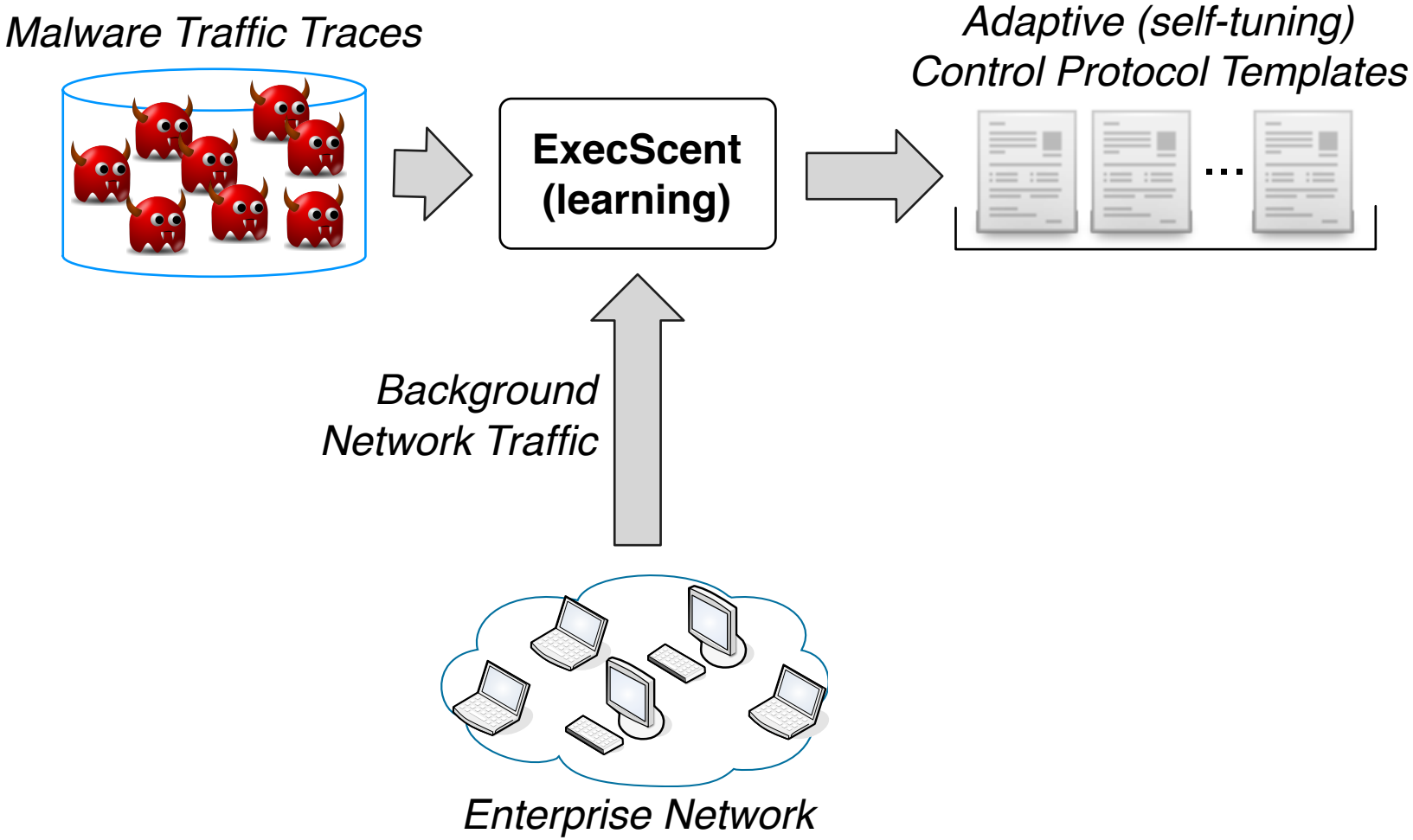
ExecScent Goals & Observations

- Goals:
  - Network detection domains & hosts.
  - Malware family attribution.
- Observations:
  - C&C protocol changes infrequently.
  - HTTP C&C application layer protocol.

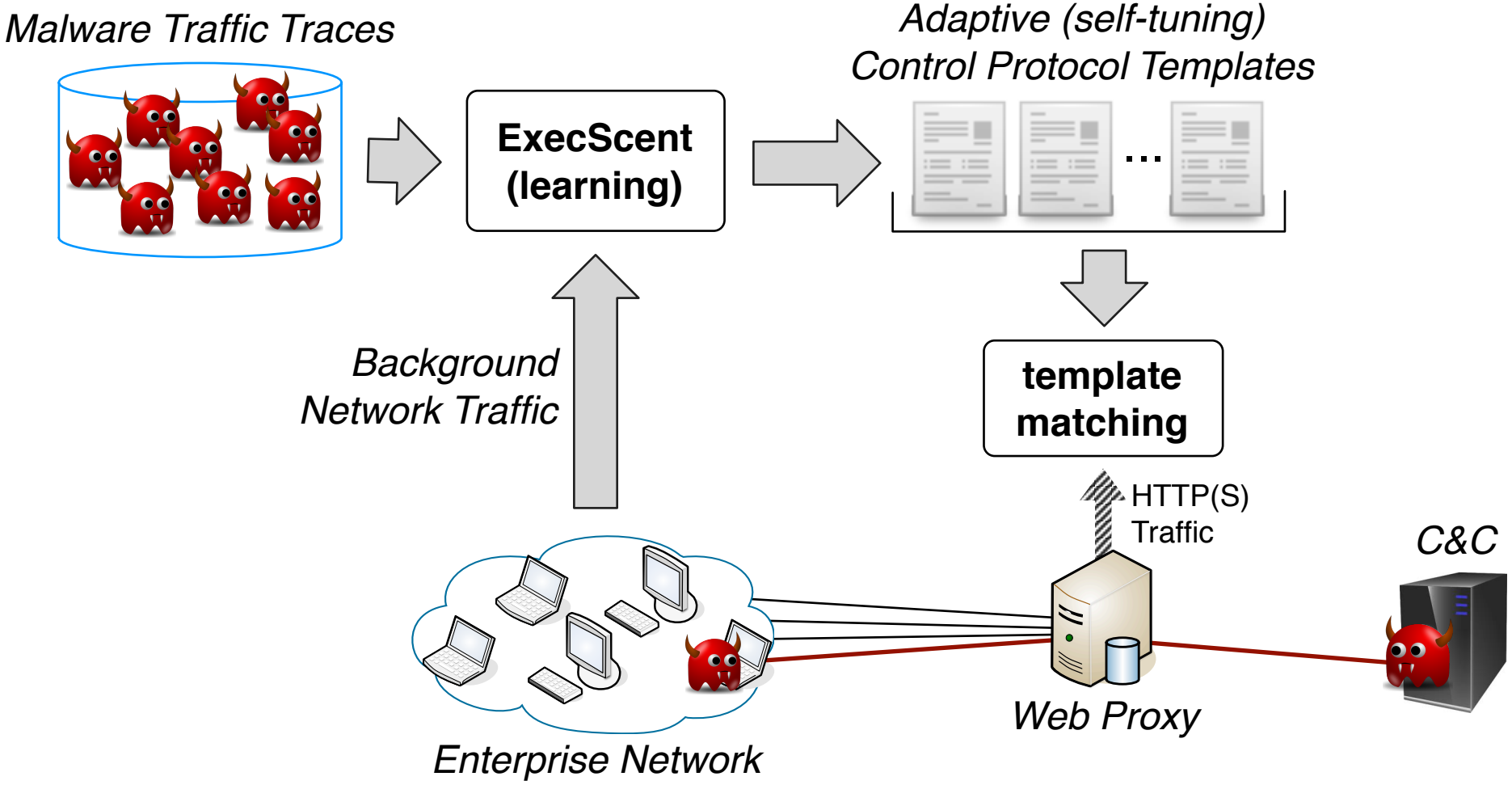
Adaptive Control Protocol Templates

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

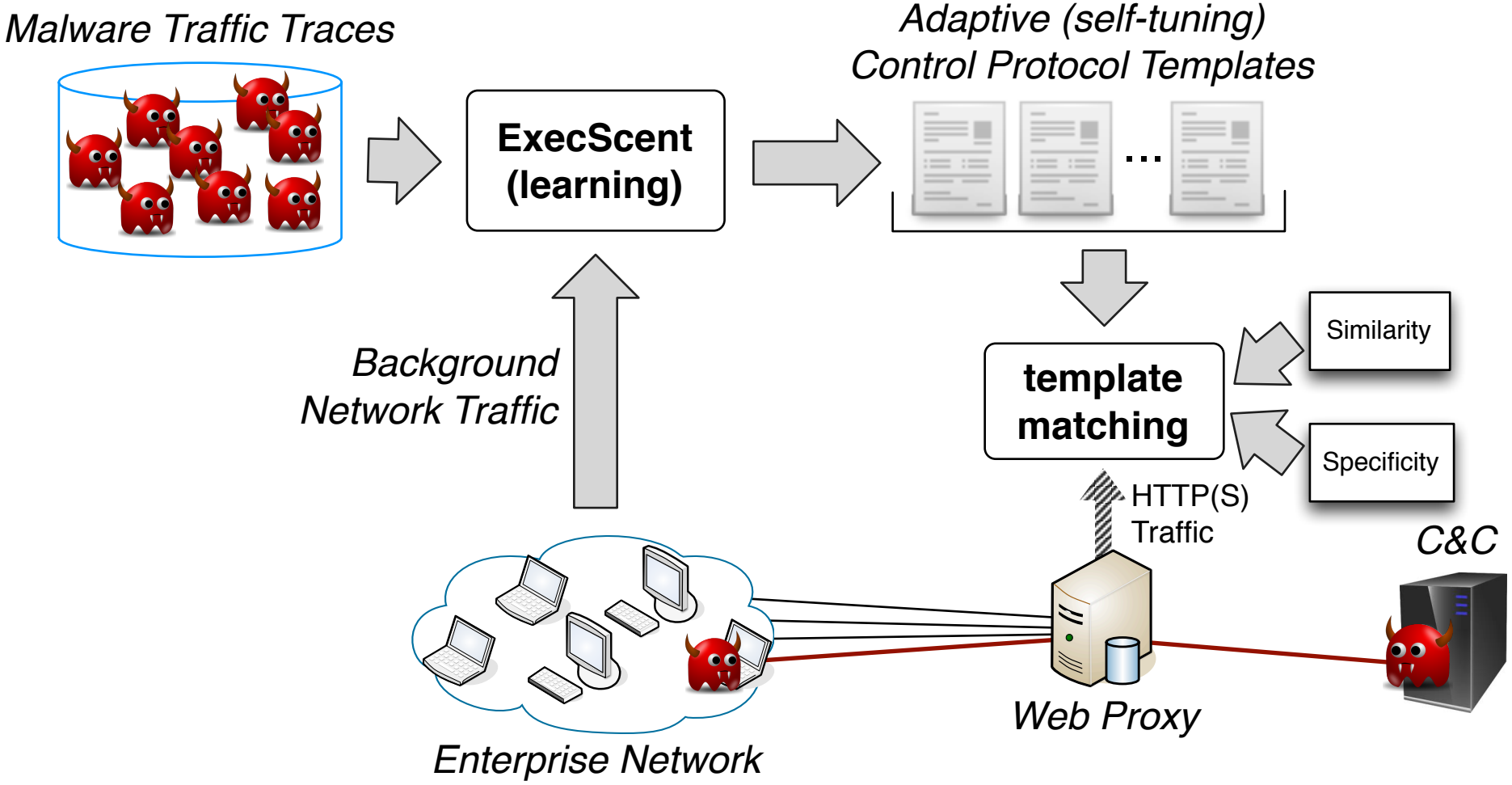
ExecScent Overview



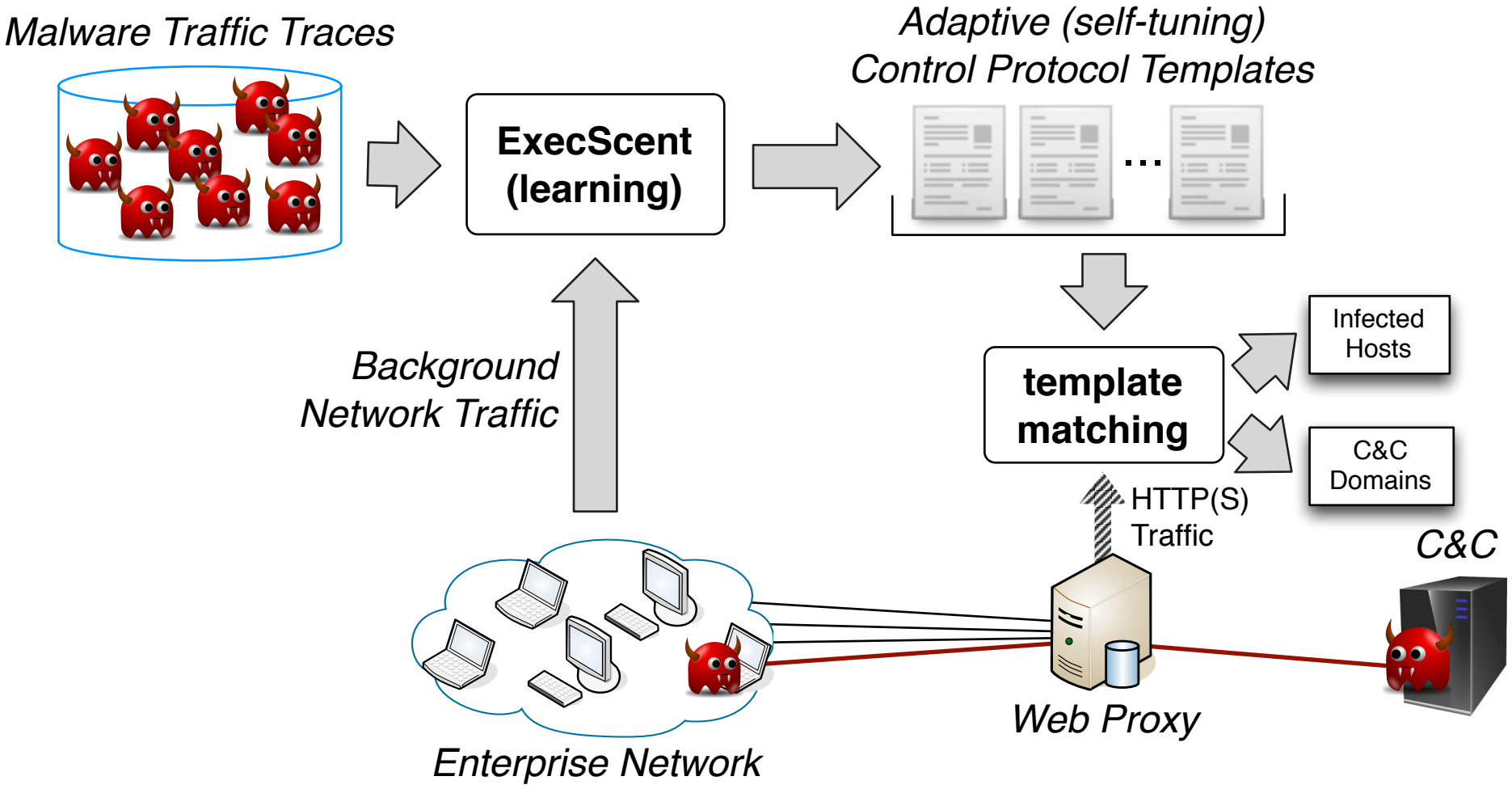
ExecScent Overview



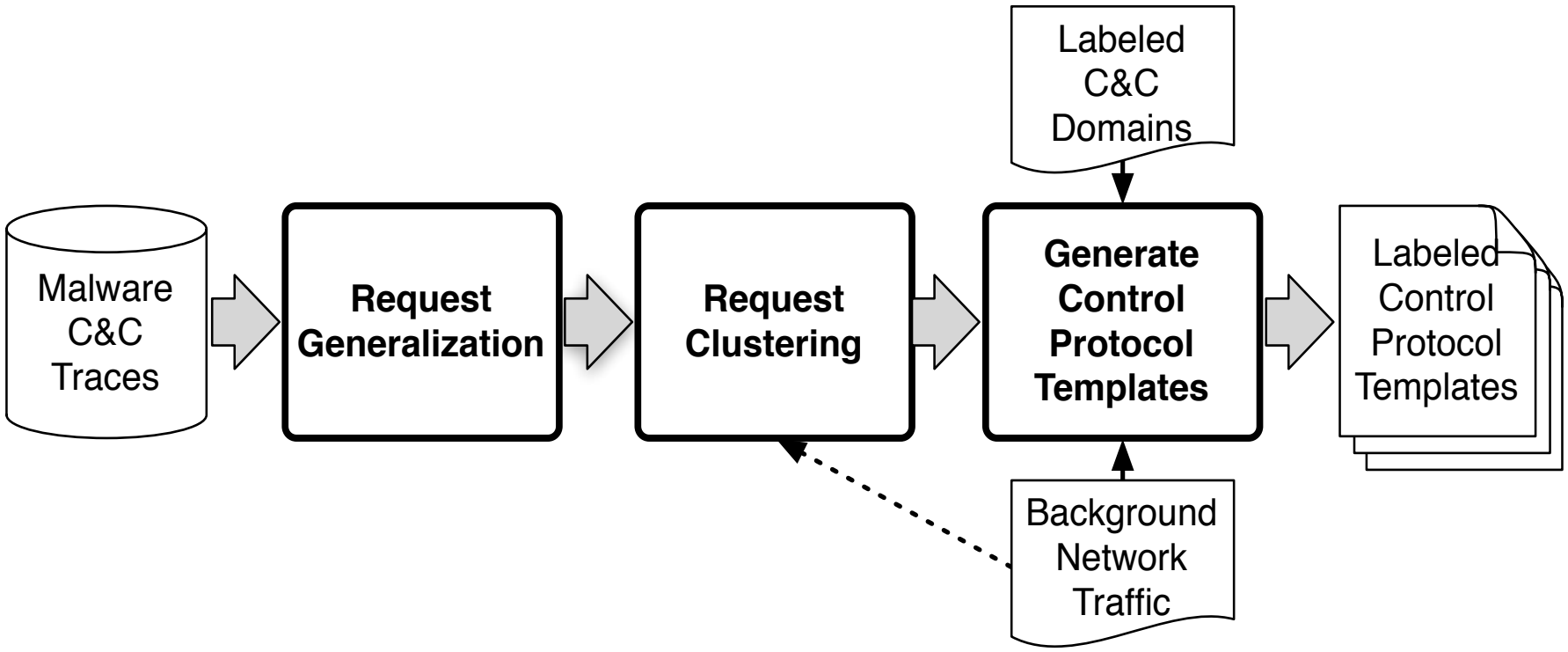
ExecScent Overview



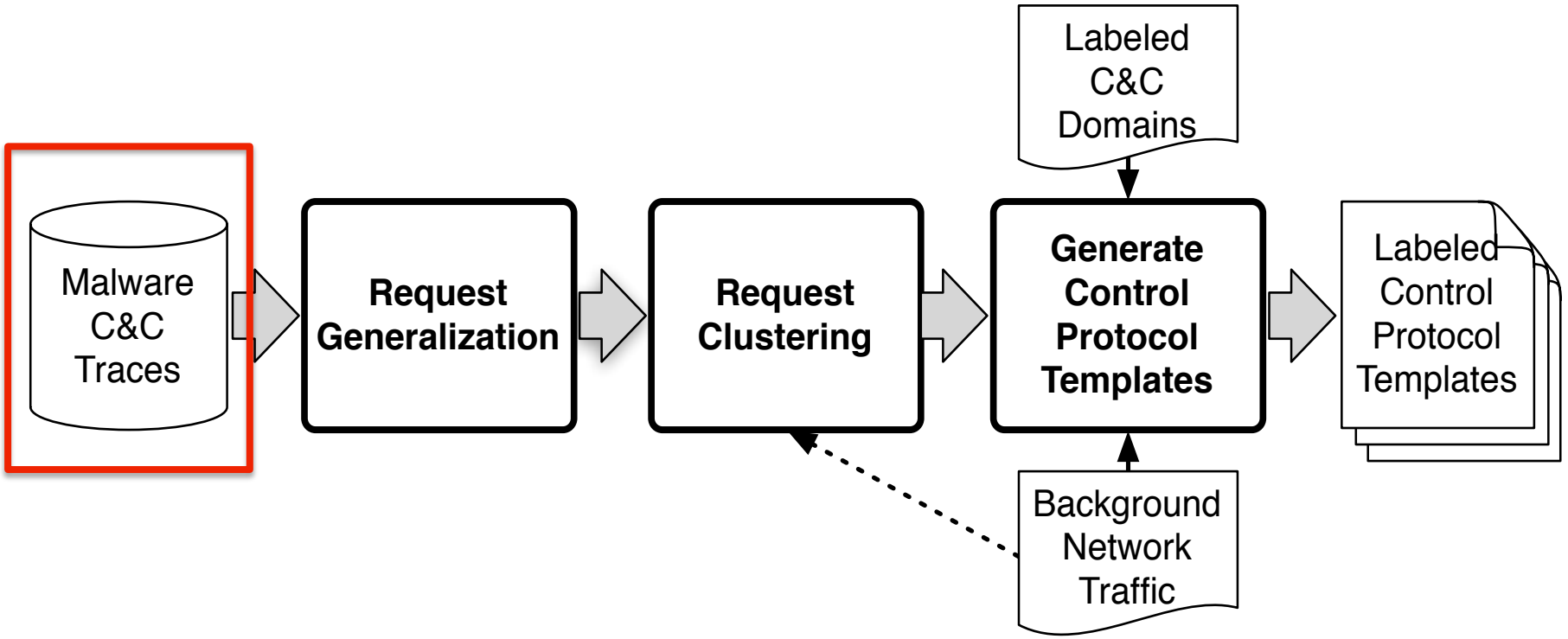
ExecScent Overview



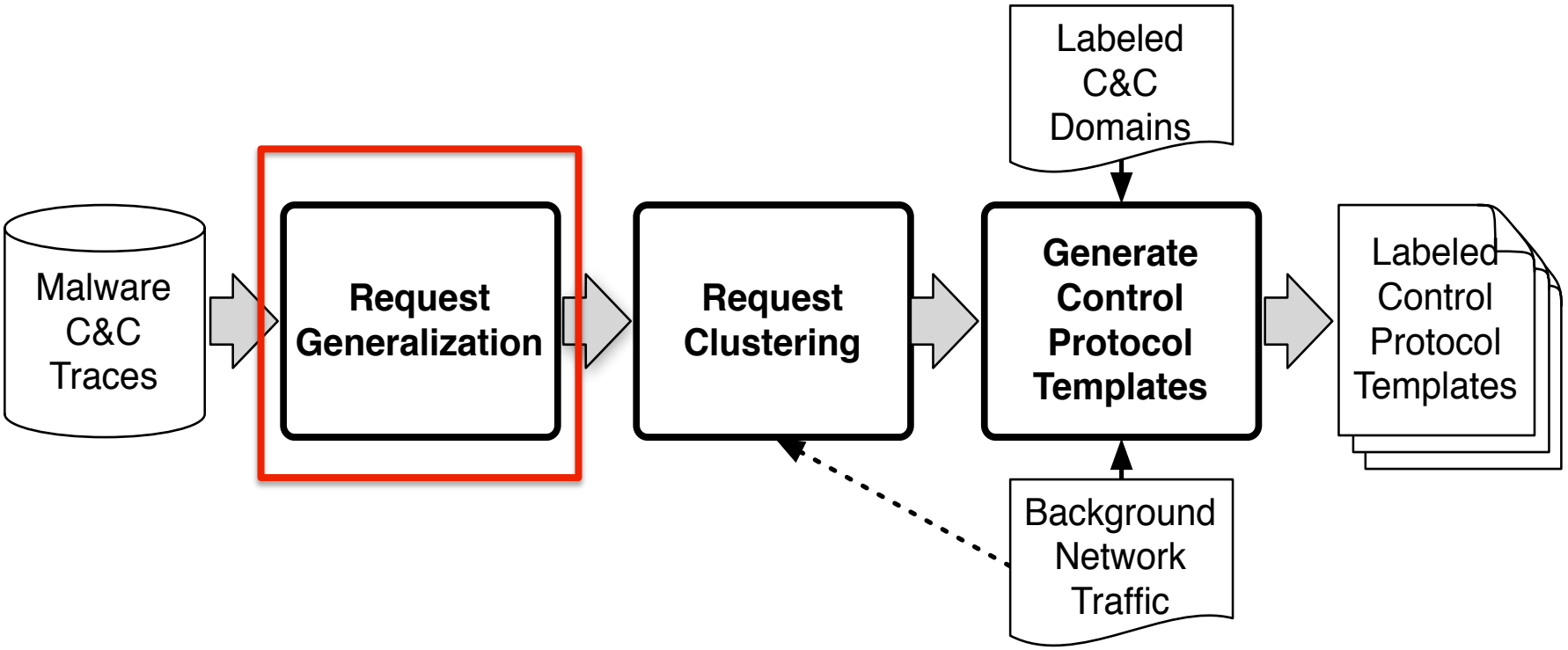
Template Learning Process



Malware C&C Traces



Request Generalization



Request Generalization

(a)

Request 1:
GET /Ym90bmV0DQo=/cnc.php?v=121&cc=IT
Host: www.bot.net
User-Agent: 680e4a9a7eb391bc48118baba2dc8e16
...

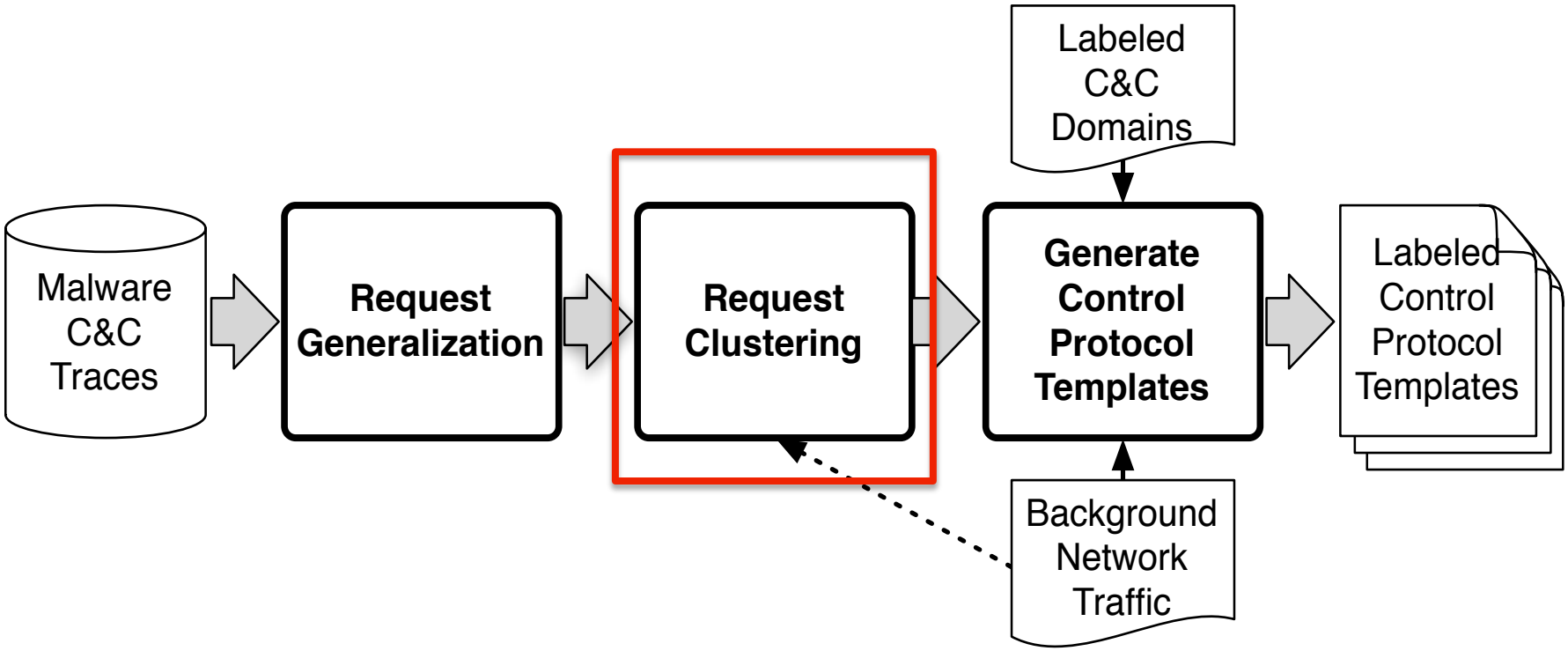
Request 2:
GET /bWFsd2FyZQ0KDQo=/cnc.php?v=425&cc=US
Host: www.malwa.re
User-Agent: dae4a66124940351a65639019b50bf5a
...

(b)

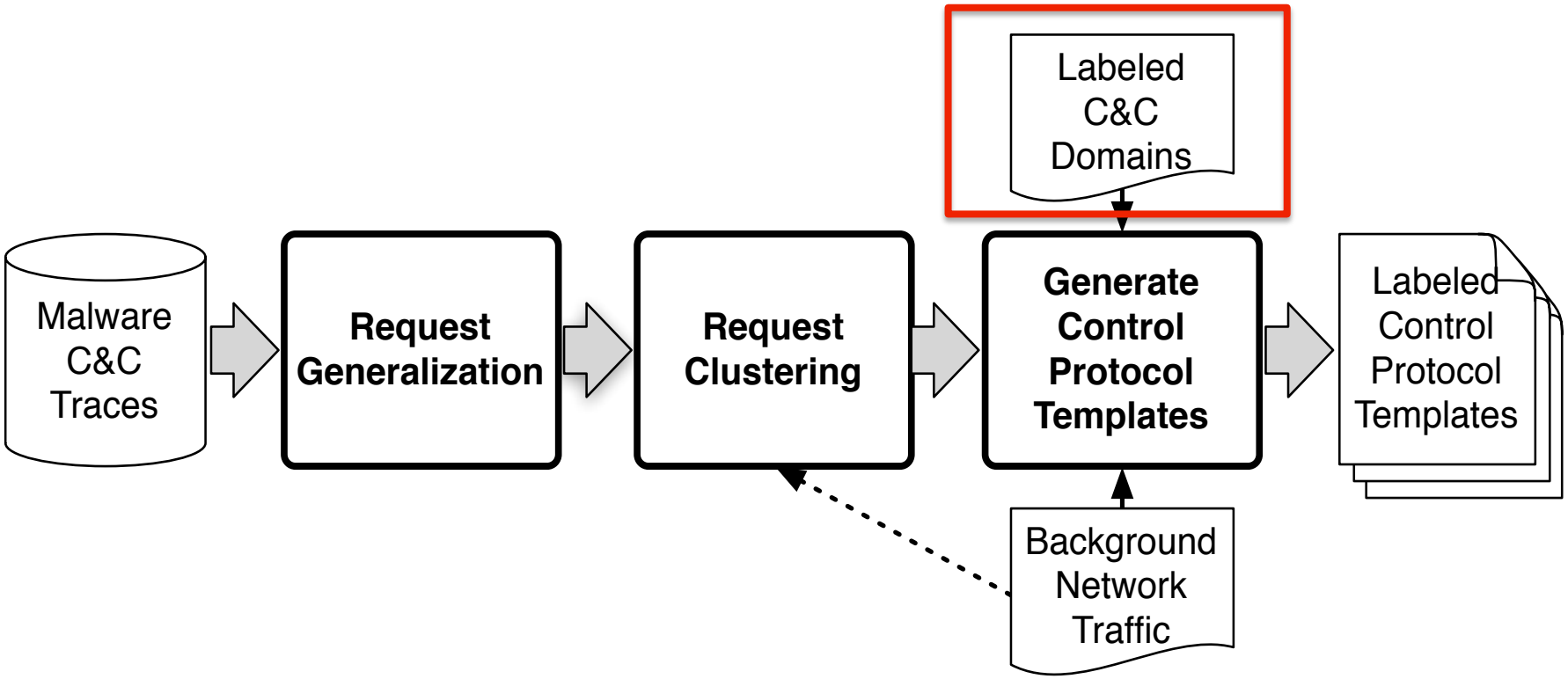
Request 1:
GET /<Base64;12>/cnc.php?v=<Int;3>&cc=<Str;2>
Host: www.bot.net
User-Agent: <Hex;32>
...

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

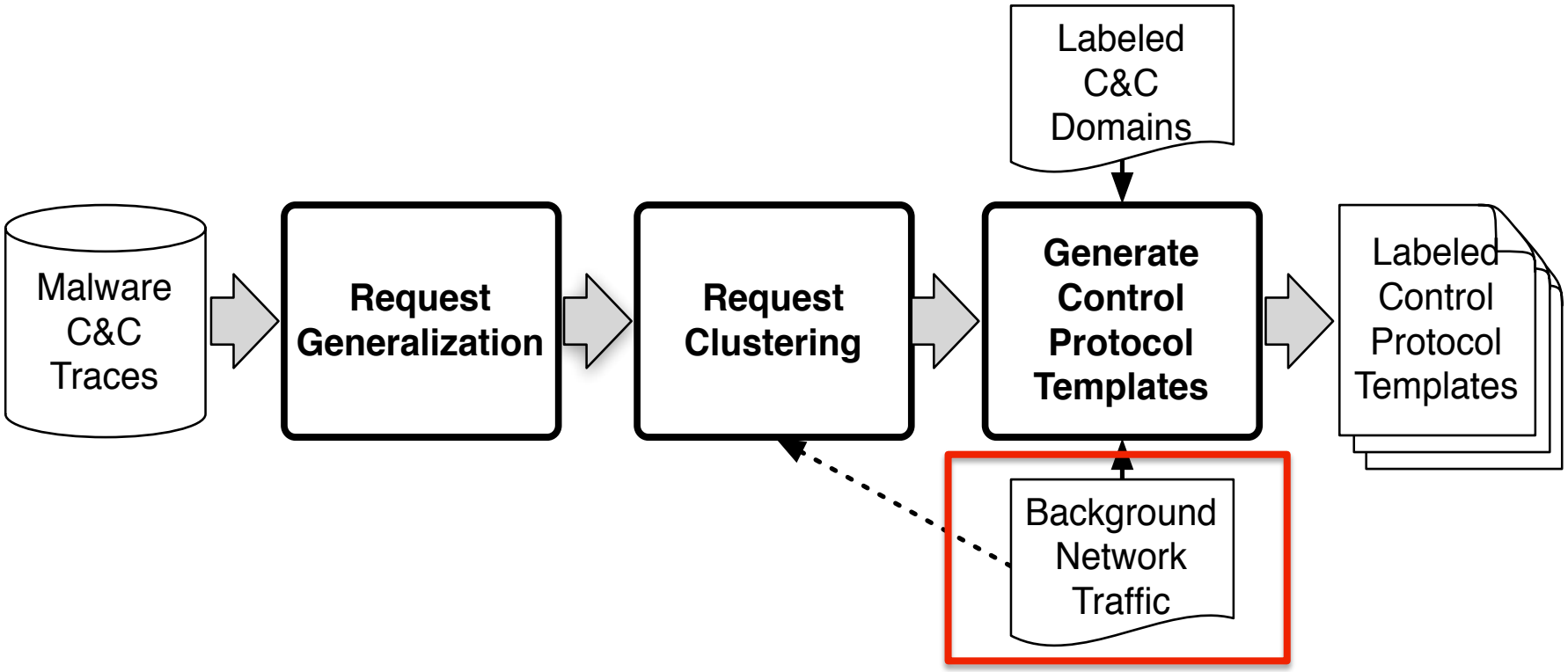
Request Clustering



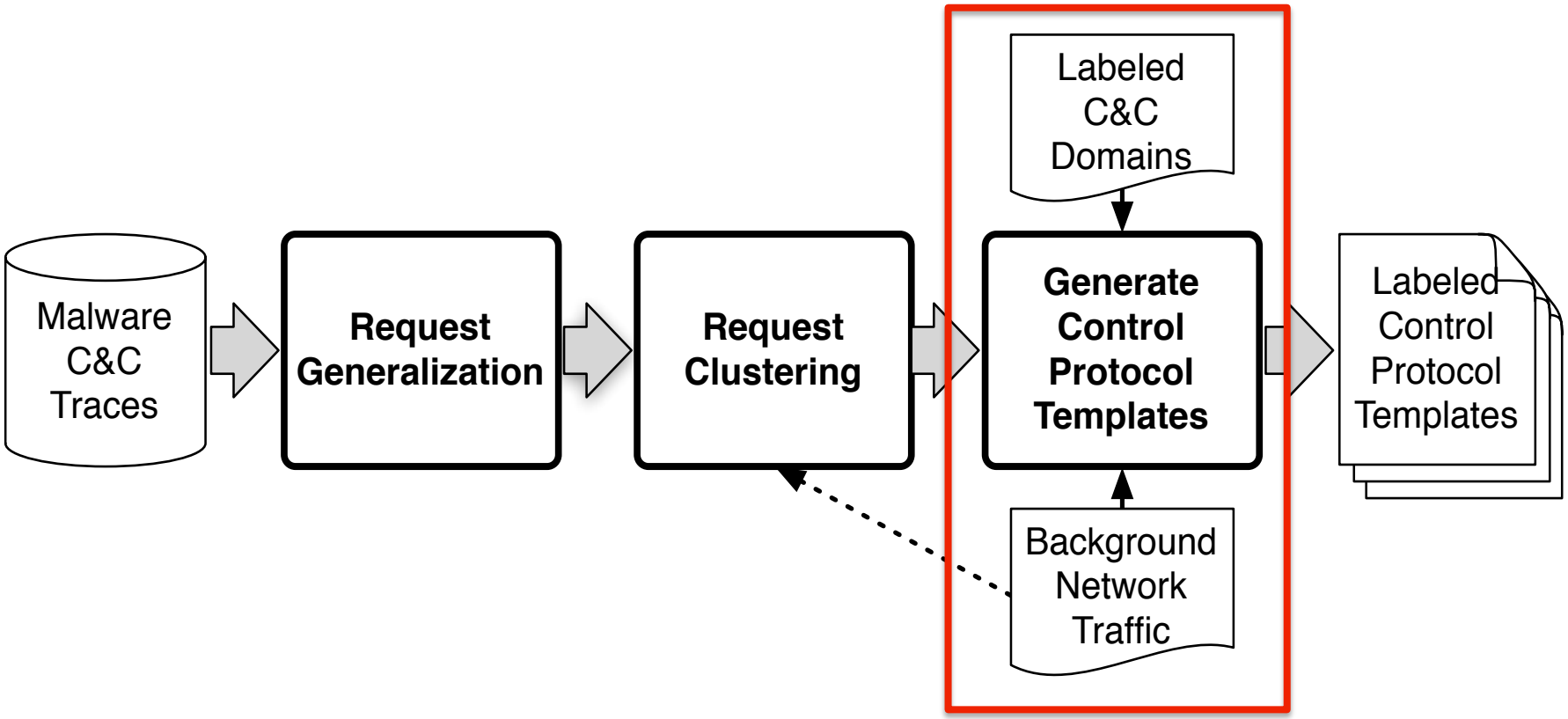
Labeled C&C Domains



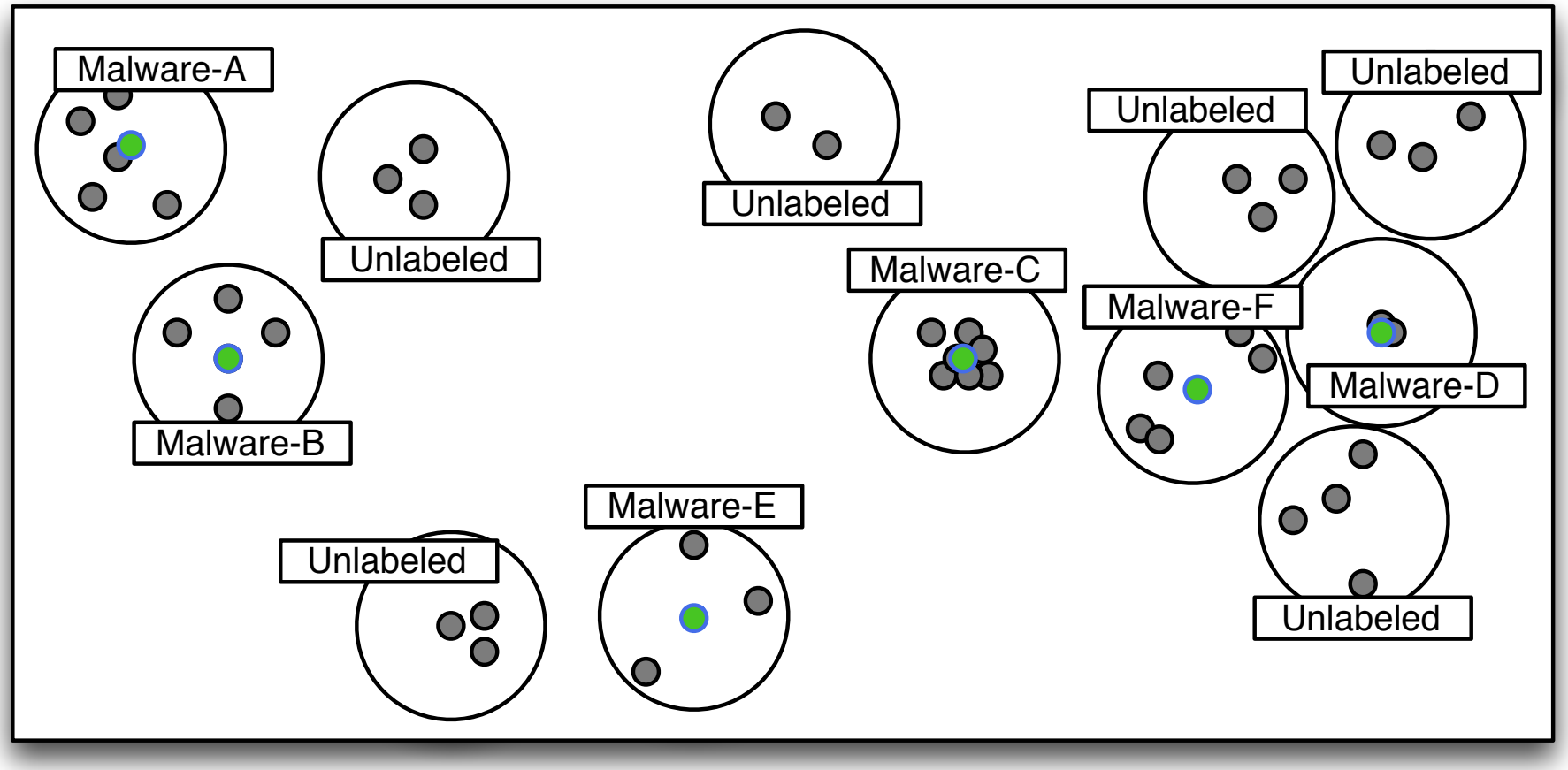
Labeled C&C Domains



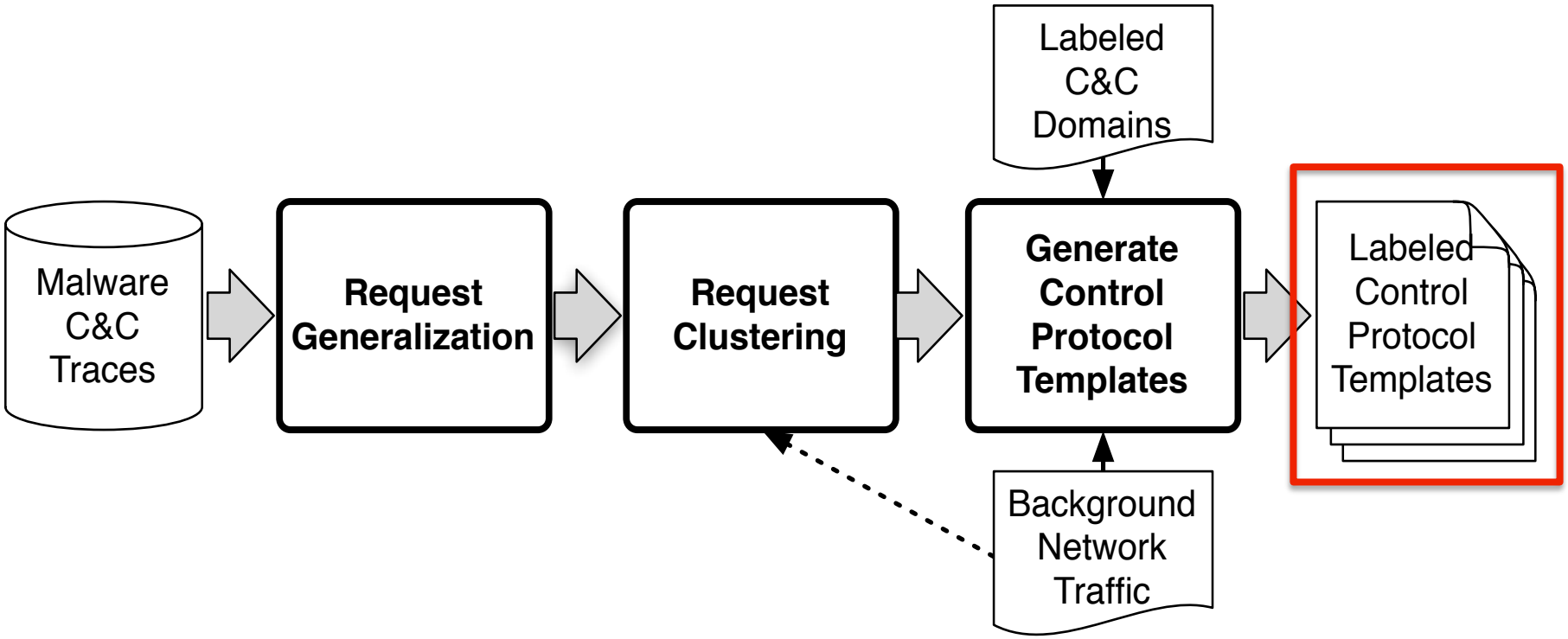
Generating CPTs



Generating CPTs



Labeled CPTs



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 /.*\?(cclv)=

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 > 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)
  - **/Ui4gUGVyZGlzY2kK (<Base64, 16>)**- Request
  - **/VGVycnkgTmVsbXMK (<Base64, 16>)**- CPT

Evaluation Deployment Networks

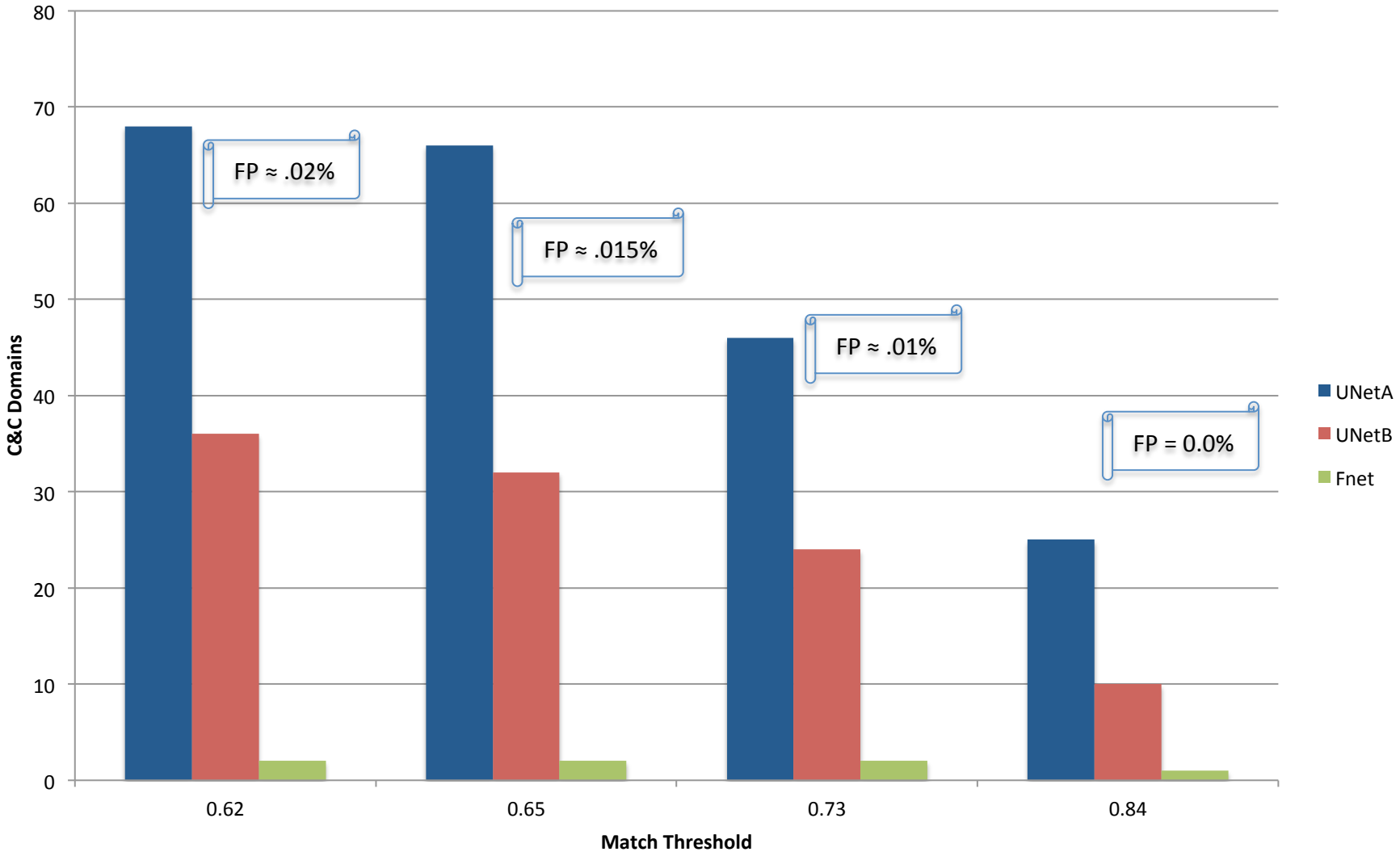
	UNetA	UNetB	FNet
Distinct Src IPs	7,893	27,340	7,091
HTTP Requests	34,871,003	66,298,395	58,019,718
Distinct Domains	149,481	238,014	113,778

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

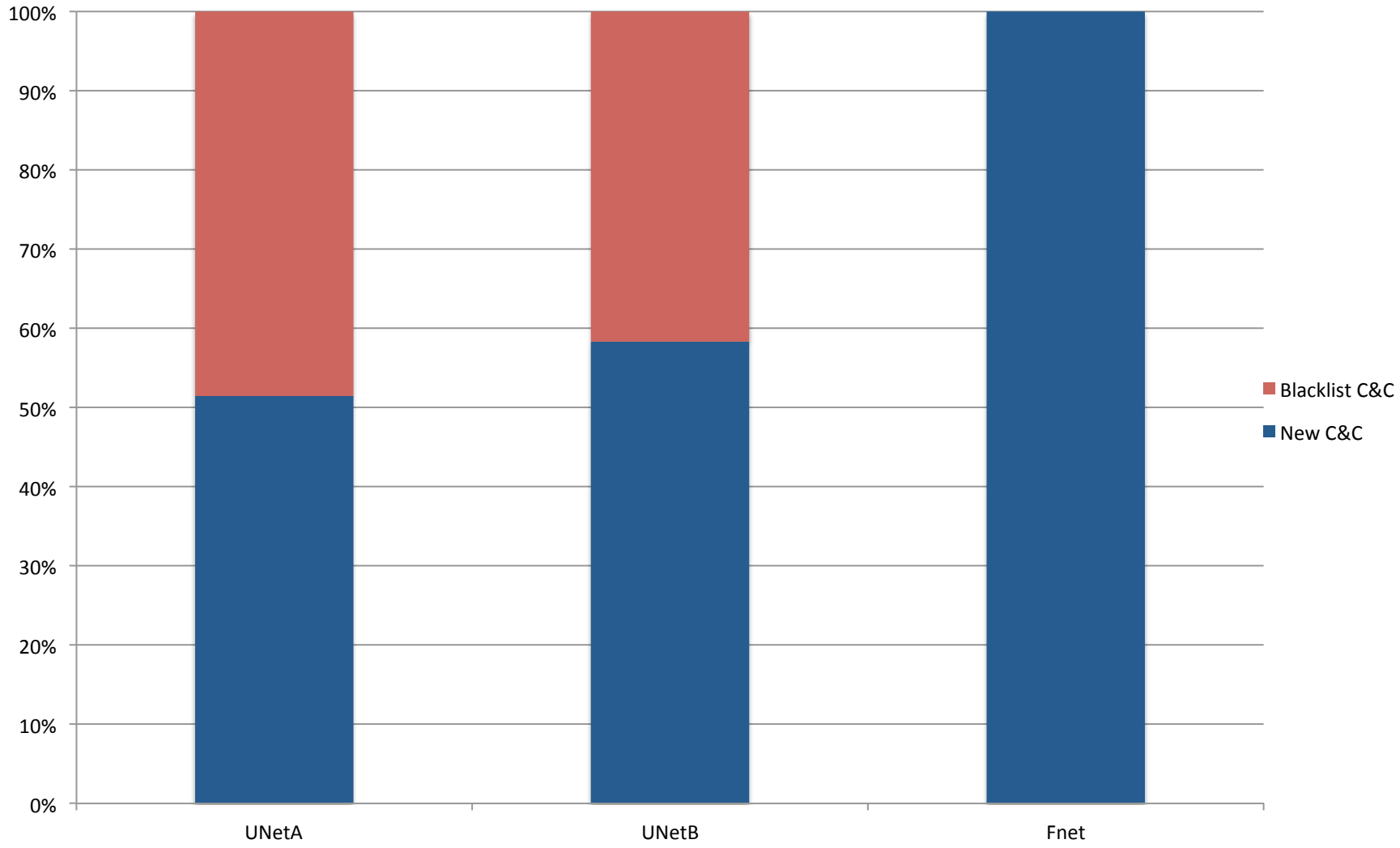
Ground Truth

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

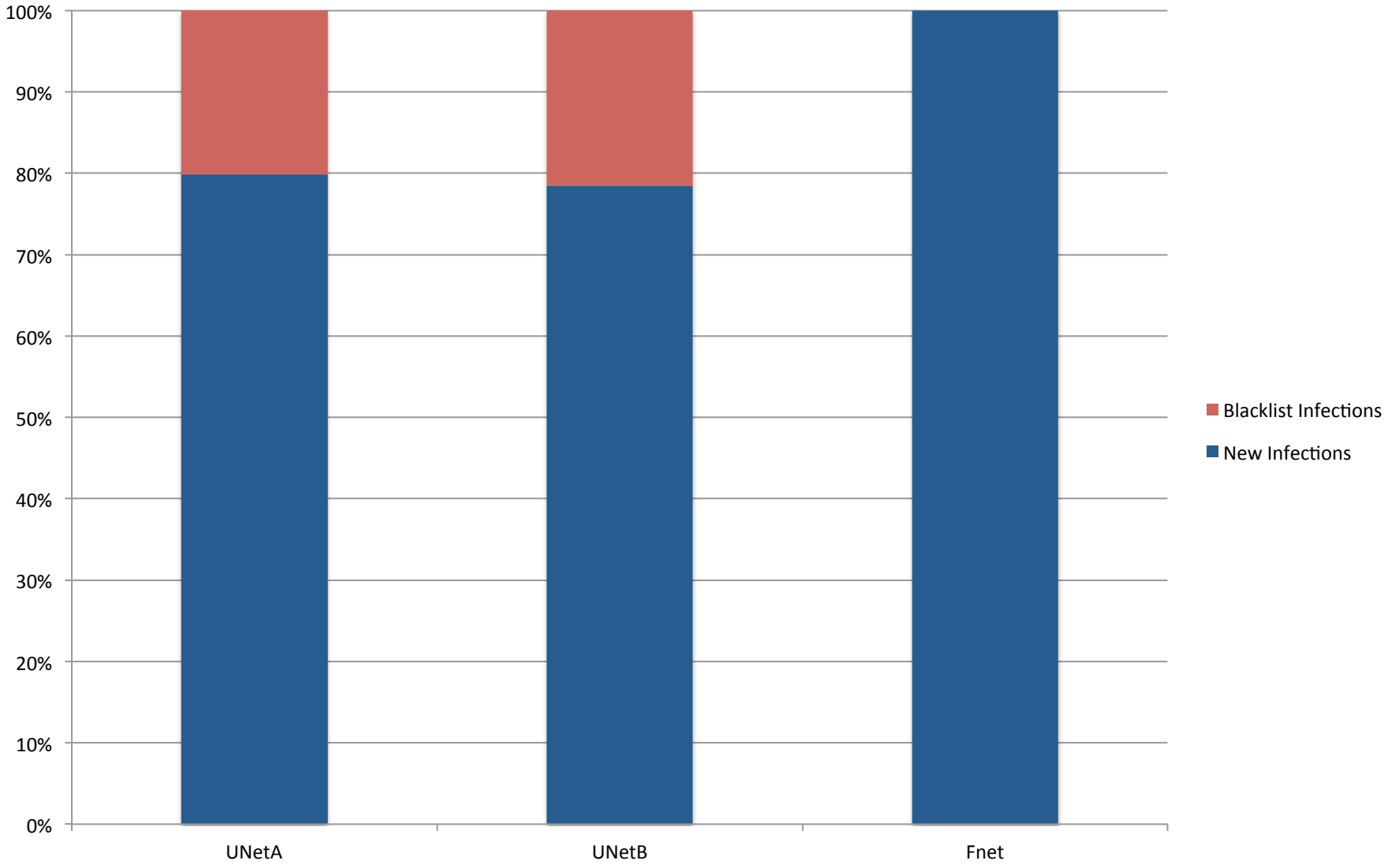
Finding C&C Domains



New vs. Blacklist Domains



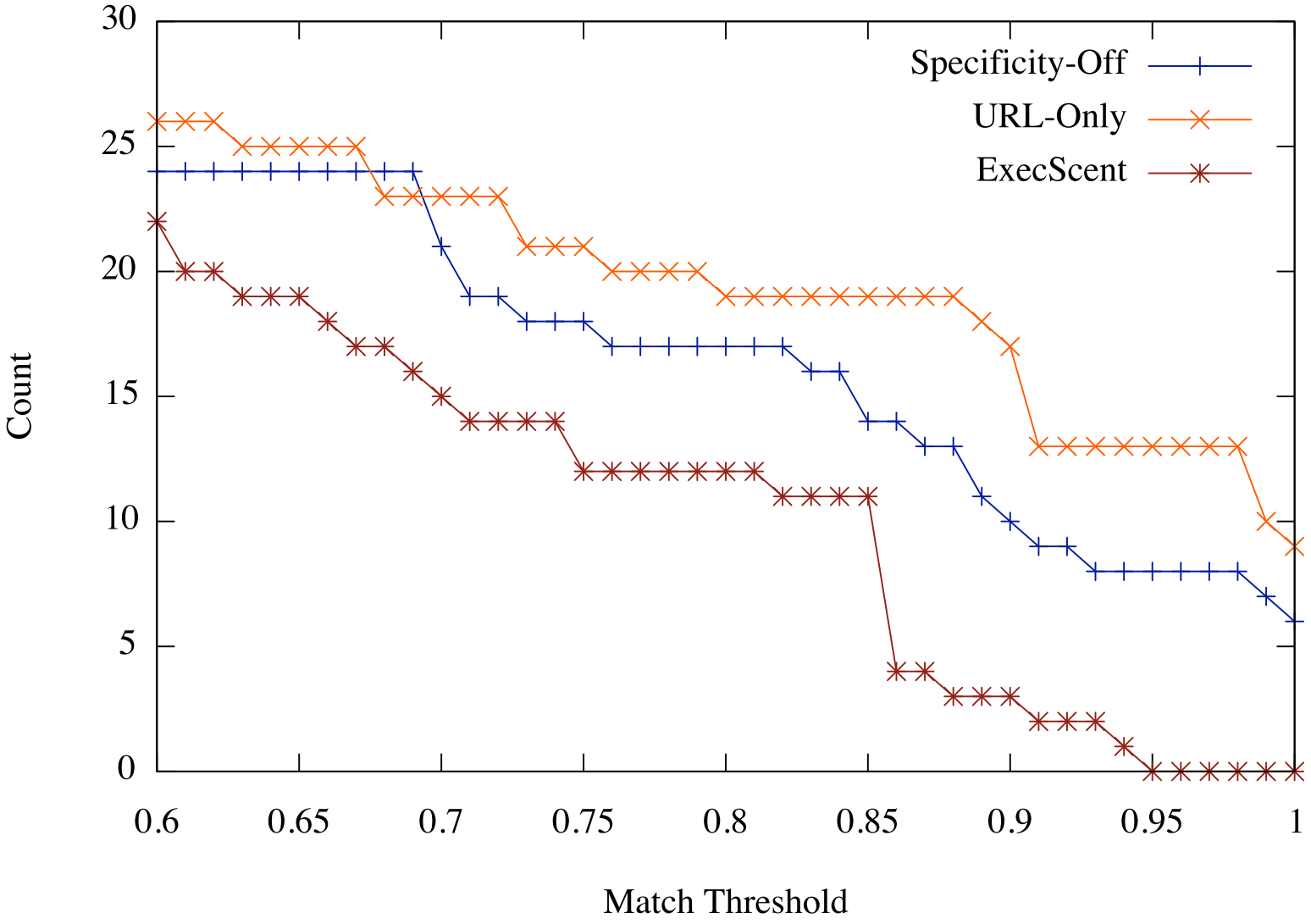
New vs. Blacklist Infected Hosts



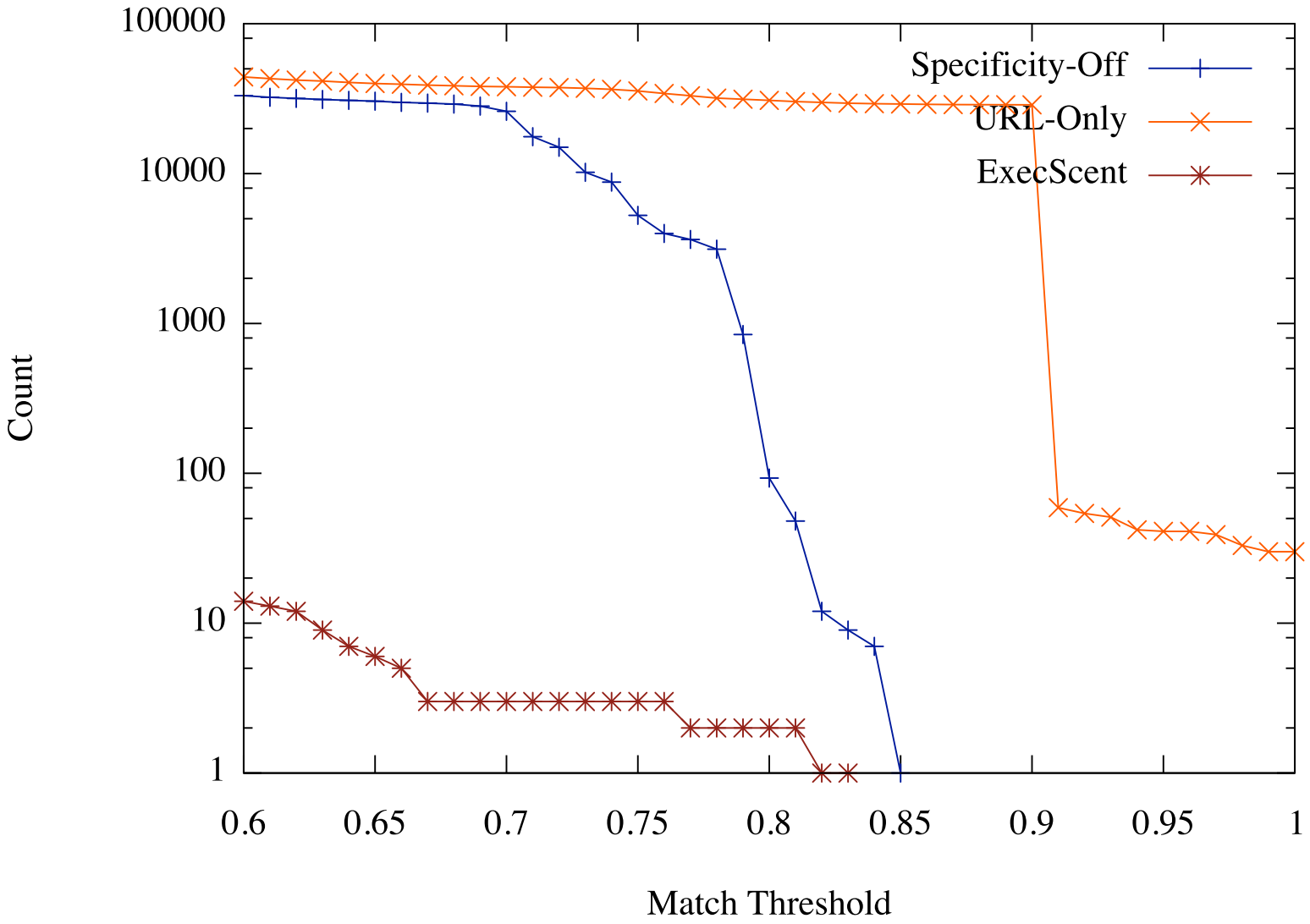
ISP Deployment

- 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.

Questions?