Project AMSEL:
Automatically Collect and Learn To Detect Malware

Martin Apel, Michael Meier

technische universität dortmund

Department of Computer Science
Chair VI – Information Systems and Security
My Office
Fraunhofer Cyber Defense Center Bonn
Cyber Defense Labs

Cyber Defense Lab Bonn

- Malware Analysis & Digital Forensics
  - Malware Analysis
  - Digital Forensics
  - Honeypots/Honeynets
  - Botnet Analysis & Countermeasures

- Mobile and Sensor Networks
  - Secure Ad hoc Routing
  - Secure Sensor Networks

- Teaching and Training
  - Student Education
  - Professional Training

Cyber Defense Lab Wachtberg

- Monitoring & Situational Awareness
  - IDS for heterogeneous Networks
  - Operational Picture & Situational Awareness
  - Intrusion Response

- Resource-efficient Cryptography
  - Efficient Key Management
  - Application Protection Protocols
  - Network Protection Protocols

- Secure Network Architectures
  - Interoperable Coalition Architectures
  - Multi-Level Security
  - Gateway Concepts
  - Protected Core Networking

http://www.enisa.europa.eu/botnets
Overview

- project overview
  - early warning systems
  - architecture of a malware early warning system
- experiences
  - early
    - detection delay
    - generalizing signatures
    - incompleteness
- summary
Early Warning Systems [1]

- aim at
  - detecting yet unclassified but potentially harmful system behavior
  - based on preliminary indications
  - establish hypotheses, predictions and advices in not yet completely understood situations
  - include two meanings of „early“
    - “fast”: start early in time in order to avoid/minimize damage
    - “incomplete”: process uncertain and incomplete information

Malware Appearance

- malware propagates in polymorphic form
  - morphing/obfuscating tools generate programs of equal/similar functionality but different (“appearance”) feature instantiations
    - 30,000 new unique (wrt. “appearance” features) malware samples a day
      - polymorphic variants of a few malware types
    - would require to handle 30,000 new signatures a day

replication

polymorphic transformation

polymorphic variant
Idea of a Malware EWS

- automatically
  - collect malware
  - analyze malware behavior
  - generate signatures
    - n-gram based vectorization of behavior reports
    - manhattan distance
    - complete linkage clustering
    - shared sub-strings of clusters (Ukkonen’s algorithm)
      - not shared with good pool
  - distribute and deploy signatures
  - report alerts centrally

Automatically Malware Sammeln und Erkennen Lernen automatically collect and learn to detect malware
Architecture

alert and threat repository

Detecting and alerting box
- Signature generator
- Clustering
- Malware collector
- Malware analysis system

Collecting and learning box
- Analysis reports
- Signatures
- Malware samples

Situation picture

detection system
Definition: Detection delay describes the time elapsing between occurrence of a malware sample at a collecting and learning box and it’s earliest possible detection at a detecting and alerting box.
Estimation of Detection Delay I
Example: known sample, signature available

0:00
min sec

alert and threat repository

malware samples
analysis reports
signatures
signatures
alerts

collecting and learning box

signature-generator
clustering
malware collector
malware analysis system

detecting and alerting box

e-mail Scanner
execution interceptor
Estimation of Detection Delay II
Example: unknown sample, no signature available

0:00 | 0:10 | 2:10 | 4:40 | 4:50
min   sec

collecting and learning box

signature generator
malware collector
clustering
malware analysis system

detecting and alerting box

e-mail scanner
execution interceptor

alert and threat repository

malware samples
analysis reports
signatures
signatures
alerts
## Summary - Detection Delay

<table>
<thead>
<tr>
<th>process</th>
<th>main influence</th>
<th>component</th>
<th>delay (min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>malware collector</td>
<td>sample download</td>
<td>collecting and learning box</td>
<td>0:10</td>
</tr>
<tr>
<td>malware analysis</td>
<td>analysis time limit</td>
<td>collecting and learning box</td>
<td>2:00</td>
</tr>
<tr>
<td>clustering and signature generation</td>
<td>number of locally available samples</td>
<td>collecting and learning box</td>
<td>0:30</td>
</tr>
<tr>
<td></td>
<td>GoodPool size</td>
<td></td>
<td>2:00</td>
</tr>
<tr>
<td>signature distribution</td>
<td></td>
<td>alert and threat repository</td>
<td>0:10</td>
</tr>
<tr>
<td>sum</td>
<td></td>
<td></td>
<td>4:50</td>
</tr>
</tbody>
</table>
# Cases: Detection Delay

<table>
<thead>
<tr>
<th></th>
<th>sample</th>
<th>signature</th>
<th>sum (delay)</th>
</tr>
</thead>
<tbody>
<tr>
<td>known malware</td>
<td>known</td>
<td>available</td>
<td>(\leq 0:10^a)</td>
</tr>
<tr>
<td>new malware</td>
<td>unknown</td>
<td>not available</td>
<td>4:50</td>
</tr>
<tr>
<td>new variant</td>
<td>unknown</td>
<td>generalizing</td>
<td>(\leq 0:10^b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>not generalizing</td>
<td>4:50^c</td>
</tr>
<tr>
<td>„fresh malware“</td>
<td>known</td>
<td>not available</td>
<td>(\leq 4:50^d)</td>
</tr>
</tbody>
</table>

---

^a signature available, but possibly not yet distributed globally

^b available signature covers variant, but possibly not yet distributed globally

^c available signature must be extended

^d signature generation still running
Conclusion on Detection Delay

• reference value for detection delay: 4 min 50 sec

• generalizing signatures allow early detection
  ♦ case „new variant“

• results of individual collecting and learning steps are available for the situation picture immediately
Detection Rate and Generalization of Signatures

- How many behavior reports of a cluster (of size N) need to be processed in order to create a signature with detection rate 1 (wrt. N cluster elements)?
Early (incomplete)

• honeypot assumption
  - all samples collected by malware collectors are malware

• incompleteness due to dynamic analysis
  - assumption: malware show malicious behavior during analysis
  - limited analysis time (about 2 minutes)
  - only simulated user interaction during malware analyse

• incompleteness of GoodPool
Time for Demo?

- Sample #10911
- Report
- Cluster
- Signature
Summary

- architecture of an automatic EWS

- focus of our ongoing research
  - new approaches for malware collection and analysis
  - clustering of malware behavior
  - generating behavior signatures
  - balancing conflicting availability and confidentiality requirements
Thank You!

- Contact
  Michael Meier
  TU Dortmund / Informatik 6
  michael.meier@udo.edu
  michael.meier@cs.uni-bonn.de
  michael.meier@fkie.fraunhofer.de
  http://ls6-www.cs.tu-dortmund.de/~meier/