Malware Analysis Through High-level Behavior

Xiyue Deng, Jelena Mirkovic
{xiyueden,mirkovic}@isi.edu
Information Sciences Institute
University of Southern California
Background

- Malware is an increasing threat to cybersecurity.

**Kaspersky Lab: 323,000 New Malware Samples Found Each Day**

Credit it to mass-produced malware and better detection through machine learning.

Antivirus provider Kaspersky Lab has revealed that around 323,000 new malware files are being identified each day by its product as opposed to 70,000 files per day in 2011. This, it claims, is an increase by 13,000 per day when compared to last year.
Background (cont’d)

- Malware is becoming more and more stealthy to evade detection and analysis
- Evasion techniques make signature-based techniques more costly and unsuccessful
- Stealth techniques make dynamic approaches less effective
Contemporary Work

- Static Binary Analysis
- Dynamic Analysis on Host
- Dynamic Analysis of Network Behavior
Static Analysis

- Identify **portions of binary code** as signature for detection
- E.g. CTPL, Generic Virus Scanner, etc.
- Pro
  - **Safe** - No malware is running
  - **More thorough** - inspect all possible branches of code
- Con
  - Susceptible from **obfuscation** techniques
    - Junk code generation
    - Encryption and oligomorphism
    - Polymorphism
    - Metamorphism
Dynamic Analysis on Host

- Analyze the **interaction** between malware and host
  - Memory access, file access, system call, etc.
- Patterns likely to present in malware -> signature
- E.g. CWSandbox, Cuckoo sandbox, etc.
- Pro
  - Overcome malware code obfuscation
- Con
  - **Not cover all behaviors** of a malware due to missing trigger
  - **Stealthy** techniques evade debuggers and VMs
    - Once detected the malware modifies its behavior
Related Work of Network Behavior

- Analyze the interaction of malware with other hosts
  - Capture network traces with all details
- Sandnet (Rossow et al., 2011)
  - Protocol popularity
  - No understanding of communication purpose
- Morales, et al., 2010
  - Defined 7 activities based on heuristics and analyzed malware for presence of these activities.
  - Limited insight into malware’s purpose
- Nari, et al. 2010
  - Perform automated classification of malware behavior
  - Flow dependency solely on IP address
Our Work

- Capturing Malware Network Behavior
  - Through our Fantasm experiment environment
- Define and generate High-Level Behavior (HLB)
  - A set of different behaviors to identify malware purpose
- Study the relations of different malwares through HLB
Fantasm Experiment Environment

- Partially contained experimental environment
  - Evaluate the risk of each outgoing flow
  - Whether the flow is essential to the malware’s functioning
  - Fake replies to the flow if possible
  - Let out essential flow while monitoring its behavior
- Decision flowchart

![Decision flowchart diagram]
Fantasm Decision Flowchart

- Essential - malware will not run without it
  - We can identify these over multiple experiment runs
  - If we can fake replies, use impersonators
    - E.g. ICMP ECHO reply, SMTP, FTP, etc.
  - Cannot fake and non-risky: let it out with strict monitoring.
    - E.g. binary download, receive commands via C&C, resolve DNS names
    - Monitor for and stop DDoS and scanning - too many suspicious, one-sided flows
- Non-essential or risky
  - Drop
## High-Level Behavior (HLB)

- Our definition: Identified network behavior patterns that are commonly used for malicious purposes
- 6 categories

<table>
<thead>
<tr>
<th>Action</th>
<th>Network Activity Features</th>
<th>Detection Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downloading</td>
<td>Connection to remote server</td>
<td>HTTP GET for a specific URL</td>
</tr>
<tr>
<td></td>
<td>Receiving payload for further action</td>
<td>Receive payload (binary or text)</td>
</tr>
<tr>
<td>Reporting</td>
<td>Connection to remote server</td>
<td>HTTP POST to remote server with parameters or payload</td>
</tr>
<tr>
<td></td>
<td>Submit data to remote server</td>
<td></td>
</tr>
<tr>
<td>Scanning</td>
<td>Detecting availability of other hosts</td>
<td>ICMP ECHO packets to other hosts</td>
</tr>
<tr>
<td></td>
<td>Detecting open ports of hosts</td>
<td>TCP SYN or UDP packets to ports</td>
</tr>
<tr>
<td>Spamming</td>
<td>Connection to public mail server</td>
<td>TCP connection to port 25, 465, 587</td>
</tr>
<tr>
<td></td>
<td>Sending spam emails</td>
<td>Text email with URL or attachment</td>
</tr>
<tr>
<td>C&amp;C Communication</td>
<td>HTTP, IRC connection to remote server</td>
<td>IRC protocol / ports</td>
</tr>
<tr>
<td></td>
<td>Receive instructions</td>
<td></td>
</tr>
<tr>
<td>Propagating</td>
<td>Trying to transfer data over to other hosts</td>
<td>FTP, samba, NFS protocol / ports</td>
</tr>
</tbody>
</table>
Experiment Setup

- Fantasm Environment Setup
  - Enable service network monitoring on Linux gateway
  - Reload operating system on Windows Node and set up necessary network configurations
  - Deploy and start the malware binary and continue running it for a given time period (5 minutes)
  - Kill the malware process and save the captured network trace
Malware Dataset

- Collections from Georgia Tech Apiary project
- With valid Virustotal report
- With 50%+ AV suites detection rate from Virustotal
- Randomly select 999 samples
Result

- Percentage distribution

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downloading</td>
<td>10.9%</td>
</tr>
<tr>
<td>Reporting</td>
<td>5.6%</td>
</tr>
<tr>
<td>Scanning</td>
<td>28.5%</td>
</tr>
<tr>
<td>Spamming</td>
<td>2.2%</td>
</tr>
<tr>
<td>C&amp;C Communication</td>
<td>0.2%</td>
</tr>
<tr>
<td>Propagation</td>
<td>11.5%</td>
</tr>
<tr>
<td>Not Detected</td>
<td>57%</td>
</tr>
</tbody>
</table>
Result (cont’d)

- Distribution of HLB count (among all showing activities)

<table>
<thead>
<tr>
<th>Count of behaviors</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67.6%</td>
</tr>
<tr>
<td>2</td>
<td>28.2%</td>
</tr>
<tr>
<td>3</td>
<td>3.4%</td>
</tr>
<tr>
<td>4</td>
<td>0.7%</td>
</tr>
<tr>
<td>5 or more</td>
<td>0%</td>
</tr>
</tbody>
</table>
## Result (cont’d)

<table>
<thead>
<tr>
<th>Behavior Combinations</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scanning + Propagating</td>
<td>43.1%</td>
</tr>
<tr>
<td>Downloading + Propagating</td>
<td>16.5%</td>
</tr>
<tr>
<td>Scanning + Spamming</td>
<td>12.2%</td>
</tr>
<tr>
<td>Downloading + Scanning</td>
<td>7.9%</td>
</tr>
<tr>
<td>Downloading + Scanning + Propagating</td>
<td>5.7%</td>
</tr>
<tr>
<td>Downloading + Uploading</td>
<td>4.3%</td>
</tr>
<tr>
<td>Uploading + Scanning</td>
<td>2.2%</td>
</tr>
<tr>
<td>Downloading + Uploading + Scanning + Propagating</td>
<td>2.2%</td>
</tr>
<tr>
<td>Downloading + Uploading + Propagating</td>
<td>1.4%</td>
</tr>
<tr>
<td>Scanning + Spamming + Propagating</td>
<td>1.4%</td>
</tr>
<tr>
<td>Uploading + Scanning + Propagating</td>
<td>1.4%</td>
</tr>
<tr>
<td>Uploading + Propagating</td>
<td>0.7%</td>
</tr>
<tr>
<td>Downloading + Uploading + Scanning</td>
<td>0.7%</td>
</tr>
</tbody>
</table>
Discussion and future work

- Malware detection
  - Based on common malware behavior combinations
- Polymorphic malware
  - Detect similar behaviors from polymorphic malware
- Malware genealogy
  - Study relations among behaviors
  - combination of malware behaviors for a common goal
- Malware evading network analysis
  - Dormant, awaiting a specific trigger
  - Interleave malicious activities with benign traffic
  - Attempt to detect analysis environment
Conclusion

- Propose using patterns of malware’s network behaviors to identify high-level behavior (HLB)
- Study prevalence of HLB in contemporary malware
- Discuss how to use HLB for future malware study
For More Information

Xiyue Deng <xiyueden@isi.edu>
Jelena Mirkovic <mirkovic@isi.edu>

Project page: http://steel.isi.edu/Projects/malwareExp/
Thank you!