Malware Classification

Coseinc Advanced research labs

Udi Shamir (tal0n)

udi@coseinc.com
Malware Classification

• During our work in the AML team we are analyzing hundreds of daily new and repeated malicious binaries, key loggers, phishing and other threats (Yeh, some are confidential :D ).

• while it is clear that without automation we could not even coverage 20% of our daily archive we had to write some new plugins for our CAMAL platform which perform the analysis section. the new plugins assist us detect new threats and better support our customers.
Malware Classification
existing products

- VxClass by Zynamics (Google)

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>use same algorithm as Bindiff</td>
<td>when new packer version arrive it fail</td>
</tr>
<tr>
<td>big signature database</td>
<td>pricey</td>
</tr>
<tr>
<td>intuitive</td>
<td>is now owned by Google:(</td>
</tr>
</tbody>
</table>
Malware Classification

existing products

• Yara, assist in identify static binary signatures, packers.
• Bindiff by Zynamics, compare and disassemble chunks of codes
• Binnavi by Zynamics, great for vulnerability detection and compare code flow
Malware Classification

- categorized (labeling) malwares into families
- identifies behavioral patterns
- identify new variants
Categorization

- the categorization process helps us distinguish betweens:
- viruses, spywares, backdoors, trojans and worms.
- Detect new variants, targeted attacks
Categorization (Ingredients)

- Collecting Samples
- Auto analysis, sets
- Defining behavioral
- Clustering method
SAMPLE COLLECTION

• **VireX Network** is isolated network our partners are using to exchange new samples with the AML team. The exchange process provides usually good binary quality.

• **Honeypots** are deployed on tier1 providers separated by different geo location and are divided into 2 main categories: low/high interaction. The honeypots provides 5% of our daily malicious binaries collection.

• **Crawlers** we developed crawlers which scan related forums and leech binaries, the collection here is minimal but often provides new mutant.

• **Intelligence** ok, no need to add here :)

Tuesday, April 26, 2011
Auto Analysis (continue ...)

- Av vendors does not agree on standard naming convention
- Some vendors use better technique to differentiates between small variants in same family while others don’t!
- we cannot rely on specific engine to help in the labeling process
- we decided to use one engine for the initial classification
Auto Analysis (continue ...)

- We used legacy malware with 2 different hash from the same family.
- Malware signature exist in all engines

<table>
<thead>
<tr>
<th>HASH</th>
<th>CLAMAV</th>
<th>FProt</th>
<th>Trend</th>
<th>McaFee</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Blaster</td>
<td>msblast</td>
<td>msblast</td>
<td>Eploit-Dcom</td>
</tr>
<tr>
<td>Y</td>
<td>Dcom.Exploit</td>
<td>msblast</td>
<td>msblast</td>
<td>W32-Blaster-W</td>
</tr>
</tbody>
</table>
Auto Analysis (continue ...)

- The inspected AV engines provides around 50% consistency
Auto Analysis

- Dealing with the enormous samples we are collecting/receiving each day is challenging and intractable.
- We are using the CAMAL framework to analyze each sample and to provide us with the initial datasets
Defining Behavioral

- Our reference data extracted from our large database located in the Camal AML
- The samples have been collected over the year taking average of 30GB a day.
- we uses one AV engine to perform the initial categorization for families
- Due to the skewed distribution we discarded classes with less than 10 samples
Defining Behavioral

<table>
<thead>
<tr>
<th>Malware Classes</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rbot</td>
<td>80</td>
</tr>
<tr>
<td>Sality</td>
<td>60</td>
</tr>
<tr>
<td>Virut</td>
<td>130</td>
</tr>
<tr>
<td>Storm</td>
<td>28</td>
</tr>
</tbody>
</table>
Defining Behavioral

- All malware are executed without user permission so we assume they are malicious.
- System call manipulations
- Loading Drivers
- Registry modification
- Modification of system files
- Geo location
### CAMAL PROFILER REPORT

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>process creation</strong></td>
<td>&quot;process&quot;,&quot;created&quot;, &quot;C:\WINDOWS\system32\rundll32.exe&quot;</td>
</tr>
<tr>
<td><strong>write opcode</strong></td>
<td>&quot;Write&quot;, &quot;C:\Documents and Settings\admin\Start Menu\Programs\Internet Explorer.url&quot;</td>
</tr>
<tr>
<td><strong>registry modification</strong></td>
<td>&quot;registry&quot;,&quot;SetValueKey&quot; &quot;{d669d1cf-0f47-11e0-928e-806d6172696f}\BaseClass&quot;</td>
</tr>
<tr>
<td><strong>delete opcode</strong></td>
<td>&quot;Delete&quot;,&quot;C:\Documents and Settings\admin\Desktop\CaptureClient\binary.exe&quot;,&quot;C:\RegTemp.txt&quot;</td>
</tr>
<tr>
<td><strong>Geo Location</strong></td>
<td>Country Edition: BY, Belarus</td>
</tr>
</tbody>
</table>
## CAMAL PROFILER REPORT

<table>
<thead>
<tr>
<th>number of opcodes (native system excluded)</th>
<th>28</th>
</tr>
</thead>
<tbody>
<tr>
<td>type of opcodes (for specific list, see Camal DB reference)</td>
<td>4, 8, 10</td>
</tr>
<tr>
<td>drop zone</td>
<td>Country Edition: BY, Belarus</td>
</tr>
<tr>
<td>Protocols Numbers (for grouping list, see camal DB reference)</td>
<td>1, 4, 6</td>
</tr>
<tr>
<td>Geo Location</td>
<td></td>
</tr>
</tbody>
</table>
K-means

- One of the simplest learning algorithm
- Solve many known clustering problems
Categorization

K-means
Categorization

K-means
Summary

- Classification is important for better understanding how malwares operate
- Detect new outbreaks, provide early warning
- Assist in detecting new variants
- Our method rely on emulation, hence we are not depend on unpackers (unless we are being detected)