Zero Day Malware Detection by Ensemble Based Hybridization for Static and Dynamic Malware Detection Techniques

Mesut Kaya, Berker BATUR, Tamer Tavaslıoğlu
COMODO, Ankara
Introduction to Malware
Malware Business
Malware Detection Techniques
Weakness of each Malware Detection Techniques
COMODO’s Solution, Valkyrie
Valkyrie Automated Static Analysis
Valkyrie Automated Dynamic Analysis
Classification Techniques Comparison
Ensemble Based Hybridization Experiments
Latest & Upcoming Researches
Malware

Malware, short for malicious software, is any software used to disrupt computer operations, gather sensitive information, gain access to private computer systems, or display unwanted advertising

Malware Types: *Virus, Worm, Spyware, Adware*, etc..

* https://en.wikipedia.org/wiki/Malware
Malware Business

- 30 billion USD/year
- %5 of computers in companies are already infected
- %24 of the enterprises are breached at least ones

* Gartner & PricewaterhouseCoopers
Malware Detection Techniques

- Signature Based / Pattern Matching
- Static Analysis
- Dynamic Analysis
- Hybrid Detections
Weakness of ind. Malware Detection Techniques

- **Signature Based / Pattern Matching**
  - Costly to produce manually
  - Insufficient for Zero-Day malware detection

- **Static Detection**
  - Packed Executable
  - Lack of runtime behavior

- **Dynamic Detection**
  - Sandbox awareness
  - Long analysis time

- **Hybrid Techniques**
  - Hard to implement & improve
  - Combination individual techniques
COMODO’s Solution, Valkyrie

- is a File Verdict System
- conducts several analysis, using hundreds of dynamic and static features
- has advanced automated malware detection components based on machine learning classifiers
- validates heuristic results with further Human Expert Analysis solution
- is a complete Cloud Solution for malware detection
Static Analysis in Valkyrie

- Anomaly Detection
  - Anti-VM present?
  - Packer detection
  - PE section anomalies
  - etc..

- Ensemble based Statistical Classification
  - Imported DLLs
  - Imported function calls
  - Hundreds of PE features
    - Machine type
    - Number of sections
    - Entropy of sections
    - Size of headers
    - etc..
Dynamic Analysis in Valkyrie

- Automated behavior extraction using Sandbox
  - Fully automated
  - No intervention while running

- Monitoring run time actions for system-wide changes
  - Modified files
  - Registry updates,
  - etc..

- Ensemble based Statistical Classification
  - Different Run time API calls and frequencies
  - Specific changes and monitored activities
    - VM awareness
    - Network activity density
    - Modified files
    - Registry updates
    - etc..
Experiments on Different Classification Techniques

- Static Analysis Malware Classification

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Overall Accuracy (10-Fold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>%92.1</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>%93.0</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>%88.8</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>%93.1</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>%94.8</td>
</tr>
<tr>
<td>Random Forest</td>
<td>%95.9</td>
</tr>
</tbody>
</table>
Experiments on Different Classification Techniques

- Dynamic Analysis Malware Classification

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Overall Accuracy (10-Fold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>%93.1</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>%94.5</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>%90.4</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>%94.1</td>
</tr>
<tr>
<td>Stochastic Gradient Descent</td>
<td>%90.1</td>
</tr>
<tr>
<td>Random Forest</td>
<td>%95.2</td>
</tr>
</tbody>
</table>
Hybridization

- Challenges
  - Missing features while using Combined Ensemble Classifier
  - Dynamic weighting based on sample
  - Trustworthy thresholds
    - Clean / Malware (or Unknown?)

- Valkyrie
  - Equally-weighted for Final Automated Analysis Result
  - Individual heuristics utilized for faulty cases
Latest & Upcoming Researches

- **Static Analysis**
  - Rich feature set
  - Op-code N-Gram analysis
  - C4.5 | J-48 algorithm

- **Dynamic Analysis**
  - Sequential Pattern Analysis
  - Hidden Markov Models
  - Enhanced behavior monitoring

- **Hybridization**
  - Boosting weak classifiers
Thank you!

https://valkyrie.comodo.com/

*Number of Zero-day malwares found by Valkyrie, listed at our web page daily*