DeepReflect: Discovering Malicious Functionality through Binary Reconstruction

Evan Downing*, Yisroel Mirsky*†◇, Kyuhong Park*◇, Wenke Lee*

* Georgia Institute of Technology
† Ben-Gurion University
◇ Equal contributions
Motivation

Malware Analyst Workflow

Given Malware Sample

- Scan File
  - Known
  - Unknown
    - Malicious
      - Create Signature
    - Non-executing

Reverse Engineer

Finished
Overview

● Analysts want to quickly identify and label malicious functions in malware
● Cannot assume or obtain labeled dataset (too expensive timewise / doesn’t exist)
  ○ Thus we identify these regions via unsupervised learning
● Cannot manually label all regions all of the time (too expensive timewise)
  ○ The analyst labels a few regions in a semi-supervised approach, which adds a bonus of labeling these identified functions
Prior Work

● ML-based solutions: Classification or detection, not behavior identification

● FireEye’s CAPA (July 2020)

● Eyeball strings and API calls indicative of behavior
Challenges & Insights

1. Need to distinguish between benign and malicious behaviors
   a. Use an unsupervised deep learning model (an autoencoder) to locate malicious functions in binaries

2. Understand the semantics of the identified malicious behavior
   a. Use a semi-supervised clustering model which classifies the identified functions
   b. Requires few labels obtained from analyst’s daily workflow
Overview of DeepReflect

Unpacked Malware Sample → Extract BB Features → Reconstruct using Autoencoder → Identify RoI → Cluster → Annotate Malware Functions

DeepReflect
Features

- Inspired from ACFG features used for bug-finding (CCS 2017)
- 18 Features:
  - **Structural**: Flow of operations (e.g., connect, send, recv, etc.)
  - **Arithmetic** Instruction Types: How mathematical operations are carried out at the higher level (e.g., encryption, obfuscation)
  - **Transfer** Instruction Types: Flow of data (arguments provided to and returned from functions)
  - **API Call** Categories: Used to execute behaviors (filesystem, registry, network, process, etc.)
# Dataset

## Benign Dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>Size</th>
<th>Category</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drivers</td>
<td>6,123</td>
<td>Business Software</td>
<td>1,692</td>
</tr>
<tr>
<td>Games</td>
<td>1,567</td>
<td>Utilities</td>
<td>1,453</td>
</tr>
<tr>
<td>Education</td>
<td>1,244</td>
<td>Developer Tools</td>
<td>1,208</td>
</tr>
<tr>
<td>Audio</td>
<td>1,023</td>
<td>Security</td>
<td>1,000</td>
</tr>
<tr>
<td>Communications</td>
<td>994</td>
<td>Design</td>
<td>844</td>
</tr>
<tr>
<td>Digital Photo</td>
<td>826</td>
<td>Video</td>
<td>787</td>
</tr>
<tr>
<td>Customization</td>
<td>778</td>
<td>Productivity</td>
<td>730</td>
</tr>
<tr>
<td>Desktop Enhancements</td>
<td>699</td>
<td>Internet</td>
<td>695</td>
</tr>
<tr>
<td>Networking</td>
<td>612</td>
<td>Browsers</td>
<td>440</td>
</tr>
<tr>
<td>Home</td>
<td>390</td>
<td>Entertainment</td>
<td>257</td>
</tr>
<tr>
<td>Itunes</td>
<td>43</td>
<td>Travel</td>
<td>17</td>
</tr>
</tbody>
</table>

## Malware Dataset

<table>
<thead>
<tr>
<th>Label</th>
<th>virut</th>
<th>vobfus</th>
<th>hematite</th>
<th>sality</th>
<th>crytex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>3,438</td>
<td>3,272</td>
<td>2,349</td>
<td>1,313</td>
<td>914</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Label</th>
<th>wapomi</th>
<th>hworld</th>
<th>pykspa</th>
<th>allaple</th>
<th>startsurf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>880</td>
<td>720</td>
<td>675</td>
<td>470</td>
<td>446</td>
</tr>
</tbody>
</table>

Top 10 most populous families
Evaluation 1: Reliability

- Ground-truth samples

- Baseline Tools
  - VGG19 model + SHAP (deep learning comparison)
  - CAPA (FireEye)
  - FunctionSimSearch (Google Project Zero)
Evaluation 1: Reliability (cont.)
Evaluation 2: Cohesiveness

- DeepReflect identified ~600k malicious functions in ~25k malware samples
- HDBSCAN produced ~22k clusters
  - Largest cluster: ~6k functions
  - Noise points: ~60k functions
- Analysts labeled 119 functions via MITRE (60% malicious, 40% benign)
- Clustering matches 89.7% of an analyst’s manually-clustered functions
Evaluation 3: Focus
Evaluation 4: Insights

- Different (unrelated) malware families share the same functions
- 1.7k clusters had at least one singleton sample
- Novel malware families form new clusters
Evaluation 5: Robustness

● Used OLLVM on Rbot and enabled combinations of obfuscations
  ○ Control-flow flattening
  ○ Instruction substitution
  ○ Bogus control-flow

● Mimicry-like attack

● DeepReflect’s results weren’t significantly affected
Discussion

- Obfuscation
- Adversarial ML attacks
- Training Data Quality
- Human Error
Questions & Comments

Email: edowning3@gatech.edu

Implementation & Dataset: https://github.com/evandowning/deepreflect