AndroOBFS: Time-tagged Obfuscated Android Malware Dataset with Family Information

Saurabh Kumar¹, Debadatta Mishra¹, Biswabandan Panda², and Sandeep K. Shukla¹

¹Indian Institute of Technology Kanpur
²Indian Institute of Technology Bombay

MSR 2022
Motivation

- Malware analysis system require labeled dataset
  - Non-obfuscated (Drebin, AMD, RmvDroid, etc..)
  - Obfuscated (PRAGuard)

- PRAGuard dataset
  - Outdated samples till March 2013
  - Does not contains family information

- Needed a new obfuscated malware dataset
  - Tagged with time
  - With family information
AndroOBFS Dataset

- Samples collected from
  - AndroZoo and VirusShare
  - For year 2018, 2019 and 2020

- Tools used
  - AVClass: Labeling with family
  - Obfuscapk: To obfuscate

- 16279 obfuscated samples in six categories
  - 14579 Familial
  - Spread across 158 families

<table>
<thead>
<tr>
<th>Sources</th>
<th>Year</th>
<th>#Samples</th>
<th>Obfuscated</th>
<th>Obfuscated (#Family≥2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AndroZoo + VirusShare</td>
<td>2018</td>
<td>6525</td>
<td>5794</td>
<td>(136)</td>
</tr>
<tr>
<td></td>
<td>2019</td>
<td>8313</td>
<td>7505</td>
<td>(94)</td>
</tr>
<tr>
<td>AndroZoo</td>
<td>2020</td>
<td>1441</td>
<td>1280</td>
<td>(44)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>16279</strong></td>
<td><strong>14579</strong></td>
<td><strong>(158)</strong></td>
</tr>
</tbody>
</table>
Dataset Creation Process

VirusShare + AndroZoo

∀ VT+ ≥ 10

Time-tagged + Family

Obfuscate Samples

Σ VIRUSTOTAL

Dex Date + AVClass

Obfuscapk
Malware Distribution

- 16279 samples distributed among six categories
  - ~50% fall under code obfuscation
  - ~23% in encryption
Usage Scenarios

- Designing robust malware detection and classification system
- Studying the efficacy of existing analysis systems
- Temporal/longitudinal study of an analysis system
Conclusion

16279 obfuscated samples in six categories

14579 familial obfuscated samples distributed among 158 families

Tagged with time

Spanning over three years from 2018 to 2020
Thank You