ABSTRACT

- Machine learning methods proposed in previous work typically reported high detection performance and fast prediction times on fixed and defective datasets.
- Based on some shortcomings, most of the datasets are not suitable for real-world deployment.
- Propose a systematic approach to generate Android malware datasets using real smartphones instead of emulators.
- Develop a new dataset namely CICAndMal2017, which covers all the shortcomings and limitations of previous datasets.
- Offer 80 network traffic features to select the best features set.
- Showed the average precision 85% and recall 88% for three classifiers namely Random Forest (RF), K-Nearest Neighbor (KNN), and Decision Tree (DT).

Previous Available Datasets

<table>
<thead>
<tr>
<th>Year</th>
<th>Dataset Title</th>
<th>Type</th>
<th>Captured Behavioral Features</th>
<th>Number of Samples</th>
<th>Shortcomings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>GENOME Project Static</td>
<td>Studied components of the malicious source code, tracked API calls and studied permission lists</td>
<td>1260 malware</td>
<td>Lack of dynamic features, Installation</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td>DREBIN Static</td>
<td>Studied malicious source code and manifest file features such as permission lists and API calls</td>
<td>5560 malware - 123,453 benign</td>
<td>Lack of dynamic features, Installation</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>AMD Static</td>
<td>Studied malicious components of code</td>
<td>405 malware</td>
<td>Static analysis</td>
<td></td>
</tr>
</tbody>
</table>

Our proposed Dataset CICAndMal2017 Static & Dynamic is completely labelled and includes network traffic, logs, API/SYS calls, phone statistics, and memory dumps of 42 malware families. Installed 429 malware - 5,065 benign. Address previous Shortcomings

User-Interaction Scenarios

- **Benign**
  - Send Message
  - Make Call
  - Enable GPS
  - Browse Internet

- **Adware**
  - Send Message
  - Make Call
  - Enable GPS
  - Browse Internet

- **Scareware**
  - Send Message
  - Make Call
  - Enable GPS
  - Browse Internet

- **Ransom**
  - Send Message and SMS
  - Make Call
  - Enable GPS
  - Browse Internet

- **SMS malware**
  - Send Message and SMS
  - Make Call
  - Enable GPS

The Network Architecture

Taxonomy of Malware Behaviors

- 20 types of attacks (A1-A20) and 4 types of C&C communications (C1-C4)

Captured & Monitored Data Sources

- Network Traffic
- Memory Dump
- Logs
- Permission
- API Calls
- Phone statistics

Conclusion and Future Works

- Reviewed serious drawbacks of available previous datasets.
- Show actual malicious behavior by installing on real devices.
- Importance of User-interaction scenarios for malware activation.
- Using real smartphones instead of emulators.
- Design different activation scenarios to trigger different families.
- Focused on the network traffic.
- Extract more than 80 network traffic features.
- **Future work:** Extract the useful features from other data.