Machine Learning Fingerprinting Methods in Cyber Security Domain: Which one to Use?

DACS Workshop
26 June 2018

Martin Laštovička
lastovicka@ics.muni.cz
Motivation

- Obsolete fingerprint databases
  - Machine learning
- Results from small static networks
  - Dynamic wireless network
- Focus on accuracy
  - Classification time & memory in a large network
OS Fingerprinting Methodology
TCP/IP Parameters

✓
- Time To Live
- TCP Window Size
- TCP SYN Size

✗
- Checksum
- Destination Port
- Maximum Segment Size
<table>
<thead>
<tr>
<th>synSize</th>
<th>winSize</th>
<th>TTL</th>
<th>OS</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>8192</td>
<td>128</td>
<td>Windows 10.0</td>
</tr>
<tr>
<td>52</td>
<td>8192</td>
<td>128</td>
<td>Windows 6.1</td>
</tr>
<tr>
<td>52</td>
<td>65535</td>
<td>128</td>
<td>Windows 10.0</td>
</tr>
<tr>
<td>60</td>
<td>65535</td>
<td>64</td>
<td>Android 6.0</td>
</tr>
<tr>
<td>60</td>
<td>14600</td>
<td>64</td>
<td>Android 4.4</td>
</tr>
<tr>
<td>60</td>
<td>29200</td>
<td>64</td>
<td>Ubuntu</td>
</tr>
<tr>
<td>64</td>
<td>65535</td>
<td>64</td>
<td>Mac OS X 10.12</td>
</tr>
<tr>
<td>64</td>
<td>65535</td>
<td>64</td>
<td>iOS 10.3</td>
</tr>
</tbody>
</table>
Dataset

- One week of flows + logs
  - 79 087 345 flows
  - 21 746 users
  - 25 642 unique MAC (1 692 vendor prefixes)
  - 253 374 Wi-Fi sessions
  - 6 104 unique IP addresses
Ground Truth

May 5 06:30:54 krakonos dhcpd: DHCPREQUEST for 147.251.x.x from 98:0c:a5:x:x:x (android-22d1bxxx) via 147.251.x.x
May 5 06:30:54 krakonos dhcpd: DHCPACK on 147.251.x.x to 98:0c:a5:x:x:x (android-22d1bxxx) via 147.251.x.x
May 5 06:31:17 krakonos dhcpd: DHCPREQUEST for 147.251.x.x from 38:a4:ed:x:x:x (Redmi3S-Redmi) via 147.251.x.x
May 5 06:31:17 krakonos dhcpd: DHCPACK on 147.251.x.x to 38:a4:ed:x:x:x (Redmi3S-Redmi) via 147.251.x.x
May 5 06:31:20 krakonos dhcpd: DHCPREQUEST for 147.251.x.x from 9c:6c:15:x:x:x (Windows-Phone) via 147.251.x.x
May 5 06:31:20 krakonos dhcpd: DHCPACK on 147.251.x.x to 9c:6c:15:x:x:x (Windows-Phone) via 147.251.x.x
May 5 06:36:24 krakonos dhcpd: DHCPREQUEST for 147.251.x.x from c0:f2:fb:x:x:x (Barboras-iPhone) via 147.251.x.x
May 5 06:36:24 krakonos dhcpd: DHCPACK on 147.251.x.x to c0:f2:fb:x:x:x (Barboras-iPhone) via 147.251.x.x
Experiment Settings
Scikit-learn Methods

- Decision tree – CART algorithm
- Naïve Bayes
- k-NN
  - $K = 3$
  - Euclidean distance metric
- SVM
  - Penalty $C = 1$
  - Gaussian kernel
Measurement

- 20 x (4 x 10) repetitions for each method
  - 4x training set size (1k – 1M)
  - 10x testing set size (1k – 10M)
Results
Performance Measures

![Graph showing performance measures for Naïve Bayes, Decision tree, SVM, and k-NN. The graph displays accuracy, precision, recall, and f-score.](image-url)
Memory Consumption

- k-NN
- Naïve Bayes
- Decision tree
- SVM

Memory [MB] vs. Number of classified flows
# Time Complexity in Seconds

<table>
<thead>
<tr>
<th>Training Size</th>
<th>10k</th>
<th>1M</th>
<th>10M</th>
<th>10k</th>
<th>1M</th>
<th>10M</th>
<th>10k</th>
<th>1M</th>
<th>10M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision tree</td>
<td>0.0004</td>
<td>0.075</td>
<td>0.810</td>
<td>0.0005</td>
<td>0.083</td>
<td>0.901</td>
<td>0.0005</td>
<td>0.087</td>
<td>0.938</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.001</td>
<td>0.229</td>
<td>4.237</td>
<td>0.001</td>
<td>0.231</td>
<td>4.366</td>
<td>0.001</td>
<td>0.232</td>
<td>4.385</td>
</tr>
<tr>
<td>k-NN</td>
<td>0.298</td>
<td>29.82</td>
<td>294.9</td>
<td>2.519</td>
<td>250.1</td>
<td>2494</td>
<td>42.71</td>
<td>4359</td>
<td>42925</td>
</tr>
<tr>
<td>SVM</td>
<td>0.374</td>
<td>37.44</td>
<td>368.6</td>
<td>3.407</td>
<td>341.8</td>
<td>3378</td>
<td>33.50</td>
<td>3314</td>
<td>34374</td>
</tr>
</tbody>
</table>
Conclusion

- Memory not a problem
- Performance measures are similar
- Time extremely dependent on model complexity
- Decision trees are best suited for OS fingerprinting in large networks
  - MU network up to 10k flow/s (6.1k avg)
Discussion

Martin Laštovička
lastovicka@ics.muni.cz

Brno Ph.D. Talent Scholarship Holder
Funded by the Brno City Municipality