On Mobile Malware Infections

N. Asokan

(joint work with Hien Thi Thu Truong, Eemil Lagerspetz, Petteri Nurmi, Adam J. Oliner, Sasu Tarkoma, Sourav Bhattacharya)
Mobile malware alarm bells

Google Search

- 2006
- 2009
- 2014

- 19 pages
- 25 pages
- 40 pages
Mobile malware alarm bells

Google Trends

[Google Trends chart showing the increase in search interest for mobile malware]
Research focus: analysis of malware

Google Scholar

A survey of mobile malware in the wild

AP Felt, M Finifter, E Chin, S Hanna... - Proceedings of the 1st..., 2011 - dl.acm.org

Abstract Mobile malware is rapidly becoming a serious threat. In this paper, we survey the current state of mobile malware in the wild. We analyze the incentives behind 46 pieces of iOS, Android, and Symbian malware that spread in the wild from 2009 to 2011. We also...

Cited by 222 Related articles All 13 versions Cite Save

About Google Scholar All About Google Privacy & Terms Give us feedback

100 pages
How prevalent is mobile malware?

We make several important observations. The mobile malware found by the research community thus far appears in a minuscule number of devices in the network: 3,492 out of over 380 million (less than 0.0009%) observed during the course of our analysis.
Gather data directly from devices

Accurately estimate malware infection rate

Identify risk factors, cheaply
Outline

- Gather data directly from devices
- Accurately estimate malware infection rate
- Identify risk factors, cheaply
Gather data directly from devices

Piggyback on a popular package

http://carat.cs.berkeley.edu

Need to be lightweight and unobtrusive
Carat (devices by continents)

Android devices: geography distribution, (April 2, 2014)

http://carat.cs.berkeley.edu
What kind of data?

- How to estimate infection rate?
  - Identify a package on device; check for match with known malware

- How to identify an Android package?
Structure of an Android Package

<package, versionCode> tuples (<p,v>) should be unique but not enforced
Structure of an Android Package

APK package

- META-INF
- AndroidManifest.xml
- ...

Packages are (self-)signed by developers. Developer certs (dc) are statistically unique.
Identifying a (malicious) package

• **Coarse-grained:**
  
  Use `<developerCert>` only
  
  • `<dc>` for short
  • upper bound for infections

• **Fine-grained:**
  
  Use `<developerCert, package, versionCode>`
  
  • `<dc, p, v>` for short
  • lower bound for infections
Carat dataset

set of tuples: \(<dc, p, v>\)
\(<\text{developerCert, pkgName, versionCode}>\)
## Carat dataset

*Mar 2013 – May 2014*

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct devices</td>
<td>99,414</td>
</tr>
<tr>
<td>Unique developer certificates &lt;dc&gt;</td>
<td>108,482</td>
</tr>
<tr>
<td>Unique &lt;dc, p, v&gt; tuples</td>
<td>512,342</td>
</tr>
</tbody>
</table>

Data
## Malware datasets

<table>
<thead>
<tr>
<th>Type</th>
<th>Mobile Sandbox</th>
<th>McAfee</th>
<th>Malware Genome</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique devcerts &lt;dc&gt;</td>
<td>3,879</td>
<td>1,456</td>
<td>136</td>
<td>4,809</td>
</tr>
<tr>
<td>Unique packages &lt;dc, p, v&gt;</td>
<td>16,743</td>
<td>3,182</td>
<td>1039</td>
<td>19,094</td>
</tr>
<tr>
<td>Unique package (.apk) files</td>
<td>96,500</td>
<td>5,935</td>
<td>1260</td>
<td>103,695</td>
</tr>
</tbody>
</table>

http://mobilesandbox.org/
http://mcafee.com
http://www.malgenomeproject.org/
Outline

- Gather data directly from devices
- Accurately estimate malware infection rate
- Identify risk factors, cheaply
Carat dataset: identifying infection

Device 1: “infected”
Device 2: “clean”

Time:

Malware dataset:

Estimates:

match found

no match found
## Incidence of infection
### Mar 2013 – May 2014

<table>
<thead>
<tr>
<th># Infected Devices</th>
<th>Mobile Sandbox</th>
<th>McAfee</th>
<th>Union</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>coarse-grained:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dc match</td>
<td>37,355 (38%)</td>
<td>32,323 (33%)</td>
<td>40,334 (40%)</td>
</tr>
<tr>
<td><strong>fine-grained:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;dc,p,v&gt; match</td>
<td>263 (0.26%)</td>
<td>255 (0.26%)</td>
<td>477 (0.48%)</td>
</tr>
</tbody>
</table>

Data collected from 99414 devices over one year
Coarse- vs. fine-grained

Discrepancy is several orders of magnitude

Mobile Sandbox

McAfee

Coarse-grained: <dc> matching
Fine-grained: <dc,p,v> matching
Re-use of signing keys

Widespread (ab)use of test keys: 544 malwares, 1948 innocuous packages signed with Android Open Source Project (AOSP) test key

Same key signing malware and non-malware: Brightest Flashlight Free v17 is malware, other versions are not.

Use fine-grained (<dc,p,v>) matching from now on

1. https://androidobservatory.org/cert/61ED377E85D386A8DFEE6B864BD85B0BFAA5AF81
Rarity of signing keys

McAfee dataset

F-Secure dataset

malware <10 keys
Rarity of signing keys: Facebook

com.facebook.katana

key 92797
key 29180
key 46574
key 778
key 81839
key 27467
key 83479
key 76751

Ratio

Months

566 568 570 572 574 576 578 580
Rarity of signing keys: Facebook

com.facebook.katana

Ratio

Months

566 568 570 572 574 576 578 580

key 92797
key 29180
key 46574
key 778
key 81839
key 27467
key 83479
key 76751

minor keys
Example: package with 2 keys

```
com.sony.smallapp.app.widget
```

On-going work: can we use key rarity to identify malware?
Malware datasets revisited

Estimates

- Mobile Sandbox: 15263
- McAfee: 2239
- Genome: 674
- Other: 597

Total: 173

27
### What is malware?

<table>
<thead>
<tr>
<th>Package name</th>
<th>No. Infected devices</th>
<th>Flagged by</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>it.evilsocket.dsploit</td>
<td>23</td>
<td>22</td>
<td>Monitoring</td>
<td>MC</td>
</tr>
<tr>
<td>com.noshufou.android.su</td>
<td>37</td>
<td>17</td>
<td>Rooting</td>
<td>MC</td>
</tr>
<tr>
<td>ty.com.android.SmsService</td>
<td>25</td>
<td>29</td>
<td>Trojan</td>
<td>MB</td>
</tr>
<tr>
<td>com.mixzing.basic</td>
<td>17</td>
<td>19</td>
<td>Adware</td>
<td>MC</td>
</tr>
<tr>
<td>pl.thalion.mobile.battery</td>
<td>10</td>
<td>12</td>
<td>Adware</td>
<td>MC</td>
</tr>
<tr>
<td>com.bslapps1.gbc</td>
<td>21</td>
<td>17</td>
<td>Adware</td>
<td>MC</td>
</tr>
<tr>
<td>com.android.antidroidtheft</td>
<td>16</td>
<td>17</td>
<td>Monitoring</td>
<td>MB</td>
</tr>
<tr>
<td>com.androidlab.gpsfix</td>
<td>7</td>
<td>9</td>
<td>Adware</td>
<td>MC</td>
</tr>
<tr>
<td>com.adhapps.QesasElanbiaa</td>
<td>7</td>
<td>18</td>
<td>Adware</td>
<td>MC</td>
</tr>
<tr>
<td>download.youtube.downloader.pro7</td>
<td>5</td>
<td>29</td>
<td>Adware</td>
<td>MB</td>
</tr>
<tr>
<td>com.android.settings.mt</td>
<td>5</td>
<td>12</td>
<td>Monitoring</td>
<td>MC</td>
</tr>
</tbody>
</table>

**Reasons for classification as “malware”**

- MC: McAfee
- MB: Mobile Sandbox

**Estimates**

- Mar 2013 – May 2014
- Number of AV tools flagging this package as malware (Total ~50 AV tools)

**Treat each dataset separately**
What is malware?

Curiously, AV vendors do take labeling by other vendors into account!

- Sometimes leads to false positives propagating
- ... and staying undetected!
Propagation of False Positives

"Google Security Tool", signed by test key

Google Security Tool, signed by legitimate Google key


Deployment of AV tools

Mar 2013 – May 2014

Anti-malware/virus tools

# devices
AV tools vs. infection  Mar 2013 – May 2014

25215 devices have some AV tool installed (25.3%)

None are infected according to any of our malware datasets
Information revealed by set of apps

Package names can be revealing:
language of device user

Can also reveal user traits:

Predicting User Traits From a Snapshot of Apps Installed on a Smartphone

Suranga Seneviratne\textsuperscript{a,b}, Aruna Seneviratne\textsuperscript{a,b}
suranga.seneviratne@nicta.com.au, aruna.seneviratne@nicta.com.au
Prasant Mohapatra\textsuperscript{c}
prasant@cs.ucdavis.edu
Anirban Mahanti\textsuperscript{d}
anirban.mahanti@nicta.com.au

\textsuperscript{a}School of EET, University of New South Wales, Australia
\textsuperscript{b}NICTA, Australia
\textsuperscript{c}Department of Computer Science, University of California, Davis

http://dx.doi.org/10.1145/2636242.2636244

Indicative of user behaviour?
Summary: infection rate estimates

Our measurement

Mar 2013 – May 2014

<table>
<thead>
<tr>
<th>Product</th>
<th>Infection Rate Estimate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeorgiaTech</td>
<td>0.0009%</td>
</tr>
<tr>
<td>Google</td>
<td>0.12%</td>
</tr>
<tr>
<td>McAfee</td>
<td>0.26%</td>
</tr>
<tr>
<td>MobileSandbox</td>
<td>0.26%</td>
</tr>
<tr>
<td>Lookout</td>
<td>2.6%</td>
</tr>
<tr>
<td>NQ Mobile</td>
<td>4.3%</td>
</tr>
</tbody>
</table>

Our measurement
Outline

Gather data directly from devices

Accurately estimate malware infection rate

Identify risk factors, cheaply

Separately for each malware dataset

See if we can detect susceptibility for infection!
“The Company You Keep”

Can the list of apps used on device detect susceptibility for infection?

Device 1  “infected”

Device 2  “clean”...
Classifying based on set of apps

- Identifying new malware requires extensive analysis of candidates
- Baseline: random sampling
  - Low infection rates imply that baseline is costly
- Using set of apps to detect susceptibility for infection is cheap
  - Lightweight instrumentation: at virtually no cost

Application: Help anti-malware vendors in the search for new malware
Classifying based on set of apps

Mar 2013 – May 2014

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Precision</th>
<th>Baseline</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detecting infection (new malware)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McAfee</td>
<td>1.2%</td>
<td>0.26%</td>
<td>4.5X</td>
</tr>
<tr>
<td>Mobile Sandbox</td>
<td>0.9%</td>
<td>0.25%</td>
<td>3.5X</td>
</tr>
<tr>
<td>Detecting infection (undetected malware)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>McAfee</td>
<td>0.16%</td>
<td>0.05%</td>
<td>3.5X</td>
</tr>
<tr>
<td>Mobile Sandbox</td>
<td>0.12%</td>
<td>0.05%</td>
<td>2.6X</td>
</tr>
</tbody>
</table>
Detecting infection: the “Real-life” case

“Original” malware set used for training;

*Training set* labeled using “Original” malware set only

“New” set used for testing

See how well we can detect infection by “New” malware set

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Precision</th>
<th>Baseline</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>McAfee</td>
<td>0.7%</td>
<td>0.19%</td>
<td>3.5X</td>
</tr>
<tr>
<td>Mobile Sandbox</td>
<td>0.3%</td>
<td>0.08%</td>
<td>4X</td>
</tr>
</tbody>
</table>

*Mar 2013 – May 2014*

Multinomial Naïve Bayes
Taking timestamps into account

Mar 2013 – Oct 2013

- Carat records have timestamps
  - At least 155 devices changed state from clean to infected during data collection period
  - can we predict likelihood of eventual infection?
Identify vulnerable devices before they are infected?

Application: Help enterprise IT admin identify users for training
Summary

• Measure Android malware infection rates directly
  • No common agreement of what is malware
  • False positives and re-classifications are common

• Identify inexpensive risk factors
  • can aid in search for new malware
  • set of apps indicative of user behaviour/traits’

http://se-sy.org/projects/malware/
Detecting infection (new)

Risk Factors

malware: “Old” (80%) “New” (20%)

devices: Clean (80%) Clean (20%) inf (old) inf (new)

Training set Test set
Detecting infection (unknown)

malware:

“Known” (80%)  “unknown” (20%)

devices:

Training set  Test set

label

flip