Identifying Unknown Android Malware with Feature Extractions and Classification Techniques

Ludovic Apvrille
ludovic.apvrille@telecom-paristech.fr

Axelle Apvrille
aapvrille@fortinet.com

Seminar at SAP
Many Android Applications (and Malware!)

Application repositories

- Google Play: 1.7 million+
- F-Droid, APPSAPK, APKTOP, …
Many Android Applications (and Malware!)

Application repositories

- Google Play: 1.7 million+
- F-Droid, APPSAPK, APKTOP, ...

Malware

- Aug. 2015. 2.7 millions+ malicious Android samples
- 2,000+ new malicious Android samples every day
Known Malware

I'm a malware
E.g. I pose as a Skype installer

We are known malware
And it's usually easy to detect us

I'm a minor variant
"Trust me, I install Opera Mini"

I'm another minor variant
E.g. I'm just repackaged (like my new dress?)

I'm a malware
E.g. I pose as a Skype installer

We are known malware
And it's usually easy to detect us

I'm a minor variant
"Trust me, I install Opera Mini"

I'm another minor variant
E.g. I'm just repackaged (like my new dress?)
Known Malware

I'm a COMPLICATED variant

- defuscated code
- encrypted strings

I'm still KNOWN malware

- native code hidden in the assets
- embedded root exploit

AV catches us usually

Sometimes, the "signature" needs to be improved

$ = detection instructions
Unknown Malware

Do they exist? YES
Malware: Android Carbon 14 Dating ;)

Shortest detection delay for some samples by all AV vendors

<table>
<thead>
<tr>
<th>Name</th>
<th>Creation date</th>
<th>Detection date</th>
<th>Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android/Wroba</td>
<td>June 16 2014</td>
<td>June 21</td>
<td>+5d</td>
</tr>
<tr>
<td>Android/Curesec</td>
<td>July 3 2014</td>
<td>July 11</td>
<td>+8d</td>
</tr>
<tr>
<td>Android/ScarePakage</td>
<td>July 13 2014</td>
<td>July 24</td>
<td>+11d</td>
</tr>
</tbody>
</table>
Malware: Android Carbon 14 Dating ;)

Shortest detection delay for some samples by all AV vendors

<table>
<thead>
<tr>
<th>Name</th>
<th>Creation date</th>
<th>Detection date</th>
<th>Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Android/Wroba</td>
<td>June 16 2014</td>
<td>June 21</td>
<td>+5d</td>
</tr>
<tr>
<td>Android/Curesec</td>
<td>July 3 2014</td>
<td>July 11</td>
<td>+8d</td>
</tr>
<tr>
<td>Android/ScarePakage</td>
<td>July 13 2014</td>
<td>July 24</td>
<td>+11d</td>
</tr>
</tbody>
</table>
| Android/Ganlet     | Nov 1 2013    | May 15 2014    | +6 months!!!
What Are We Interested In?

- **KNOWN (and detected)**
  - MALWARE → **NOTHING TO DO (BORING...)**

- **MINOR VARIANT (usually detected)**
  - → **NOT VERY INTERESTING FOR RESEARCH**

- **MAJOR VARIANT**
  - detected → **AH HA?!**
  - not detected

- **UNKNOWN MALWARE** → **TOP INTEREST**
  - RESEARCH/PR INNOVATION
Problems with Manual Search

Too many apps and marketplaces to crawl
Waste time on clean apps
Even a team of 100 analysts is insufficient
Problems with Manual Search

Too many apps and marketplaces to crawl
Waste time on clean apps
Even a team of 100 analysts is insufficient

We need an automated system that helps identifying unknown malware with less effort

→ SherlockDroid
SherlockDroid to the Rescue!

Crawl Android marketplaces

Spot suspicious apps

Focus on major variants and unknown malware
SherlockDroid Architecture
SherlockDroid (Unbiaised) Benefits

WE WASTE OUR PRECIOUS TIME ON CLEAN SAMPLES (and usually don’t have time to find nasty samples)

WE DON’T HAVE TIME TO ANALYZE MANUALLY MORE THAN THIS

HIGHER CHANCES TO SPOT INTERESTING MALWARE (it can’t be perfect, though)
Remarks on SherlockDroid

It is not an AV scanner because SherlockDroid does not handle known malware / minor variants
Remarks on SherlockDroid

It is not an AV scanner because SherlockDroid does not handle known malware / minor variants

We will miss some malware
We’re not (yet) perfect ;-)
Remarks on SherlockDroid

It is not an AV scanner
because SherlockDroid does not handle known malware / minor variants

We will miss some malware
We’re not (yet) perfect ;-)  
But we would have missed them without SherlockDroid too
Crawlers - Evading Detection

Easy to implement
but constantly needs to be maintained :( 

- Search Limit
- Download activity per IP address
- User Agent verification
- Android ID verification [https://github.com/Akdeniz/google-play-crawler](https://github.com/Akdeniz/google-play-crawler)

- Your IP: [hidden]
- URL: [www.appsapk.com/android/all-apps](http://www.appsapk.com/android/all-apps)
- Your Browser: libwww-perl/6.03
- Block ID: BNP002
- Block reason: Scanning tool access attempt.
- Time: Fri, 20 Jun 2014 05:30:21 -0400
- Server ID: cp76
### DroidLysis - Extracting Properties/Features

**Static extraction of 289 properties**

#### 54 File-related properties
- Permissions, certificate, ...

#### 22 Resource properties
- Native code, resource risky calls (*su, mount*, etc.), Javascript, URLs, ...

#### 70 Dalvik code properties
- API usage, actions, intents, constants, implementation techniques (e.g., *junk bytecode injection*)

#### 143 Third party kits properties
- Advertisements, statistics reporting, error reporting
Ruling out Third Party Code

**DIFFICULT TO SPOT THE DIFF**

- **GOOD**
  - The actkit requests `INTERNET & GPS` that's frequent (whether you like it or not)
- **SUSPICIOUS**
  - The payload app requests `INTERNET & GPS` why ???

**DISCOUNT TRIP TO LUXEMBOURG**
Alligator - Classification (1/2)

Usual classification approach

Data to be classified

Already classified data

1
2

Classifier (SVM or k-NN or ...)

1
2

d

Classified data

Alligator

Data to be classified

Classifier (SVM and k-NN and ...)

1
2

d
d
d
d

Classified data

Alligator automatically combines classification algorithms in order to obtain better classification results
**Alligator - Classification (2/2)**

### Other capabilities

- Favor a cluster over another
- Forget/boost too abnormal elements

### Other

- Shown to better classify than other classifiers (e.g., SVM) in various application domains (e.g., image classification)
- Free and open-source, easy to install and configure, scriptable

[alligator.telecom-paristech.fr]
SherlockDroid: Hall of “Fame”

- Android/MisoSMS.A!tr.spy
- Android/Odpa.A!tr.spy
- Adware/Geyser!Android
- Riskware/Flexion!Android
- Riskware/SmsControlSpy!Android
- Riskware/Zdchial!Android
- Riskware/SmsCred!Android
- Riskware/Blued!Android
- Riskware/SneakFont!Android

SherlockDroid: Unknown Malware Identified

Do you known any other framework who identified real unknown malware?
Do you known any other framework who identified real unknown malware?

**Answer:** DroidRanger: 2
SherlockDroid: Unknown Malware Identified

Do you known any other framework who identified real unknown malware?

Answer: DroidRanger: 2

AAS, Andromaly, CopperDroid, Crowdroid, Drebin, MADAM, MAST, pBMDS, PUMA... tested on artificial or known malware
Into Android/MisoSms Trojan Spyware

Android/MisoSms.A!tr.spy

- Poses as Google Settings app
- Sends 1 initial email with phone number of victim
- Listens to incoming SMS
- Forwards them by email to attackers
Into Geyser Adware

Adware/Geyser!Android

Posts GPS location in clear text


LOL - In falsepositives.txt:

"Reputable companies including banks, US Government/ Military sector are using our tools"
Learning and Classification Results

Typical results we expect

- FP/FN shall be as low as possible (Obviously)
- FP shall be much lower than FN (Missing a malware is not a big deal w.r.t. wasting time on false alerts)

Samples

- **Learning clusters**: 500k samples used in the learning clusters
  - ~487k malware, ~12k clean
  - Gathered before June 2014
- **Testing clusters**: 1.5k clean and 3k malware gathered after Sept. 2014
### Learning and Classification Results (Cont.)

<table>
<thead>
<tr>
<th>Learning cluster size</th>
<th>Learning time</th>
<th>Classification time</th>
<th>FP</th>
<th>FN</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>480,000</td>
<td>~ 11 h</td>
<td>6 mn</td>
<td>1.78%</td>
<td>0.52%</td>
<td>0.93%</td>
</tr>
<tr>
<td>50,000</td>
<td>20 mn</td>
<td>~ 34 s</td>
<td>0.72%</td>
<td>2.1%</td>
<td>1.67%</td>
</tr>
</tbody>
</table>
Learning and Classification Results: Comparison

- Alligator learning time
- Alligator classification time
- SVM learning time
- SVM Classification time

Number of samples in the learning clusters vs. Learning time (in seconds) and Classification time (in seconds)

Regular cluster size vs. Malware cluster size
Learning and Classification Results: Comparison (Cont.)

Alligator vs. SVM - False Positive and False Negative rates depending on learning cluster size

- Alligator FP
- Alligator FN
- Alligator FP+FN
- SVM FP
- SVM FN
- SVM FP+FN

24/28 Nov., 2015

Institut Mines-Telecom
Identifying Unknown Android Malware
What About the SAP Android App?

- Property extraction: 17 seconds
- Classification: 4 seconds
What About the SAP Android App? (Cont.)

→ Classified as regular ;-)

Filename: b1.mobile.android.apk_38ccc092.apk
SHA256: 2890685023c151dc35ee98d7cd4dd93772531e69a38bad0a505d28f98dda971f

This sample looks clean
Analysis time: 318 seconds

Sample suspiciousness

Malware: 27
Clean: 77

→ Classified as regular ;-)

26/28 Nov., 2015
Institut Mines-Telecom
Identifying Unknown Android Malware
Conclusion and Future Work

- SherlockDroid is operational, tested on a huge number of applications from various application markets
- 9 unknown malware identified (Actually: 10!)
- For classification purpose, it relies on the Alligator meta-classifier

What’s next?

- Feature extraction: mix contextual information (e.g., call stack) and the related features
  - Sending an email is not the same if it is for a bug report or for a connection to a Command&Control server
- Differentiate between malware and Potentially Unwanted Applications
Thank You

Contact info

SherlockDroid: aapvrille at fortinet dot com
Alligator: ludovic dot apvrille at telecom minus paristech dot com

References

Alligator Release: alligator.telecom-paristech.fr


Powerpoint slides? No way! This is \texttt{\LaTeX} - Beamer!