Bug Fixes, Improvements,... and Privacy Leaks: A Longitudinal Study of PII Leaks Across Android App Versions

Jingjing Ren*, Martina Lindorfer†, Daniel J. Dubois*, Ashwin Rao‡, David Choffnes* And Narseo Vallina-Rodriguez§

*Northeastern University †UC Santa Barbara  
‡University Of Helsinki §IMDEA Networks Institute And ICSI

Sponsored by:
Outline

Motivation
Threat Model
Methodology
Macroscopic Trends in Privacy
Conclusion
Are there any changes in privacy when I update the app?
The Evolution of Privacy in Mobile Devices

How does mobile privacy evolve over time? (😊 or 😞)
Outline

Motivation

Threat Model
  ◦ Privacy definition
  ◦ Leak definition

Methodology

Macroscopic Trends in Privacy

Conclusion
What Do I Mean by “Privacy” in This Work?

What information is shared?  

- Personally Identifiable Information (PII)
  - Tracking ID
  - User information
  - Location
  - Contact
  …

How is it being shared?  

- Transport security
  - Encrypted (HTTPS)
  - or
  - Plaintext (HTTP)

Where is it going?  

- Destination:
  - First party (app owner)
  - or
  - Third party (advertising & analytics)
Threat Model

Eavesdropping

First/Third Party

Data aggregation
What is a “Leak” in This Work?

Goal: Understand user information shared with other parties

- A leak may or may not be a violation

PII

Third party

First party

Not a Leak

HTTP

HTTPS

HTTP

HTTPS

Username
Password
Email Address
Location
Birthday
Gender

Name
Phone Number
Unique IDs
- Advertising IDs
- IMEI
- MAC Address

Username
Password
Email Address
Outline

Motivation
Threat Model

**Methodology**
- Controlled Experiments
- Detecting PIILeaks
- Privacy Attributes

Macroscopic Trends in Privacy
Conclusion
Controlled Experiments

App selection criteria
- Multiple versions [1]
- Popularity
- Amenable to traffic analysis
  - MITM TLS connections

Privacy measurements
1. Interact with apps
2. Detect privacy leaks
3. Validate manually

512 Android apps
7,665 versions (APKs)
8 years

Interaction

Inducing privacy leaks requires interaction
- Real, controlled user interactions are good, … but not scalable 😞

Automated and scripted interaction
- Monkey: randomly generated events with good coverage
- login and replay across the versions
- ~10 minutes per experiment

Test environment
- 5 Android phones
- MITM proxy to intercept both plain-text and encrypted network traffic
Detecting PII Leaks

Network traffic analysis

- PII leaks, by definition, leak over Internet
- ReCon: using ML to detect without prior knowledge of PII values [Mobisys’16]
- Manual validation
Privacy Attributes

PII Types
- Password
- Gender
- Location
- Android ID
- GSF ID
- Ad ID

1st Party
- Pinterest.com

3rd Party
- Tracking.com

HTTP
- 203
- 271
- 301
- 402
- 502
- 602
- 603
- 604
- 606

HTTPS
- 13
Case study: Pinterest

Takeaway: High variance in privacy risks across versions
- password leaks, PII types, frequency, encryption

Sending password to a third party in 2 out of 12 versions

Increased frequency for Android ID

More types (gender, location, android ID etc.) are leaked

HTTPS is used only sometimes

Jan. 5, 2017: disclosed
Feb. 7, 2017: fixed
Privacy Leaks For Individual Apps

https://recon.meddle.mobi/appversions/
Outline

Motivation
Threat Model
Methodology
**Macroscopic Trends in Privacy**
Conclusion
Macroscopic Trends in Privacy

Summary of Results

Variations in PII Leaks
HTTPS Adoption Trends
Third-Party Characterization
Multidimensional analysis
Summary of Results

Not all versions leak PII

Tracking ID and location leaks dominate

6 apps (48 APKs): leaked in plaintext
7 apps (36 APKs): leaked to third party
Macroscopic Trends in Privacy

Summary of Results

Variations in PII Leaks
  ◦ How different are leaking frequencies

HTTPS Adoption Trends

Third-Party Characterization

Multidimensional analysis
5.6% apps see a several orders of magnitude difference
- fine-grained location tracking
- increased opportunities for network eavesdroppers to invade user privacy
Macroscopic Trends in Privacy

Summary of Results

Variations in PIILeaks

HTTPS Adoption Trends

- Extremely slow, for half of the domains:
  - 10% apps, 2 years
  - 50% apps, 5 years

Third-Party Characterization

Multidimensional analysis
Macroscopic Trends in Privacy

Summary of Results

Variations in PII Leaks

HTTPS Adoption Trends

Third-Party Characterization

Multidimensional analysis
High-risk Tracking Across Apps

Tracking ID: IMEI, Android ID, advertising ID, MAC address etc.
Other PII: location, gender, name, email

Might permanently link individuals/personal information to a tracking ID

>100 domains

<table>
<thead>
<tr>
<th>Third Parties</th>
<th>(Tracking ID + ) Other PII</th>
<th>#Apps</th>
<th>#APKs</th>
</tr>
</thead>
<tbody>
<tr>
<td>google[*]</td>
<td>Location, Gender, First/Last Name, Email</td>
<td>124</td>
<td>387</td>
</tr>
<tr>
<td>kochava.com</td>
<td>Email, Gender</td>
<td>8</td>
<td>36</td>
</tr>
<tr>
<td>vungle.com</td>
<td>Location, Gender</td>
<td>7</td>
<td>34</td>
</tr>
<tr>
<td>mopub.com</td>
<td>Location</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56txs4.com</td>
<td>Gender</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Macroscopic Trends in Privacy

Summary of Results
Variations in PII Leaks
HTTPS Adoption Trends
Third-Party Characterization

Multidimensional analysis
Combined privacy worsens over time

- mainly due to more PII types and more domains
Outline

Motivation
Threat Model
Methodology
Macroscopic Trends in Privacy

Conclusion
Conclusion

• Privacy has *worsened* over time
  • PII can *change* substantially across versions
  • HTTPS Adoption is *slow*
  • Third-party tracking is *pervasive and broad*

• Need for *continuous* monitoring
  • Using systems: ReCon, Lumen, AntMonitor etc.

Disclaimer: we recommend updating apps for security reasons
SHOULD YOU UPDATE YOUR APP?

What's this? | Back to ReCan | Learn more details

Lifestyle

Privacy Score | PrivacyLeaks | Sites Contacted

Privacy Risk Score
(Version Number)
Privacy Risk Trend

PRI

VAC

OVER TIME

Preferences | Settings

Which PII do you care about the most? (Toggle slider higher for more importance)

PASSWORD | GENDER | LOCATION | TRACKING ID

Variance 0.45 | Threshold 1.5 | Slope 0

Submit

https://recon.meddle.mobi/appversions/