Mobile and Sensor Systems

Lecture 7: Mobile Privacy

Prof Cecilia Mascolo
Mobile Privacy

• In this lecture we will discuss some issues and solutions related to mobile systems and data privacy
Mobile Phone Data

• Mobile phones generate user data which can reveal a lot about the user
• Where does this data go?

• Initially collected by a few “trusted” companies
• More and more smaller companies have built their business model around data
App Data Gathering

• More often than not apps collect sensing data beyond its technical needs (for advertising purposes)
Where does the data go?

- The app collects location. Does it then go to:
  - Location services
  - Advertisers
  - Developers

- The OS does not offer support to know this…
Permissions: Android

- **Storage**: Modify or delete the contents of your USB storage
- **Phone calls**: Read phone status and identity
- **Network communication**: Full network access
- **Network communication**: Receive data from Internet, view Wi-Fi connections, view network connections
- **Your accounts**: Find accounts on the device
- **Affects Battery**: Prevent phone from sleeping
Permissions: iOS

• Apple doing the vetting
• Now notifying users about location data usage
Privacy Data Breach Detection

• Monitor behavior of apps to determine when privacy sensitive information leaves the phone.

• Data Flow Analysis (DFA): looks for routes between data sources and sinks
  – Any of these routes without user consent is classified as leak

• Capability Leak:
  – Explicit: malicious app manages to hijack permissions granted to other trusted apps
  – eg apps can have sharedUserID: apps by same developer have same ID so permissions for that ID are shared
## Sources and Sinks

### Sources

<table>
<thead>
<tr>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location Data: GPS, last base station location, WLAN</td>
</tr>
<tr>
<td>Unique, Identifiers: IMEI, IMSI</td>
</tr>
<tr>
<td>Authentication, Data: Cashed password data</td>
</tr>
<tr>
<td>Contact and Calendar, Contacts, address and schedule</td>
</tr>
<tr>
<td>Call State, Start and end of incoming call, number of incoming call</td>
</tr>
</tbody>
</table>

### Sinks

<table>
<thead>
<tr>
<th>Sinks</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMS, Communication: data can be transferred by SMS</td>
</tr>
<tr>
<td>File Output: Applications can write data to files that are globally readable</td>
</tr>
<tr>
<td>Network: Applications can access network by sockets or HTTP</td>
</tr>
<tr>
<td>Intents, objects: applications can send data objects to other apps</td>
</tr>
<tr>
<td>Content, Resolver Apps can use API to edit shared memory of device</td>
</tr>
</tbody>
</table>
Dynamic Analysis: TaintDroid

- **Challenges** …
  - Smartphones are resource constrained
  - Third-party applications are entrusted with several types of privacy sensitive information
  - Context-based privacy information is dynamic
  - Applications can share information
  - TaintDroid is a modification to the Android OS which allows for dynamic tracking of sensitive data movements from an app to other apps and sinks
TaintDroid

• TaintDroid automatically labels data from privacy-sensitive sources and transitively applies labels as sensitive data propagates through program variables, files, and interprocess messages.

• When tainted data are transmitted over the network, or otherwise leave the system, TaintDroid logs the data labels, the application responsible for transmitting the data, and the data destination.
Static Analysis

• Static Analysis covers all the paths from sources to sinks
  – Takes much longer than dynamic analysis
  – No time or efficiency overhead (done offline)
  – Example (LeakMiner):
Cellular Network Leaks

- Various parts of a cellular network are prone to leaks and attacks...
Cellular Network Leaks

• Mobile devices roam around and register with BTSs: they can be identified from these records and pre-existing location profiles. One paper identifies 80% of users.

• the network interface itself can allow listening attacks/leaks (GSM).
Performance

Table 5: Performance: all four identification periods

<table>
<thead>
<tr>
<th></th>
<th>Dec 2004</th>
<th>Jan 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>one hour</td>
<td>40.91%</td>
<td>48.98%</td>
</tr>
<tr>
<td>one day</td>
<td>65.15%</td>
<td>65.31%</td>
</tr>
<tr>
<td>one week</td>
<td>74.24%</td>
<td>75.51%</td>
</tr>
<tr>
<td>one month</td>
<td>77.27%</td>
<td>87.76%</td>
</tr>
</tbody>
</table>

WiFi based Leaks

• WLAN fingerprints can be used to infer social relation between the users.
• Mobile devices broadcast their Wi-Fi information that contain device ID or MAC address, it is possible for adversaries to track locations of devices.
• Users’ names can also be detected by analyzing applications, websites and ad content in traffic data through WiFi hot-spots
WiFi based Leaks (2)

• By having WiFi connection info, applications can get the MAC address of devices, which is a unique ID. Since this ID is permanent, it allows third parties to track users.

• Applications can also learn about the last scanned list of WiFi hot spots. This includes MAC address, name, signal strength, operating channels and so on. This information can later be used to geo-locate user positions.

• It is also possible for applications to determine configured network lists on devices. Moreover, by comparing these lists, social relationship between individuals can also be inferred, such as professional, family, interest groups and the like.
Mobile Sensing Leaks

• Sensors are powerful (eg mood automatic recognition)
• Accelerometer is different on all devices (so it can be used as a “device signature”).
• Or it can be used to identify different users of same device
• Touch sensor usage behaviour, keying behaviour can give away user identity
Mobile Sensing De-anonymization

• Even in large anonymized datasets there is a risk that a user can be identified.
  – Example of the Netflix challenge
  – Mobile sensing datasets exhibit data “sparsity”

• With some user information it is possible to single out the user in a dataset that just contains activity profiles of users.
De-anonymization

• Netflix type datasets contain narrow range of behaviour (preferences of movies)
• Sensing data contain a wider range:
• Correlation of activities can be a “key”
  – Eg you go to the gym at a certain time after a train ride
  – This increases data sparsity
  – Broad range of auxiliary information which can be used by an adversary
Example

• Auxiliary information of the adversary is a collection of activities by a user.

• Aim: Identify which of the anonymized activity streams belongs to the target user.

• Auxiliary information may be collected by observing the user or from available public sources.
Location Data Privacy
Location Data Leaks

• Location aware apps collect data which is (sometimes) provided to third parties
  – Advertisement?

• Location data can be used to infer more personal information (or user identity from anonymous datasets)

• Profile of users can be built from solely location tracks
Examples

• GPS of 172 users find their home location with median error of 60m.
  – Features (last destination of day, long stay, time..)
• Similar features have been used to detect “significant places” for a user (work, gym etc)
• Speed of travelling/transport modality
• Data can be cross correlated with social network data for more deanonymization
And more

- If the approximate locations of a user’s home and workplace can be deduced from a location trace, then the median size of the individual’s anonymity set in the U.S. working population is 1, 21 and 34,980, for locations known at the granularity of a census block, census tract and county respectively.

![Diagram showing size of anonymity set under disclosure of work location (red circles), home location (green squares) or both (black triangles). Location granularity is either census tract (left graph) or county (right graph). Note the different scales on the Y-axes.](image)
Mobility Prediction

• Given user patterns are highly regular, a user can not only be identified in a dataset but his previous patterns can be used to predict her future movements

• In addition her friends patterns if known can considerably aid this prediction
Different Types of Prediction

• Location can be:
  – Logical location (workplace, home)
  – Geographic location
    • Discrete areas (e.g., square in a grid)
    • GPS locations

• Not only *where* but also *when*
  – Spatio-temporal prediction is hard
Transition Matrix

\[
T = \begin{pmatrix}
0.1 & 0.3 & 0.2 & 0.4 \\
0.4 & 0.2 & 0.15 & 0.25 \\
0.1 & 0.3 & 0.4 & 0.2 \\
0.1 & 0.2 & 0.2 & 0.5 \\
\end{pmatrix}
\]
Various Prediction Techniques

- Techniques forecast
  - the next location of a user
  - his/her arrival and residence time, i.e., the interval of time spent at that location.
Prediction through correlated movement

• Analysis of the correlation of mobility patterns of the users

• Movement correlation is measured through mutual information

• The resulting network can be considered as a network of “movements”
Users are shape.
Same color indicates temporal “similarity”
Example

Blue dotted line is prediction with multivariate non linear predictor
Red line real data
Black line ARMA linear modelling
Summary

• We have described various privacy issues related to mobile devices and shown how they could be addressed.

• Challenges we discussed were
  – Systems related
  – Sensing related
References