Are You Spying on Me?
Large-Scale Analysis on IoT Data Exposure through Companion Apps

Yuhong Nan\textsuperscript{1*}, Xueqiang Wang\textsuperscript{2*}, Luyi Xing\textsuperscript{3}, Xiaojing Liao\textsuperscript{3}, Ruoyu Wu\textsuperscript{4}, Jianliang Wu\textsuperscript{4}, Yifan Zhang\textsuperscript{3}, and XiaoFeng Wang\textsuperscript{3}

Sun Yat-sen University\textsuperscript{1}, University of Central Florida\textsuperscript{2}, Indiana University Bloomington\textsuperscript{3}, Purdue University\textsuperscript{4}
Are You Spying on Me?
Large-Scale Analysis on IoT Data Exposure through Companion Apps

Home automation

Health monitoring
Are You Spying on Me?  
Large-Scale Analysis on IoT Data Exposure through Companion Apps

Your Sex Toy Might Be Spying on You
As more people reach for "smart" bedroom devices, experts worry about flawed security

BY JESSICA DUFFIN WOLFE
ILLUSTRATION BY JANET MAC
Published 14m, Oct. 27, 2021
This article was published over a year ago. Some information may no longer be current.

Amazon’s Alexa isn’t just AI — thousands of humans are listening
One of the only ways to improve Alexa is to have human beings check it for errors
By Nick Statt | @nickstatt | Apr 10, 2019, 8:25pm EDT

Google calls Nest's hidden microphone an 'error'
The tech giant didn't inform customers that the home security hub had a microphone.
Key questions to address privacy concerns

- What types of data are being collected by IoT devices?
- How is the data collected, and to which party is it shared?
Are You Spying on Me?
Large-Scale Analysis on IoT Data Exposure through Companion Apps

Prior research & limitations

<table>
<thead>
<tr>
<th>Device</th>
<th>IP Addr</th>
<th>MAC Addr</th>
<th>Network Chipset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnamed Device 1</td>
<td>172.24.1.82</td>
<td>AC:xxxx:xxxxxxx</td>
<td>Hi-flying electronics technology Co.</td>
</tr>
<tr>
<td>Unnamed Device 2</td>
<td>172.24.1.138</td>
<td>D0:xxxx:xxxxxxx</td>
<td>LI FI LABS MANAGEMENT PTY LTD</td>
</tr>
<tr>
<td>Unnamed Device 3</td>
<td>172.24.1.99</td>
<td>B0:xxxx:xxxxxxx</td>
<td>D-Link International</td>
</tr>
<tr>
<td>Unnamed Device 4</td>
<td>172.24.1.98</td>
<td>D8:xxxx:xxxxxxx</td>
<td></td>
</tr>
<tr>
<td>Unnamed Device 5</td>
<td>172.24.1.133</td>
<td>D0:xxxx:xxxxxxx</td>
<td></td>
</tr>
</tbody>
</table>

Monitoring network traffic
[IMC19, CCS19, PETS19]

- Small device set, lab-environment 😞
- Data encryption, not scale-well 😞

Crowd source
[SEC19, Ubicomp2020]

- Collecting real user data and leads to big privacy concern 😞
Our goal: Large-scale, fine-grained understanding of IoT device data exposure in the wild.
Observation 1 - Two typical data management modes in IoT ecosystem

(a) **Device-to-mobile data management** (using NFC, Bluetooth, BLE, etc.)

(b) **Cloud-based data management** (Google Home, Amazon Echo, etc.)
Observation 2 – Semantics-rich IoT companion apps

- **IoT Code Block**
  - A cluster of texts (in a method) which includes meaningful labels describing IoT device data.

- **IoT Data Point**
  - Individual IoT data if the text label indicates information related to IoT Devices.
Are You Spying on Me?
Large-Scale Analysis on IoT Data Exposure through Companion Apps

IoTProfiler

Device Companion apps

Taxonomy of IoT Sensitive Data

IoT Data Identifier
- Locating IoT Code Blocks
- Identifying IoT Data Points

Leakage Tracker
- Semantics-driven Variable Binding
- Privacy Risks Analysis

Reported Results
A taxonomy of privacy-sensitive IoT data

- Challenge: Diversity of privacy-sensitive IoT data.
- Construct an IoT taxonomy by analyzing known IoT reports, papers, documents and industry standards, etc.
- 550 data items, with 8 sub-categories
  - Released at https://sites.google.com/view/iotprofiler.
**IoT data point identification**

- **Challenge:** IoT data handled in companion apps are often kept together with the app local data.
- **Solution:** A two-stage classifier
  - Only considers **clustered IoT data labels**

---

**Diagram Description:**
- IoT Taxonomy
  - sleep_length
  - battery_level
  - walk_distance
- IoT Code Block Identification (IDI)
- Non-IoT Code Block
- Text similarity
- IoT Data Point(s)
Variable binding and data exposure detection

- **Alias data labels**

  ```java
  intent.putExtra("voltage_val", dev.battery_level)
  ```

- **Link data labels to program variables**

  ```java
  JsonObject.put("walk_distance", a)
  a = JsonObject.get("walk_distance")
  ```

E.g., are these data flows disclosed within the application’s privacy policy, or data transmitted securely?

Please check our paper for further technical details.
Datasets

- IoT companion apps collected in the wild.
  - 6,208 IoT apps as of Aug. 2020
IoT data exposure in the wild

<table>
<thead>
<tr>
<th>App Store</th>
<th>Data Type</th>
<th>Device Identifier</th>
<th>Network Identifier</th>
<th>Subtotal</th>
<th>Biometric Data</th>
<th>Location Data</th>
<th>Environmental Data</th>
<th>Subtotal</th>
<th>Device Metadata</th>
<th>Device Usage and Status</th>
<th>Timing Data</th>
<th>Subtotal</th>
<th>Any Data Type</th>
<th>Exposure w/o Disclosure</th>
<th>Insecure Transmission</th>
<th>Share to Third Party</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Store</td>
<td>Device Tracking Data</td>
<td>1.2</td>
<td>525 (96.4%)</td>
<td></td>
<td>1.2</td>
<td>823 (13.3%)</td>
<td>1.6</td>
<td>96 (1.5%)</td>
<td>1.2</td>
<td>123 (2.3%)</td>
<td>1.2</td>
<td>568 (9.1%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Store</td>
<td>Sensor Data</td>
<td>1.8</td>
<td>1,078 (17.4%)</td>
<td></td>
<td>1.6</td>
<td>197 (3.2%)</td>
<td>1.6</td>
<td>197 (3.2%)</td>
<td>1.8</td>
<td>334 (5.6%)</td>
<td>1.8</td>
<td>813 (13.4%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Store</td>
<td></td>
<td>2.1</td>
<td>711 (11.5%)</td>
<td></td>
<td>2.2</td>
<td>172 (2.8%)</td>
<td>2.2</td>
<td>172 (2.8%)</td>
<td>2.3</td>
<td>275 (4.6%)</td>
<td>2.3</td>
<td>735 (11.8%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US Store</td>
<td></td>
<td>5.5</td>
<td>1,560 (29.2%)</td>
<td></td>
<td>5.6</td>
<td>1,973 (36.8%)</td>
<td>5.6</td>
<td>1,973 (36.8%)</td>
<td>5.8</td>
<td>1,992 (32.1%)</td>
<td>5.8</td>
<td>1,579 (29.6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese Store</td>
<td></td>
<td>6.2</td>
<td>413 (75.5%)</td>
<td></td>
<td>5.5</td>
<td>1,128 (21.3%)</td>
<td>5.5</td>
<td>1,128 (21.3%)</td>
<td>6.3</td>
<td>502 (8.8%)</td>
<td>6.3</td>
<td>413 (75.0%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Any Store</td>
<td></td>
<td>5.6</td>
<td>1,973 (31.8%)</td>
<td></td>
<td>5.6</td>
<td>1,973 (36.8%)</td>
<td>5.6</td>
<td>1,973 (36.8%)</td>
<td>5.8</td>
<td>1,992 (32.1%)</td>
<td>5.8</td>
<td>1,579 (29.6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 50,667 IoT code blocks and 174,943 IoT data points from 5,795/6,208 (93.3%) apps.
Are You Spying on Me?
Large-Scale Analysis on IoT Data Exposure through Companion Apps

IoT data exposure in the wild

- 1,973 apps (31.8%) from at least 1,559 unique device vendors are found to collect sensitive IoT data without proper disclosure.
- Each app exposes 5.6 IoT data items on average.
IoT data exposure in the wild

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Data Item</th>
<th># Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>US Store</td>
</tr>
<tr>
<td>Device Tracking Data</td>
<td>device id</td>
<td>318 (6.0%)</td>
</tr>
<tr>
<td></td>
<td>wifi password</td>
<td>247 (4.6%)</td>
</tr>
<tr>
<td></td>
<td>mac address</td>
<td>154 (2.9%)</td>
</tr>
<tr>
<td></td>
<td>ssid</td>
<td>154 (2.9%)</td>
</tr>
<tr>
<td>Sensor Data</td>
<td>body weight</td>
<td>135 (2.5%)</td>
</tr>
<tr>
<td></td>
<td>temperature</td>
<td>69 (1.3%)</td>
</tr>
<tr>
<td></td>
<td>altitude</td>
<td>39 (0.7%)</td>
</tr>
<tr>
<td></td>
<td>humidity</td>
<td>37 (0.7%)</td>
</tr>
<tr>
<td>Device Attached Data</td>
<td>start/end time</td>
<td>251 (4.7%)</td>
</tr>
<tr>
<td></td>
<td>model name</td>
<td>244 (4.6%)</td>
</tr>
<tr>
<td></td>
<td>device name</td>
<td>210 (3.9%)</td>
</tr>
<tr>
<td></td>
<td>duration</td>
<td>162 (3.0%)</td>
</tr>
</tbody>
</table>

IoT data exposure for users of different regions

Cross-region IoT data flows

For more findings, please check out our paper.
Example

- Health monitoring device

Own server:
https://d37se***.37bit.net:8**3

healthlink.cn
Case study

- Cigarette holder – use expected data
  - Harmful substances (e.g., nicotine)
  - Number of cigarettes smoked

- Data without the user’s awareness
  - Smoking habits: **smoking times (puff), locations**
  - Health conditions: **breathing capacity**
Summary

- New techniques to enable fine-grained analysis of IoT data exposure.
- Large-scale understanding of IoT data exposure.

- Potential applications include the auto-generation of privacy labels, which can help IoT apps/devices become privacy compliant.

Thank You!