Using the CI tuple for Auditing Data Collection Practices (from the Edge)

Athina Markopoulou
Rahmadi Trimananda, Hao Cui
UC Irvine

http://properdata.eng.uci.edu
Problem Space

**END SYSTEMS**
Code Analysis
Network Traffic

**PRIVACY POLICY**
Analysis, NLP
Synthesis

**LAW:**
GDPR
CCPA
ADPPA?
...

Need for unified/auditable specification: **opportunity for CI tuple**
Network Point of View

Application Domains

Web/Browsers

Mobile devices & apps

Smart TV & apps

Smart Speakers
Voice Assistants

Smart Home
IoT

VR/AR devices
& apps

Networking View

App
App
... App

Platform
Hardware & Software

Network Traffic

Malicious
3rd parties: Advertisers, Trackers...

1st parties

Implementation Challenges:
- Capture packets in real-time; on-device, on the WiFi router, in the middle of a network
- Encryption → visibility into protocols: IP, HTTP/HTTPS, DNS, TLS/SNI
- Exercise apps automatically, and at scale
- Low level → difficult to infer high level properties
Network Point of View

**Application Domains**

- Web/Browsers
- Mobile devices & apps
- Smart TV & apps
- Smart Speakers
- Voice Assistants
- Smart Home
- IoT
- VR/AR devices
- & apps

**Networking View**

- Malicious
- 3rd parties: Advertisers, Trackers...
- 1st parties

**Goals:**
- Diagnosis: who sent it (e.g., app, platform, SDK) what data type (e.g. which PII) goes, to what destination (e.g. ATS), for what purpose?
- Control: can we do something about it (block, obfuscate, add noise etc)?
End Systems/Networking View

CVInspector [NDSS'21]
Your Echos are Heard, [2022]

AntMonitor
NoMoAds [PETS'19]
NoMoATS [PETS'20]

Mobile devices & apps

PingPong [NDSS'20]

Rokustic, Firetastic [PETS'20]
SmartTV Fingerprinting [PETS'22]

Smart TV & apps

OVRSeen [SEC'22]
**Example: Results from OVRseen**

**Data type sent out**

<table>
<thead>
<tr>
<th>Data Types (21)</th>
<th>Apps 1st</th>
<th>Apps 3rd</th>
<th>Apps Pl.</th>
<th>FQDNs 1st</th>
<th>FQDNs 3rd</th>
<th>FQDNs Pl.</th>
<th>% Blocked 1st</th>
<th>% Blocked 3rd</th>
<th>% Blocked Pl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PII</td>
<td>6</td>
<td>64</td>
<td>2</td>
<td>6</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>38</td>
<td>100</td>
</tr>
<tr>
<td>Device ID</td>
<td>6</td>
<td>65</td>
<td>0</td>
<td>5</td>
<td>13</td>
<td>0</td>
<td>20</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>User ID</td>
<td>6</td>
<td>31</td>
<td>18</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>17</td>
<td>43</td>
<td>50</td>
</tr>
<tr>
<td>Android ID</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>50</td>
</tr>
<tr>
<td>Serial Number</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Person Name</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>Email</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>-</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>Geolocation</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>-</td>
<td>50</td>
<td>1</td>
</tr>
</tbody>
</table>

**Fingerprint**

| SDK Version     | 23       | 69       | 20       | 34        | 28        | 4         | 6            | 46           | 0             |
| Hardware Info   | 21       | 65       | 19       | 25        | 23        | 3         | 4            | 39           | 33            |
| System Version  | 16       | 62       | 19       | 20        | 21        | 3         | 5            | 43           | 33            |
| Session Info    | 7        | 66       | 2        | 7         | 13        | 1         | 14           | 46           | 100           |
| App Name        | 4        | 65       | 2        | 4         | 10        | 1         | 25           | 40           | 100           |
| Build Version   | 0        | 61       | 0        | 0         | 3         | 0         | -            | 100          | -             |
| Flags           | 6        | 53       | 2        | 6         | 8         | 1         | 0            | 50           | 100           |
| Usage Time      | 2        | 59       | 0        | 2         | 4         | 0         | 0            | 50           | -             |
| Language        | 5        | 28       | 16       | 5         | 9         | 1         | 0            | 56           | 0             |
| Cookies         | 5        | 4        | 2        | 5         | 3         | 1         | 0            | 33           | 100           |

**VR Sensory Data**

| VR Play Area    | 0        | 40       | 0        | 0         | 1         | 0         | -            | 100          | -             |
| VR Movement     | 1        | 24       | 2        | 1         | 6         | 1         | 0            | 67           | 100           |
| VR Field of View| 0        | 16       | 0        | 0         | 1         | 0         | -            | 100          | -             |
| VR Pupillary    | 0        | 16       | 0        | 0         | 1         | 0         | -            | 100          | -             |

**Total**

|               | 33       | 70       | 22       | 44        | 39        | 5         | 5            | 36           | 20            |

**Top -10 destinations**

- eSLD
- ATS FDN

**Centralized ecosystem: FB/Oculus, unity**

Driven by tracking & social/analytics, not by ads
Example: Results from OVRseen

- Purpose can be (partly) inferred from network data
  
  - Heuristics: key-value pairs (manually labeled purpose based on keys; “adid” → advertising, “passwd” → security etc); also looked at name of apk and compare destination; information about destination domain (organization/ATS: lookup DuckDuckGo, CrunchBase) [Mobipurpose,’19 Purpliance’21]

- Purpose Stated in Privacy Policy
  
  - Purpose from [Polisis’18] matched for data flows [Policheck] with consistent disclosures [OVRseen’22]
Directly extract CI params

CI tuple: (sender, recipient, data type; [subject]; (purpose; other TP))

**Sender:**
- Application (dev)
- 3rd party library
- Platform, device
- Malware

**Recipient:**
- 1st, 3rd parties, platform, cloud
- Advertisers & trackers (ATS)
- Organization

**Data Type:**
- Personally Identifiable Information (PII)
- Fingerprinting
- Activity Data
- Sensor data

**[Subject:]**
- Typically the user of the app and platform

**Purpose:**
- Functionality
- Analytics
- Tracking
- Ads
- Personalization
- Security

Network Packets

"data flow"
We find that Amazon processes voice data to infer user interests and uses it to serve targeted ads on-platform (Echo devices) as well as off-platform (web). Smart speaker interaction leads to as much as 30X higher ad bids from advertisers. Finally, we find that Amazon’s and skills’ operational practices are often not clearly disclosed in their privacy policies.
Auditing Network Traffic

Sender:
- Application (dev)
- 3rd party library
- Platform, device
- Malware

Recipient:
- 1st, 3rd parties, platform, cloud
- Advertisers & trackers (ATS)
- Organization

Data Type:
- Personally Identifiable Information (PII)
- Fingerprinting
- Activity Data
- Sensor data

[Subject:] Typically the user of the app and platform

Purpose:
- Functionality
- Analytics
- Tracking
- Ads
- Personalization
- Security

Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])
Auditing Network Traffic vs. Privacy Policy

CI tuple: \( (\text{sender}, \text{recipient}, \text{data type}; [\text{subject}]; (\text{purpose}; \text{other TP})) \)

- **Sender:** Application (dev) - 3rd party library - Platform, device - Malware
- **Recipient:** 1st, 3rd parties, platform, cloud - Advertisers & trackers (ATS) - Organization
- **Data Type:** Personally Identifiable Information (PII) - Fingerprinting - Activity Data - Sensor data
- **[Subject:]:** Typically the user of the app and platform
- **Purpose:** Functionality - Analytics - Tracking - Ads - Personalization - Security
- **Other Aspects:** With Notice? - With Consent? - Consistent Disclosure?

Network: \( (\text{sender}=\text{app/platform/SDK}, \text{destination}=\text{domain/org}, \text{data type}, [\text{purpose}]) \)

Policy: \( (\text{sender}, \text{destination}=\text{entity}, \text{data type}, \text{purpose}, [\text{other}]) \)
Privacy Policy Analysis

- **Early trend: analysis by experts**
  - within CI: [Shvartzshnaider et al. ‘19]

- **Recent development: automated via NLP**

- **Today, NLP-based privacy policy analysis can successfully:**
  - extract (data types, recipient=entity) [PolicyLint’19]
  - extract purpose [Polisys’18], [MobiPurpose’19], [Purpliance’21]
    - Although this is more tricky; and unclear how to connect to other parameters
  - check the consistency of network vs. policy side [PoliCheck’20, OVSeen’22]
    - Ontologies are necessary for that
  - be applied at scale to a large number of policies and application domains
    - Mobile, amazon skills, VR apps
Example: Extracting Collection Statements [PolicyLint+]

NLP pipeline

- Dependencies
- Part of Speech
- Named Entities

Heuristic information extraction:
- Pruning dependency tree
- Sentence identification

Collection Statement
- "Entity" = first_party
- Action = collect
- Data = email_address
### Collection Statements vs. Data Flows

#### From privacy policy: $P=$ (data type, entity)

<table>
<thead>
<tr>
<th>Disclosure Type</th>
<th>Privacy Policy Text</th>
<th>Action: Data Collection Statement (P)</th>
<th>Data Flow (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consistent</strong></td>
<td>Clear</td>
<td><code>collect : (com.cvr.terminus, usage time, we)</code></td>
<td><code>(usage time, we)</code></td>
</tr>
<tr>
<td></td>
<td>Vague</td>
<td><code>collect : (com.HomeNetGames.WW1oculus, pii, third party)</code></td>
<td><code>(serial number, oculus)</code></td>
</tr>
<tr>
<td></td>
<td>Omitted</td>
<td><code>collect : (com.kluge.SynthRiders, -, -)</code></td>
<td><code>(system version, oculus)</code></td>
</tr>
<tr>
<td></td>
<td>Ambiguous</td>
<td><code>collect : (com.SDI.TWD, pii, third party)</code></td>
<td><code>(serial number, oculus)</code></td>
</tr>
<tr>
<td></td>
<td>Incorrect</td>
<td><code>not_collect : (com.downpourinteractive. onward, pii, third party)</code></td>
<td><code>(device id, unity)</code></td>
</tr>
</tbody>
</table>

#### From network traffic: $F=$ (data type, destination)

<table>
<thead>
<tr>
<th>Disclosure Type</th>
<th>Privacy Policy Text</th>
<th>Action: Data Collection Statement (P)</th>
<th>Data Flow (F)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Inconsistent</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Ontologies taken into account when checking for consistency of collection statements vs. data flows**

- Data Ontology
- Entity Ontology
Example: VR Ontologies

- Ontologies are necessary to check consistency
- Today heuristically defined: a combination of data-driven and expert curation
Extracting Purpose

Polisis [Sec’2018]

- Older NLP models: DNN, RNN; trained on OPP-115
- Segmented text in paragraphs, used 9 categories
- Granularity usually per phrase, e.g., for first/third-party collection/use
- Service available online

MobiPurpose [Ubicomp’2019]:

- Network Traffic of android apps:
  - key-value pairs: manually labeled purpose based on keys; “adid”→advertising, “passwd”→security etc.
  - also looked at name of apk and destination domain/entity

Purpliance [CCS’2021]:

- Network side: built classifiers based on MobiPurpose
- Policy side: New NLP models (BERT); looking for: “to”, ”in order to”, “for…–ing”
- Hierarchy in purposes → 5 main purposes
- Applications: (1) contradictions, including purposes; tuple 5 vs. nested; (2) consistency of data flow-to-privacy policy, with purpose
Auditing Consistency of Network Traffic vs. Privacy Policies

CI tuple: (sender, recipient, data type; [subject]; (purpose; other))

Sender: - Application (dev) - 3rd party library - Platform, device - Malware
Recipient: - 1st, 3rd parties, platform, cloud - Advertisers & trackers (ATS) - Organization
Data Type: - Personally Identifiable Information (PII) - Fingerprinting - Activity Data - Sensor data
[Subject:] - Typically the user of the app and platform
Purpose: - Functionality - Analytics - Tracking - Ads - Personalization - Security
Other Aspects: - With Notice? - With Consent? - Consistent Disclosure?

Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])
Policy: (sender, destination=entity, data type, purpose, [other])
Examples of disclosure requirements (right to know)

<table>
<thead>
<tr>
<th>Category</th>
<th>CCPA Sections</th>
<th>GDPR Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories of personal information collected, used, or shared</td>
<td>1798.130(a)(5)(B-C)</td>
<td>14(1)(d)</td>
</tr>
<tr>
<td>Source (GDPR) / Categories of sources (CCPA) of the personal information</td>
<td>1798.110(c)(2)</td>
<td>14(2)(f)</td>
</tr>
<tr>
<td>Purposes for the collection, use, and sharing of personal information</td>
<td>1798.110(c)(3)</td>
<td>13(1)(c), 14(1)(c)</td>
</tr>
<tr>
<td>Categories of third parties with whom personal information is shared</td>
<td>1798.110(c)(4)</td>
<td>13(1)(e), 14(1)(e)</td>
</tr>
</tbody>
</table>

This affects how policies are written:
- “Collect” (1st party) vs. “share” (3rd party);
- “Use” vs. “collect” vs. purpose? How does it fit in information flow?
- Parameters of CI tuple can be “bloated”
Implication: “bloating” CI parameters

Example of good privacy policies following that format of sections

Information We Collect ...
- Name
- Age or date of birth ...

How We Use the Information We Collect or Receive
- To create, administer and troubleshoot accounts, ...
- To credit or accept payments; ...

Sharing Information ...
- Our affiliates located all over the world ...
- Third-party service providers: ...

We collect the following categories of personal information:
- Device information... such as IP address...
- Location. We use this information to provide features...

We use your personal information... to:
- Provide the Services...
- Authenticate your account...

We disclose the personal information... as follows:
- With our travel partners...
- With social networking services...

Example of bad FB privacy policy [Shvartzshnaider et al. 2019]

Figure 3: Example of CI parameter bloating in privacy policy text (top) and mapped into possible interpretations (bottom).
Open Problems and Directions

Q1: Auditing Network traffic vs Policies (NLP)

Network: (sender=app/platform/SDK, destination=domain/org, data type, [purpose])

Policy: (sender, destination=entity, data type, purpose, [other])

- Dealing with parameters obtained from different sources
- Propose the full CI tuple to be used for this particular auditing.

Q2: Rethink the tuple data structure

- Hierarchy
- Purpose

Q3: Be proactive

- Beyond assessing/evaluating practices as appropriate or not
- Participate in defining laws, standards and open interfaces
Q2. Extend/Refine the Tuple data structure?

Hierarchy is necessary for consistency checking
Hierarchy is happening already
Hierarchy is more scalable
Hierarchies are hard to define:
local (within a policy) vs global (by experts)

Law: currently encourages “bloating”; ongoing discussion in CPPA

Law: distinguishes collect/share/use

Re-consider purpose:
• part of another parameter (data type, purpose), (entity, purpose)?
• or part of TP?
• its own parameter?
CI tuple for Auditing from the Edge

(sender, recipient, data type, subject, TP)

END-SYSTEM
Network Traffic

PRIVACY POLICY
NLP Analysis

LAW:
GDPR
CCPA
ADPPA?
...

Need for unified/auditable specification:
• opportunity for CI tuple to define the data structure for auditing and data rights requests
Thank you!

CI for Auditing (from the edge)

athina@uci.edu

http://properdata.eng.uci.edu
https://athinagroup.eng.uci.edu/projects/ovrseen/