POLICYCOMP: Counterpart Comparison of Privacy Policies Uncovers Overbroad Personal Data Collection Practices

Lu Zhou, Chengyongxiao Wei, Tong Zhu, Guoxing Chen, Xiaokuan Zhang, Suguo Du, Hui Cao, and Haojin Zhu
Privacy Laws and Privacy Policy

General Data Protection Regulation (GDPR)

California Consumer Privacy Act (CCPA)

provide legal frameworks on how to collect/use personal data

- Purpose limitation
- Data minimization
- ...

1. About this Privacy Policy

1.1. This privacy policy explains how information about you is collected, used, and disclosed by when you play one of our Games. This Privacy Policy is applicable where acts as a data controller with respect to your data. This is the case where we determine the purposes and means of the data processing in our Games.

3. What information does collect?

3.1. collects certain information when you play our Games. We may also collect information from ad network providers and other

App’s privacy policy: Explain how to collect and use personal data

- personal data collection practices (PDCPs) (e.g., name, phone number)
- clear purposes for processing them...
Privacy Laws and Privacy Policy

Users inattentively click “YES” without a complete understanding of privacy policies (Just 13% read them in full [European Commission])

Privacy Laws
provide legal frameworks on how to collect/use personal data

Privacy Policy
App’s privacy policy: Explain how to collect and use personal data
Related Work

- Privacy Policy Understanding
  - ACL’ 21, PETs’ 21, USENIX Security’ 14, USENIX Security’ 18

- Consistency: Whether a privacy policy is written logically sound
  - USENIX Security’ 19, DSN’ 16, RE’ 13

- Consistency: Whether a privacy policy is consistent with app’s behaviors
  - ICSE’ 16, NDSS’ 17, USENIX Security’ 20, ICSE’ 18, PETs’ 19, CCS’ 21
Related Work

- **Privacy Policy Understanding**
  - *ACL’ 21, PETs’ 21, USENIX Security’ 14, USENIX Security’ 18*

- **Consistency: Whether a privacy policy is written logically sound**
  - *USENIX Security’ 19, DSN’ 16, RE’ 13*

- **Consistency: Whether a privacy policy is consistent with app’s behaviors**
  - *ICSE’ 16, NDSS’ 17, USENIX Security’ 20, ICSE’ 18, PETs’ 19, CCS’ 21*

Whether PDCPs are **necessary** for given purposes
(comply with the principle of data minimization?)

**PDCP:** Personal Data Collection Practice
Research Question

**Overbroad collection:** the app developers claim more PDCPs in privacy policies than actually needed for desired services of users.
Research Question

Overbroad collection: the app developers claim more PDCPs in privacy policies than actually needed for desired services of users.

(Privacy Policy) In the exceptional circumstance that we collect any special category information (information about your health, sexual orientation, racial or ethnic profile, political opinions...)

Stock trading app
# Installs: 500K+

- This app collects sensitive personal data without stating the specific purposes
Research Question

**Overbroad collection:** the app developers claim more PDCPs in privacy policies than actually needed for desired services of users.

(Privacy Policy) In the exceptional circumstance that we collect any special category information (information about your health, sexual orientation, racial or ethnic profile, political opinions...)

- This app collects sensitive personal data without stating the specific purposes

Stock trading app

# Installs: 500K+

Our Goal: Identify overbroad collections of PDCPs from privacy policies
Challenge I: The lack of detailed standards

- Privacy is a context-dependent concept
  - different kinds of services
  - different personal data necessary for the services

This hinders the lawmakers from defining a clear and widely applicable privacy boundary between “necessary” and “unnecessary” on a wide range of apps.
Challenge II: Unclear purposes in privacy policies

- use a separate section to explain the purposes of all collected data
  - e.g., “We may use collected personal data for any purpose as below ...”

- use unclear language to describe purposes
  - e.g., “We may use your personal data to develop new services”

It is difficult to determine **exact purposes for each PDCP** since many privacy policies only specify purposes at the app level or explain purposes using unclear language.
Existing method

Challenge I: The lack of detailed standards

Challenge II: Unclear purposes in privacy policies

Existing method: relies on legal experts to manually audit each PDCP

Not scalable for

Millions of apps

An automated tool is needed to preliminarily screen out overbroad PDCPs for the legal experts to review
Our Idea - Counterpart Comparison

Basic intuition

Similar apps ➔ Similar functionalities ➔ Similar privacy contexts ➔ Similar PDCPs

<table>
<thead>
<tr>
<th>APPs</th>
<th>Email</th>
<th>Device ID</th>
<th>Name</th>
<th>Photo</th>
<th>Location</th>
<th>Audio</th>
<th>Gender</th>
<th>SSN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target app</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
</tr>
<tr>
<td>counterparty I</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td></td>
<td>★</td>
</tr>
<tr>
<td>counterparty II</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td></td>
<td>★</td>
</tr>
<tr>
<td>counterparty III</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td></td>
<td>★</td>
</tr>
<tr>
<td>counterparty IV</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td></td>
<td></td>
</tr>
<tr>
<td>counterparty V</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td></td>
<td>★</td>
</tr>
</tbody>
</table>
Our Idea- Counterpart Comparison

- **Basic intuition**
  - Similar apps → Similar functionalities → Similar privacy contexts → Similar PDCPs

- **Counterpart (similar app) comparison**
  - Leverage the PDCPs in counterpart apps’ privacy policies as potential standards
  - A PDCP in the target app’s privacy policy is more likely to be necessary if it is also in counterpart apps’ privacy policies.
System Overview

1. Standard
   find top-k counterpart apps

2. Extract
   extract PDCPs from privacy policies

3. Audit
   Risk Estimation + Reasoning
System Design

**Input: Target app A**

- **Candidate sourcing**
  - Google Play and alternative app recommendation websites
  - *A’s similar apps*

- **Semantic similarity-based ranking**

**Output: Top-k counterparts (similar apps) of A**
System Design

- **NLP-based PDCP extraction**
  - Sentence parsing
  - Potential PDCPs extraction

- **PDCP regularization**

---

- Candidate sourcing
- Candidate ranking
- Top-k counterparts

- **PDPC Extraction & Regularization Module**
  - PDCP Extraction
  - Regularization

- **Overbroad Collection Analysis Module**
  - Name
  - Contacts
  - PDCPs in A

- **Contact list**
  - Your contacts
  - Address book

[Diagram showing extraction and analysis processes]
System Design

- **Overbroad likelihood**
  the ratio of counterparts that do not collect \( d_i \):
  \[
  L(d_i) = \frac{1}{k} \sum_{m=1}^{k} \left\{ \begin{array}{ll}
  1, & \text{if } d_i \notin D(P_m) \\
  0, & \text{otherwise}
  \end{array} \right.
  \]

- **Overbroad PDCF risk estimation**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Risk level</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L(d_i) &gt; \alpha ) &amp; ( d_i \in S_\Omega )</td>
<td>High</td>
<td>Overbroad collection of Class-I personal data</td>
</tr>
<tr>
<td>( L(d_i) &gt; \alpha ) &amp; ( d_i \notin S_\Omega )</td>
<td>Medium</td>
<td>Overbroad collection of Class-II personal data</td>
</tr>
<tr>
<td>( L(d_i) \leq \alpha )</td>
<td>Low</td>
<td>Mostly agreed personal data collection</td>
</tr>
</tbody>
</table>
Evaluation

- **Sources**
  - Google Play ("similar apps")
  - three alternative-app recommendation websites (e.g., AlternativeTo)

- **10,042 target apps**
  - 72.85% of which have over 100,000 downloads

- **30,281 distinct counterpart apps**
Evaluation

- Extracted **57,993** PDCPs from 10,042 target apps
- **48.29%** of PDCPs have the risk of overbroad collection
  - high-risk: \(\frac{871}{57,993} = 1.5\%\)
  - medium-risk: \(\frac{27,132}{57,993} = 46.79\%\)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Risk level</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>(L(d_i) &gt; \alpha \land d_i \in S_\Omega)</td>
<td>High</td>
<td>Overbroad collection of <strong>Class-I personal data</strong></td>
</tr>
<tr>
<td>(L(d_i) &gt; \alpha \land d_i \notin S_\Omega)</td>
<td>Medium</td>
<td>Overbroad collection of <strong>Class-II personal data</strong></td>
</tr>
<tr>
<td>(L(d_i) \leq \alpha)</td>
<td>Low</td>
<td>Mostly agreed personal data collection</td>
</tr>
</tbody>
</table>

**Class-I personal data**: highly protected personal data expressly stated under a privacy protection law
Notification to Developers

- **2,000 target apps**
  - 1,661 emails are successfully delivered
- Receive **52** responses, **39** of which acknowledge our findings
- The privacy policies of **74 apps** have been updated by removing 180 overbroad PDCPs we sent without replying to us.

<table>
<thead>
<tr>
<th>Table 7: The responses from developers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledge our findings</td>
</tr>
<tr>
<td>all findings</td>
</tr>
<tr>
<td>partial findings</td>
</tr>
<tr>
<td>Disagree with our findings</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Case Study I

Privacy Policy Generators

- tools for generating coarse-grained privacy policy
- cannot cover all requirements of an app
Case Study I

Privacy Policy Generators

- tools for generating coarse-grained privacy policy
- cannot cover all requirements of an app

(Email response) “We actually do not collect these personal data, this policy was generated by an online tool and had it by default.”

(Email response) “the privacy policy are automatically generated by privacy policy generator so its content does not represent the data that the app collects.”
Case Study II

One-to-many privacy policies

- Developers owning multiple apps used one single privacy policy
- **One-to-many privacy policies** tend to include more overbroad PDCPs

**PDCPs in the privacy policy of a browser**

- Email address, Name, Phone number, Payment Info, Browsing History
- Audio, Contact list, Location, IP address, Device ID
- ...
Case Study II

One-to-many privacy policies

- Developers owning multiple apps used one single privacy policy
- **One-to-many privacy policies** tend to include more overbroad PDCPs

(Email response) “This is a common privacy policy which is currently used for all apps in our account. **But for sure we will revise the policy soon and update it accordingly.**”

PDCPs in the privacy policy of a browser

- Email address, Name, Phone number, Payment Info, Browsing History
- Audio, Contact list, Location, IP address, Device ID
- ...
Limitation

- The same overbroad PDCPs may be shared by the target app and its counterparts
- Lack of highly similar counterpart apps for some target apps
- Inaccurate PDCPs extraction and regularization due to the limitations of existing NLP tools.

Future directions

- More detailed standards + state-of-the-art NLP tools
POLICYCOMP: Counterpart Comparison of Privacy Policies Uncovers Overbroad Personal Data Collection Practices

Thank you!

lead author’s email: zhoulu@xidian.edu.cn