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

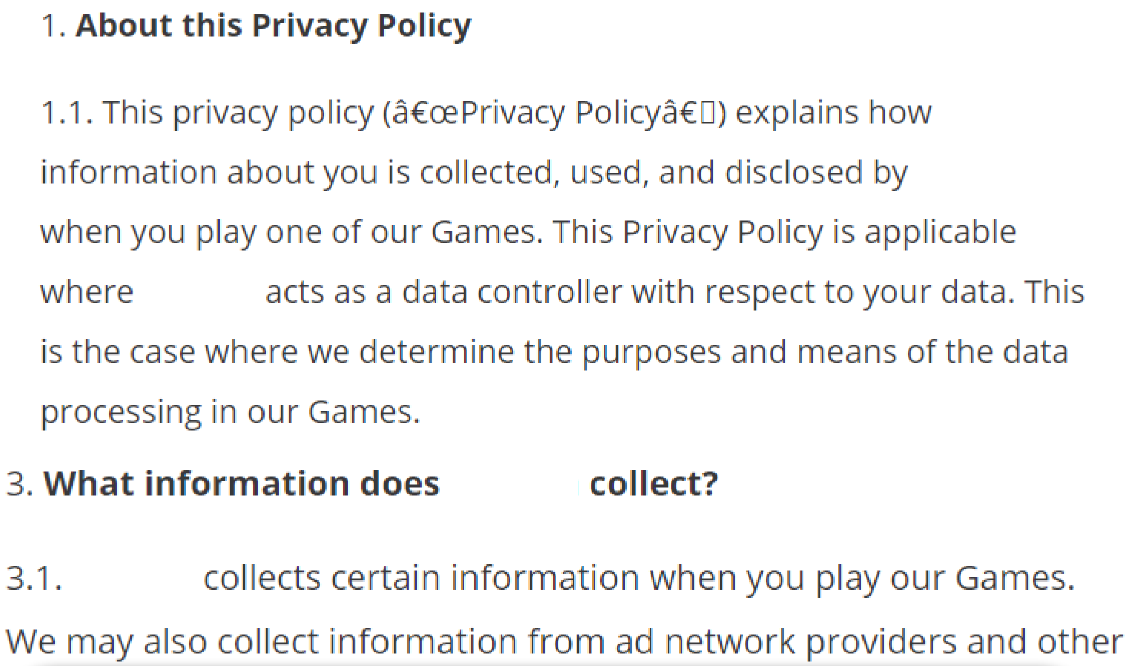
Privacy Laws and Privacy Policy

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 (â€œPrivacy Policyâ€) 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

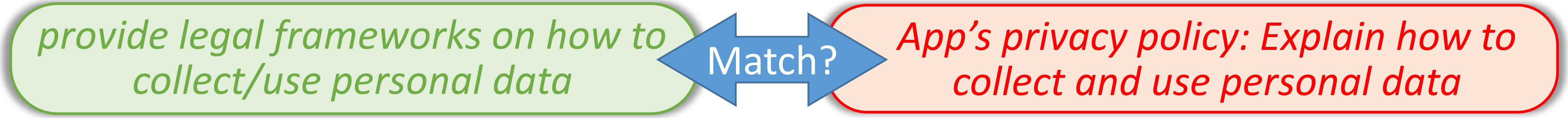
- personal data collection practices (**PDCPs**) (e.g., name, phone number)
- clear purposes for processing them…

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 **Privacy Policy**



provide legal frameworks on how to collect/use personal data

Match?

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…)

Stock trading app

# Installs: *500K+*

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

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

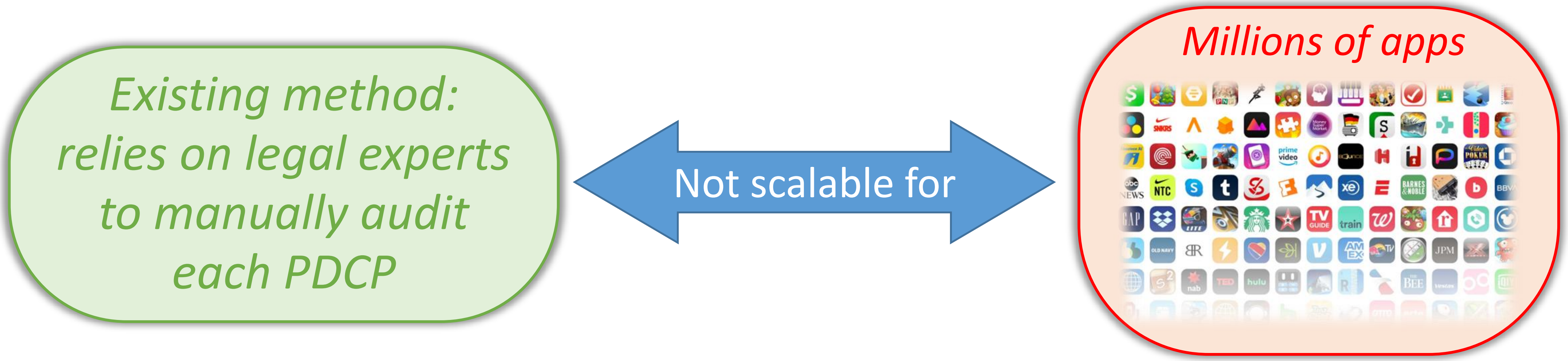
- e.g., “*We may use your personal data to develop new services*”

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

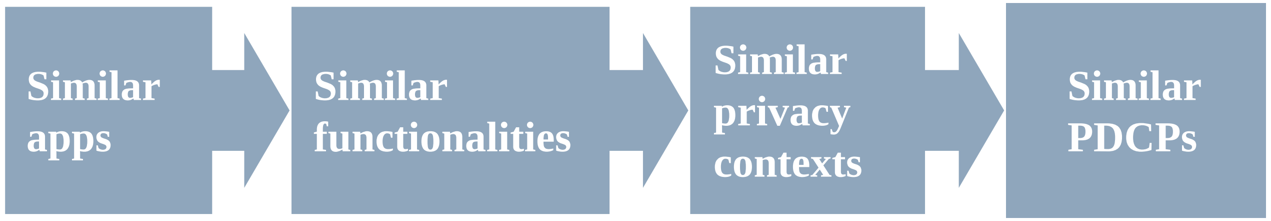
Not scalable for

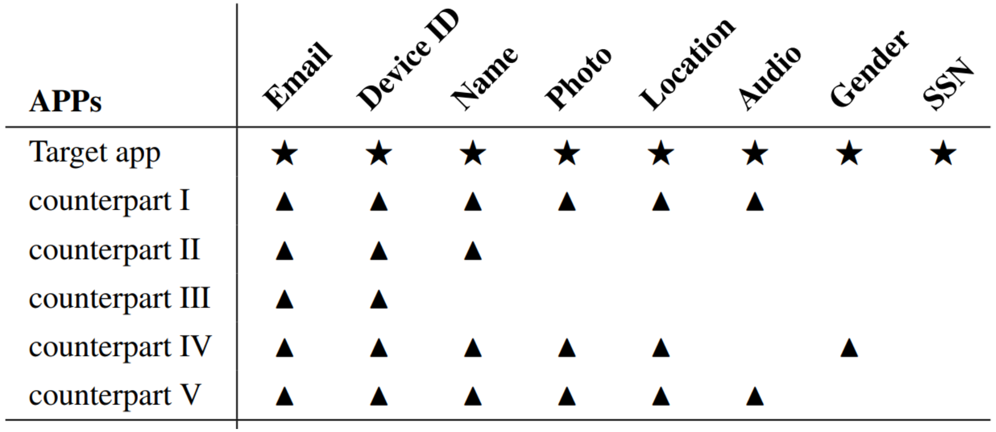
Millions of apps

An automated tool is needed to preliminarily screen out overbroad PDCPs for the legal experts to review

Our Idea- Counterpart Comparison

□ **Basic intuition**

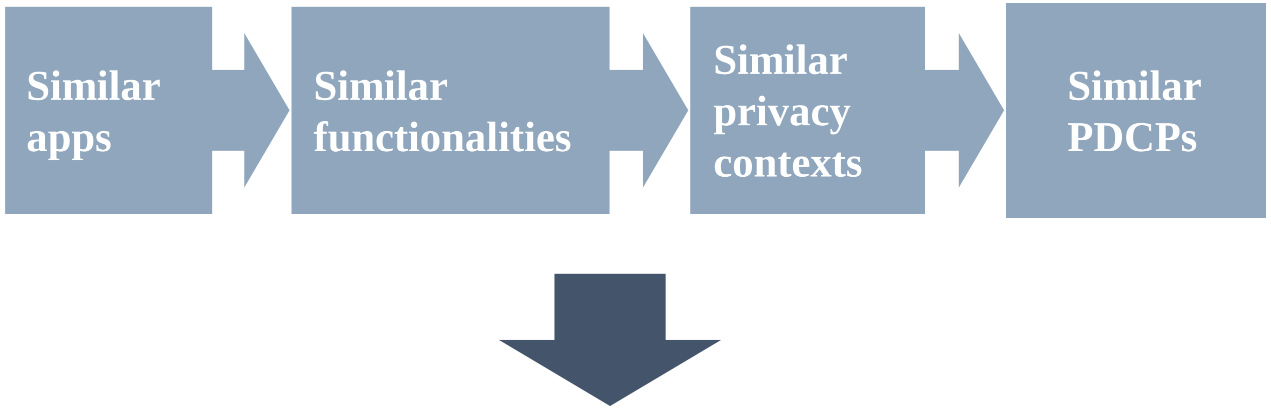




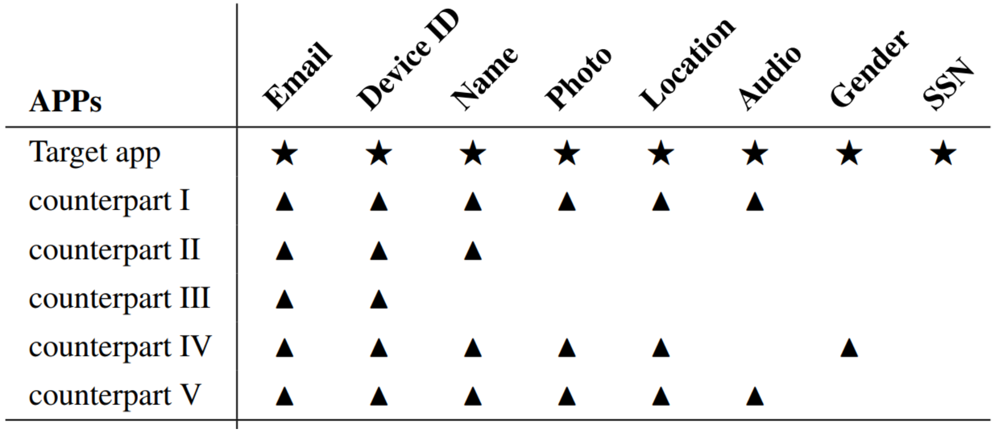
APPs	Email	Device ID	Name	Photo	Location	Audio	Gender	SSN
Target app	★	★	★	★	★	★	★	★
counterpart I	▲	▲	▲	▲	▲	▲		
counterpart II	▲	▲	▲					
counterpart III	▲	▲						
counterpart IV	▲	▲	▲	▲	▲		▲	
counterpart V	▲	▲	▲	▲	▲	▲		

Our Idea- Counterpart Comparison

□ Basic intuition



Similar apps → Similar functionalities → Similar privacy contexts → Similar PDCPs

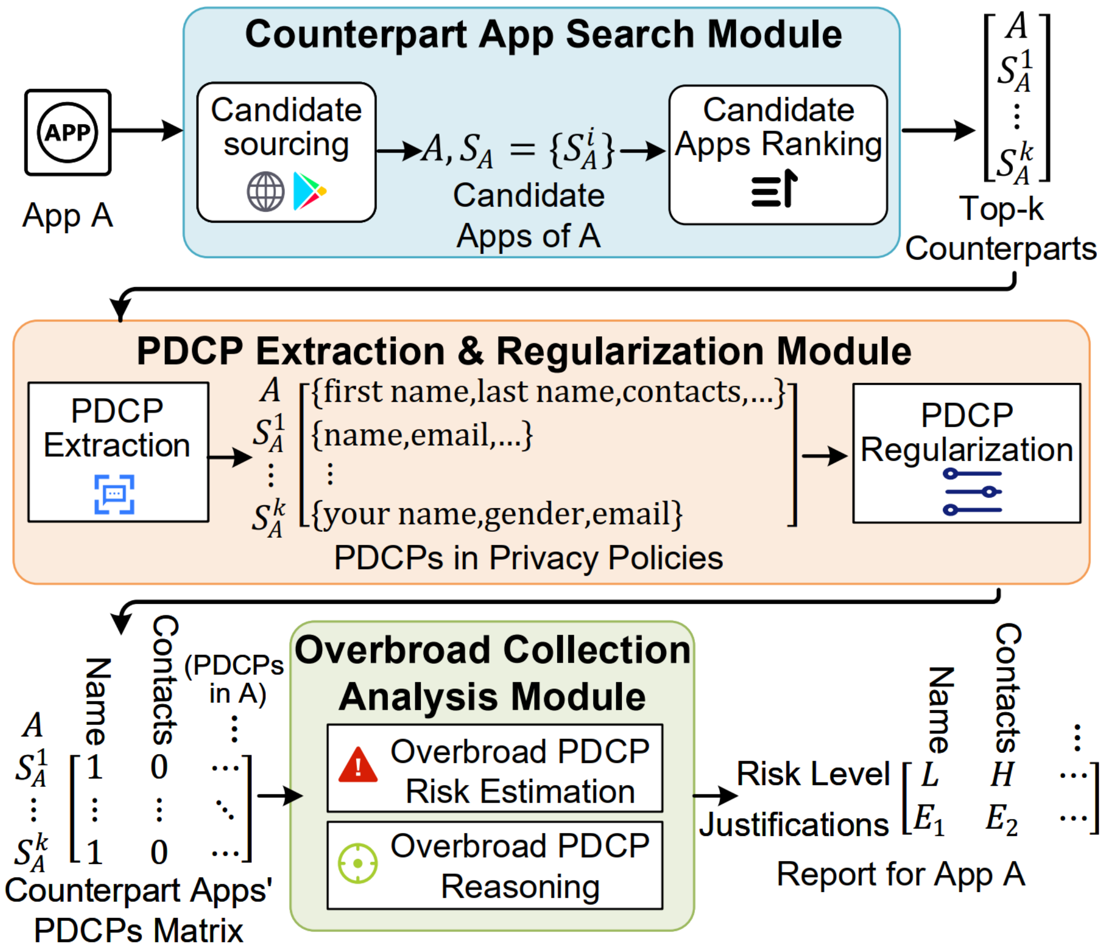


APPs	Email	Device ID	Name	Photo	Location	Audio	Gender	SSN
Target app	★	★	★	★	★	★	★	★
counterpart I	▲	▲	▲	▲	▲	▲		
counterpart II	▲	▲	▲					
counterpart III	▲	▲						
counterpart IV	▲	▲	▲	▲	▲		▲	
counterpart V	▲	▲	▲	▲	▲	▲		

□ 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



Counterpart App Search Module
APP
App A
Candidate sourcing
$A, S_A = \{S_A^i\}$
Candidate Apps of A
Candidate Apps Ranking
$\begin{bmatrix} A \\ S_A^1 \\ \vdots \\ S_A^k \end{bmatrix}$
Top-k Counterparts
PDCP Extraction & Regularization Module
PDCP Extraction
A {first name,last name,contacts,...}
S_A^1 {name,email,...}
S_A^k {your name,gender,email}
PDCPs in Privacy Policies
PDCP Regularization
Name
Contacts
(PDCPs in A)
A
S_A^1 1 0 ⋯
S_A^k 1 0 ⋯
Counterpart Apps' PDCPs Matrix
Overbroad Collection Analysis Module
Overbroad PDCP Risk Estimation
Overbroad PDCP Reasoning
Risk Level L H ⋯
Justifications E_1 E_2 ⋯
Report for App A

1. Standard

find top-k counterpart apps

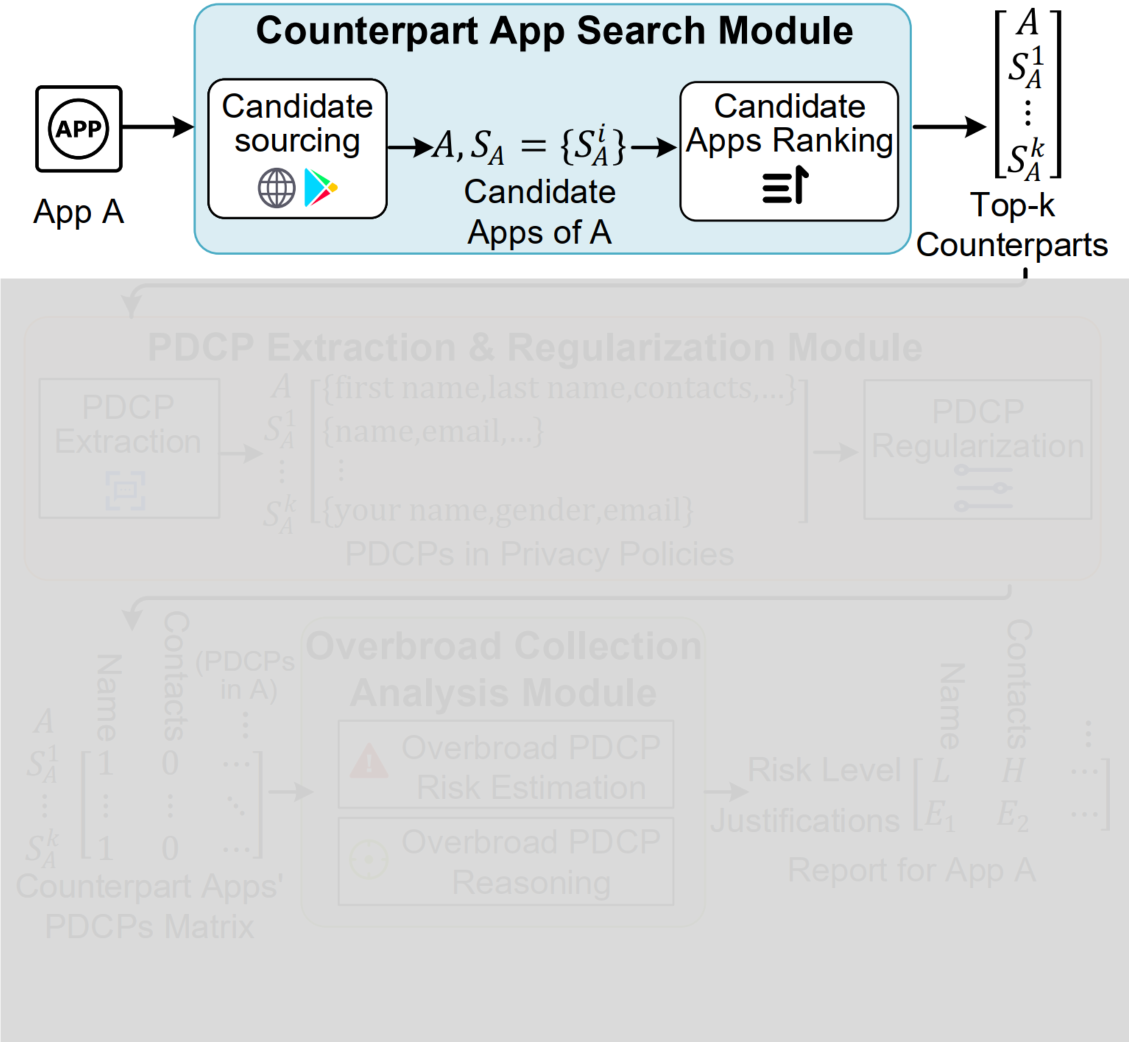
2. Extract

extract PDCPs from privacy policies

3. Audit

Risk Estimation + Reasoning

System Design


Counterpart App Search Module
APP
App A
Candidate sourcing
$A, S_A = \{S_A^i\}$
Candidate Apps of A
Candidate Apps Ranking
$\begin{bmatrix} A \\ S_A^1 \\ \vdots \\ S_A^k \end{bmatrix}$
Top-k Counterparts
PDCP Extraction & Regularization Module
PDCP Extraction
{first name,last name,contacts,...}
{name,email,...}
{your name,gender,email}
PDCPs in Privacy Policies
PDCP Regularization
Name
Contacts
(PDCPs in A)
Counterpart Apps' PDCPs Matrix
Overbroad Collection Analysis Module
Overbroad PDCP Risk Estimation
Overbroad PDCP Reasoning
Risk Level
Justifications
L
H
E_1
E_2
Report for App A

Input: *Target app A*

↓

□ **Candidate sourcing**

- Google Play and alternative app recommendation websites

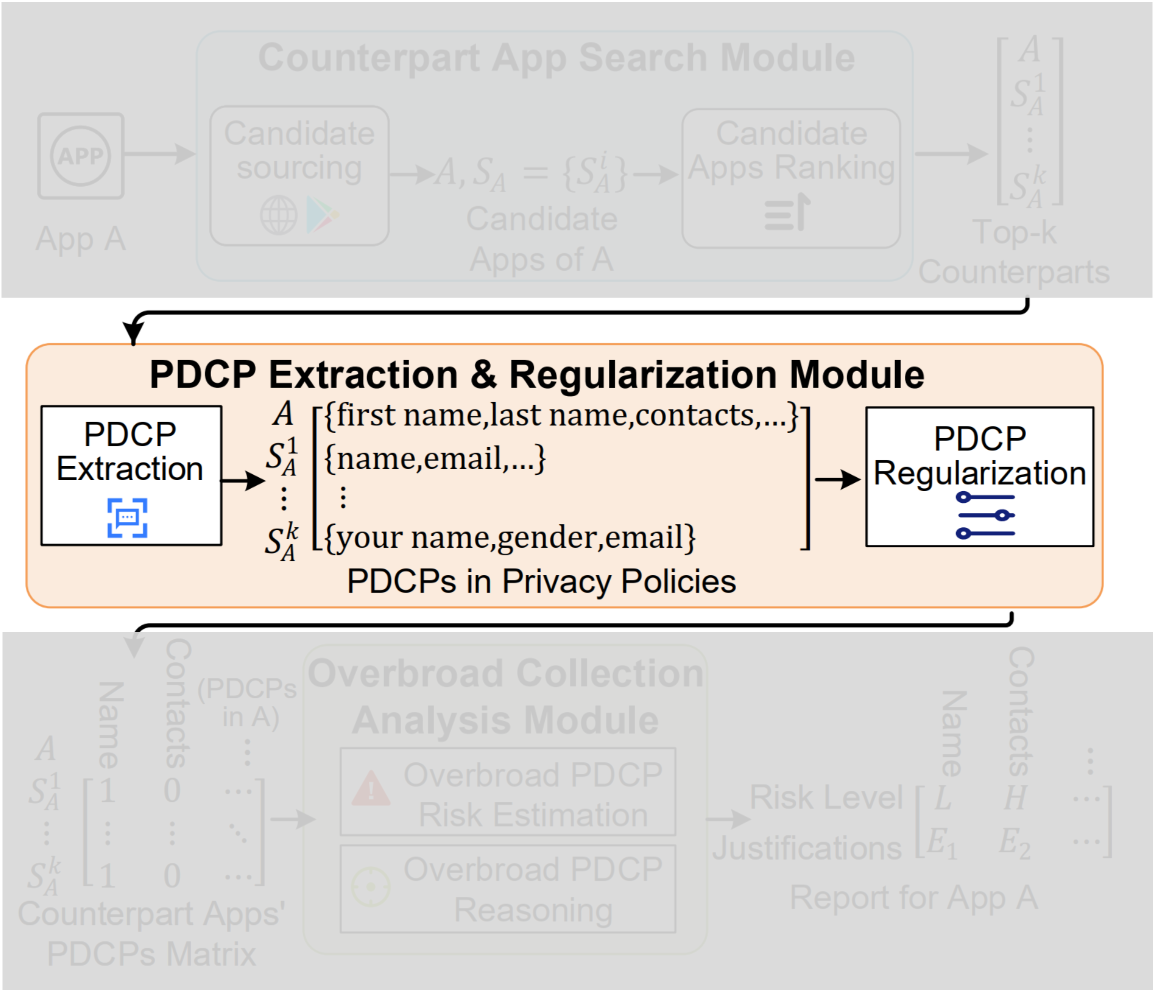
↓ *A's similar apps*

□ **Semantic similarity-based ranking**

↓

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

System Design

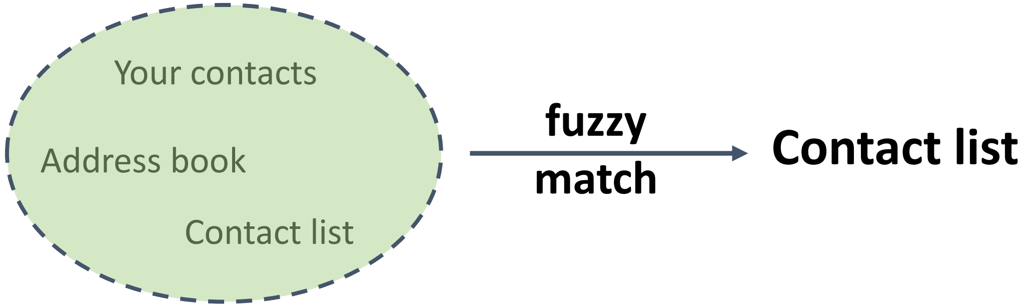


Counterpart App Search Module
APP
App A
Candidate sourcing
$A, S_A = \{S_A^i\}$
Candidate Apps of A
Candidate Apps Ranking
$[A, S_A^1, \vdots, S_A^k]$
Top-k Counterparts
PDCP Extraction & Regularization Module
PDCP Extraction
A {first name,last name,contacts,...}
S_A^1 {name,email,...}
S_A^k {your name,gender,email}
PDCPs in Privacy Policies
PDCP Regularization
Overbroad Collection Analysis Module
(PDCPs in A)
Name
Contacts
Counterpart Apps' PDCPs Matrix
Overbroad PDCP Risk Estimation
Overbroad PDCP Reasoning
Risk Level
Justifications
L H
E_1 E_2
Report for App A

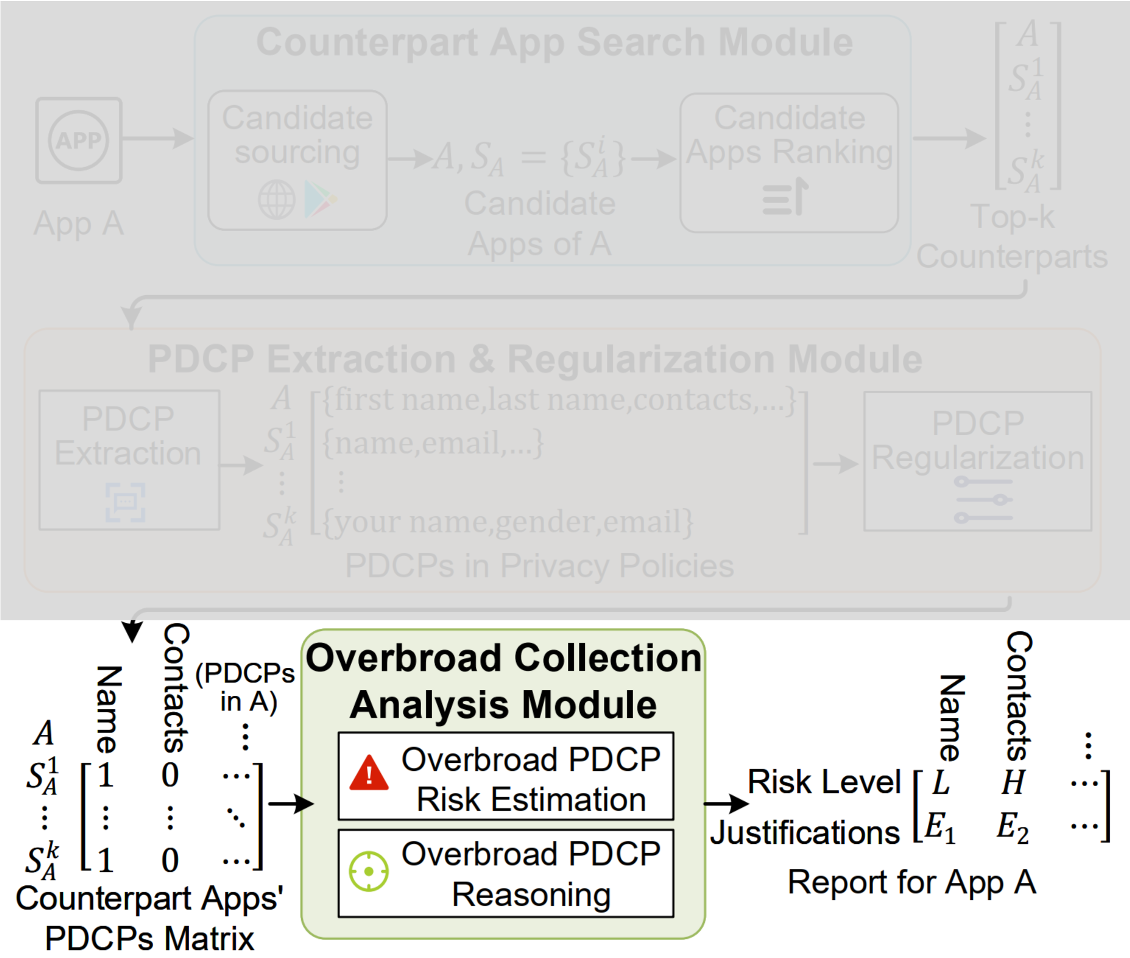
□ NLP-based PDCP extraction

- Sentence parsing
- Potential PDCPs extraction

□ PDCP regularization



System Design

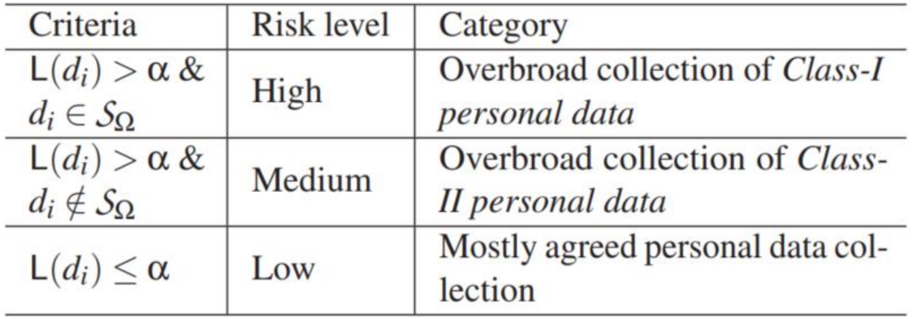


☐ Overbroad likelihood

the ratio of counterparts that do not collect d_i:

$$\mathsf{L}(d_i) = \frac{1}{k} \sum_{m=1,\cdots,k} \begin{cases} 1, & \text{if } d_i \notin \mathsf{D}(P_m) \\ 0, & \text{otherwise} \end{cases}$$

☐ Overbroad PDCP risk estimation



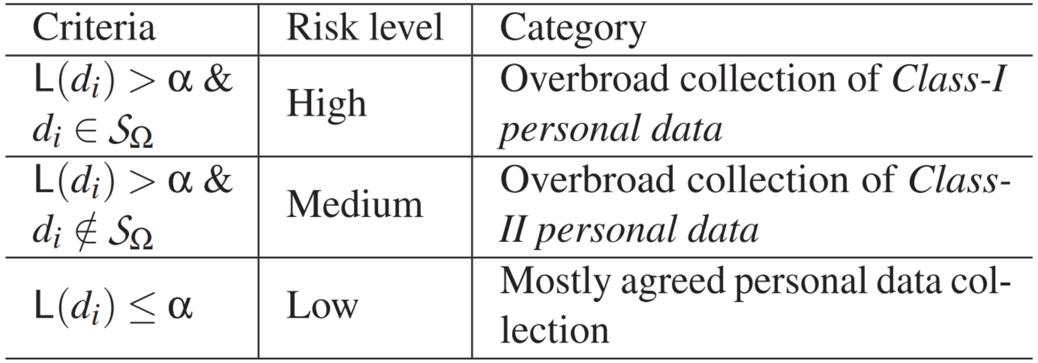
Criteria	Risk level	Category
$\mathsf{L}(d_i) > \alpha$ & $d_i \in \mathcal{S}_\Omega$	High	Overbroad collection of *Class-I personal data*
$\mathsf{L}(d_i) > \alpha$ & $d_i \notin \mathcal{S}_\Omega$	Medium	Overbroad collection of *Class-II personal data*
$\mathsf{L}(d_i) \leq \alpha$	Low	Mostly agreed personal data collection

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: 871 / 57,993 = 1.5%
  - medium-risk: 27, 132 / 57,993 = 46.79%



Criteria	Risk level	Category
$\mathsf{L}(d_i) > \alpha$ & $d_i \in \mathcal{S}_\Omega$	High	Overbroad collection of *Class-I personal data*
$\mathsf{L}(d_i) > \alpha$ & $d_i \notin \mathcal{S}_\Omega$	Medium	Overbroad collection of *Class-II personal data*
$\mathsf{L}(d_i) \leq \alpha$	Low	Mostly agreed personal data collection

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.

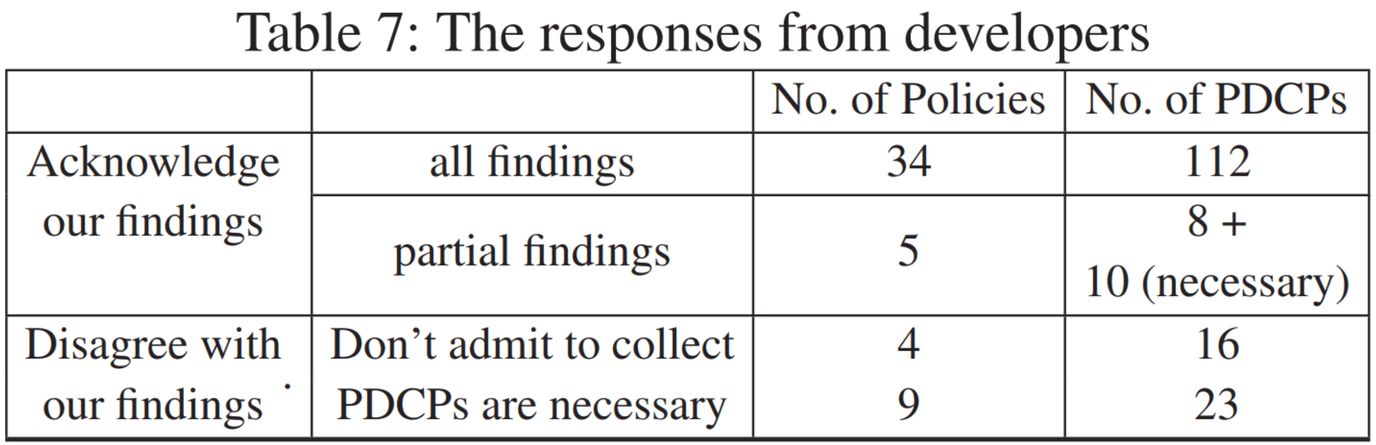


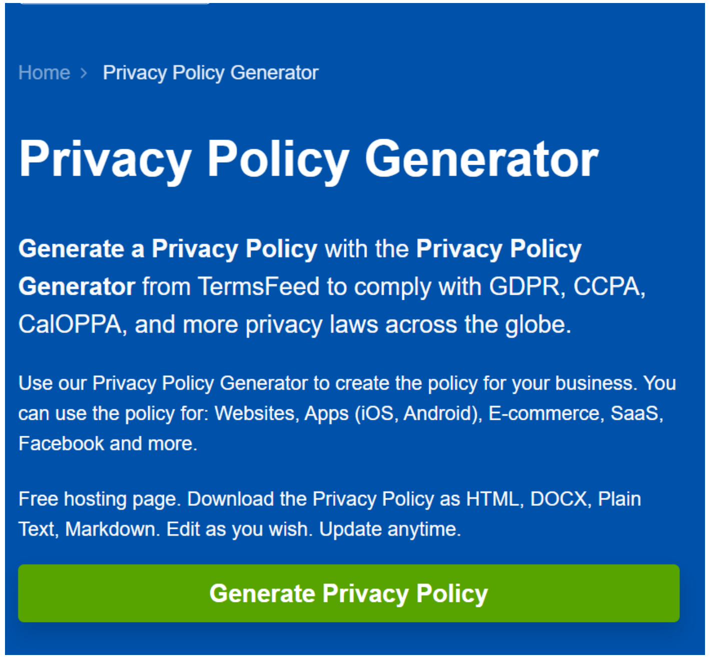
Table 7: The responses from developers

		No. of Policies	No. of PDCPs
Acknowledge our findings	all findings	34	112
	partial findings	5	8 + 10 (necessary)
Disagree with our findings	Don’t admit to collect	4	16
	PDCPs are necessary	9	23

Case Study I

Privacy Policy Generators

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



Home › Privacy Policy Generator

Privacy Policy Generator

Generate a Privacy Policy with the **Privacy Policy Generator** from TermsFeed to comply with GDPR, CCPA, CalOPPA, and more privacy laws across the globe.

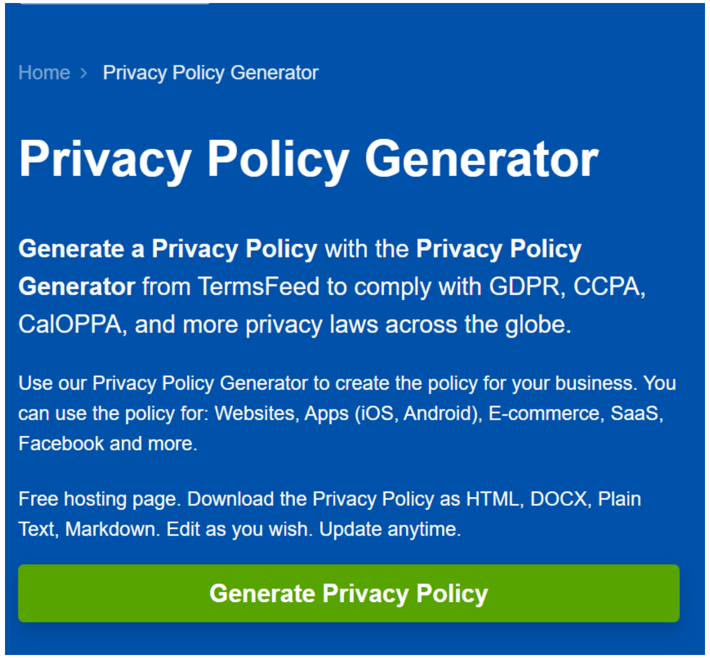
Use our Privacy Policy Generator to create the policy for your business. You can use the policy for: Websites, Apps (iOS, Android), E-commerce, SaaS, Facebook and more.

Free hosting page. Download the Privacy Policy as HTML, DOCX, Plain Text, Markdown. Edit as you wish. Update anytime.

Case Study I

Privacy Policy Generators

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



Privacy Policy Generator

Generate a Privacy Policy with the **Privacy Policy Generator** from TermsFeed to comply with GDPR, CCPA, CalOPPA, and more privacy laws across the globe.

Use our Privacy Policy Generator to create the policy for your business. You can use the policy for: Websites, Apps (iOS, Android), E-commerce, SaaS, Facebook and more.

Free hosting page. Download the Privacy Policy as HTML, DOCX, Plain Text, Markdown. Edit as you wish. Update anytime.

(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