PoliGraph: Automated Privacy Policy Analysis using Knowledge Graphs

Hao Cui, Rahmadi Trimananda, Athina Markopoulou, Scott Jordan
University of California, Irvine

Motivation

Privacy laws, such as the CCPA and GDPR, require organizations to provide **privacy policies** to explain their data collection practices. The CCPA gives consumers the *right to know*:

- The categories of personal information being collected (*data types*)
- The categories of third parties with whom personal information is shared (*entities*)
- The business / commercial purpose for collecting personal information (*purposes*)

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

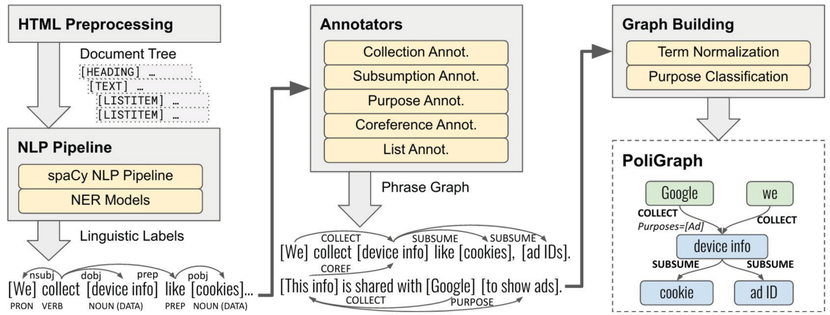
We use your ***personal information***... to:
- ***Provide the Services***...
- ***Authenticate your account***...

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

(An excerpt from KAYAK's privacy policy)

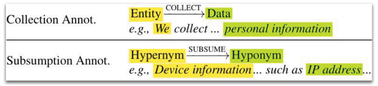
Privacy policy analysis is vital for auditing data collection practices. Manual analysis requires experts and is hard to scale up. Thus, researchers apply natural language processing (NLP) to extract information in a privacy policy.

Implementation



PoliGraph-ER is our NLP-based system that generates PoliGraph from policy text.

- The HTML preprocessor turns privacy policies in HTML format into a tree structure which preserves headings and lists. The NLP pipeline utilizes the tree structure to concatenate complete sentences and assign English linguistic features to words.
- Annotators match syntactic patterns to identify relations between phrases. The modular design allows each to focus on particular relations or syntactic patterns.

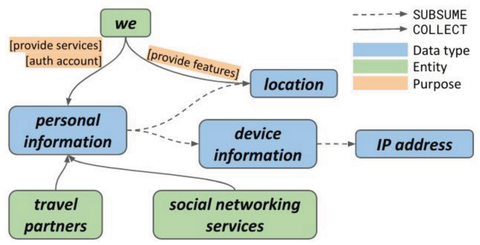


- Resolving coreferences (e.g., *this information* → *location*) is a difficult NLP task. The coreference annotator uses our own approach that is optimized for privacy policies.
- Term normalization and purpose classification turn phrases into canonical forms (e.g., *contact details*→*contact info.*, *to provide ads*→*advertising*) to allow automatic analysis.

Methodology

PoliGraph: We propose to extract and encode the information disclosed in a privacy policy into a *knowledge graph* that represents relations between terms.

- Two types of nodes: *data types* and *entities*.
- *COLLECT* edges: An entity is disclosed to collect a data type.
- *SUBSUME* edges: A generic term subsumes a more specific term.
- *Purposes* as attributes of a *COLLECT* edge.

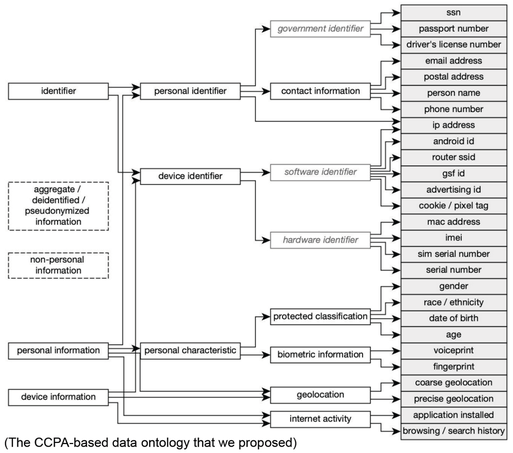


Beyond what is captured by individual nodes, edges, and attributes, PoliGraph allows to make inferences about indirect relations:

- Subsumption: **subsume**(*personal info.*, *location*)
- Collection: **collect**(*social networking service*, *location*)
- Set of purposes: **purposes**(*we*, *location*) = {*provide features*}

Ontologies are hierarchical data structures that define subsumption relations between terms (data types or entities).

- In a PoliGraph, *SUBSUME* edges encode subsumption relations as defined within a particular policy – **local ontologies**. However, the internal definitions in a policy can be *incomplete* or even *misleading*.
- We designed **global ontologies** based on authoritative sources to encode external knowledge as ground truth.



(The CCPA-based data ontology that we proposed)

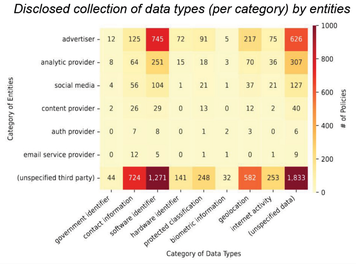
Applications

Evaluation: PoliGraph-ER identifies 61% more collection statements (i.e. *who* collects *what*) than prior work (i.e. PolicyLint), with over 90% precision.

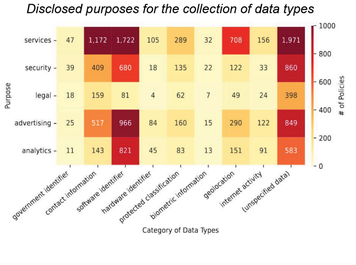
Policies summarization: We generate PoliGraphs for 5,518 privacy policies of Android apps and reveal common patterns across them:

- We find that 64% of policies disclose the collection of *software identifiers* (in particular, *cookies*).
- *Advertisers* and *analytics providers* are *major entities* that collect such data.
- Half of the policies disclose *data usage* for *advertising* and *analytics*. Potential use of sensitive data types for non-core purposes is concerning.
- The prevalent use of generic terms for data types (e.g., "personal information") without more precise definitions reduces the transparency and leaves the specific data types being collected *unknown*.

Disclosed collection of data types (per category) by entities



Disclosed purposes for the collection of data types



Definitions of terms: By comparing local ontologies with the CCPA-based global data ontology, we find that many policies use definitions that are misleading.

- e.g., *"… ask to share non-personal information … which may include your age or date of*

Hypernym	Hyponym (# Policies)
non-personal info.	geolocation (112), ip address (112), device identifier (97), gender (76), application installed (68), age (67), advertising id (55), imei (20), cookie / pixel tag (20), coarse geolocation (19), android id (19), internet activity (18), mac address (14), date of birth (14)
aggregate/deidentified/ pseudonymized info.	ip address (88), device identifier (84), geolocation (74), browsing / search history (14)
contact information	gender (12)
internet activity	ip address (18)
geolocation	ip address (70), router ssid (12), postal address (10)
personal identifier	advertising id (76), cookie / pixel tag (52), device identifier (45), age (27), geolocation (25), date of birth (21)

Other applications: We revisit two applications that were studied by prior work:

- **Contradicting statements**: We use PoliGraph to detect contradictions in a privacy policy and show false positives by prior work.
- **System audit**: We use PoliGraph to check the consistency between policies and network traffic, where we identify many more clear disclosures than prior work (i.e., PoliCheck).

ProperData Frontier 2nd Annual Symposium
Dec 1-2, 2022 | Irvine, California

SaTC Frontiers: Collaborative: *1956393* [1], *1955227* [2], *2103439* [3], *1956435* [4]
Institutions: UC Irvine [1], Northeastern Univ. [2], UC Davis [3], USC [4]
Lead PIs: A. Markopoulou [1], D. Choffnes [2], Z. Shafiq [3], K. Psounis [4]