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 type).
- The categories of third parties with whom personal information is shared (entities).
- The business/commercial purpose for collecting personal information (purposes).

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.

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.
- Subsumes edges: A generic term subsumes a more specific term.

Purposes as attributes of a_collect edge.

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

- In a Poligraph, subsumes 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.

Implementation

HTML Preprocessing

NLP Pipeline

Annotators

Graph Building

- HTML preprocess removes HTML tags and text (DOM) to expose entities (names, numbers, terms).

- NER Models identify instances of data type and entities.

- Phrase Graphs are constructed to capture relations.

- Annotators match syntactic patterns to identify relations between phrases. The modular design allows each to focus on particular relations or syntactic patterns.

- Graph building creates an object graph by aligning the facts.

Poligraph-er is our NLP-based system that generates Poligraph from policy text.

- The HTML preprocess removes HTML formatted into a tree structure which preserves headings and lists. The NLP pipeline utilizes the tree structure to concatenate complete sentences and assigns 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.

Applications
Evaluation: Poligraph-er identifies 61% more collection statements (i.e., who collects what) than prior work (i.e., PolicyLinter), 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.

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 age or date of birth…”

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 disclaimers than prior work (i.e., PoliCheck).

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 age or date of birth…”

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 disclaimers than prior work (i.e., PoliCheck).