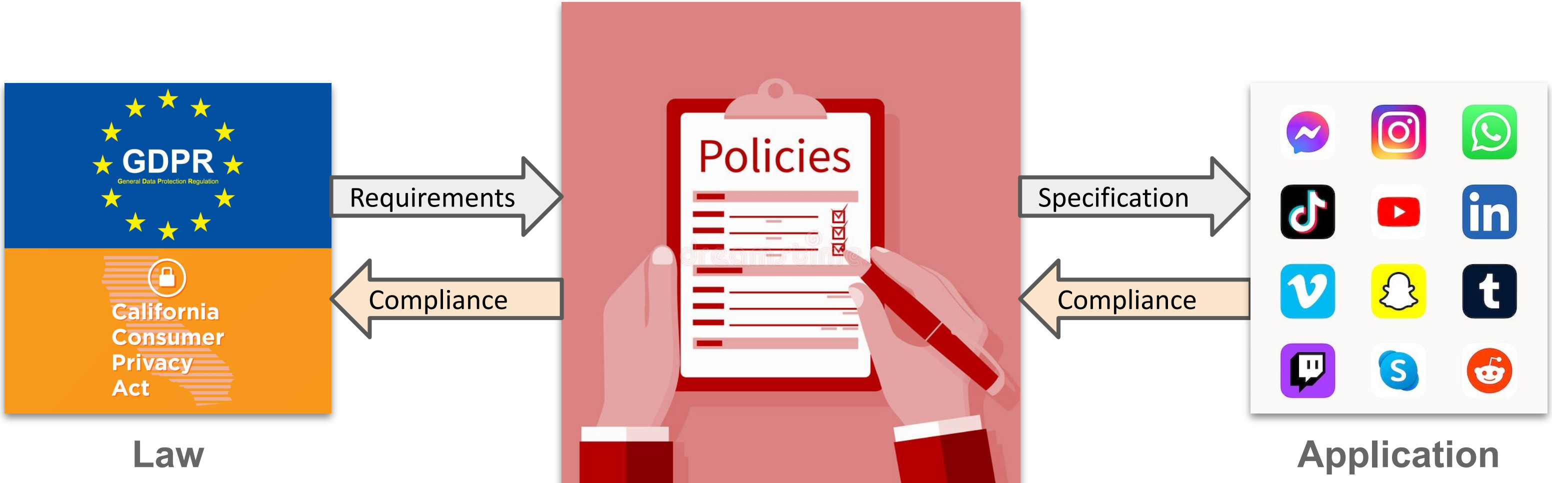
PoliGraph: Automated Privacy Policy Analysis using Knowledge Graphs

Hao Cui, Rahmadi Trimananda,
Athina Markopoulou, Scott Jordan

ProperData
PROTECTING PERSONAL DATA FLOW ON THE INTERNET

UCI University of California, Irvine

Privacy Policy



Privacy Policy

important for understanding & auditing data collection, sharing and use

Privacy Policy

Privacy policies are lengthy and complicated...

Opinion | THE PRIVACY PROJECT

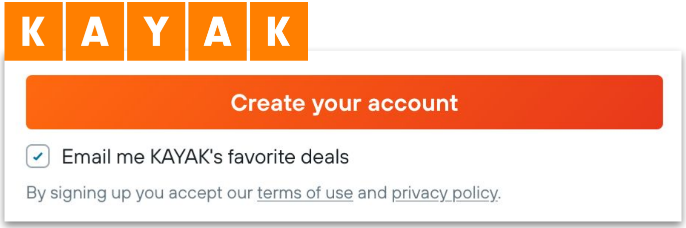
We Read 150 Privacy Policies. They Were an Incomprehensible Disaster.

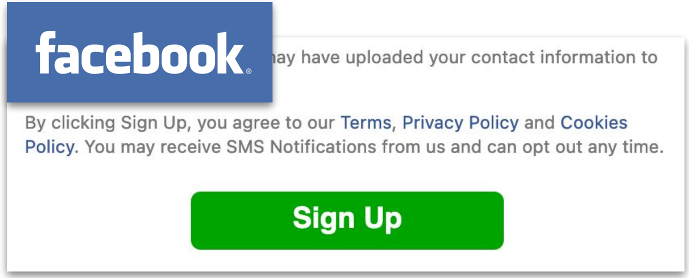
By Kevin Litman-Navarro

In the background here are several privacy policies from major tech and media platforms. Like most privacy policies, they're verbose and full of legal jargon — and opaquely establish companies' justifications for collecting and selling your data. The data market has become the engine of the internet, and these privacy policies we agree to but don't fully understand help fuel it.

Source: NYTimes

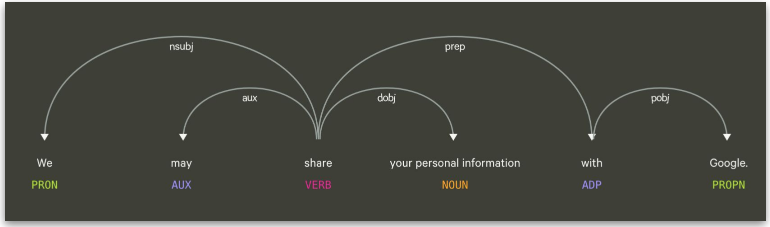
No one reads them...

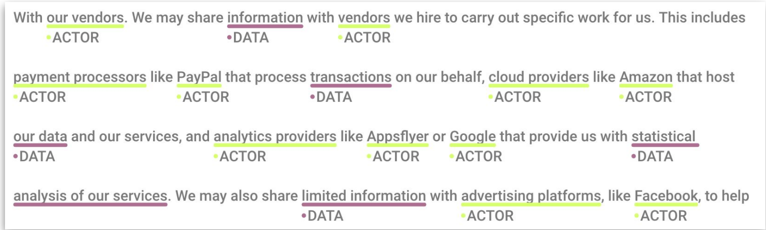


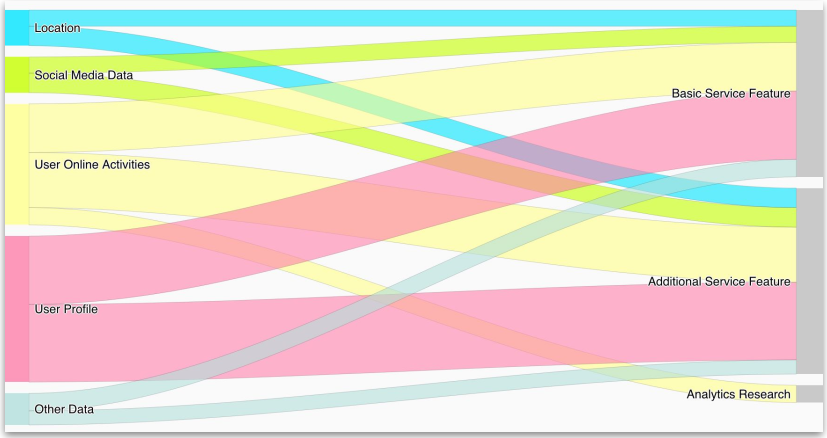


Related Work

NLP Analysis - Automated







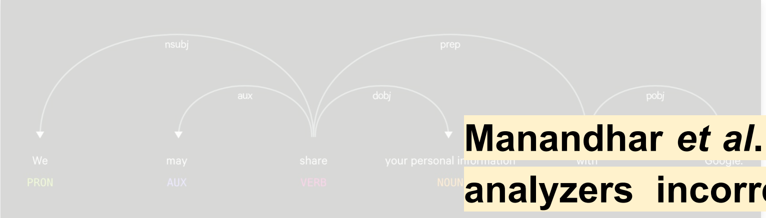
Polisis (USENIX 18)

(*we*, collect, *personal information*)
(*google*, collect, *personal information*)
(*we*, collect, *ip address*)
(*we*, collect, *location*)
(*we*, collect, *this information*)
……

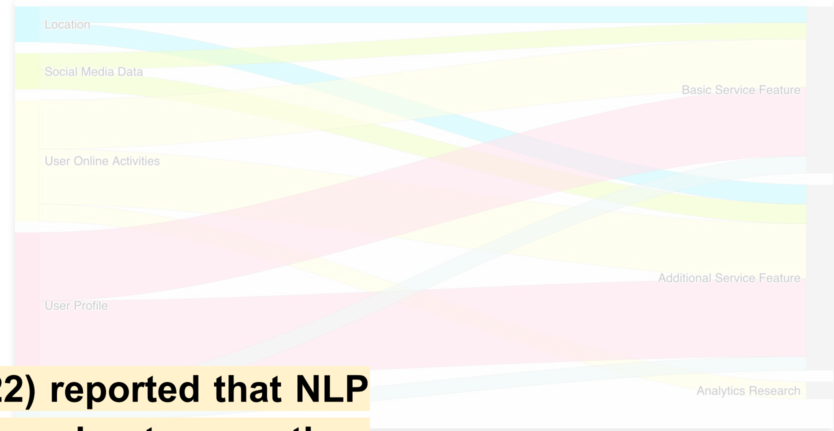
PolicyLint, PoliCheck (USENIX 19/20)

Related Work

NLP Analysis - Automated



Manandhar *et al*. (USENIX 22) reported that NLP analyzers incorrectly reason about more than half of the privacy policies they analyzed.



(*we*, collect, *personal information*)
(*google*, collect, *personal information*)
(*we*, collect, *ip address*)
(*we*, collect, *location*)
(*we*, collect, *this information*)
……

PolicyLint, PoliCheck (USENIX 19/20)

Outline

- ❑ PoliGraph Framework
- ❑ PoliGraph-ER Implementation
- ❑ Evaluation
- ❑ Applications

Privacy Policy Example

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 info. is shared (*entities*)
- The business / commercial purpose for collecting personal info. (*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***...

Limitation of Prior Work

Limitation #1: Missing context.

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***...

Labels (Polisis)

<location> *<generic personal info>* *<device info>*
<basic services> *<security>* *<third-party use>*

What data, shared with whom, for what purposes?

Disconnected Tuples (PolicyLint, PurPliance)

(*we*, collect, *personal information*)
(*we*, collect, *ip address*)
(*we*, collect, *location*)
(*we*, collect, *this information*) [*provide features*]
(*travel partners*, collect, *personal information*) ...

What specific data type do travel partners collect?
What information is used to provide features?
What do "personal / this info." mean?

PoliGraph Framework

Key Idea: Analyze each privacy policy as a whole. Extract and encode information disclosed in one privacy policy (data types / entities / purposes) into a *knowledge graph* - PoliGraph.

We collect the following categories of ***personal information***:

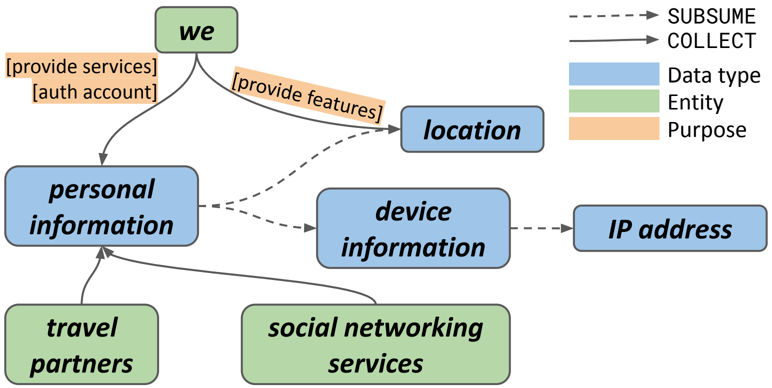
- ***Device information***... such as ***IP address***...
- ***Location***. ***We*** use ***this information*** to ***provide features***...

We use your ***personal information***... to:

- ***Provide the Services***...
- ***Authenticate your account***...

We disclose the ***personal information***... as follows:

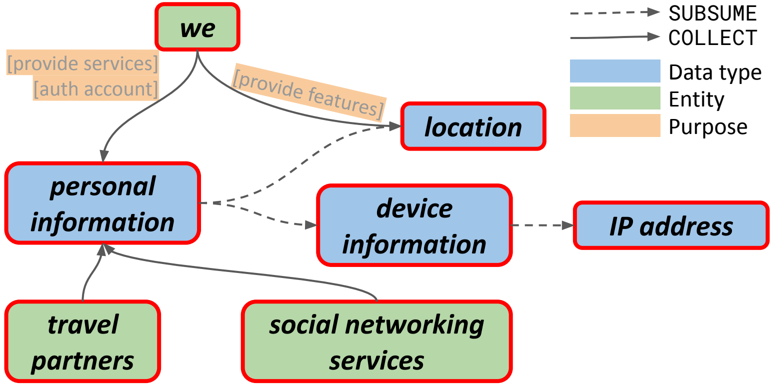
- With our ***travel partners***...
- With ***social networking services***...



PoliGraph Framework

PoliGraph

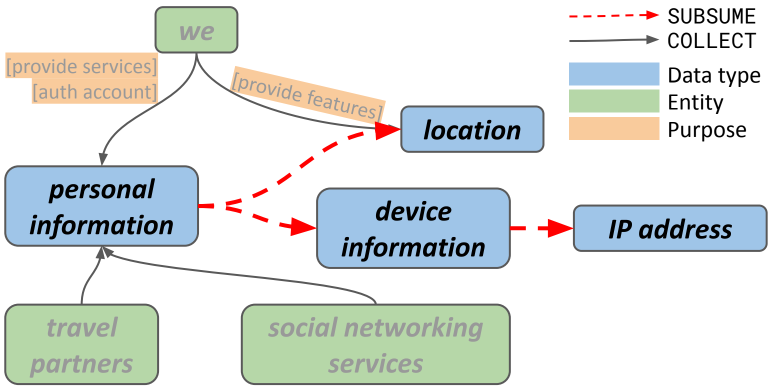
- Data types, entities as nodes.



PoliGraph Framework

PoliGraph

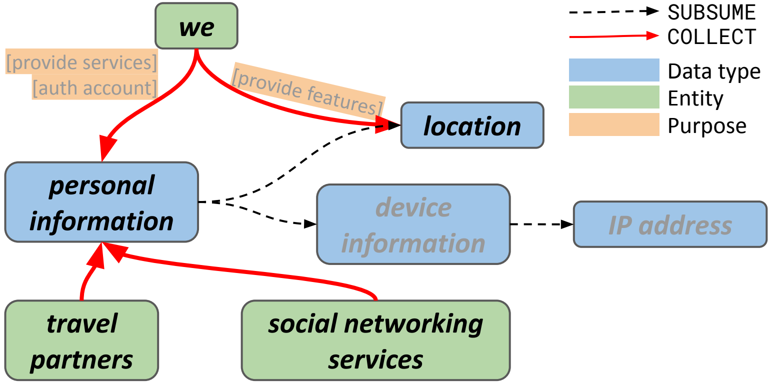
- Data types, entities as nodes.
- Two kinds of relations as edges:
    - SUBSUME (generic term -> specific term)



PoliGraph Framework

PoliGraph

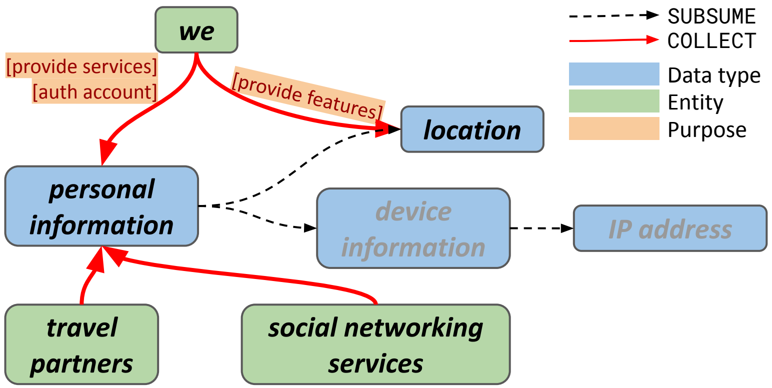
- Data types, entities as nodes.
- Two kinds of relations as edges:
  - SUBSUME (generic term -> specific term)
  - COLLECT (entity -> data type)



PoliGraph Framework

PoliGraph

- Data types, entities as nodes.
- Two kinds of relations as edges:
    - SUBSUME (generic term -> specific term)
    - COLLECT (entity -> data type)
- Purposes (of collection) as edge attributes.



PoliGraph Framework

PoliGraph

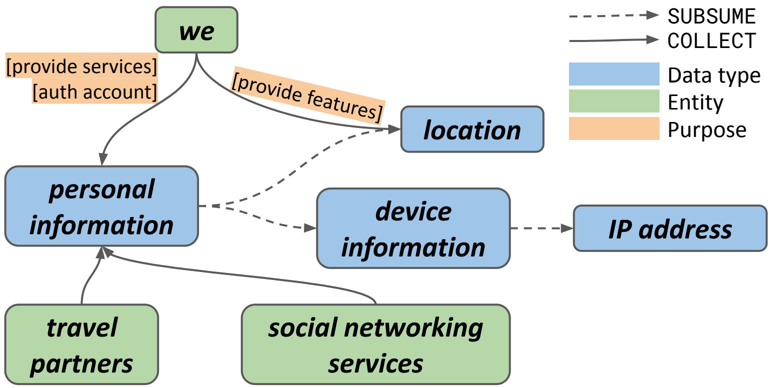
- Data types, entities as nodes.
- Two kinds of relations as edges:
  - SUBSUME (generic term -> specific term)
  - COLLECT (entity -> data type)
- Purposes (of collection) as edge attributes.

Inferences on a PoliGraph

Definition 2.2. ***Subsumption Relation.*** In a PoliGraph G, we say that a term t_1 (hypernym) *subsumes* another term t_2 (hyponym), denoted as $subsume(t_1, t_2)$, *iff* there exists a path from t_1 to t_2 in G where every edge is a *SUBSUME* edge.[4]

Definition 2.3. ***Collection Relation.*** In a PoliGraph G, we say an entity $n \in N$ *collects* a data type $d \in D$, denoted as $collect(n, d)$, *iff* there exists an entity $n' \in N$ and a data type $d' \in D$ where $subsume(n', n) \wedge subsume(d', d)$[5] and edge $n' \xrightarrow{\text{COLLECT}} d'$ exists.

Definition 2.4. ***Set of Purposes.*** Following Definition 2.3, if a purpose $p \in Purposes(n' \xrightarrow{\text{COLLECT}} d')$, we say n collects d for the purpose p. We denote the set of all instances of such p in G as a set $purposes(n, d)$.



PoliGraph Framework

PoliGraph

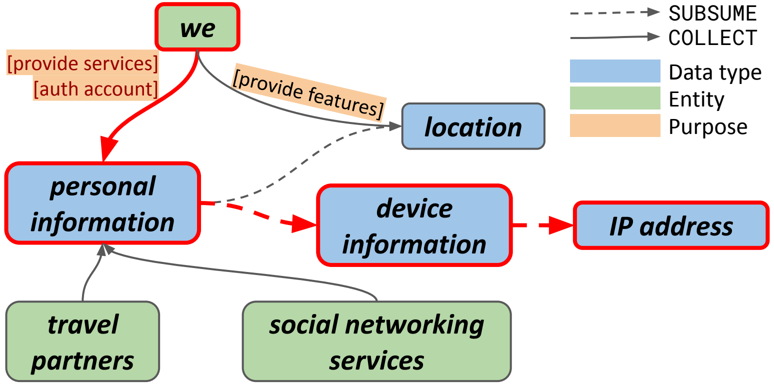
- Data types, entities as nodes.
- Two kinds of relations as edges:
  - SUBSUME (generic term -> specific term)
  - COLLECT (entity -> data type)
- Purposes (of collection) as edge attributes.

Inferences on a PoliGraph

Definition 2.2. ***Subsumption Relation.*** In a PoliGraph G, we say that a term t_1 (hypernym) *subsumes* another term t_2 (hyponym), denoted as $subsume(t_1,t_2)$, *iff* there exists a path from t_1 to t_2 in G where every edge is a *SUBSUME* edge.[4]

Definition 2.3. ***Collection Relation.*** In a PoliGraph G, we say an entity $n \in N$ *collects* a data type $d \in D$, denoted as $collect(n,d)$, *iff* there exists an entity $n' \in N$ and a data type $d' \in D$ where $subsume(n',n) \wedge subsume(d',d)$[5] and edge $n' \xrightarrow{\text{COLLECT}} d'$ exists.

Definition 2.4. ***Set of Purposes.*** Following Definition 2.3, if a purpose $p \in Purposes(n' \xrightarrow{\text{COLLECT}} d')$, we say n collects d for the purpose p. We denote the set of all instances of such p in G as a set $purposes(n,d)$.



collect(we, IP address)

purposes(·) = {provide features, auth account}

PoliGraph Framework

PoliGraph

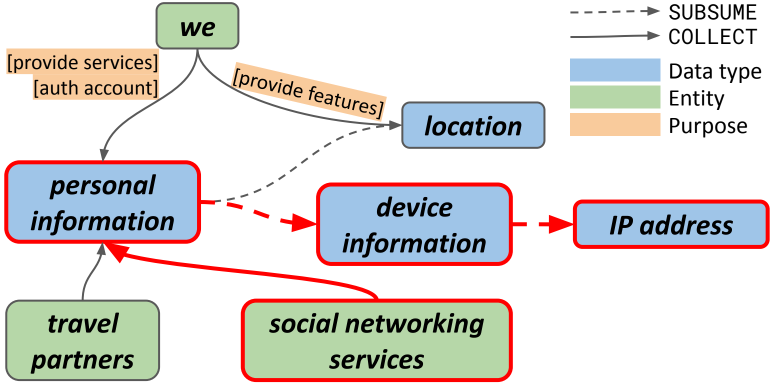
- Data types, entities as nodes.
- Two kinds of relations as edges:
  - SUBSUME (generic term -> specific term)
  - COLLECT (entity -> data type)
- Purposes (of collection) as edge attributes.

Inferences on a PoliGraph

Definition 2.2. ***Subsumption Relation.*** In a PoliGraph G, we say that a term t_1 (hypernym) *subsumes* another term t_2 (hyponym), denoted as $subsume(t_1, t_2)$, *iff* there exists a path from t_1 to t_2 in G where every edge is a *SUBSUME* edge.[4]

Definition 2.3. ***Collection Relation.*** In a PoliGraph G, we say an entity $n \in N$ *collects* a data type $d \in D$, denoted as $collect(n, d)$, *iff* there exists an entity $n' \in N$ and a data type $d' \in D$ where $subsume(n', n) \land subsume(d', d)$[5] and edge $n' \xrightarrow{\text{COLLECT}} d'$ exists.

Definition 2.4. ***Set of Purposes.*** Following Definition 2.3, if a purpose $p \in Purposes(n' \xrightarrow{\text{COLLECT}} d')$, we say n collects d for the purpose p. We denote the set of all instances of such p in G as a set $purposes(n, d)$.

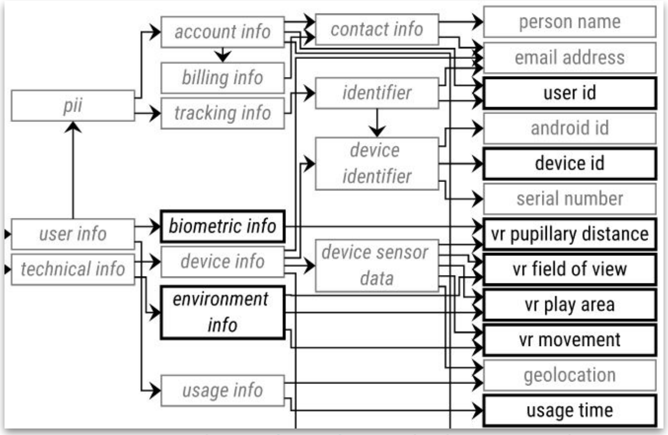


collect(social networking services, IP address)

Limitation of Prior Work

Limitation #2: Ontologies are used to relate generic and specific terms. But these ontologies are not universal, resulting in ambiguous or wrong interpretations.

(*we*, collect, ***personal information***)



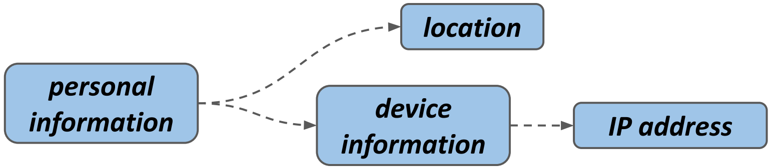
Data ontology, from OVRSeen (USENIX 22)

personal information
-> *device information* -> *IP address*
-> *location*
-> ……
policy 1

personal information
-> *contact information* -> *phone number*
non-personal information
> *device information* -> *IP address*
policy 2

Revisit Ontologies

- SUBSUME edges naturally induce ontologies – *Local ontologies*

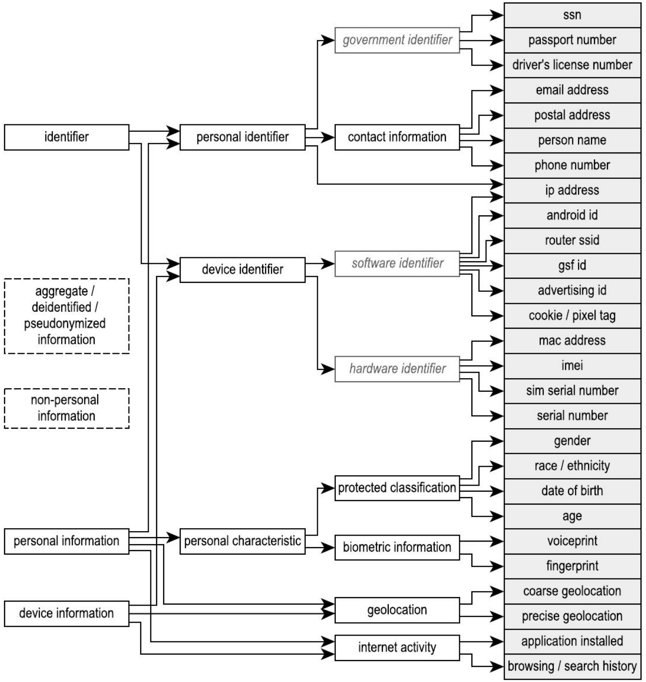


  - But privacy policies are often not clear enough: what if just “device information”?
  - Some definitions can be misleading: “non-personal information”.

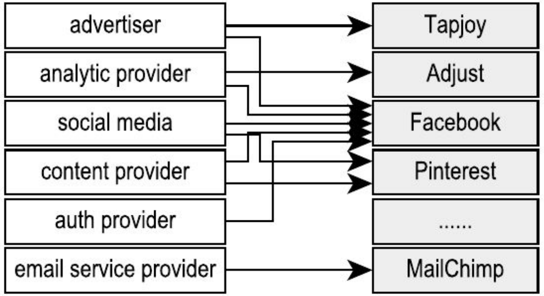
- External knowledge or ground truth - *Global ontologies*
  - Like the ones in prior work, but based on authoritative sources.

Global Ontologies

The CCPA-based data ontology



The entity ontology based on public datasets



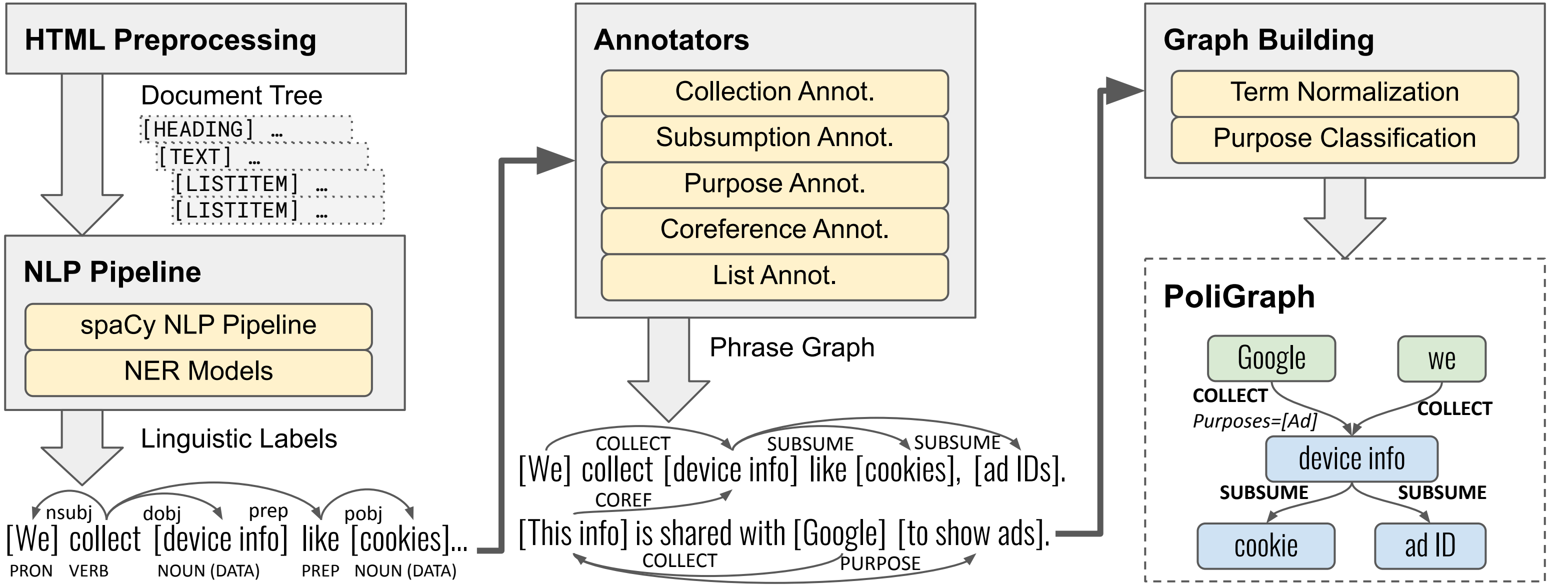
Other designs can be used with PoliGraph as well.

Outline

- PoliGraph Framework
- PoliGraph-er Implementation
- Evaluation
- Applications

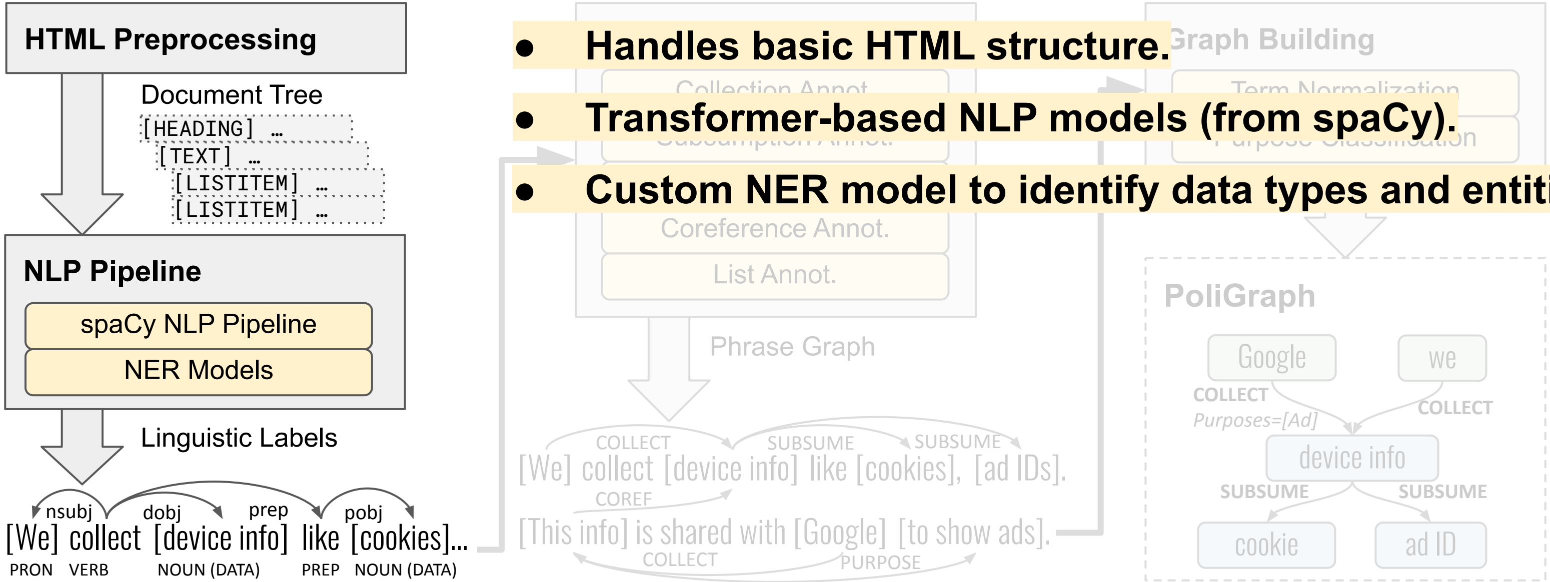
PoliGraph-er Implementation

The NLP-based system to generate PoliGraph.



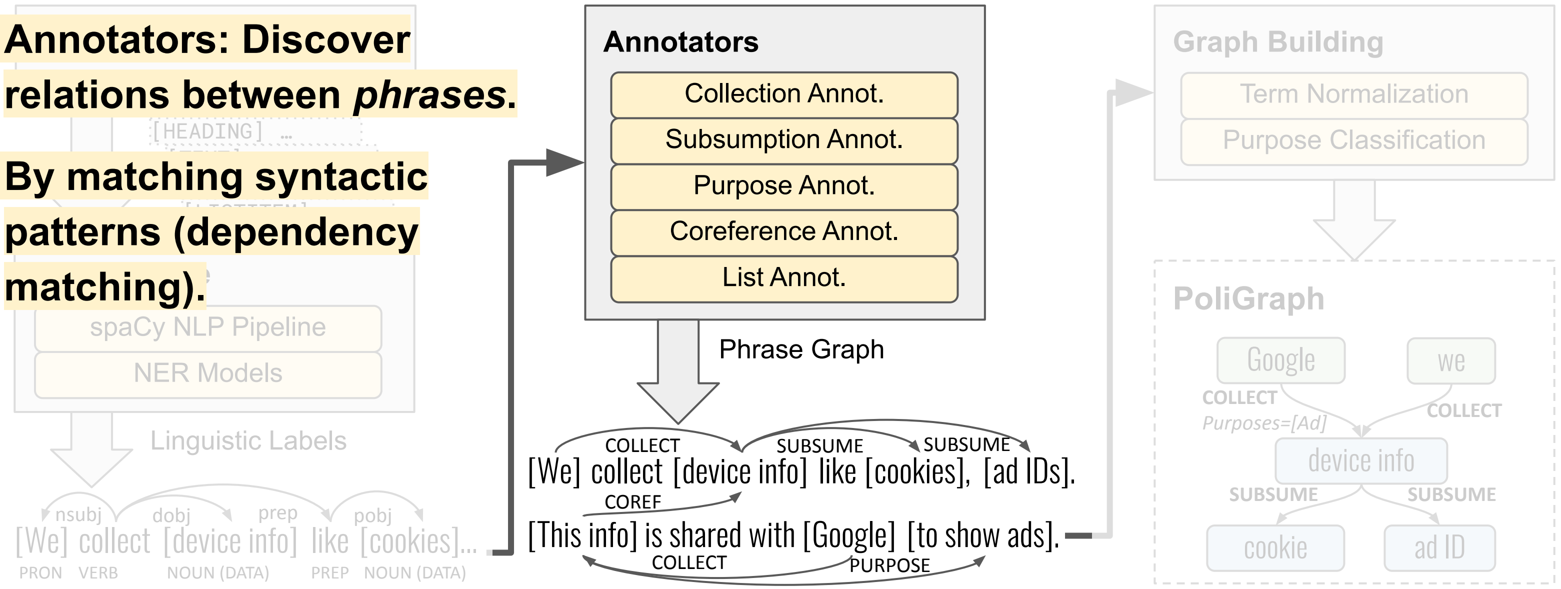
PoliGraph-ER Implementation

- **Handles basic HTML structure.**
- **Transformer-based NLP models (from spaCy).**
- **Custom NER model to identify data types and entities.**



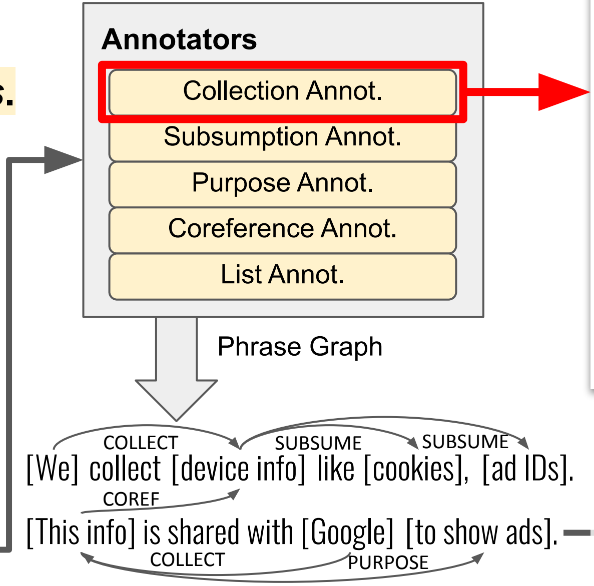
PoliGraph-ER Implementation

- **Annotators: Discover relations between *phrases*.**
- **By matching syntactic patterns (dependency matching).**



PoliGraph-ER Implementation

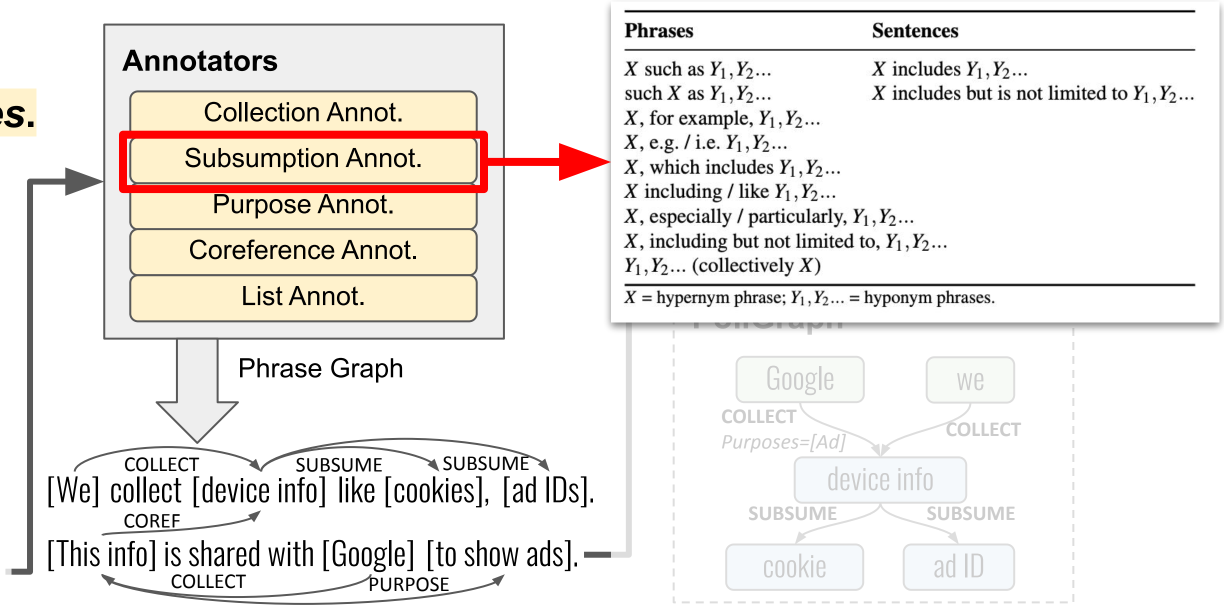
- **Annotators: Discover relations between *phrases*.**
- **By matching syntactic patterns (dependency matching).**



Root Verbs (*Examples: ENTITY* $\xrightarrow{\text{COLLECT}}$ *DATA*)	**Syntatic Patterns**
share, trade, exchange (*We share your device IDs with Google.*)	ENTITY:nsubj DATA:dobj with,ENTITY:pobj
collect, gather, obtain, get, receive, solicit, acquire (*Google may collect your device IDs.*)	ENTITY:nsubj DATA:dobj
provide, supply (*We provide Google with your device IDs.*)	ENTITY:nsubj ENTITY:dobj with,DATA:pobj
provide, supply, release, disclose, transfer, transmit, sell, give, pass, divulge (*We may transmit device IDs to Google.*)	ENTITY:nsubj DATA:dobj to,ENTITY:pobj
use, keep, access, analyze, process, store, save, log, utilize, record, retain, preserve, need (*Google may use your device IDs.*)	ENTITY:nsubj DATA:dobj
have, get (*Google has access to your device IDs.*)	ENTITY:nsubj access,to,DATA:pobj
make (*Google makes use of device IDs.*)	ENTITY:nsubj use:dobj of,DATA:pobj
enable, allow, permit, authorize, ask, require, permit (*This enables Google to collect your device IDs.*) (*You authorize Google to collect your device IDs.*)	(compounded with above patterns)

PoliGraph-ER Implementation

- **Annotators: Discover relations between *phrases*.**
- **By matching syntactic patterns (dependency matching).**



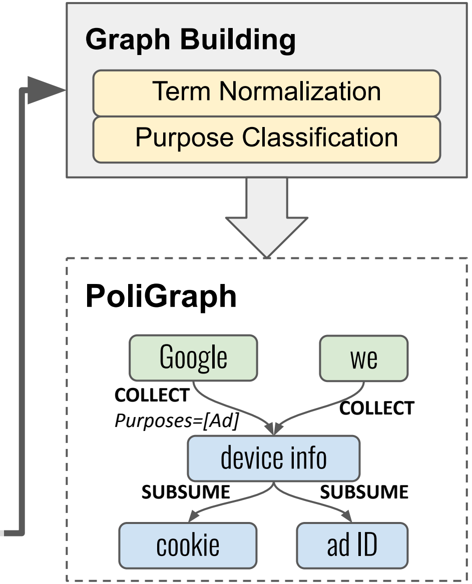
Phrases	Sentences
X such as Y_1, Y_2...	X includes Y_1, Y_2...
such X as Y_1, Y_2...	X includes but is not limited to Y_1, Y_2...
X, for example, Y_1, Y_2...	
X, e.g. / i.e. Y_1, Y_2...	
X, which includes Y_1, Y_2...	
X including / like Y_1, Y_2...	
X, especially / particularly, Y_1, Y_2...	
X, including but not limited to, Y_1, Y_2...	
Y_1, Y_2... (collectively X)	

X = hypernym phrase; Y_1, Y_2... = hyponym phrases.

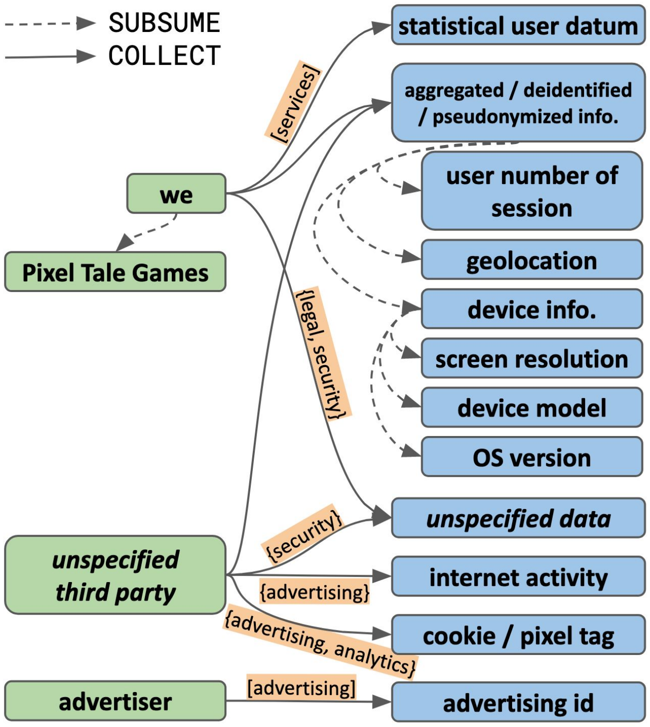
PoliGraph-ER Implementation

Normalize and merge *phrases* into PoliGraph nodes

- **Term normalization**
  - e.g. contact details / contact data / … -> contact information
- **Purpose classification**
  - Text classification:
  - Labels: services, security, legal, advertising, analytics



PoliGraph Example



A PoliGraph for the privacy policy of *Pixel Tale Games*

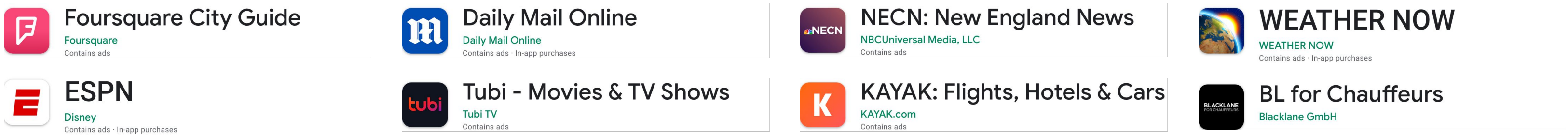
A typical PoliGraph can contain up to **hundreds** of nodes and edges.

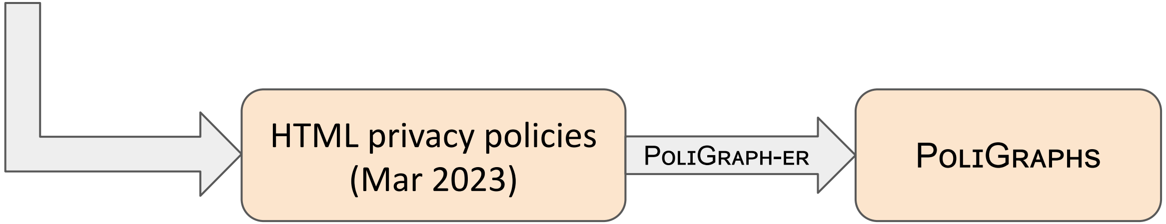
Outline

- ❑ PoliGraph Framework
- ❑ PoliGraph-ER Implementation
- ❑ Evaluation
- ❑ Applications

Dataset

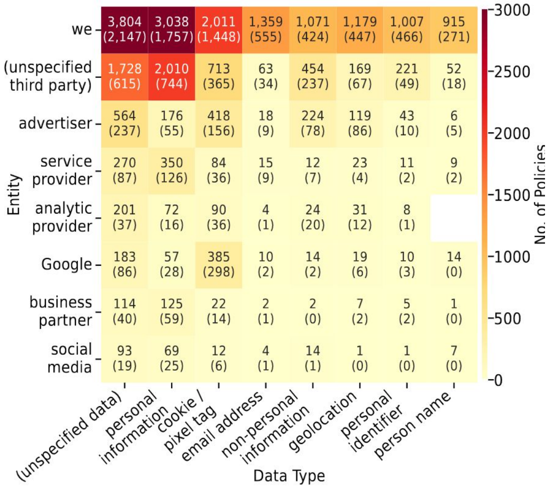
PoliCheck dataset: 6,084 unique privacy policies used by 13,626 Android apps.



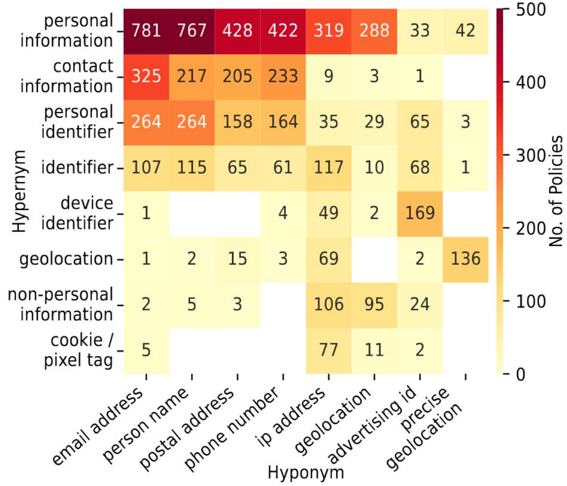


PoliGraph Edges

COLLECT edges: 90.4% precision



SUBSUME edges: 87.7% precision



Most false positives are caused by **NLP errors** (e.g. recognizing irrelevant phrases).

Comparison w/ Prior Work

- **Method**: PoliGraph collect(n, d) relations v.s. PolicyLint tuples (n, collect, d)

	# tuples	precision	recall
Ground truth	878		
PolicyLint	291	91.8%	30.4%
PoliGraph-ER	640	**96.9%**	**70.6%**

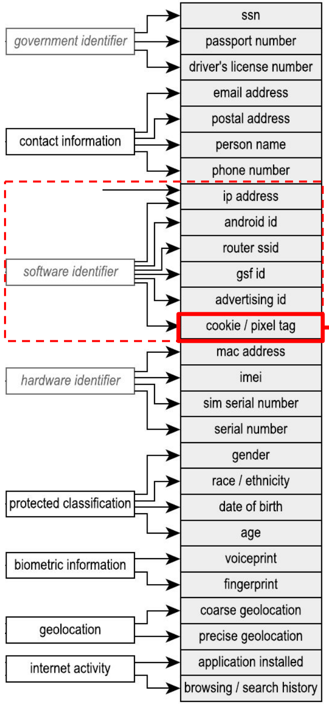
- More relations (30%→70%) are covered.

Outline

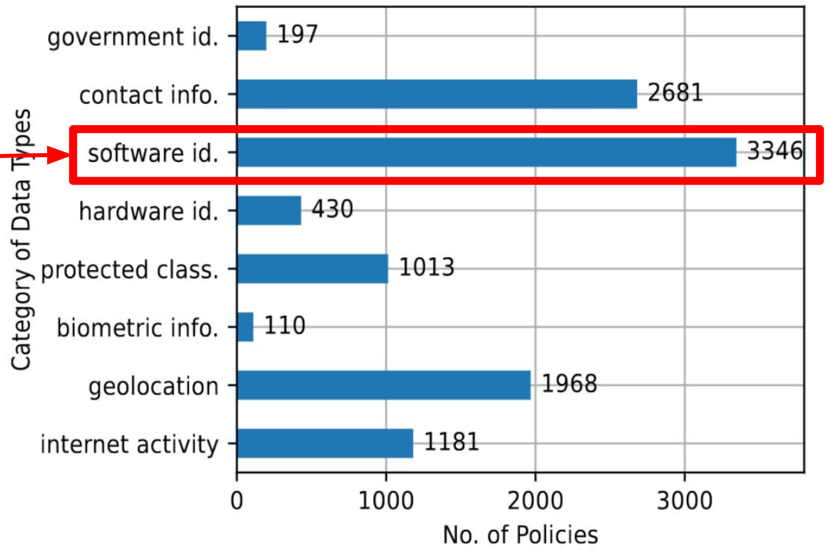
- PoliGraph Framework
- PoliGraph-ER Implementation
- Evaluation
- Applications

Application #1: Policies Summarization

- Reveal common patterns across policies in the dataset.
    - Using global ontologies to categorize data types / entities.

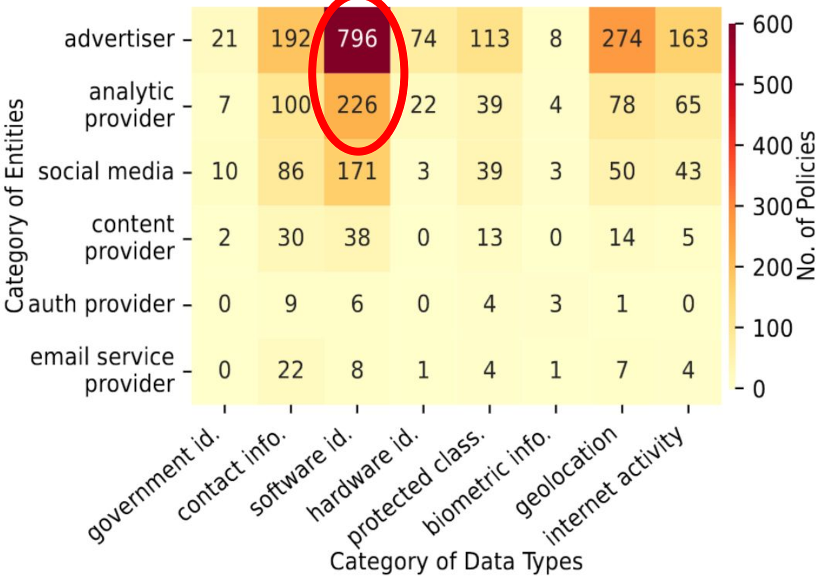


2,688 collect cookies



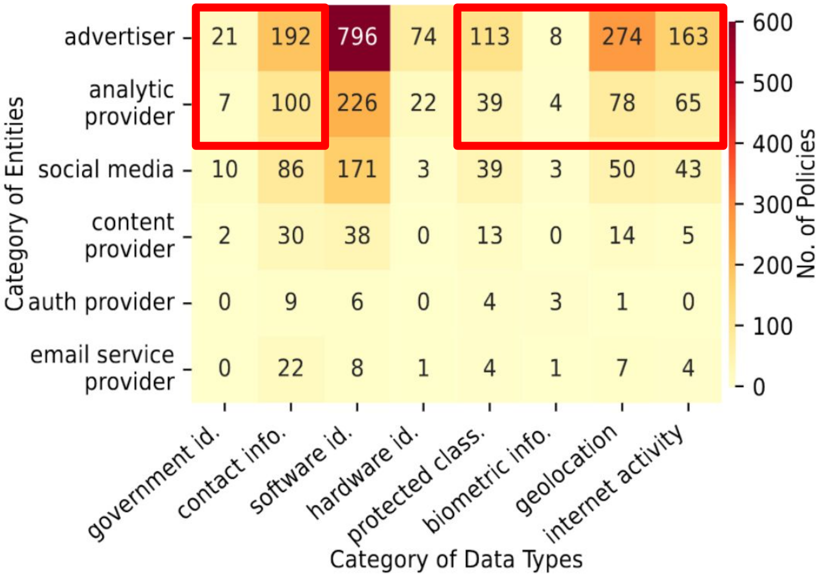
Application #1: Policies Summarization

Frequently shared with
advertisers / analytic providers



Application #1: Policies Summarization

Alarming, often due to the use of blanket terms,
e.g., sharing "personal information".



Application #2: Correct Definitions of Terms

- Identify possible misleading definitions
  - e.g., "We collect non-personal information, such as geolocation..."
  - By comparing the local ontology against the global ontology.

Hypernym	**Hyponym (# Policies)**
non-personal info.	ip address (126), geolocation (123), device identifier (108), gender (76), application installed (72), age (70), identifier (46), internet activity (44), device information (38), coarse geolocation (35) ...
aggregate/deidentified/ pseudonymized info.	ip address (122), device identifier (89), geolocation (78), browsing / search history (16) ...
internet activity	ip address (151), device identifier (107), geolocation (40), advertising id (13), cookie / pixel tag (10) ...
geolocation	ip address (76), postal address (15), router ssid (10) ...

Application #2: Correct Definitions of Terms

- "Non-standard" terms
  - Data types that are frequently used, but not in our global data ontology.

Missing definitions in many privacy policies.

Term (# Policies)	**Possible definitions found in policies**
technical info. (311)	*From 126 policies:* advertising id, age, android id, browsing / search history, cookie / pixel tag, device identifier, email address, geolocation, imei, ip address, mac address ...
profile info. (178)	*From 17 policies:* age, contact information, date of birth, email address, gender, geolocation, person name, phone number ...
demographic info. (315)	*From 112 policies:* age, browsing / search history, date of birth, email address, gender, geolocation, ip address, postal address, precise geolocation, race / ethnicity, router ssid ...
log data (81)	*From 52 policies:* advertising id, android id, cookie / pixel tag, coarse geolocation, cookie / pixel tag, email address, geolocation, imei, ip address, mac address, person name ...

Broad and varied definitions across different policies.

Revisiting Known Applications

Contradiction Analysis

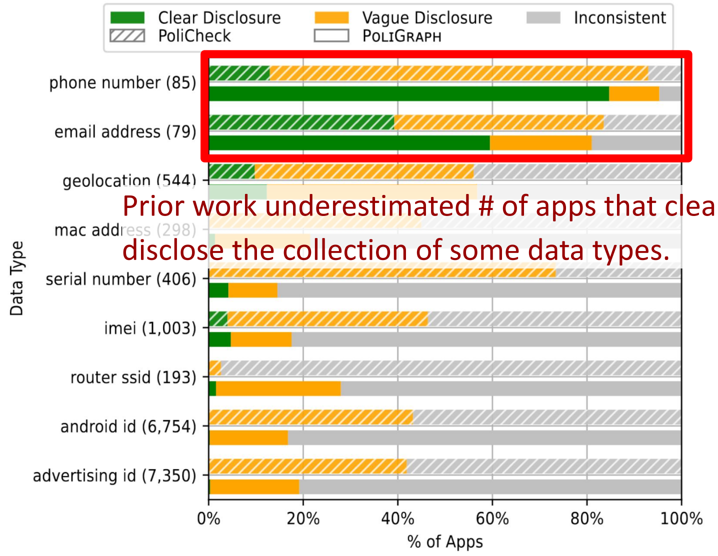
- Studied by PolicyLint (USENIX 19)

	# pairs of edges	
Invalid*	183	(11.7%)
Non-conflicting parameters	731	(46.7%)
Different purposes	*114*	*(7.3%)*
Different data subjects	*121*	*(7.7%)*
Different actions	*624*	*(39.8%)*
Contradictions according to PolicyLint's ontologies	441	(28.2%)
Conflicting edges	211	(13.5%)
Total	[illegible]	

By taking additional contexts into account, we avoided false alarms in prior work.

Data Flow-to-Policy Consistency

- Studied by PoliCheck (USENIX 20)



Prior work underestimated # of apps that clearly disclose the collection of some data types.

Summary

- **PoliGraph Framework** – Encoding a privacy policy as a knowledge graph.
- **PoliGraph-er Implementation** – The NLP system to generate PoliGraphs.
- **Evaluation** – Significantly higher recall than prior work.
- **Applications**
  - Policies summarization: Revealing common patterns across many privacy policies.
  - Term definitions: Assessing the correctness of definitions w.r.t. global ontologies.
  - Revisiting known applications.

Open Source: https://github.com/UCI-Networking-Group/PoliGraph
Extended Paper: arxiv:2210.06746
Funding Ack.: NSF (ProperData), UC Noyce Initiative

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED