Automated Expansion of Privacy Data Taxonomy for Compliant Data Breach Notification

Yue Qin*
Indiana University Bloomington &
Central University of Finance and Economics
qinyue@cufe.edu.cn

Yue Xiao*
Indiana University Bloomington &
IBM Research
xiaoyue@ibm.com

Xiaojing Liao
Indiana University Bloomington
xliao@iu.edu

***Abstract*—In privacy compliance research, a significant challenge lies in comparing specific data items in actual data usage practices with the privacy data defined in laws, regulations, or policies. This task is complex due to the diversity of data items used by various applications, as well as the different interpretations of privacy data across jurisdictions. To address this challenge, privacy data taxonomies have been constructed to capture relationships between privacy data types and granularity levels, facilitating privacy compliance analysis. However, existing taxonomy construction approaches are limited by manual efforts or heuristic rules, hindering their ability to incorporate new terms from diverse domains. In this paper, we present the design of GRASP, a scalable and efficient methodology for automatically constructing and expanding privacy data taxonomies. GRASP incorporates a novel hypernym prediction model based on granularity-aware semantic projection, which outperforms existing state-of-the-art hypernym prediction methods. Additionally, we design and implement *Tracy*, a privacy professional assistant to recognize and interpret private data in incident reports for GDPR-compliant data breach notification. We evaluate *Tracy* in a usability study with 15 privacy professionals, yielding high-level usability and satisfaction.**

I. Introduction

In the field of privacy compliance, a key research question is: how can we check whether the actual data usage practices comply with those stated in privacy law, regulation, or policies? The challenge researchers must address lies in *comparing* the specific data items in the actual data usage practices with the private data defined in the privacy law, regulation, or policies. This task is nontrivial due to the considerable diversity in data items employed by different applications. Examples include security-critical data such as "password" and "social security number", as well as Software Development Kit (SDK)-specific sensitive data, such as "device topic", which is used in IoT for identifying user devices, and "IDFA", defined by Facebook for advertisers to precisely track users. Privacy data outlined in the privacy statements, on the other hand, also differ in law, regulation, or policies across different jurisdictions. For instance, the Children's Online Privacy Protection Act (COPPA) explicitly classifies "screen names" and "usernames" as personal data to be protected, while the General Data Protection Regulation (GDPR) acknowledges online identifiers as personal data, which can include information like IP addresses, cookies, and device IDs, and does not explicitly mention "screen names" or "usernames" as specific categories of personal data. In our study, we refer to both protected data in data usage practice and privacy law, regulation, or policies as "restricted data".

Such divergent interpretations of restricted data add complexity to the privacy compliance checking process. To tackle this challenge, several privacy data taxonomies [1], [2], [3] have been constructed to capture the relationships between the restricted data regarding their types (e.g., "location information" vs. "financial information") and granularity levels (e.g., "coarse location" vs. "precise location"). Specifically, a private data taxonomy establishes a hierarchical architecture where each restricted data is linked with its subconcepts or instances through hypernym semantic relationships, as shown in Figure 1. This hierarchical architecture provides a semantically rich interpretation of restricted data, enabling a better understanding of their relationships, types, and levels of granularity to facilitate privacy compliance analysis [4], [5], [6], [1], [2], [7]. However, to the best of our knowledge, existing privacy data taxonomies have been primarily constructed based on manual efforts [4], [6], [8], [9], [5] or heuristic rules [10], [1], [11], limiting their ability to incorporate previously-unknown terms into the taxonomy, especially from corpora in different domains. Therefore, a scalable and efficient methodology is in need for automatically constructing and expanding privacy data taxonomies, facilitating a more robust and accurate assessment of privacy data practices.

Restricted data identification and integration. In this paper, we present the first automatic method to identify restricted data and integrate it into an existing privacy data taxonomy. This goal has been made possible through the application of hypernym prediction techniques, which evaluate the existence of hypernym relationships between a data object and the existing restricted data in the privacy data taxonomy.

*These authors contributed equally to this work, ordered alphabetically. This work was completed while Yue Qin and Yue Xiao were Ph.D. students at Indiana University Bloomington.

Network and Distributed System Security (NDSS) Symposium 2025
24-28 February 2025, San Diego, CA, USA
ISBN 979-8-9894372-8-3
https://dx.doi.org/10.14722/ndss.2025.230230
www.ndss-symposium.org

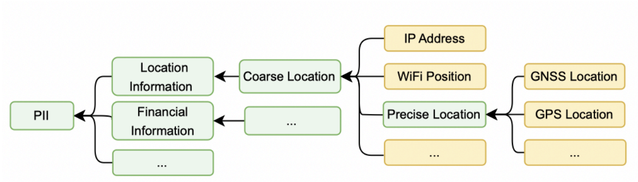


Fig. 1: Example of a privacy data taxonomy. The arrow represents hypernym semantic relationship.

While hypernym prediction techniques have been extensively studied in the Machine Learning community, existing hypernym prediction techniques *cannot* be directly applied for restricted data identification and integration. Specifically, we observed that existing approaches overlook the granularity levels between different hypernym relationships. For instance, this problem occurs when a hypernym prediction model (e.g., lexical pattern-based methods [12], [13], [10], [14], [11], distributional representation-based methods [15], [16], [17], [18], [19] and projection-based methods [20], [21], [22]) might learn that a specific data type, such as "location" or "financial information", is a strong indicator, but it fails to capture the varying granularity levels between restricted data, e.g., "precise location" and "coarse location" when determining their hypernym relationships. In the context of restricted data identification and integration, granularity plays a critical role in accurately characterizing the sensitivity and significance of restricted data. It provides a finer distinction between different restricted data, allowing for a more precise comparison and analysis during privacy compliance analysis [4], [5], [6], [1], [2], [7], [23], [24], for example, ensuring the limited and appropriated collection of personal information for specified purposes (i.e., data minimization principle [25], [26]).

To address this challenge, we present a novel granularity-aware hypernym prediction model called GRASP, which captures granularity information from taxonomy structure to guide hypernym prediction. GRASP recognizes that in the privacy data taxonomy, upper-level restricted data (e.g., "location information", "coarse location") generally represent more overarching concepts, compared to lower-level terms (e.g., "precise location", "GPS location"), resulting in larger granularity differences between corresponding pairs. This insight enables GRASP to separate the learning process for different granularity levels, such as pairs involving "coarse location" and "precise location" as the hypernym. Our evaluations on a prototype we built show that this new model is effective: GRASP achieves the precision of 97.4% and 91.0%, the recall of 95.2% and 89.1%, and F1 of 96.3% and 89.9%, in the evaluation experiments on two existing privacy data taxonomies (a privacy data taxonomy built on privacy policies [1] and a domain-specific privacy data taxonomy built on IoT privacy documentation [3], § IV-A), respectively. In the meanwhile, GRASP outperforms state-of-the-art hypernym prediction methods.

***Tracy*: a privacy professional assistant for GDPR-compliant data breach notification**. We further demonstrate the effectiveness of GRASP via a legal-technical application *Tracy* for privacy-compliant data breach notification task. Privacy laws and regulations (e.g., GDPR, CCPA) mandate that companies experiencing a data breach must notify the affected individuals at risk [25], [26]. In the data breach notification process, privacy professionals should identify the type of data involved in the data breach and determine whether it falls under the category of private data as defined in the privacy laws and regulations. However, prior studies [27], [28], [29] have highlighted challenges faced by privacy professionals in this task, primarily due to semantic heterogeneity: the incident reports generated by technical professionals during data breach notification, and the regulatory requirements outlined in privacy laws, may employ diverse terminologies and notations, making it difficult to establish clear mappings between the two, thus complicating the identification of privacy data for privacy law and regulation compliance.

To enhance the efficiency and explainability of data breach notification process, we design and implement *Tracy*, a privacy professional assistant to facilitate private data recognition for GDPR-compliant data breach notification. The key functionality of *Tracy* involves leveraging GRASP to identify restricted data from the data breach incident reports. The identified terms are then presented to privacy professionals through a graphical visualization of a privacy data taxonomy that connects them to the corresponding privacy data defined in GDPR. *Tracy* aids privacy professionals in determining and understanding private data within the breach reports and their compliance implications with GDPR, thereby streamlining the reporting process and promoting compliance with privacy regulations.

The usability of *Tracy* was evaluated through a user study involving 15 privacy professionals to assess 12 real-world data breach incident reports. Using *Tracy*, a total of 89 restricted data were successfully extracted from the incident reports, and for each restricted data, the corresponding partial taxonomy was presented. The results of the user study demonstrated that *Tracy* significantly contributed to the participants' understanding and evaluation of the identified restricted data in data breach notification process.

Contributions. The main contributions of this paper are summarized below.

- We design and implement GRASP for automatically constructing and expanding privacy data taxonomy and facilitating privacy compliance analysis. GRASP incorporates a novel hypernym prediction model based on granularity-aware semantic projection, which outperforms existing state-of-the-art hypernym prediction methods.
- We design and implement *Tracy*, a privacy professional assistant to recognize and interpret private data in incident reports for GDPR-compliant data breach notification. We evaluate *Tracy* in a usability study with 15 privacy professionals, yielding high-levels of product usability and satisfactions.
- We release the code and data at [30].

II. Background

Privacy-compliant data breach notification. Privacy-compliant data breach notification refers to the process of handling and notifying affected individuals about data breaches in a manner that adheres to relevant privacy laws, regulations, and policies. It involves understanding and adhering to the specific requirements and guidelines set forth in privacy laws such as GDPR and CCPA, or other relevant data protection regulations in different jurisdictions. For instance, under GDPR Art. 33, organizations are obligated to report data breaches promptly and transparently when personal data has been compromised, resulting in a risk to individuals. In GDPR Art. 4(1), personal data are broadly defined as any information that is related to an identified or identifiable natural person, typically including confidential data or personally identifiable information (PII).

In the process of data breach reporting, privacy professionals play a primary role in identifying and categorizing privacy data, assessing potential privacy harms (e.g., financial and reputation harm, or identity theft), and determining their compliance implications. However, a challenge they face is that the private data contained in data breach incident reports are sometimes described using technical and domain-specific language [27], [28], [29], which differs from the privacy concepts defined in regulations such as GDPR and CCPA. For example, GDPR Art. 4(1) broadly defines personal data as any information that is related to an identified or identifiable natural person, typically including confidential data or personally identifiable information (PII). On the other hand, data breach incident reports often describe domain-specific leaked data, such as “IDFA” in iOS SDKs, “MQTT device topics” in IoT (see § V-A), creating a disconnect between the terminology used in the reports and the legal definitions in privacy regulations. This semantic heterogeneity makes it difficult for privacy professionals to establish clear mappings between the technical language used in the incident reports and the legal terminology defined in privacy regulations. As a result, accurately identifying and interpreting privacy-sensitive data becomes a complex task, leading to potential gaps in compliance and privacy protection. In our study, we develop *Tracy* to aid privacy professionals in automatically identifying and interpreting privacy data in incident reports, facilitating the data breach reporting process, and ensuring compliance with privacy regulations.

Privacy data taxonomy. As illustrated in Figure 1, privacy data taxonomy is a hierarchical structure that categorizes and organizes different restricted data [31], [1], [2], [3]. In our study, we model the Privacy Data Taxonomy as a rooted tree $\mathcal{T}$ constructed from a set of known hyponym-hypernym pairs (u, v). In the Privacy Data Taxonomy $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, $\mathcal{V} = \{v\}$ is a set of known restricted data and $\mathcal{E} = \{(u, v)|u \in \mathcal{V}, v \in \mathcal{V}\}$ are directed edges representing the hypernymy relationships, each connecting a hyponym u to its hypernym v. To enhance an existing taxonomy $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ by incorporating additional restricted data q from a corpus $\mathcal{D}$, the process involves inserting restricted data q into $\mathcal{T}$ and expanding it to form a more comprehensive taxonomy $\mathcal{T}' = (\mathcal{V}', \mathcal{E}')$, where $\mathcal{V}' = \mathcal{V} \cup \{q\}$ and $\mathcal{E}' = \mathcal{E} \cup \{(q, v)\}$ with (q, v) being the newly-discovered hypernymy relationship between q and known restricted data v in $\mathcal{T}$.

Privacy data taxonomy plays a crucial role in privacy compliance analysis. It facilitates the detection of non-compliance by employing an inconsistency detection logic to identify discrepancies between the data flow associated with data object d and its corresponding privacy statement linked to data d' in privacy laws, regulations, and policies. The correlation between d and d' is modeled using semantic relations, including synonym ($d \equiv d'$), hyponym ($d \sqsubset d'$), and hypernym ($d \sqsupset d'$). These semantic relationships have been captured by the privacy data taxonomy, enabling a comprehensive privacy compliance analysis.

Existing privacy data taxonomies have been predominantly constructed through manual efforts [4], [6], [8], [9], [5] or heuristic rules [10], [1], [11]. However, these approaches have limitations in incorporating previously-unknown restricted data into the taxonomy, especially when dealing with new and emerging data practices or domains. In this paper, we design and implement Grasp, which leverages hypernym prediction techniques to enable dynamic and scalable taxonomy construction. Note that our investigation takes the first step to explore the hierarchy between technical data terminology and the corresponding legal definition of data category. Identifying the hierarchy between various laws and regulations is out of the scope of this study.

Hypernym prediction techniques. Hypernymy refers to the linguistic relation between a concept or a general category of things (i.e., hypernyms), and its subcategories or instances (i.e., hyponyms) [32], [33]. Formally, if u is a general term and v is a specific term whose semantic meaning is included in u, we regard u as a hypernym of v (denoted by $u \sqsupset v$) and v as a hyponym of u (denoted by $v \sqsubset u$). Hypernym prediction is a natural language processing task that involves predicting the hypernym relationship between two words or terms. As in [21], [34], we model the hypernym prediction as a binary classification problem $F : \{(x, y)\} \rightarrow \{0, 1\}$, where 1 represents the hypernymy relation and 0 represents the non-hypernymy relation. Hypernym prediction is a fundamental component of taxonomy construction and expansion, as it helps establish hierarchical relationships between concepts. In our study, we present a novel hypernym prediction model specifically tailored for constructing privacy data taxonomies, addressing the unique challenges and requirements of this domain.

III. Grasp: Design and Methodology

A. Overview

The Grasp system consists of two main components: a hypernym prediction model (§ III-B), which determines whether a hypernym relationship exists between a data object and a known restricted data in the privacy data taxonomy, and a taxonomy expansion module (§ III-C), which integrates new restricted data into the existing privacy data taxonomy.

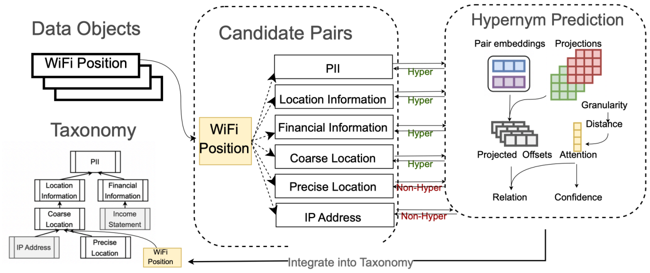


Fig. 2: Workflow of GRASP for restricted data identification and interpretation.

As illustrated in Figure 2, when a new data object is extracted from the corpus (e.g., data breach incident reports), GRASP assesses the hypernym relationship between the data object and existing restricted data in the taxonomy using the trained hypernym prediction model. If GRASP identifies at least one hypernym of the data object with high confidence, the data object will be recognized as a restricted data and aligned to the hypernym with the smallest granularity level to maintain the hierarchy of the taxonomy.

Example. We take `WiFi Position` as an example, which is an instance of coarse location[1] that approximates device location using the characteristics of nearby wireless access points (APs) connected by the device. To begin, GRASP traverses the taxonomy using Breadth First Search (BFS), generating and evaluating candidate pairs. Given the privacy data taxonomy in Figure 1, GRASP will first evaluate whether a hypernym relationship exists between the pair ⟨PII, WiFi Position⟩. If a hypernym relationship is identified, GRASP proceeds to generate candidate pairs between the successors of "PII", i.e., "location information", "financial information". Subsequently, the model recognizes that "location information" is the only hypernym of `WiFi Position` at the second level, and its successor "coarse location" is also the only hypernym at the third level. Also, none of the successors of "coarse location" (e.g., "precise location", "IP address") is determined to be a hypernym of `WiFi Position`. As a result, the model outputs `WiFi Position` as a newly-discovered restricted data, with its hypernym `Coarse Location`, and expands the taxonomy by the new hypernym pair.

B. Granularity-aware Hypernym Prediction

In our study, we identify the significance of granularity level information present in the privacy data taxonomy, which is not adequately captured by existing hypernym prediction models (§ IV-D). To address this limitation, we introduce the Granularity-aware Hypernym Prediction model, which leverages the structure of the privacy data taxonomy to capture and incorporate granularity information, enabling more accurate and informed hypernym prediction.

Specifically, the architecture of the proposed hypernym prediction model is shown in Figure 3. It first groups hypernym-hyponym pairs in the privacy data taxonomy into several clusters based on different granularity levels. After that, the model learns the projection matrix M from the hyponyms $\{x\}$ to the hypernyms $\{\mathrm{y}\}$ for each cluster, and then aggregates the projected offsets $Mx - y$ regarding the projections derived from all clusters. Finally, GRASP applies an attention-based classifier to evaluate the existence of the hypernym relationship between a candidate pair.

Granularity-aware clustering. The coherent hierarchy of privacy data taxonomy ensures that upper-level restricted data (e.g., "PII", "location information") represent more overarching concepts, compared to lower-level ones (e.g., "precise location", "IP address") that are close to the leaf nodes. Therefore, the granularity level of a restricted data v in the privacy data taxonomy can be implied by its position on the taxonomy. Considering that one restricted data can be involved in several paths in the taxonomy, we design the granularity of a restricted data as a value between $(0, 1)$ that captures its relative position

$$s_v = \frac{1}{\exp\{d_v - d'_v\} + 1},$$

where d_v is the depth (i.e., distance to the root) of v and d'_v represents the maximum length of the paths between v and the leaf nodes connected to v. This will assign a larger granularity to the restricted data who are close to the root of the taxonomy and a smaller granularity to those close to the leaf nodes. We further define the granularity of a hypernym semantic relationship between a restricted data v and a term as a tuple $(s_v, a_{(u,v)})$, where s_v is the granularity level of the restricted data and $a_{(u,v)})$ is the cosine distance between embeddings of token pairs.

GRASP guides the clustering process using the granularity to distinguish different hypernym or non-hypernym relations. Let $D^{(+)} = \{(u, v)^+\}$ denote the set of positive pairs with hypernym relation and $D^{(-)} = \{(u, v)^-\}$ denote the set of negative pairs with non-hypernym relation. We assume there are $K^{(+)}$ and $K^{(-)}$ granularities in hypernymy and non-hypernym relations, respectively. The positive pairs and the negative pairs are divided into $K^{(+)}$ and $K^{(-)}$ clusters, by applying KMeans [35] on $\{(s_v, a_{(u,v)^+})\}$ and $\{(s_v, a_{(u,v)^-})\}$, respectively.

Given the learned clusters, for each specific data pair (x_i, y_i), GRASP computes the granularity level of the relation $(s_{y_i}, a_{(x_i,y_i)})$ and its distance between the relation granularity to the centers of all clusters. With the distance $\eta^+(i, k)$ to the center of k-th positive cluster and the distance $\eta^-(i, k)$ to the center of k-th negative cluster, we define the attention weight that pair (x_i, y_i) is associate with k-th positive or negative

[1] The precision or coarseness of location data is determined based on the specific data entry it provides. Other specific aspect such as device signals is not within the scope of our current taxonomy. For example, WiFiLocation (i.e., geolocation of Wifi AccessPoint's MAC address) is categorized as coarse location information for identifying device users, and regulated by the ACCESS_COARSE_LOCATION permission in Android developer's guide.

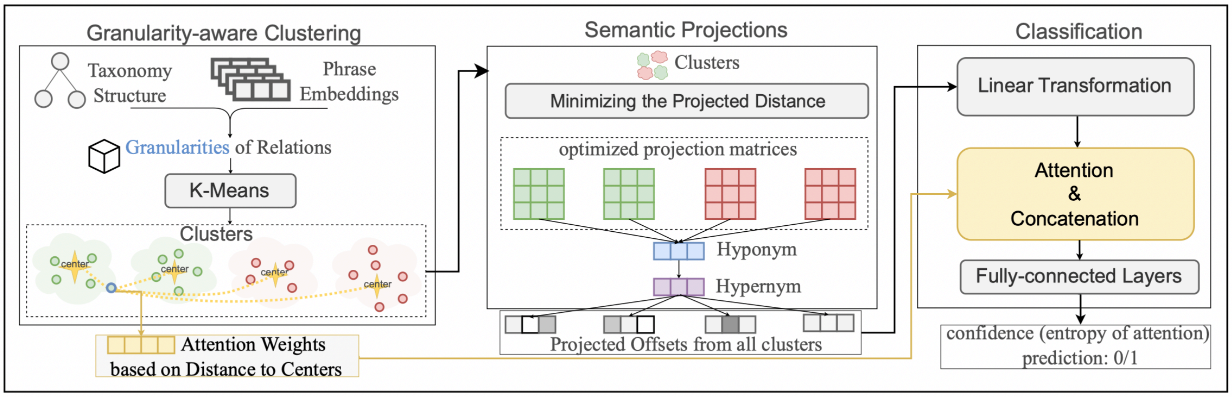

Granularity-aware Clustering
Taxonomy Structure
Phrase Embeddings
Granularities of Relations
K-Means
Clusters
center
center
center
center
Attention Weights based on Distance to Centers
Semantic Projections
Clusters
Minimizing the Projected Distance
optimized projection matrices
Hyponym
Hypernym
Projected Offsets from all clusters
Classification
Linear Transformation
Attention & Concatenation
Fully-connected Layers
confidence (entropy of attention)
prediction: 0/1

Fig. 3: Architecture of hypernym prediction model.

cluster as:

$$\omega^{+}_{i,k} = \frac{\frac{1}{\eta^{+}(i,k)}}{\sum_{j=1}^{K+} \frac{1}{\eta^{+}(i,j)}}, \qquad \omega^{-}_{i,k} = \frac{\frac{1}{\eta^{-}(i,k)}}{\sum_{j=1}^{K-} \frac{1}{\eta^{-}(i,j)}}.$$

These attention weights are used later in the classifier of GRASP to soften the projected offsets for each data pair.

Semantic projection. Prior research on hypernym prediction has shown the effectiveness of projection-based approaches [20], [21], [22], where the hypernym prediction model learns a projection matrix $\boldsymbol{M}$ from the hyponym x to the hypernym y such that $\boldsymbol{M}\,\vec{\boldsymbol{x}} \approx \vec{\boldsymbol{y}}$. Building on this line of research, GRASP learns an orthogonal projection matrix that maps the terms' embeddings from the hyponyms to the hypernyms or non-hypernyms while capturing the granularity of the relation. Specifically, let $D_k^{(+)} = \{(x_1, y_1)^+, (x_2, y_2)^+, \ldots, (x_{|D_k^{(+)}|}, y_{|D_k^{(+)}|})^+\}$ denote all token pairs in the k-th positive cluster. For $k \in \{1, 2, \ldots, K^{(+)}\}$, GRASP learns the projection matrix $\boldsymbol{M}_k^{(+)}$ associated with the k-th granularity in hypernym relation by minimizing the sum of the projected distance between each token pair in the k-th cluster:

$$\min_{\boldsymbol{M}_k} \sum_{i=1}^{|D_k^{(+)}|} \|\boldsymbol{M}_k^{(+)} \vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i^{(+)}\|_2^2 \quad \text{s.t. } \boldsymbol{M}_k^{(+)T} \boldsymbol{M}_k^{(+)} = \boldsymbol{I}.$$

This is equivalent with the Orthogonal Procrustes problem [36]

$$\boldsymbol{M}_k^{(+)} = \operatorname*{argmin}_{\Omega:\Omega^T\Omega=\boldsymbol{I}} \|\Omega \boldsymbol{X}_k^{(+)} - \boldsymbol{Y}_k^{(+)}\|_F,$$

where $\boldsymbol{X}_k^{(+)} = [\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, ..., \vec{\boldsymbol{x}}_{|D_k^{(+)}|}]$ and $\boldsymbol{Y}_k^{(+)} = [\vec{\boldsymbol{y}}_1, \vec{\boldsymbol{y}}_2, ..., \vec{\boldsymbol{y}}_{|D_k^{(+)}|}]$. The solution is $\boldsymbol{M}_k^{(+)} = U^{(+)} V^{(+)T}$, where $U^{(+)}$and $V^{(+)}$ are the unitary matrices in the singular value decomposition (SVD) [37] of $Y^{(+)} X^{(+)T}$, i.e., $Y^{(+)} X^{(+)T} = U^{(+)} \Sigma V^{(+)T}$. Similarly, for the negative pairs, we learn the projection for the k-th granularity of non-hypernym relation as:

$$\boldsymbol{M}_k^{(-)} = \operatorname*{argmin}_{\Omega:\Omega^T\Omega=\boldsymbol{I}} \|\Omega \boldsymbol{X}_k^{(-)} - \boldsymbol{Y}_k^{(-)}\|_F,$$

where $\boldsymbol{X}_k^{(-)} = [\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, ..., \vec{\boldsymbol{x}}_{|D_k^{(-)}|}]$ and $\boldsymbol{Y}_k^{(-)} = [\vec{\boldsymbol{y}}_1, \vec{\boldsymbol{y}}_2, ..., \vec{\boldsymbol{y}}_{|D_k^{(-)}|}]$.

After learning the projection matrices, GRASP calculates the projected offsets for each specific data pair (x_i, y_i), based on the projections derived from all clusters, i.e., $\{\boldsymbol{M}_k^{(+)} x_i - y_i\}$ and $\{\boldsymbol{M}_k^{(-)} x_i - y_i\}$ for $k = 1, 2, \ldots, K$. Next, GRASP uses the projected offsets along with the attention weights for all clusters as inputs to a classifier to determine the existence of the hypernym relationship as below.

- *Compare with previous semantic projection.* The previous projection-based approaches [20], [21], [22] overlook granularity information, leading to a mixing of coarse-grained relations (e.g., "coarse location" and "GPS location") and fine-grained relations (e.g., "precise location" and "GPS location") into the same cluster. This limitation reduces the precision of the learned projection (§ IV-D). In contrast, our proposed method in GRASP fully leverages the taxonomy architecture, enabling separate projection learning processes for different granularity levels. This approach avoids the issue of mistakenly recognizing "precise location" as the hypernym of a new instance of "coarse location", such as "WiFi position," because they share the same projection matrix. Instead, GRASP can accurately distinguish between different granularity levels. Importantly, trivially increasing the number of clusters for previous projection-based approaches [20], [21], [22] is not promising to address this issue due to the highly approximated contextual features of coarse/precise location; also, this may introduce data sparsity and overfitting issues [21]. Note that the number of clusters required by GRASP is not necessarily larger than previous methods. This is because relations under different subdomains (e.g., location information and financial information) with similar granularity can be assigned to the same cluster. This approach does not introduce extra false positives since the contextual features of the hypernyms under different subdomains are diverse. In § IV-D, we experimentally demonstrate that the projection matrices learned by GRASP outperform others by achieving smaller projected distances for both hypernym and non-hypernym relations. This indicates that our method effectively captures the granularity information and results in improved hypernym prediction accuracy.

Classification. Given the set of projected offsets and attention weights regarding all granularity levels, the classifier first applies a linear transformation on each projected offset to enhance the discriminative power of the features while preserving the structure of a vector space:

$$z^{+}_{i,k} = W^{(+)}_{k} \cdot (M^{(+)}_{k} \vec{x}_i - \vec{y}_i), \quad z^{-}_{i,k} = W^{(-)}_{k} \cdot (M^{(-)}_{k} \vec{x}_i - \vec{y}_i).$$

It then multiplies $\{z_{i,k}\}$ to the attention weights associated with the same granularity level, and concatenates the two weighted sums to make the decision on hypernym relation and non-hypernym relation independently:

$$h_i = \sigma \left(\textsc{Concat} \left[\sum_{k=1}^{K^{+}} \omega^{+}_{i,k} z^{+}_{i,k} \;||\; \sum_{k=1}^{K^{-}} \omega^{-}_{i,k} z^{-}_{i,k} \right] \right)$$

Finally, GRASP feeds $\{h_i\}$ into a 2-layer Multi-layer Perceptron (MLP), and the entire model is jointly trained by the cross entropy loss: $L = -\frac{1}{N} \sum_i [t_i \log p_i + (1-t_i) \log(1-p_i)]$, where t_i labels whether y_i is the hypernym of x_i.

C. Privacy Data Taxonomy Expansion

Given a previously unknown candidate term u, GRASP identifies its privacy sensitivity by determining whether a hypernym relationship exists between u and arbitrary restricted data in the taxonomy. Specifically, we start from the root node and traverse the taxonomy by Breadth-first search (BFS) to generate candidate hypernym-hyponym pairs and query GRASP. If a restricted data t is not determined as the hypernym of the candidate, the subtree rooted by t is pruned from the searching space. Meanwhile, if at least one restricted data along a path in the taxonomy (e.g., PII → location information → coarse location) is identified as the hypernyms of the candidate, we integrate the candidate into the taxonomy by linking it only to the hypernym with the smallest granularity. This helps maintain the coherent hierarchy of the privacy data taxonomy.

Filtering. When analyzing the query response, we adopt the entropy of attention weights on positive clusters

$$H(\omega_i) = \sum_{k=1}^{K} -\omega^{+}_{i,k} \log \omega^{+}_{i,k}$$

to characterize the how well the candidate pair (x_i, y_i) can fit into the learned clusters of hypernym relations. A large entropy indicates that the candidate pair is far from all clusters, and the relation granularity is unfamiliar to the model. In this case, the candidate pair is less likely to have a hypernym relation and the weighted sum of the projected distance computed from each cluster can be less representative. Therefore, the entropy on attention weights of positive clusters can measure the confidence of the hypernym relationship between the pair, where low entropy indicates high confidence. We set a threshold τ to determine whether to accept the prediction. Specifically, positive predictions with $H(w) > \tau$ are *not* considered as having a hypernym relationship.

IV. EVALUATION

A. Datasets

We use two pre-existing privacy data taxonomies [1], [3] to construct datasets for the evaluation experiments of GRASP. More specifically, we build two ground-truth datasets of hypernym-hyponym pairs from two taxonomies, i.e., a general privacy data taxonomy in PolicyLint [1], namely *PrivacyPolicy* (PP) Taxonomy, and a domain-specific privacy data taxonomy in IoTProfiler [3], namely *IoT* Sensitive Data Taxonomy, respectively.

PrivacyPolicy (PP) Taxonomy. We refined PolicyLint's ontologies [1] as done in PoliCheck [2]. Specifically, we remove irrelevant hypernymy pairs (e.g., ⟨content, contribution⟩) and add hyponyms of restricted data already on the taxonomy (e.g., ⟨biometric information, blood type⟩). The high-level data categories in this taxonomy are manually extracted based on the definition in GDPR [25]. In total, this taxonomy is constructed with 680 restricted data and 2,176 hypernymy relations.

IoT Sensitive Data Taxonomy. We apply the taxonomy of privacy-sensitive IoT data items in IoTProfiler [3]. This taxonomy is constructed by inspecting 29 research papers and 54 news reports covering disparate privacy threats associated with IoT-sensitive data, ranging from device identifiers to IoT sensor data and usage data attached to the IoT devices, etc [3]. This taxonomy contains 76 restricted IoT data items and 138 hypernymy relations.

Ground truth of hypernym-hyponym pairs. For each hypernym-hyponym pair in the above two taxonomies, we create negative pairs by randomly sampling 5 non-hyponyms from the taxonomy for the hypernym in each pair. In total, we establish two groudtruth sets of hypernym-hyponym pairs: the *PP dataset* containing 2,176 hypernym-hyponym pairs and 10,880 non-hypernym-hyponym pairs, and the *IoT dataset* consisting of 138 hypernym-hyponym pairs and 690 non-hypernym-hyponym pairs. We measured Cohen's kappaa [38] for inter-annotator agreement on the correctness of the hypernym-hyponym pairs in the groundtruth sets, resulting in a score of 0.87, which indicates nearly perfect agreement [39]. For disagreed cases (e.g., ⟨user information, guess-create hashtag⟩), the annotators discussed their reasoning behind their annotations. If a consensus cannot be reached through the discussion, we involve a third privacy expert and resolve the disagreement through a majority vote.

B. Baseline Approaches

We implement six baseline approaches of hypernym prediction models, including four naive baselines, two state-of-the-art hypernym discovery models, and large language models (LLMs), for comparison, as elaborated below.

- Naive Baselines: We implemented four naive baselines in the form of *Feature* + *Classifier*. Here, we use the concatenation $[x||y]$ and the offset $x - y$ of token representations for each candidate pair as *Feature*, and apply Logistic Regression [40] and Multi-layer Perceptron [41] to construct the naive baseline models, respectively.

- Hypernym Discovery Models: we comprehensively evaluate two state-of-the-art hypernym discovery models: (1) SphereRE [34], which enhances hypernym prediction by learning lexical relation representations within hyperspherical embedding space; and (2) MWP [21], which uses Multi-Wahba Projection (MWP) to learn a fuzzy orthogonal mapping for each latent hypernymy or non-hypernymy relationship. Moreover, we involve three variants of each method: (a) methods followed by (N) learn the projection without the orthogonal constraint; (b) models followed by (O) learn the projection with the orthogonal constraint; (c) models followed with [P] use *only* the projected offsets (i.e., $Mx - y$) as features while removing other features (e.g., x, y) to identify the relation between tokens. The third kind of variants are included as baselines since GRASP only uses the projected offsets as features. We apply MLP as the classifier of GRASP and all state-of-the-art baselines.
- Large Language Models (LLMs): we also evaluate the performance of LLMs by directly employing the text generation models to predict the existence of hypernym relations. Specifically, we adopt the latest GPT model recommended for fine-tuning [42], `gpt-3.5-turbo-0125`, and craft dialogues to inquire if a hypernym relationship exists between each candidate pair in our testing set. We carefully refine the prompts based on previous studies utilizing LLM for relation extraction and knowledge graph construction [43], [44]. The hypernym and non-hypernym pairs in our training set are used for both model fine-tuning and prompt engineering. We provide a detailed account of the fine-tuning and prompt engineering strategy in Appendix B.

C. Experimental Settings

Implementation. We implement GRASP and all naive baselines using PyTorch [45] and Python 3.6 [46]. For token representations, we integrate the cutting-edge GPT-3 embedding model, text-embedding-ada-002 [47], in our system. To identify privacy-sensitive data objects from the wild (e.g., attack reports), we use SpaCy [48] to chunk noun phrases, acquire phrase embedding by GPT-3, and query GRASP with each noun phrase to determine its sensitivity. For each groundtruth dataset of hypernym-hyponym pairs, We use 80% pairs for training and 20% pairs for testing.

Hyperparameters. We use 10-fold cross-validation with grid search on the training set to find the best hyperparameters. The confidence threshold is set as $\tau = 0.7$. For the IoT dataset, we set the number of clusters K^+ and K^- as 5 and the learning rate as 0.02. For the PP dataset, we set K^+ and K^- as 16 and the learning rate as 0.08. We use the same number of clusters for MWP [21], which also learns a projection matrix for the token pairs within a cluster. We also select the best learning rate by cross-validation for each baseline model on the two datasets. For both datasets, we set the step scheduler of learning with gamma as 0.1, step size as 5, and the hidden layer dimension as 500. In all experiments, we add 3 copies of the minority class (e.g., pairs with hypernymy relation) to balance the training data and keep the testing data unchanged.

D. Evaluation of Hypernym Prediction

Effectiveness. The results of hypernym prediction are shown in Table I. We can conclude that GRASP outperforms all baseline approaches. The IoT dataset yields a large margin over the second-best method, MWP(N). The underlying reason might be that the sparsity of IoT privacy taxonomy poses challenges for the models to learn the appropriate projections for different hypernym/non-hypernym relations. However, the sparsity in the taxonomy architecture also yields large granularity differences which can be captured by GRASP to boost projection and prediction. Overall, GRASP outperforms the second-best model by 7.65% F1 on average (i.e., mean on the two datasets). The precision-recall curves are presented in Figure 4a and Figure 4b . Comparing the naive baselines, we observe that the concatenation of the token representation performs relatively well as the features for hypernym prediction. Comparing the projected-offsets-only models (i.e., SphereRE(N)[P], SphereRE(O)[P], MWP(N)[P], MWP(O)[P]) to the best naive baseline method (i.e., Concat+MLP), we observe that the projected offsets itself learned by these state-of-the-art methods can hardly capture the relation between tokens. Specifically, on the PP dataset, the performance of all four methods is worse than the best naive baseline. In addition, we surprisingly find that the non-orthogonal MWP model slightly outperforms the orthogonal version on our two datasets, which is contrary to the results on two general-domain and two domain-specific hypernymy datasets in [21]. The underlying reason might be that MWP(N) generates a better projection matrix than MWP(O) on our privacy-related datasets. We will discuss this using the projection matrix analysis experiments based on clustering in the section below.

Ablation Study. We perform an ablation study to gain a better insight into how the proposed mechanisms affect the effectiveness of GRASP. Table II shows the improvement by each proposed mechanism, respectively. The first and second rows refer to the result of clustering token pairs only according to their attribute similarity or structural similarity (i.e., the granularity of the hypernym). The third row refers to the result without assigning token pairs into different granularity clusters. In this case, we learn one positive/negative projection matrix for all positive/negative pairs, as $M^{(\cdot)} = \underset{\Omega:\Omega^T\Omega=\boldsymbol{I}}{\operatorname{argmin}} \|\Omega\boldsymbol{X}^{(\cdot)} - \boldsymbol{Y}^{(\cdot)}\|_F$. In the fourth row, instead of being the weighted sum along with the attention to each granularity cluster, the aggregated hidden representation is computed as the average mean of the projected distance. From the results, we observe that the clustering mechanism utilizing granularity information contributes most to the performance of GRASP. Specifically, the structural similarity plays a more critical role (10.9% and 25.8% gain in F1 on PP and IoT datasets) than the attribute similarity (4.6% and 3.9% gain in F1) in the clustering, indicating the significance of introducing the taxonomy structure information in the task of hypernym prediction. In addition, the attention mechanism, which controls the use of granularity-aware projection learned from each cluster, has a large impact on the model performance (13.5% and 21.4% gain in F1).

TABLE I: Evaluation results of Hypernymy Relation Identification. Each result is in the format of mean ± std of 5 trials.

Method	PrivacyPolicy Dataset				IoT Dataset			
	Precision	Recall	F1	Micro F1	Precision	Recall	F1	Micro F1
Concat + LR	0.765 ± .00	0.871 ± .00	0.815 ± .00	0.947 ± .00	0.701 ± .00	0.701 ± .00	0.701 ± .00	0.933 ± .00
Offset + LR	0.671 ± .00	0.843 ± .00	0.747 ± .00	0.929 ± .00	0.624 ± .00	0.756 ± .00	0.684 ± .00	0.910 ± .00
Concat + MLP	0.889 ± .02	0.927 ± .02	0.907 ± .01	0.972 ± .00	0.701 ± .02	0.894 ± .04	0.786 ± .05	0.943 ± .02
Offset + MLP	0.866 ± .03	0.871 ± .02	0.868 ± .01	0.962 ± .01	0.788 ± .06	0.741 ± .09	0.764 ± .07	0.955 ± .02
SphereRE (N)[P]	0.774 ± .05	0.764 ± .03	0.768 ± .02	0.934 ± .01	0.778 ± .05	0.822 ± .10	0.794 ± .03	0.955 ± .01
SphereRE (O)[P]	0.870 ± .03	0.863 ± .03	0.866 ± .02	0.962 ± .01	0.660 ± .11	0.800 ± .12	0.715 ± .08	0.931 ± .03
SphereRE (N)	0.902 ± .02	0.890 ± .02	0.896 ± .02	0.971 ± .01	0.721 ± .09	0.822 ± .06	0.765 ± .06	0.945 ± .02
SphereRE (O)	0.904 ± .02	0.901 ± .02	0.903 ± .02	0.972 ± .00	0.741 ± .03	0.822 ± .13	0.776 ± .06	0.950 ± .01
MWP (N)[P]	0.817 ± .03	0.920 ± .03	0.865 ± .02	0.959 ± .01	0.664 ± .05	**1.000** ± .00	0.798 ± .03	0.945 ± .01
MWP (O)[P]	0.721 ± .03	**0.990** ± .00	0.834 ± .02	0.944 ± .01	0.597 ± .06	**1.000** ± .00	0.746 ± .05	0.926 ± .02
MWP (N)	0.917 ± .03	0.893 ± .03	0.905 ± .02	0.973 ± .01	0.771 ± .10	0.844 ± .06	0.802 ± .07	0.955 ± .02
MWP (O)	0.907 ± .03	0.899 ± .01	0.903 ± .01	0.972 ± .00	0.733 ± .06	0.844 ± .06	0.784 ± .05	0.950 ± .01
GPT-3.5 [Finetune+Prompt]	0.867 ± 0.00	0.867 ± 0.00	0.867 ± 0.00	0.962± 0.00	0.667 ± 0.00	0.889 ± 0.00	0.762 ± 0.00	0.940 ± 0.00
Grasp	**0.974** ± .01	0.952 ± .00	**0.963** ± .00	**0.990** ± .00	**0.910** ± .09	0.889 ± .08	**0.899** ± .05	**0.976** ± .01

TABLE II: Ablation study: each experiment evaluates Grasp while disabling a corresponding component.

Component	PrivacyPolicy Dataset				IoT Dataset			
	Precision	Recall	F1	Micro F1	Precision	Recall	F1	Micro F1
w/o. structural	0.871 ± .01	0.838 ± .02	0.854 ± .02	0.959 ± .00	0.523 ± .07	0.829 ± .06	0.641± .05	0.897 ± .04
w/o. attribute	0.892 ± .01	0.943 ± .01	0.917 ± .01	0.976 ± .00	0.801 ± .08	0.929 ± .05	0.860± .07	0.967 ± .02
w/o. clustering	0.748 ± .01	0.876 ± .02	0.807 ± .01	0.941 ± .00	0.592 ± .05	0.768 ± .07	0.669± .06	0.903 ± .01
w/o. attention	0.809 ± .02	0.848 ± .02	0.828 ± .01	0.950 ± .00	0.672 ± .02	0.698 ± .03	0.685± .02	0.931 ± .01

TABLE III: Analyze the projections under different clustering methods. *Smaller* average projected distance implies *better* quality of the projection.

Data / Method	Positive Relation		Negative Relation	
	PP	IoT	PP	IoT
SphereRE (N)	60.246	3.178	39.811	3.757
SphereRE (O)	1.176	1.193	1.324	1.221
MWP (N)	0.658	0.625	0.819	0.836
MWP (O)	1.000	1.000	1.000	1.000
Grasp	**0.360**	**0.393**	**0.592**	**0.534**

Analysis of projection matrix based on clustering. Table III shows the projected distance between token pairs with different semantic projections. The projection matrix M aims at minimizing the distance between the token representations by a linear transformation. Specifically, the first row of Table III shows the value of $\frac{1}{|D^{(\cdot)}|}\sum_{i=1}^{|D^{(\cdot)}|}||\boldsymbol{M}^{(\cdot)}\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i^{(\cdot)}||_2^2$, where $(\cdot)$ represents the positive (i.e., hypernym) relation or the negative (i.e., non-hypernym) relation. Each of the second to the last row shows the value of $\frac{1}{|D^{(\cdot)}|}\sum_{k=1}^{K}\sum_{i=1}^{|D_k^{(\cdot)}|}||\boldsymbol{M}_k^{(\cdot)}\vec{\boldsymbol{x}}_i - \vec{\boldsymbol{y}}_i^{(\cdot)}||_2^2$, with $M_k(\cdot)$ learned under different clustering algorithm. To make the results comparable, we normalize all token representations $\{w\}$ by $||w||_2 = 1$. From the results in Table III, we observe that Grasp learns the best projection matrices with the smallest average projected distance. This owes to that Grasp learns the projection between token pairs in a more interpretable and reasonable manner. It clusters data pairs using taxonomy structure information in addition to the attribute information to ensure the token pairs used to optimize the projection matrix have similar transformations. However, the baseline methods perform no clustering (SphereRE), or cluster data with only attribute information (MWP).

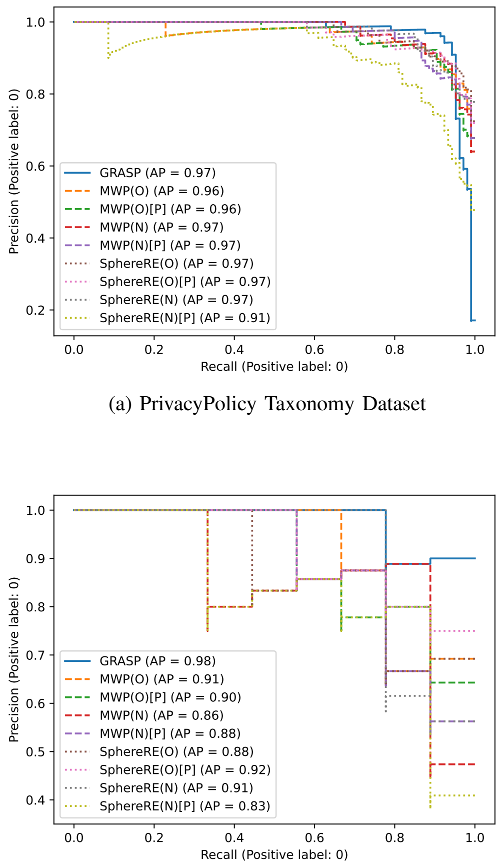
Parameter Sensitivity. We measure the hyperparameter sensitivity of Grasp from three aspects: (1) the number of clusters K, (2) the hidden dimension size of the linear transformation layer, and (3) the number of copies on positive pairs to balance the training data. Details and results of this analysis are shown in Appendix A.

V. Applications: GDPR-compliant Data Breach Notification

In this section, we present the design, implementation and evaluation of *Tracy*, which serves as a privacy professional assistant specifically tailored for GDPR-compliant data breach notification. This application highlights the potential capabilities of Grasp in assisting privacy professionals with privacy compliance assessment.

A. Design and Implementation

Design goals and overivew. In our study, we design *Tracy*, which incorporates Grasp to help privacy professionals comprehend the type of data involved in the data breach incident reports and determine whether it falls under the category of



(a) PrivacyPolicy Taxonomy Dataset

(b) IoT Taxonomy Dataset

Fig. 4: Precision-recall-curves of Hypernyn Prediction.

private data as defined by GDPR. *Tracy* is designed to achieve the following goals:

- *Traceability. Tracy* should allow privacy professionals to trace a restricted data to its hypernym and siblings, providing a clear understanding of why it is categorized as a type of private data according to GDPR.
- *Effectiveness. Tracy* should accurately identify the hypernym of a restricted data and assess its privacy sensitivity.
- *Efficiency. Tracy* should enhance the user's efficiency in identifying private data for data breach notification.
- *Usability. Tracy* should offer a positive user experience, fulfilling the usability requirements of its intended audience.

The pipeline of *Tracy* is shown in Figure 5. Specifically, given an uploaded data breach incident report (❶), *Tracy* starts by extracting noun phrases from the report as potential candidates of restricted data (❷) and retrieving the phrase embeddings from the fine-tuned GPT model (❸). Next, using GRASP, *Tracy* identifies those candidates that exhibit hypernym relationships with those private data explicitly defined in GDPR (e.g., telephone number, credit card number) (❹). Once a noun phrase is recognized as a restricted data, *Tracy* highlights the mention of the identified noun phrase in the report and displays the associated sub-graph of its PP taxonomy, with a root node representing private data defined in GDPR (❺). An illustration of this display is shown in Figure 6.

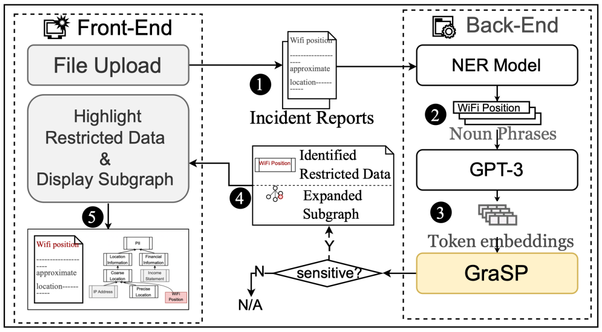


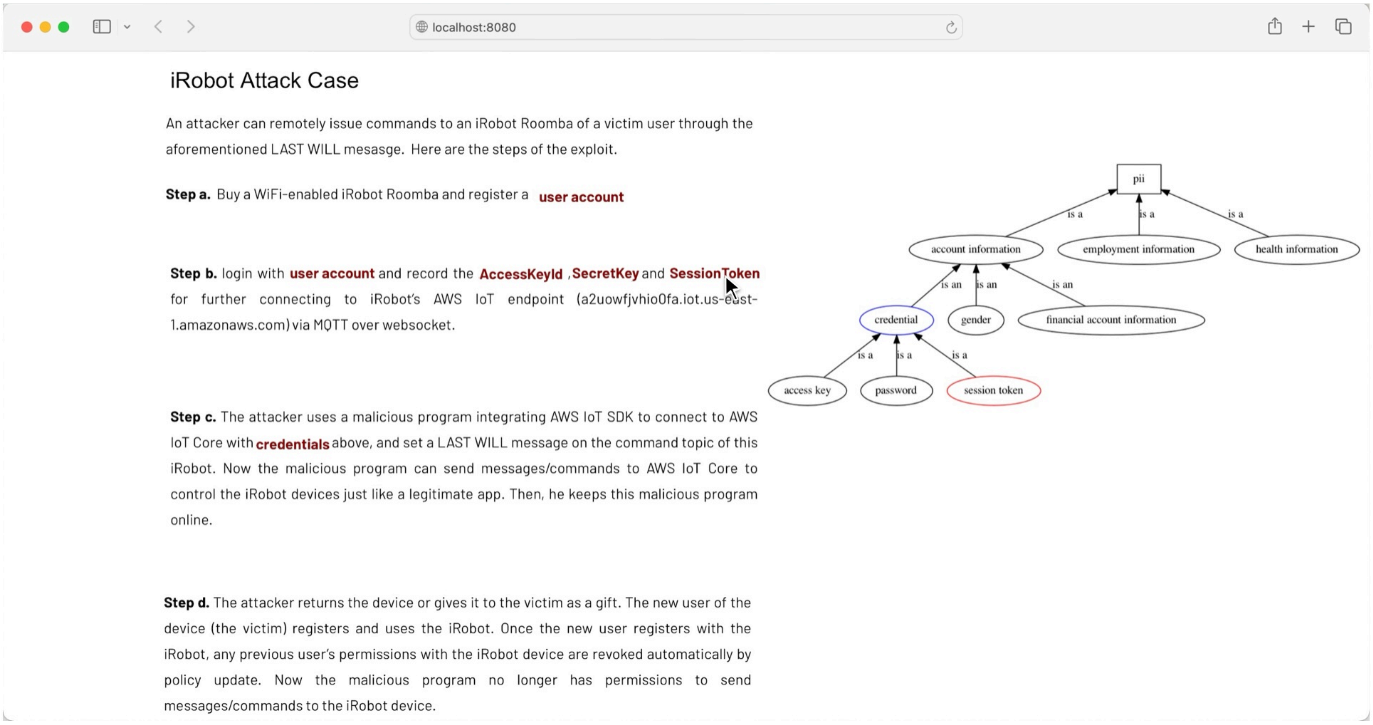
Fig. 5: Pipeline of *Tracy*.

Implementation. To realize our design, we implement *Tracy* as a public online web service hosted on the cloud service platform PythonAnywhere [49]. The implementation includes the front-end and back-end design to provide a seamless user experience. In the back end, we created a Flask environment to encapsulate GRASP trained on PrivacyPolicy dataset, GPT-3, and SpaCy NER model. The environment receives inputs through GET requests over HTTP/HTTPS. The Flask scripts, along with the models, are uploaded to PythonAnywhere [49] for hosting. In the front end, we design a file upload interface that allows users to submit a data breach incident report to *Tracy*. After the back-end processes the report using GRASP to identify restricted data, *Tracy* highlights the identified restricted data within the report for easy reference. Additionally, when the user hovers over an identified restricted data (e.g., `Session Token` in Figure 6), *Tracy* displays the associated hypernym-hyponym graph, providing explanatory information to help the user assess its sensitivity. By examining this graph, users can trace the hypernym and siblings of the identified restricted data, gaining insights into its potential compliance risk. To avoid overwhelming users with excessive information, the displayed graph only includes a partial taxonomy associated with the identified restricted data. Default configurations include showing the path from the identified restricted data to a root node representing personal data defined in GDPR and limiting each hypernym node to exhibit a maximum of three hyponym nodes. Meanwhile, *Tracy* offers flexible configuration options, allowing users to adjust the width and depth of the graph. Overall, the front-end implementation contains 2k+ lines of JavaScript, CSS, and HTML.

B. Evaluation of Tracy

We conducted a human subject study to evaluate the traceability, effectiveness, efficiency, and user satisfaction of *Tracy* from the perspectives of privacy professionals. In particular, we aimed to answer the following questions:

Q1: To what extent can hypernym-hyponym graph facilitate a better interpretation of restricted data by allowing participants to trace its hypernym and siblings in the privacy data taxonomy?



iRobot Attack Case

An attacker can remotely issue commands to an iRobot Roomba of a victim user through the aforementioned LAST WILL mesasge. Here are the steps of the exploit.

Step a. Buy a WiFi-enabled iRobot Roomba and register a **user account**

Step b. login with **user account** and record the **AccessKeyId** ,**SecretKey** and **SessionToken** for further connecting to iRobot's AWS IoT endpoint (a2uowfjvhio0fa.iot.us-east-1.amazonaws.com) via MQTT over websocket.

Step c. The attacker uses a malicious program integrating AWS IoT SDK to connect to AWS IoT Core with **credentials** above, and set a LAST WILL message on the command topic of this iRobot. Now the malicious program can send messages/commands to AWS IoT Core to control the iRobot devices just like a legitimate app. Then, he keeps this malicious program online.

Step d. The attacker returns the device or gives it to the victim as a gift. The new user of the device (the victim) registers and uses the iRobot. Once the new user registers with the iRobot, any previous user's permissions with the iRobot device are revoked automatically by policy update. Now the malicious program no longer has permissions to send messages/commands to the iRobot device.

Fig. 6: A screenshot of *Tracy* UI.

Q2: How do participants perceive the results output by *Tracy* in terms of restricted data identification? To what degree do they agree with the accuracy and relevance of the identified hypernyms relationships?

Q3: How efficient is *Tracy* in supporting privacy professionals with privacy data identification?

Q4: How well does *Tracy* meet user expectations regarding its practical usage?

Participant recruitment. We recruited privacy professionals, including security and privacy engineers specializing in risk management areas [50], as well as legal professionals specialized in security policies, protocols, and privacy laws, for two main reasons: (1) they represent the target users of *Tracy* and play a critical role in privacy compliance assessment, making their feedback invaluable for the evaluation metrics (**Q1**, **Q4**); (2) specialized in interpreting privacy data within their respective fields, these professionals possess credibility in privacy assessments from both technical and legal perspectives. Their expertise and assessment results are crucial in evaluating the effectiveness **(Q2)** and efficiency **(Q3)** of *Tracy*.

With IRB approval, we recruited participants through flyers distributed to incident response teams of Big Tech companies, as well as through the email lists of interdisciplinary programs in cyber risk and law at US universities (see the detailed requirements for recruitment in Appendix § C1). The individuals, who are interested in participating in the user study, are required to fill in a screening survey (see Appendix § C2), including contact information, demographics information, education/working background, and online consent form (administered through Qualtrics, see Appendix § C3). In total, we recruited 15 participants, including 8 security professionals in the industry who have real-world experience in data breach risk management areas and 7 legal professionals who have a background in privacy laws, regulations, and policies. All participants, including 4 students recruited from a university's online program, have working experience in incident response teams within IT companies or security consulting firms, practicing law in the US jurisdictions of California, Illinois, and Indiana. Specifically, they specialized in general international privacy laws and regulations such as the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), the Personal Information Protection Act (PIPA), the Indiana Consumer Data Protection Act (ICDPA), etc. We followed a standard and ethical way [51], [52], [53] to reward participants ($50/hour for each participant) in the user study. Table IV details the demographics of participants. The number of participants is substantial, falling in the range as suggested by qualitative research best practices literature [54] and aligning with related works [55], [56], [57], [58]. Each incident report underwent review by 4-5 professionals, in line with median legal team sizes reported in industry surveys [59]. Theoretical saturation [60] was achieved after the 13th interview, with no new themes emerging thereafter.

Procedure. In our study, we gathered 12 distinct real-world data breach incident reports from collaborating organizations. We organized these reports into four groups, with each group consisting of three incident reports. Each participant was assigned to review two different groups of the reports, one with the assistance of *Tracy* and the other without, in a within-subjects experiment. We followed the procedure of

TABLE IV: Participant demographics

	Gender	Age	Role	Industry experience/background	Work years
P1	Female	35-44	Legal professional	*Full time lawyer and part-time data scientist*	5+
P2	Female	25-34	Legal professional	*Full time lawyer in house of institution, wrote policies on daily basis*	1+
P3	Male	25-34	Security professional	*Responsible for risk management in IT security*	3+
P4	Male	25-34	Legal professional	*Specialized in cybersecurity and public policy*	-
P5	Male	25-34	Security professional	*Responsible for product quality, truthworthy assurance*	3+
P6	Female	18-24	Legal professional	*Specialized in laws and policy*	-
P7	Female	35-44	Security professional	*responsible for preventing data breaches, securing information systems*	7+
P8	Male	35-44	Legal professional	*responsible for the necessary regulations and protocols enforcement*	10+
P9	Male	25-34	Security professional	*work as Network security engineer for the federal government*	3+
P10	Male	25-34	Security professional	*responsible for Technical Support and Systems Administration*	5+
P11	Male	54+	Security professional	*responsible for Troubleshooting technical issues. Risk mitigation planning*	15+
P12	Male	18-24	Legal professional	*Specialized in cybersecurity and privacy law*	-
P13	Male	25-34	Security professional	*responsible for technical reviews for new features and identify privacy concerns*	3+
P14	Male	25-34	Security professional	*responsible for privacy reviews*	1+
P15	Female	25-34	Legal professional	*responsible for writing and reviewing privacy policies*	3+

comparative usability study [61], [62], [63] where the review tasks with the target tool are tested after manual assessment. Specifically, for the first group, participants were asked to manually review the reports and identify any restricted data they could find. We recorded their speed of restricted data identification and the total time spent on the review process. After that, the participants were assigned a different group of reports to eliminate any biases. They were then required to upload each incident report to *Tracy* and conduct a real-time test session to identify restricted data with the tool's assistance. For each restricted data highlighted by *Tracy*, we asked participants (1) whether the semantic relationship between each hypernym and hyponym pair holds true from their perspective; (2) whether each phrase identified by *Tracy* belongs to restricted data as they interpret it. Finally, the participants were asked to answer six qualitative questions (as shown in § C4) to evaluate the traceability and their overall satisfaction with the tool based on their usage experience.

Evaluation metric. Among 12 data breach incident reports used in our study, *Tracy* identified 89 restricted data and 103 hypernyms associated with them (i.e., 103 hypernym-hyponym pairs). In our study, based on Fleiss' kappa [38], a measure to determine the level of agreement between two or more raters/annotators, we calculate the agreement rate p_i regarding the identified subject i by *Tracy*, as well as the agreement rate P_i regarding the inter-rater annotation for the subject i, using the following formula:

$$p_i = \frac{n_{i1}(n_{i1}-1)}{N(N-1)}$$

$$P_i = \frac{1}{N(N-1)}\sum_{j=1}^{2} n_{ij}(n_{ij}-1),$$

where n_{ij} represents the number of annotators who assign the i-th subject to the j-th annotation, $i \in \{1, 2, ...S\}$ is the index of identified subjects, $j \in \{1, 2\}$ is the index of annotations (e.g., 1 = is a restricted data/hypernym-hyponym pair; 2 = not a restricted data/hypernym-hyponym pair), N is the total number of annotators per subject.

Results. This study spanned over two months for survey design, participant recruitment, data collection, and data analysis. Below we elaborate on the answers to the aforementioned questions.

• *The restricted data are traceable in the hypernym-hyponym graph, enabling the interpretation of its sensitivity in privacy data taxonomy.* We assessed the traceability by asking the participants whether the hypernym-hyponym graph allowed them to better understand the sensitivity of the identified data. As a result, all participants (N=15) confirmed the graphical representation significantly enhanced the explainability of the identified restricted data. For instance, P2, a legal professional who conducts daily legal reviews, provided the following feedback:

> *"When reviewing those technical reports, I feel like a hamster running in a wheel when it comes to keeping up with privacy, law, and privacy terms. For example, I even do not know what does "user topic" means. However, with the help of the graph, I am able to infer "user topic" is a type "user identifier" in the IoT domain and I can say user topic is sensitive, through a hypernym-hyponym relationship. Such linkage helps me better understand sensitive by surrounding other privacy data."*

• *Tracy achieves a high agreement rate among participants on the identified restricted data and hypernym-hyponym pairs.* In our study, we presented all the restricted data (N=89) and hypernym-hyponym pairs (N=103) identified by *Tracy* and asked each participant to label those they believed to be accurate. We use the agreement rate p_i and P_i in Fleiss' kappa to evaluate the degree of agreement regarding the identified results of *Tracy* and inter-rater annotation among participants (see *Evaluation Metric*).

Among 89 restricted data identified by *Tracy*, 44 subjects achieve 100% agreement rates (i.e., $p_i = 1$ and $P_i = 1$), indicating that the participants fully agree with the model predictions as well as the annotations of each other on these subjects. We observe that *Tracy* helps participants to successfully recognize some restricted data that are not obviously related to privacy concerns and thus often overlooked. For example, *"Siri recording"* is considered a restricted data by all participants, given the hypernym *"audio recording"*, which falls under the category of *"activity information"*.

By looking into the subjects with a rather low agreement (i.e., small pi), we observe that these disputed cases also yield disagreement among participants (i.e., small Pi). For example, *"device topic"* and *"IOSDevicePairingID"* have the lowest Pi as 0.33 among all identified subjects and both of their p_i is 0.17. Such disagreement mainly results from the inadequate or ambiguous context for explaining the usage scenarios of certain data. For example, P3 and P5 with data breach risk management background consider that *IOSDevicePairingID* can be non-sensitive in a specific scenario when it is supposed to be shared to allow others to connect

> *" If you're trying to connect to something, that's like public; say like you're pairing to an Alexa device or your car, the ID does show up typically publicly. If you wanted to make it private you could but I feel like a lot of people don't necessarily do that. "*

P10 with system management background questioned the sensitivity of *"device topic"* as

> *" ...device topic is possibly not depending on the device information that necessarily yield information about a particular user... "*

This finding emphasizes the importance of ensuring the quality and clarity of incident reports for privacy-compliant data breach notification. Providing detailed and explicit information about the data items and their usage contexts can effectively reduce the uncertainty and ambiguity surrounding the sensitivity of the data, leading to more accurate privacy compliance analysis for legal professionals, regardless of whether they use *Tracy* or not.

Regarding 103 hypernym-hyponym pairs identified by *Tracy*, 73 pairs achieve 100% agreement rates in both p_i and P_i. The pair with the lowest agreement is ⟨Identifier, Device Topic⟩, with p_i being 0.3 and P_i being 0.4. This disputed case mainly results from the UI design of *Tracy*, which displays only a partial graph to prevent overwhelming users with excessive information (§ V-A). As a consequence, crucial information that could guide decision-making might be missing from the presentation, leading to lower agreement in this specific case. More specifically, *Tracy* illustrates the hypernym (i.e., "Identifier") and siblings (i.e., "Serial number", "IP address") of "Device Topic" to the participants. However, the sibling "Account Topic", which has closer semantics, was not displayed to the participants. This limited information is not sufficient to help participants fully recognize the sensitivity of "Device Topic". However, in the post-interview process, when we offer the configurable option for participants to adjust the size of the graph based on their preferences and requirements, with additional siblings displayed to the users, they believe that the hypernym relationship ⟨Identifier, Device Topic⟩ is also true. P1 commented,

> *"I could not understand the phrase meaning but if the other two siblings are all identifiers, then the device topic is definitely identifier."*

• *Tracy can reduce 75.17% time cost for restricted data assessment.* We demonstrate the efficiency of *Tracy* by comparing the review time with and without *Tracy*. Specifically, we record the time participants spent to identify the restricted data in reports with/without *Tracy*. The results show that participants only spent 3.6 minutes on average to identify all restricted data with the help of *Tracy*, compared to an average of 14.5 minutes without *Tracy*. In practice, given the large amount of incident reports received by a company (e.g., Meta has received 150,000 data breach incident reports since 2011 [64]) and the limited staff resources, privacy professionals are facing a massive demand for privacy compliance assessment, which raises great demand to reduce the workload.

P15, a participant doing privacy review on a daily basis commented that

> *"such a visual-friendly display attracts my attention to privacy-related data in the first place, which helps me quickly recognize whether there existing potential privacy leakage risk."*

When asked how *Tracy* can assist data breach incident report review, P13 commented that

> *"The terms created by white engineering boys are bizarre and not inclusive enough for legal people. I need to google those terms, but it's not an easy search and time-consuming because a lot of it means something else. This tool kind of creates privacy vocab, and you just look up a technical term and it will show up its context, which greatly improves my work efficiency."*

• *The vast majority of participants expressed satisfaction with the usability of Tracy and showed their will to use Tracy in the future.* To understand how *Tracy* satisfies the potential users, we measure the satisfaction metrics by asking the following qualitative questions: (1) who can benefit from *Tracy*; (2) what usage scenario *Tracy* can serve for; (3) whether potential users are willing to utilize *Tracy* for privacy assessment.

Among all 15 participants, 11 of them believe that legal professionals can benefit from *Tracy*. P5 with legal background commented that

> *"It can help me identify the privacy harm of data breaches by identifying the data sensitivity and fill-*

> *ing my knowledge gap of those technical terminologies. For example, the terminologies in The California Privacy Rights Act, consumer Data Protection Act, and safe harbor laws are very high-level and general, which are never covering those detailed, specific, technical terminologies."*

Further, the security and privacy engineers are considered as the beneficiaries by the second (9 participants), and third (7 participants) most participants. P1 commented that

> *"The graph issues a citation from specific data to the general terms governed by laws and regulations, which is particularly useful for engineers who are not entirely familiar with legal stuff."*

Finally, product developers and system analysts are considered as the beneficiaries by 5 and 4 participants, although their job duties are less involved with privacy assessment. P11 commented that

> *"System analysts, product developers who are not that familiar with security and security, terminology, security or privacy, information, security may need more help from this tool."*

Further, we asked participants what usage scenario our tool can serve. We summarized the key use cases, as shown in Figure 7. The most mentioned assistance scenario is to conduct data privacy compliance reviews and audits, followed by identifying privacy harms when reviewing attack incidents. Four participants believed that *Tracy* can help alert or avoid sensitive data exposure (e.g., returned internally by the low-level system) through inappropriate public API design due to the negligence of data sensitivity. Six participants, all of whom are with a legal background, mentioned that the tool can be used in a privacy law class to educate future lawyers about the definition and scope of privacy-sensitive data.

Finally, we asked the participants about their interest in using our tool in the future. The results show that an overwhelming 93.33% (N=14) of participants expressed their willingness to incorporate our tool into their daily workflow. Participant P8, however, had a different perspective by stating, *"Probably not. I think the tool can provide a lot of inferences, which is helpful. But without this tool, I can also infer them with my specialized knowledge."* We acknowledge that *Tracy* is less supportive for those experts who already possess a strong interdisciplinary understanding of technology, social, and legal domains, along with rich experiences in privacy compliance analysis. Overall, the positive response from the majority of participants indicates the potential value our tool holds for enhancing restricted data assessment and privacy compliance analysis.

Baseline Study without *Tracy*. During the baseline study where participants manually reviewed reports without Tracy, we tracked data items receiving high agreement to be sensitive (i.e., $P_i > 0.5$) among participants but were not flagged by Tracy, yielding 5 instances as false negatives. Additionally, we observed that 2, 3, and 5 participants overlooked 4, 3, and 1 instances of personal data in the reports, which Tracy was able to identify. In the post-review phase, we presented the results identified by Tracy but overlooked by the participants, and the participants agreed that these were indeed personal data.

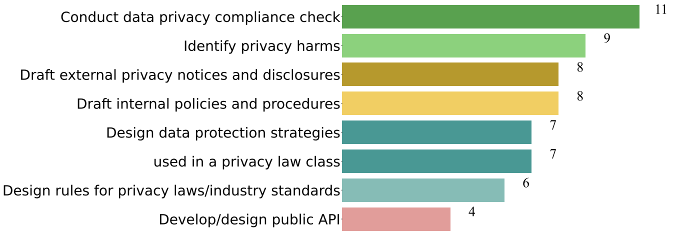


Fig. 7: The usage scenario our tool can serve for

VI. Discussion

GRASP demonstrates the proposed granularity-aware hypernym prediction model significantly contributes to the privacy data taxonomy expansion. However, we acknowledge that the effectiveness of the proposed mechanisms can be affected by the size of the taxonomy. To explore how taxonomy size impacts hypernym prediction, we evaluate GRASP on the adjusted groundtruth dataset of hypernym-hyponym pairs (§ IV-A) based on scaled taxonomies. More specifically, we randomly select nodes and remove them from the taxonomy. If the selected node is a leaf node, we directly remove it. Otherwise, we remove the selected node and link its successors to its predecessor. With the new, scaled taxonomy, we generate the hypernym-hyponym pairs using the approach described in § IV-A. With 10%, 20%, and 30% nodes removed, the performance of GRASP remains consistent, yielding 93.5%, 88.6%, 87.9% precision, and 92.6%, 89.7%, 88.4% recall for hypernym prediction, respectively, based on the average of 5 tests. This indicates that our proposed hypernym prediction technique is not significantly affected by the taxonomy size.

Regarding *Tracy*, we have observed that the effectiveness of *Tracy* depends on the corpora quality. *Tracy* may encounter challenges in accurately identifying data items with obscure or incomplete descriptions in their context. The sensitivity of certain data items can vary across different usage scenarios, making their identification ambiguous without sufficient context or explanations. For example, the term "Shared User Certificate" can refer to either a non-sensitive public user certificate or a sensitive user certificate shared with privileged permissions. However, some incident reports lack detailed descriptions, leading to data quality issues [65], and making it challenging for *Tracy* to determine the sensitivity of such data items. To address this limitation, expert knowledge and manual efforts from domain experts familiar with the protocol, code implementation, and system architecture of the device may be required to accurately identify the sensitivity of such data items. As a future direction, we plan to explore methods to incorporate expert knowledge and context-specific information into *Tracy* to enhance its accuracy in such cases.

VII. Related Works

Privacy Data Taxonomy Construction. The goal of privacy data taxonomy construction is to define hypernym relationships between restricted data to allow reasoning, e.g., privacy compliance analysis, over different data types and granularities. Traditional methods paid significant manual efforts in constructing privacy data taxonomies [31], [4], [6], which are widely used to facilitate privacy data protection measures as well as compliance checks [8], [9], [5].

Recent works propose automatic generation for privacy data taxonomy by capturing the semantic relationships between restricted data from a large corpus of privacy policies [10], [1], [11]. For example, Hosseini et al. summarize 14 heuristics to extract hypernym relation between information type phrases (e.g., "user content" is a kind of "user information") in privacy policies [10]. However, the application scope of such heuristics is limited because they cannot recognize semantic relations between phrases without specifically defined tokens (e.g., "information") or co-occurred tokens (e.g., "user"). Evans et al. [11] and Andow et al. [1] proposed automatic construction of privacy data taxonomy by extracting hypernym relationship between restricted data based on lexical-syntactic patterns (e.g., x is a y or y such as x). Such pattern-based methods suffer from sparsity issues as they require the hypernym and the hyponym to co-occur in the surrounding sentences, which may not hold in the divergent context of privacy data. All existing methods face challenges in automatically incorporating previously-unknown terms into the taxonomy, especially from corpora in different domains. In this paper, we propose a novel hypernym prediction model to address these limitations and enable the automatic construction and expansion of privacy data taxonomies.

Hypernym prediction techniques. Traditional pattern-based approaches leverage regular expressions [11], lexical-syntactic path rules [12], [13], [10], [14], [11], or semantic patterns captured by neural networks [66] to extract hypernymy relations from texts. Distributional methods include unsupervised hypernymy measures modeling the degree of the existence of a hypernymy relation [67], [68], [69], [70] and supervised methods predicting the relation between each pair modeled as an embedding vector [15], [16], [17], [18]. However, these methods are limited in handling complicated linguistic regularities and may encounter the "lexical memorization" problem [71] that the model memorizes specific word patterns or lexical cues from the training data.

Projection-based approaches [19], [72], [73], [34], [21] improve previous methods by explicitly learning the relation as the mapping from the hyponyms to their hypernyms or non-hypernyms. Fu et al. [19] employ a clustering algorithm to split the relations into several groups where each group learns a piecewise linear projection model. This method is further improved by learning a number of linear projections [20] as well as negative sampling [72], [73]. Wang et al. propose a projection model that constrains the projection matrix to be orthogonal [21], and a transductive model that learns relation representations by mapping them to the hyperspherical embedding space to separate different relation triples [34]. These methods employ embedding-based clustering to split the relations into several groups where each group learns a specific projection matrix by minimizing the summation over the projected distance. Although the clustering helps combat the relation complexity, for example, to distinguish concept-subconcept relation and concept-instance relation, it overlooks that hypernyms with similar contexts can be either coarse-grained or fine-grained, leading to different granularities of the relations. This may improperly mix token pairs under different granularities into one cluster, limiting the precision of the learned projection. In this paper, we propose a novel granularity-aware hypernym prediction model that captures granularity information from taxonomy structure to guide the clustering, which achieves more precise hypernym prediction.

VIII. Conclusion

In this paper, we propose Grasp for automatically constructing and expanding privacy data taxonomy to facilitate privacy compliance analysis. The system employs a novel hypernym prediction model that utilizes granularity information to guide the projection learning from hyponyms to hypernyms. Experimental results demonstrate the effectiveness of Grasp in the evaluation experiments on two real-world privacy data taxonomies. Also, Grasp outperforms state-of-the-art hypernym prediction models. Furthermore, we design and implement *Tracy*, a privacy professional assistant to recognize and interpret private data in incident reports for GDPR-compliant data breach notification. A user study involving 15 privacy professionals has confirmed that *Tracy* significantly enhances the efficiency and explainability of restricted data assessment.

Acknowledgment

We would like to express our gratitude to the anonymous reviewers for their valuable and constructive feedback. This work is supported in part by the National Science Foundation (CNS-1850725, 2343618) and Luddy Faculty Fellowship.

References

[1] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie, "Policylint: Investigating internal privacy policy contradictions on google play." in *USENIX Security Symposium*, 2019, pp. 585–602.

[2] B. Andow, S. Y. Mahmud, J. Whitaker, W. Enck, B. Reaves, K. Singh, and S. Egelman, "Actions speak louder than words:{Entity-Sensitive} privacy policy and data flow analysis with {PoliCheck}," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 985–1002.

[3] Y. Nan, X. Wang, L. Xing, X. Liao, R. Wu, J. Wu, Y. Zhang, and X. Wang, "Are you spying on me? large-scale analysis on iot data exposure through companion apps," 2023.

[4] L. Elluri, A. Nagar, and K. P. Joshi, "An integrated knowledge graph to automate gdpr and pci dss compliance," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1266–1271.

[5] A. Tauqeer, A. Kurteva, T. R. Chhetri, A. Ahmeti, and A. Fensel, "Automated gdpr contract compliance verification using knowledge graphs," *Information*, vol. 13, no. 10, p. 447, 2022.

[6] K. P. Joshi, L. Elluri, and A. Nagar, “An integrated knowledge graph to automate cloud data compliance,” *IEEE Access*, vol. 8, pp. 148 541–148 555, 2020.

[7] J. Wang, Y. Xiao, X. Wang, Y. Nan, L. Xing, X. Liao, J. Dong, N. Serrano, H. Lu, X. Wang *et al.*, “Understanding malicious cross-library data harvesting on android.” in *USENIX Security Symposium*, 2021, pp. 4133–4150.

[8] L. Elluri and K. P. Joshi, “A knowledge representation of cloud data controls for eu gdpr compliance,” in *2018 IEEE World Congress on Services (SERVICES)*. IEEE, 2018, pp. 45–46.

[9] A. Kurteva, “Implementing informed consent with knowledge graphs,” in *The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers 18*. Springer, 2021, pp. 155–164.

[10] M. B. Hosseini, S. Wadkar, T. D. Breaux, and J. Niu, “Lexical similarity of information type hypernyms, meronyms and synonyms in privacy policies,” in *2016 AAAI Fall Symposium Series*, 2016.

[11] M. C. Evans, J. Bhatia, S. Wadkar, and T. D. Breaux, “An evaluation of constituency-based hyponymy extraction from privacy policies,” in *2017 IEEE 25th International Requirements Engineering Conference (RE)*. IEEE, 2017, pp. 312–321.

[12] R. Snow, D. Jurafsky, and A. Ng, “Learning syntactic patterns for automatic hypernym discovery,” *Advances in neural information processing systems*, vol. 17, 2004.

[13] R. Snow, D. Jurafsky, and A. Y. Ng, “Semantic taxonomy induction from heterogenous evidence,” in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 801–808.

[14] S. Roller and K. Erk, “Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment,” *arXiv preprint arXiv:1605.05433*, 2016.

[15] J. Weeds, D. Clarke, J. Reffin, D. Weir, and B. Keller, “Learning to distinguish hypernyms and co-hyponyms,” in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 2249–2259.

[16] Z. Yu, H. Wang, X. Lin, and M. Wang, “Learning term embeddings for hypernymy identification,” in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[17] A. T. Luu, Y. Tay, S. C. Hui, and S. K. Ng, “Learning term embeddings for taxonomic relation identification using dynamic weighting neural network,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 403–413.

[18] K. A. Nguyen, M. Köper, S. S. i. Walde, and N. T. Vu, “Hierarchical embeddings for hypernymy detection and directionality,” *arXiv preprint arXiv:1707.07273*, 2017.

[19] R. Fu, J. Guo, B. Qin, W. Che, H. Wang, and T. Liu, “Learning semantic hierarchies via word embeddings,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1199–1209.

[20] J. Yamane, T. Takatani, H. Yamada, M. Miwa, and Y. Sasaki, “Distributional hypernym generation by jointly learning clusters and projections,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1871–1879.

[21] C. Wang, Y. Fan, X. He, and A. Zhou, “A family of fuzzy orthogonal projection models for monolingual and cross-lingual hypernymy prediction,” in *The World Wide Web Conference*, 2019, pp. 1965–1976.

[22] Y. Bai, R. Zhang, F. Kong, J. Chen, and Y. Mao, “Hypernym discovery via a recurrent mapping model,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 2912–2921.

[23] Y. Xiao, Z. Li, Y. Qin, X. Bai, J. Guan, X. Liao, and L. Xing, “Lalaine: Measuring and characterizing non-compliance of apple privacy labels,” in *32th USENIX Security Symposium (USENIX Security 23)*, 2023.

[24] H. Lu, Q. Zhao, Y. Chen, X. Liao, and Z. Lin, “Detecting and measuring aggressive location harvesting in mobile apps via data-flow path embedding,” *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 7, no. 1, pp. 1–27, 2023.

[25] “GDPR: The EU General Data Protection Regulation. ,” https://eugdpr.org.

[26] “CCPA: California Consumer Privacy Act. ,” https://oag.ca.gov/privacy/ccpa.

[27] M. Altman, A. Cohen, K. Nissim, and A. Wood, “What a hybrid legal-technical analysis teaches us about privacy regulation: The case of singling out,” *BUJ Sci. & Tech. L.*, vol. 27, p. 1, 2021.

[28] O. Klymenko, O. Kosenkov, S. Meisenbacher, P. Elahidoost, D. Mendez, and F. Matthes, “Understanding the implementation of technical measures in the process of data privacy compliance: A qualitative study,” in *Proceedings of the 16th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2022, pp. 261–271.

[29] M. Usman, M. Felderer, M. Unterkalmsteiner, E. Klotins, D. Mendez, and E. Alégroth, “Compliance requirements in large-scale software development: An industrial case study,” in *Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21*. Springer, 2020, pp. 385–401.

[30] “Data and code:,” https://sites.google.com/view/grasphypernym/home.

[31] R. Slavin, X. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T. D. Breaux, and J. Niu, “Toward a framework for detecting privacy policy violations in android application code,” in *Proceedings of the 38th International Conference on Software Engineering*, 2016, pp. 25–36.

[32] R. Green, C. A. Bean, and S. H. Myaeng, *The semantics of relationships: an interdisciplinary perspective*. Springer Science & Business Media, 2002, vol. 3.

[33] L. J. Brinton, *The structure of modern English: A linguistic introduction*. John Benjamins Publishing, 2000, vol. 1.

[34] C. Wang, X. He, and A. Zhou, “Spherere: Distinguishing lexical relations with hyperspherical relation embeddings,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1727–1737.

[35] H.-H. Bock, “Clustering methods: a history of k-means algorithms,” *Selected contributions in data analysis and classification*, pp. 161–172, 2007.

[36] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*. OUP Oxford, 2004, vol. 30.

[37] S. Banerjee and A. Roy, *Linear algebra and matrix analysis for statistics*. Crc Press Boca Raton, 2014, vol. 181.

[38] J. Cohen, “Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit.” *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[39] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[40] M. P. LaValley, “Logistic regression,” *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008.

[41] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[42] “Fine-tuning OpenAI text generation models,” https://platform.openai.com/docs/guides/fine-tuning/create-a-fine-tuned-model.

[43] J. H. Caufield, H. Hegde, V. Emonet, N. L. Harris, M. P. Joachimiak, N. Matentzoglu, H. Kim, S. A. Moxon, J. T. Reese, M. A. Haendel *et al.*, “Structured prompt interrogation and recursive extraction of semantics (spires): A method for populating knowledge bases using zero-shot learning,” *arXiv preprint arXiv:2304.02711*, 2023.

[44] M. Trajanoska, R. Stojanov, and D. Trajanov, “Enhancing knowledge graph construction using large language models,” *arXiv preprint arXiv:2305.04676*, 2023.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[46] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[47] “GPT3.5 and GPT4 Models,” https://platform.openai.com/docs/models.

[48] M. Honnibal and I. Montani, “spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing,” 2017, to appear.

[49] “pythonanywhere,” https://www.pythonanywhere.com/.

[50] J. More, A. J. Stieber, and C. Liu, “Chapter 2.m - tier 2—lateral: Business analyst,” in *Breaking Into Information Security*, J. More, A. J. Stieber, and C. Liu, Eds. Boston: Syngress, 2016, pp. 148–149. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012800783900046X

[51] D. Saha, A. Chan, B. Stacy, K. Javkar, S. Patkar, and M. L. Mazurek, “User attitudes on direct-to-consumer genetic testing,” in *2020 IEEE*

European Symposium on Security and Privacy (EuroS&P). IEEE, 2020, pp. 120–138.

[52] M. Wei, M. Stamos, S. Veys, N. Reitinger, J. Goodman, M. Herman, D. Filipczuk, B. Weinshel, M. L. Mazurek, and B. Ur, “What twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users’ own twitter data,” in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 145–162.

[53] A. McDonald, C. Barwulor, M. L. Mazurek, F. Schaub, and E. M. Redmiles, “” it’s stressful having all these phones”: Investigating sex workers’ safety goals, risks, and practices online,” in *30th {USENIX} Security Symposium ({USENIX} Security 21)*, 2021.

[54] G. Guest, A. Bunce, and L. Johnson, “How many interviews are enough? an experiment with data saturation and variability,” *Field methods*, vol. 18, no. 1, pp. 59–82, 2006.

[55] D. Votipka, E. Zhang, and M. L. Mazurek, “Hacked: A pedagogical analysis of online vulnerability discovery exercises,” in *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021, pp. 1268–1285.

[56] D. Votipka, S. Rabin, K. Micinski, J. S. Foster, and M. L. Mazurek, “An observational investigation of reverse engineers’ processes,” in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 1875–1892.

[57] K. R. Fulton, A. Chan, D. Votipka, M. Hicks, and M. L. Mazurek, “Benefits and drawbacks of adopting a secure programming language: Rust as a case study,” in *Seventeenth Symposium on Usable Privacy and Security ({SOUPS} 2021)*, 2021, pp. 597–616.

[58] R. Stevens, D. Votipka, E. M. Redmiles, C. Ahern, P. Sweeney, and M. L. Mazurek, “The battle for new york: a case study of applied digital threat modeling at the enterprise level,” in *27th {USENIX} Security Symposium ({USENIX} Security 18)*, 2018, pp. 621–637.

[59] “Most legal departments are small in size,” https://www.legaldive.com/news/legaldepartments-smallsize-benchmarkingreport-ACC-MLA-2022-legaldepartments/626878/.

[60] P. I. Fusch Ph D and L. R. Ness, “Are we there yet? data saturation in qualitative research,” 2015.

[61] J. Kaiser and M. Reichenbach, “Evaluating security tools towards usable security,” in *IFIP 17th World Computer Congress, Montreal, Canada*. Citeseer, 2002.

[62] Q. Wu, Y. Xiao, X. Liao, and K. Lu, “{OS-Aware} vulnerability prioritization via differential severity analysis,” in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 395–412.

[63] E. Zeng and F. Roesner, “Understanding and improving security and privacy in multi-user smart homes,” in *Proceedings of the 28th USENIX Security Symposium*, pp. 159–176.

[64] “vulnerability report,” https://engineering.fb.com/2021/12/15/security/bug-bounty-scraping/.

[65] C. Macrae, “The problem with incident reporting,” *BMJ Quality & Safety*, vol. 25, no. 2, pp. 71–75, 2016. [Online]. Available: https://qualitysafety.bmj.com/content/25/2/71

[66] V. Shwartz, Y. Goldberg, and I. Dagan, “Improving hypernymy detection with an integrated path-based and distributional method,” *arXiv preprint arXiv:1603.06076*, 2016.

[67] A. Lenci and G. Benotto, “Identifying hypernyms in distributional semantic spaces,” in ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 2012, pp. 75–79.

[68] S. Roller, K. Erk, and G. Boleda, “Inclusive yet selective: Supervised distributional hypernymy detection,” in *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, 2014, pp. 1025–1036.

[69] E. Santus, A. Lenci, Q. Lu, and S. S. Im Walde, “Chasing hypernyms in vector spaces with entropy,” in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, 2014, pp. 38–42.

[70] H.-S. Chang, Z. Wang, L. Vilnis, and A. McCallum, “Distributional inclusion vector embedding for unsupervised hypernymy detection,” *arXiv preprint arXiv:1710.00880*, 2017.

[71] O. Levy, S. Remus, C. Biemann, and I. Dagan, “Do supervised distributional methods really learn lexical inference relations?” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 970–976.

[72] C. Wang and X. He, “Chinese hypernym-hyponym extraction from user generated categories,” in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1350–1361.

[73] D. Ustalov, N. Arefyev, C. Biemann, and A. Panchenko, “Negative sampling improves hypernymy extraction based on projection learning,” *arXiv preprint arXiv:1707.03903*, 2017.

Appendix

A. Parameter Sensitivity Analysis

To understand parameter sensitivity, we vary three parameters while keeping the others at default values: (1) the number of clusters for each hypernym relation, (2) the hidden dimension size of the linear transformation layer, and (3) the number of copies on positive pairs to balance the training data.

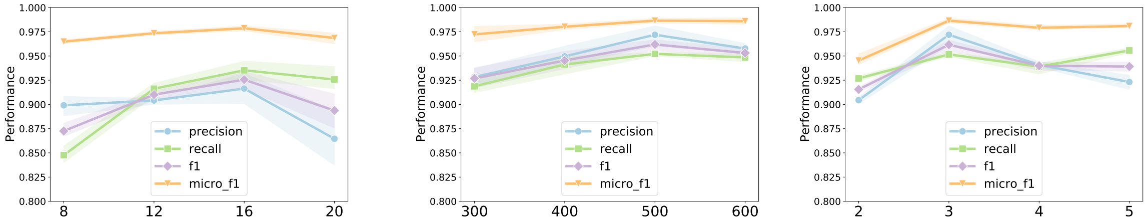
The results of GRASP’s performance on the PP dataset are shown in Appendix Figure 8. We observe that the number of clusters (K) has the most significant impact on model’s performance (see Figure 8a). Smaller values of K (i.e., $K \leq 8$) limit the models’ capability to distinguish the granularity between token pairs, leading to imprecise projection matrices and transformed distance. On the other hand, setting K too large (e.g., larger than 16) can result in data sparsity issues and overfitting, which reduces the generalizability of the model. Regarding the effect of the hidden dimension size of the linear transformation layer (Figure 8b), interestingly, we observe that the model’s performance remains similar under different hidden dimension sizes. This suggests that GRASP is not prone to overfitting and does not heavily rely on a large number of parameters. Figure 8c shows the effect of over-sampling by adding copies of the minority class to balance the training data. We selected the best hyperparameter values based on these experiments and applied them in all the experiments reported in Table I.

B. Model Fine-tuning and Prompt Engineering for Hypernym Prediction

In the API of GPT 3.5, the prompt for a chat conversation is a list of messages, each marked as one of three roles: system, user, or assistant. The system message encapsulates internal instructions system developer for the conversation. The user message and assistant message serve as the inquiry from the user and the response generated by the model, respectively.

To let the model better understand the task of hypernym prediction, we model each token pair in our dataset into a triplet T in the form of *(entity$_1$, relation, entity$_2$)*. The system message consists of a task instruction, the expected format of the response, and an explanation. The user message provides the triplet and the assistant message provides the response in the format as stated in the system message. Below is an example of fine-tuning data entry:

```
{"messages": [
{"role": "system",
"content": "Determine whether the following triplet
is correct in the format:
Triplet: ⟨the triplet⟩
Answer: ⟨the answer⟩
```



(a) Varying number of clusters: K

(b) Varying hidden dimensions

(c) Varying #copies

Fig. 8: Parameter sensitivity of GRASP on PP dataset. Each figure shows the result of varying the x-axis parameter.

```
The triplet is in the format of (entity 1, hyponym
of, entity 2), where the hyponym refers to a
subclass or an instance of a class.
The answer should be either ’Yes’ or ’No’"},

{"role": "user",
"content": "Triplet: (operating system, hyponym of,
device information) "},

{"role": "assistant",
"content":"Triplet: (operating system, hyponym of,
device information),
Answer: Yes"}]}
```

When evaluating model performance, we append examples for reference to the system message to let the model better understand the entities involved in the queried triplet. More specifically, we retrieve all positive and negative hypernym pairs containing the entity from the training data. For example, while querying

```
Triplet:(vibration mode, hyponym of, device
metadata),
```

the following response examples are appended to the *system message*, with all pairs containing `device metadata` in the training set as a reference :

```
Here are some examples for reference:
---
Triplet: (device metadata, hyponym of, device
attached data) Answer: Yes
---
Triplet: (intensive setting, hyponym of, device
metadata), Answer: Yes
---
Triplet: (bluetooth info, hyponym of, device
metadata), Answer: Yes
---
Triplet: (screen resolution, hyponym of, device
metadata), Answer: Yes
---
Triplet: (temperature setting, hyponym of, device
metadata), Answer: Yes
---
Triplet: (paired device, hyponym of, device
metadata), Answer: Yes
---
Triplet: (device model, hyponym of, device
metadata), Answer: Yes
---
Triplet: (smoke time, hyponym of, device metadata),
Answer: No
---
Triplet: (credential, hyponym of, device metadata),
Answer: No
---
Triplet: (home temperature, hyponym of, device
metadata), Answer: No
---
```

C. User study materials

1) Recruitment Email: Dear participants, We are writing to see if you would like to participate in a new research study being conducted at the University. This research plays an important role in advancing our understanding of privacy-related data types/items in the real world and testing the traceability and usability of our tool. We recruit participants who care about their privacy. The following information summarizes the study and what it involves:

• **Study topic**. GrASP: Identify Privacy-Sensitive Information via Granularity-Aware Hypernym Discovery

• **Study Purpose**. This study is to advance our understanding of privacy-related data types/items in the real world and test the traceability and usability of results outputted by our tool. The tool employs hypernym discovery to achieve a thorough and delicate identification.

• **Participation Requirements** The study lasts approximately eight (8) weeks. You may need 75 mins to finish the user study including experiments and surveys. The following criteria were used to select participants for the user study:

- participants must be legal professionals, security/privacy engineers, or have a background in risk management, laws, and regulations
- participants must be familiar with privacy risk assessments,

- participants must be familiar with privacy compliance checks,
- participants must have prior experience reading and writing incident reports

  (5) participants with industrial experience are preferred.

• **Compensation**. The study will take 75 mins. Upon completion of the study, you will be paid $50/hour with Amazon gift card.

2) Screen Survey:

1) What is your name?
2) What is your frequently used email address? Please also provide other contact information (like phone number) if you don't use email often. Note: Email will be the primary contact method
3) Are you a U.S. person or resident of the United States for tax purposes? (this includes U.S. citizens and Resident Aliens)
   a) Yes
   b) No
4) What's your job title?
5) Please describe your job responsibility.
6) How much working experience do you have?
7) Do you have experience reading and writing incident reports before?
8) Please choose ALL available time periods so we could arrange traffic.

3) Online Consent Form: You are being asked to participate in a research study. Scientists do research to answer important questions that might help change or improve the way we do things in the future. This document will give you information about the study to help you decide whether you want to participate. Please read this form, and ask any questions you have, before agreeing to be in the study.

All research is voluntary. You can choose not to take part in this study. If you decide to participate, you can change your mind later and leave the study at any time. You will not be penalized or lose any benefits if you decide not to participate or choose to leave the study later.

This research is intended for individuals 18 years of age or older. If you are under age 18, do not complete the survey. The purpose of this study is to understand the importance of privacy-related data type identification and testing the usability and traceability of results output from our tool. We are asking you if you want to be in this study because you received our recruitment materials and filled in our contact form. If you agree to be in the study, you will do the following things.

- You will complete a survey to evaluate the participants' knowledge about privacy-related data types/noun-phrase. (5 mins)
- You will complete a survey to manually check 30 pairs of phrases in the privacy domain to determine if phrases in each pair (email address, personal information) have a Hypernym-hyponym relationship. (5-10 mins)

What are the risks and benefits of taking part in this study? The risks of participating in this research are:

- You may feel confused for some questions
- Your background knowledge of the identification of privacy-related data will be evaluated

You may be uncomfortable while answering the survey questions. While completing the survey, you can skip any questions that make you uncomfortable or that you do not want to answer. There is a risk someone outside the study team could get access to your research information from this study. More information about how we will protect your information to reduce this risk is below. From this study, you will learn more about how to identify privacy-related data in the real-world.

We will protect your information and make every effort to keep your personal information confidential, but we cannot guarantee absolute confidentiality. No information which could identify you will be shared in publications about this study.

Your personal information may be shared outside the research study if required by law. We also may need to share your research records with other groups for quality assurance or data analysis. These groups include the Indiana University Institutional Review Board or its designees, and state or federal agencies who may need to access the research records (as allowed by law).

For questions about your rights as a research participant, to discuss problems, complaints, or concerns about a research study, or to obtain information or to offer input, please contact Human Research Protection Program office.

4) Post Survey:

1) How do you think the is-a graph can effectively help you identify sensitive phrases?
   a) Extremely effective
   b) Very effective
   c) Moderately effective
   d) Slightly effective
   e) Not effective at all
2) How accurately do you think *Tracy* can align a noun phrase to privacy data taxonomy?
   a) Extremely accurately
   b) Very accurately
   c) Moderately accurately
   d) Slightly accurately
   e) Not accurately at all
3) Will you consider using *Tracy* in the future when you need to identify whether a phrase is sensitive or not or perform other privacy engineering-related task?
   □ Yes
   □ No
   □ Might or might not
4) Who do you think can benefit from *Tracy*?
   □ Privacy Engineer
   □ Product Developer
   □ Legal Professional
   □ Security Engineer
   □ System Analyst
   □ Other
5) What're use scenarioes do you think this tool can be applied? (open-question)
6) If any, please describe the shortcoming of *Tracy* and anything that can be improved for*Tracy* (open-question).