HHS Public Access

Author manuscript

IEEE Symp Visual Cyber Sec (VIZSEC). Author manuscript; available in PMC 2023 November 10.

Published in final edited form as:

IEEE Symp Visual Cyber Sec (VIZSEC). 2022 October ; 2022: . doi:10.1109/vizsec56996.2022.9941431.

PRIVEE: A Visual Analytic Workflow for Proactive Privacy Risk Inspection of Open Data

Kaustav Bhattacharjee,
NJIT

Akm Islam,
NJIT

Jaideep Vaidya,
Rutgers University

Aritra Dasgupta[*]
NJIT

Abstract

Open data sets that contain personal information are susceptible to adversarial attacks even when anonymized. By performing low-cost joins on multiple datasets with shared attributes, malicious users of open data portals might get access to information that violates individuals' privacy. However, open data sets are primarily published using a release-and-forget model, whereby data owners and custodians have little to no cognizance of these privacy risks. We address this critical gap by developing a visual analytic solution that enables data defenders to gain awareness about the disclosure risks in local, joinable data neighborhoods. The solution is derived through a design study with data privacy researchers, where we initially play the role of a red team and engage in an ethical data hacking exercise based on privacy attack scenarios. We use this problem and domain characterization to develop a set of visual analytic interventions as a defense mechanism and realize them in PRIVEE, a visual risk inspection workflow that acts as a proactive monitor for data defenders. PRIVEE uses a combination of risk scores and associated interactive visualizations to let data defenders explore vulnerable joins and interpret risks at multiple levels of data granularity. We demonstrate how PRIVEE can help emulate the attack strategies and diagnose disclosure risks through two case studies with data privacy experts.

Index Terms:

Human-centered computing; Visualization; Visualization application domains; Visual analytics

1 Introduction

Accessibility of open data portals (e.g., NYC open data [41]) is like a double-edged sword. On the one hand, they make institutions and organizations accountable by providing public

[*] aritra.dasgupta@njit.edu .

access to proprietary information. On the flip side, inadvertent data leaks could compromise the privacy of data subjects. Recent research has shown how the lack of checks and balances in the conventional release-and-forget model [45] makes it surprisingly easy to breach privacy. An underlying reason for such a high privacy risk is the joinability of multiple open data sets that contain information about people. However, data owners and custodians (hereafter referred to as defenders) lack effective ways in which joinability risks can be summarized and communicated at the time of data set release or whenever a vulnerability is detected online.

Several recent examples of privacy breach scenarios emphasize the urgent need to address this problem. The Australian Department of Health released *de-identified* medical records for 2.9 million patients (10% of the population), but researchers were able to reidentify the patients and their doctors using other open demographic information [13]. Passengers' private information might be disclosed through the public transportation open data released by the city municipal of Riga, Latvia [36]. Researchers were also able to re-identify the details for 91% of all the taxis in NYC using an anonymized open taxi dataset and an external dataset [23].

Complete automation of the risk evaluation process is not feasible due to several reasons, like the presence of noisy metadata and the requirement for human expertise. Noisy metadata hinders the automatic profiling of these datasets. The various definitions and temporal nature of privacy risks, owing to the intermittent release of new datasets, point to the necessity for a human-in-the-loop approach, where defenders can configure and update risk computation techniques based on evolving compliance needs.

To address this critical need, we conducted a design study with urban informatics and data privacy researchers to develop a **proactive risk inspector** that is privy to the sensitive information that can be leaked before and after dataset release in urban, open data portals. PRIVEE, the visual analytic workflow resulting from this design study process, acts as a data-driven risk confidante and informer for the defender in the analysis loop. PRIVEE emulates potential attack scenarios and enables defenders to triage risky dataset combinations and ultimately diagnose the severity of disclosed information through dataset joins. A defender can thus proactively check for risks while releasing a dataset or depend on PRIVEE to be alerted when new vulnerabilities emerge owing to newly available, joinable data.

As the first contribution of this design study, we characterize the problem of disclosure evaluation and develop a set of visual analytic tasks that can be executed in a workflow to detect, calibrate, and inform data defenders about disclosure risks (Sections 3, 4). These tasks, developed in collaboration with privacy experts, emerged when we analyzed the problem through the lens of an adversary and developed several attack scenarios. We observed that it is possible to breach the privacy of open datasets using these scenarios, thus corroborating the findings of NYC taxi data in a larger scope where we can find information about data subjects [23]. As our second contribution, we designed the visualizations required for implementing the PRIVEE workflow and let defenders explore and interpret risks at the **metadata level**, triage vulnerable dataset groups and corresponding high-risk **joinable**

dataset pairs, and ultimately reason about the severity of the information disclosed at a **record-level** (Section 5). The design of these techniques is rooted in the idea of automation with transparent explanations which are responsive to user-controlled risk configurations (Sections 6, 7, 8). Finally, we present an interactive interface to help data defenders execute the workflow and demonstrate its effectiveness in the end-to-end diagnosis of disclosure (Sections 9, 10) through two case studies with domain experts.

2 Background & Related Work

The Open Data Charter was signed by the leaders of the G8 nations in 2013, leading to the increasing adoption of datasets that can be freely used, re-used, and redistributed by anyone, commonly referred to as *open datasets* [8, 31]. Though these are generally anonymized before release, joining two anonymized datasets using protected attributes can lead to the disclosure of sensitive information. In this context, *direct identifiers* are those protected attributes that can directly link to and identify an individual from a dataset (like name, id, SSN), while *quasi-identifiers* are those protected attributes that individually do not uniquely identify an individual but when combined with others, can identify an individual (e.g., age, race, gender, location). Disclosure risks can be mainly of two types: *identity disclosure*, where the data consumer knows who the individuals are, and *attribute disclosure*, where the values of different quasi-identifiers or sensitive attributes (like disease, salary, etc.) are revealed. Figure 1 shows two examples of identity and attribute disclosures using open datasets related to traffic stop-search, police citation, mobile clinic, and county health records.

A suite of anonymization methods [29] exists to address the problem of linking among public and private datasets, for example, between Census data and hospital records. The most promising among those methods is the notion of differential privacy [25] that the US Census has recently adopted [24, 48]. However, besides US Census data, which is just one of the sources of openly available data about people and their behavior, there are now a plethora of open data portals. As mentioned earlier, the adoption of open data is based on the promise of transparency and utility, as depicted by the FAIR principles [61], and at the same time, on the need for adherence to emerging privacy laws [40]. The unrestricted availability of open data [30] naturally raises the question: what if datasets within the open data ecosystem are linked even without other sensitive information from private datasets? Recent studies have demonstrated how even heavily anonymized datasets can be used to re-identify about 99% of Americans [45]. Re-identification or the disclosure of sensitive information is a challenge that has been previously explored by multiple researchers [19, 20, 62].

Data owners often practice the **release-and-forget model** where datasets, once released, are not analyzed further for potential privacy risks concerning the newly released datasets [43, 47]. However, the risk of re-identification can be considered a temporal function [51], thus requiring proactive monitoring of the risks. We developed the PRIVEE workflow and visualization interface with the specific goal of realizing a defender-in-the-loop analytical framework that can be privy to the disclosure risks or possibility of accidental leakage of sensitive information whenever new datasets are released. Though the target users

of PRIVEE are mainly data custodians or data owners, even data subjects [5] can use the workflow to inspect how vulnerable their identity or personal information might be in the presence of multiple, linked data sets. We use the concept of dataset joinability [38] in the presence of quasi-identifiers as a means for calibrating disclosure risks that are communicated using interactive visualizations throughout the PRIVEE workflow. Commercial tools like Google Cloud Data Loss Prevention (DLP) also help visualize the disclosure risk of a particular dataset using the quasi-identifiers [59]. While we do find other examples of visualization techniques for expressing disclosure risks of individual datasets [15, 33] and sensitive information [16, 35], *interactively visualizing disclosure risk among joinable open datasets* is essentially an open problem that we address in PRIVEE.

3 Problem Characterization

To understand the requirements for addressing the disclosure risks through the linking of open datasets, we decided to conduct a *red-team exercise* with the help of researchers in data privacy and urban informatics. A red-team exercise can be generally defined as a structured process to better understand the capabilities and vulnerabilities of a system by viewing the problems through the lenses of an adversary [63]. We engaged in a cold-start exploration process, followed by a more focused exploitation of datasets with privacy-related attributes, to develop a shared mental model of the problems related to the vulnerabilities and understand the functional requirements of a system addressing these vulnerabilities. We used the data sketches method and shared ideas about the different strategies with our collaborators [37] and explored multiple attack scenarios.

Red-team exercises generally follow the cyber kill chain, which starts with the initial reconnaissance step, where attackers try to *find vulnerable entry points* into any target system [32]. Following this step, we bootstrapped our red-teaming activity by first defining an initial set of privacy-related attributes, like age, race, gender, and location, to name a few. During our initial exploration, we collected 39, 507 datasets from around 500 data portals and observed through an automated analysis that about 5404 datasets have some combinations of quasi-identifiers. We filtered out datasets related to non-human objects, leading to the retrieval of a seed set of 426 datasets, including 151 individual record-level (e.g., records of people committing crimes) and 275 aggregated record-level (e.g., college records) datasets [6]. Analysis of these datasets led to interesting observations where some of the datasets have a highly skewed distribution of records across different categories of the quasi-identifiers. For example, the dataset *Whole Person Care Demographics 2* [60] from the *County of San Mateo Datahub portal* [54] has only one record for a 26-year-old Hawaiian female, similar to the example shown in Figure 1b. This can lead to identity disclosure and may leak sensitive information when joined with other datasets.

After building an initial collection of vulnerable datasets, we aimed to understand the consequence of an attacker joining them and accessing sensitive information. In this context, we would like to highlight that join is a fundamental operation that connects two or more datasets, and joinability is the measure to determine if two datasets are linkable by any number of join keys [9, 22]. When these *join keys coincide with protected attributes* like age, race, location, etc., the outcome of the join can potentially reveal sensitive information about

an individual or even disclose the individual's identity. As a next step in the red-teaming exercise, we randomly selected vulnerable pairs of datasets from multiple open data portals [10, 41, 44] and analyzed them for *joinability risks*, in terms of what kind of sensitive information may be leaked by these joins.

Several iterations of the selection of joinable pairs and join keys led to the discovery of disclosure between the datasets *Juvenile Arrests* and *Adult Arrests* from the *Fort Lauderdale Police Open Data Portal* [28]. We observed that two individuals, aged 15 and 21, mentioned separately in these datasets, were involved in the same incident of larceny on 20th March 2018, at the Coral Ridge Country Club Estate, Fort Lauderdale, similar to the example in Figure 1a. We repeated this exercise and found other examples where dataset joins ultimately led to disclosures.

4 Visual Analytic Goals and Tasks

The results from the red-teaming exercise confirmed our intuition that datasets with quasi-identifiers, when linked together, can potentially divulge sensitive information. Analyzing the functional requirements, we, together with our collaborators, concluded that totally automating the risk evaluation process is infeasible as human intervention is necessary at multiple stages of risk definition, interpretation, and subsequent exploration of the dataset combinations at high risk. To formulate a solution, we collaboratively developed PRIVEE, a visual risk inspection workflow in which defenders can proactively engage to stay one step ahead of the attackers (Figure 2).

PRIVEE is motivated by protecting the most vulnerable data sets against data join attacks. The workflow serves the dual purpose of: i) observing the open datasets to detect potential privacy vulnerabilities and ii) being a trusted informer for the data defenders that can visually explain and communicate disclosure risks while encouraging a deeper exploration of the attack and defense strategies. Automating the analysis of the disclosures directly at the record level can be an alternative, but this may lead to a seemingly infinite number of combinations to explore. Our streamlined workflow, developed from the experience gained during this design study process, will help the data defenders focus on a set of highly vulnerable datasets, thus reducing the number of combinations to be explored. In this section, we first describe the inputs and then define the high-level goals of the PRIVEE workflow in order to map them to the corresponding visual analytic tasks ultimately realized in a web-based interface.

Inputs to the workflow:

We initiate our defense strategy on the seed set of privacy-related datasets, which are about people as the data subjects, that we collected during the red teaming activity. While collecting these datasets, we followed the universally accepted common quasi-identifiers like age, race, gender, etc., with the notion that an open data ecosystem should, at a minimum, protect against attacks using these well-known quasi-identifiers.

After carefully curating the metadata from the seed datasets, we observed that there is no standard nomenclature for the attributes across the different data portals. This lack of

standardization established the importance of creating a metadata dictionary, focusing on the well-known quasi-identifiers while providing defenders the guidance and flexibility to define other privacy-related attributes. These attributes and the datasets selected based on their metadata serve as the inputs to the PRIVEE workflow (Figure 2a).

G1: Triage Joinable Groups: Candidate datasets for inspection selected from the initial input can be of the order of tens or hundreds. Finding all possible combinations of dataset joins among them is computationally expensive. Moreover, the large set of join outcomes will not lend well to human interpretation of risk. Also, during the red-teaming exercise, we observed that the risky datasets could also be construed from the datasets with vulnerable data distributions. Therefore, the next tasks in the defender's workflow are to focus on groups of datasets that can be joined and then triage those groups based on risk indicators:

T1: Explore cluster signatures: As shown in Figure 2b, this task lets defenders explore cluster signatures in terms of presence (clusters c1, c3) or absence (cluster c2) of the privacy-related attributes and their overall semantics. Involving the defender ensures that their inputs influence the algorithms used for grouping, using weighted clustering. They can thus control the triaging process by judging the groups' risks and privacy relevance. This task ultimately helps them select clusters of interest for further inspection of joinability risks.

T2: Find vulnerable datasets based on data distributions: The red-teaming exercise highlighted the presence of disclosure risk in datasets with a highly skewed records distribution across different categories of the quasi-identifiers. This task helps to distinguish between the most vulnerable and other datasets by inspecting a high likelihood of finding unique records for given quasi-identifiers.

G2: Compare Joinability Risks: Once a cluster of datasets is prioritized for inspection as part of G1, defenders would like to compare joinable pairs of datasets in this group that may potentially disclose sensitive information. To achieve this goal, we use disclosure risk metrics to automatically suggest risky pairs based on their feature profiles and then visualize those suggestions so defenders can interpret the metrics. The following task achieves this:

T3: Explore and Explain Disclosure Risks: This task focuses on pairs of datasets that can be ranked using multiple disclosure risk metrics. Within those rankings, we want to use visual cues that directly explain: which features are responsible for high risk, the differences between high and low-risk pairs, and if other features should augment the defender's definition of privacy relevance.

G3: Identify cases of disclosure: Once dataset pairs are selected as part of G2, defenders would like to understand the severity of the join outcomes. Fully automating this process may lead to many scenarios where the disclosures are less concerning and do not warrant any significant change in the defense strategies. To provide more control to defenders in their diagnosis of cases of actual disclosure, the tasks required to accomplish this goal are:

T4: Detect matching records across data sets: Matching records are the records present in both datasets in a pair. The main objective of this task is to detect lower frequencies of matching records, which may lead to the disclosure of sensitive information about an individual or disclose their identity.

T5: Augmenting the risky feature set with suggestions: One way of discovering disclosures is finding attributes that have the same values for all the records of the joined datasets. For example, joining two hospital datasets may reveal that all the patients common in both the hospitals are treated for cancer, leading to attribute disclosure for these patients. In this task, we suggest a set of attributes that may be highly related to the joining attributes, thus helping the users augment the feature set for the dataset join.

5 Design Overview

The design of PRIVEE is motivated by the need for a transparent explanation and evaluation of the risk inspection process. We implemented a web-based interface that enables data defenders to iterate between multiple entry points, evaluate the reasons for the dataset joinability and analyze disclosure risks for different combinations of datasets and attributes. In this section, we provide an overview of the design requirements for realizing the aforementioned visual analytic goals and tasks. **An interactive version of the interface** for PRIVEE may be accessed through the Chrome browser at http://privee.dataopen.online/.

Risk Profiling at metadata level:

PRIVEE helps to analyze the datasets' risk profiles through a filter bar, located conveniently at the top of the interface (Figure 3a), which contains a search option for the different tags and options to select the data portals and the dataset granularity. During the initial page load, this filter bar is positioned at the center of the page in order to avoid overwhelming the user with the search results. Defenders can select any combination of the tags from the tags search option, which is enriched with a modified bar chart showing the frequency distribution of the tags. Though the tags are sorted in descending order, the grey bar in the background (achieved by tweaking a linear-gradient bar) provides an idea of the frequency distribution of these tags among all the collected datasets. Privacy-related attributes can also be selected using filters.

Triaging joinable groups:

In order to fulfill G1, PRIVEE employs a set of visualizations to help the data defenders triage the joinable groups from the datasets selected using their metadata. This includes a projection plot, a word cloud, and a bar chart depicting the attributes' frequency, as illustrated in Figure 3b. This combination of visualizations is repeated for the different groups of joinable datasets. Though PRIVEE automates the grouping of the datasets, these visualizations provide the data defender a transparent method to understand the group signatures and update the groups based on their domain knowledge and definition of privacy relevance.

Finding vulnerable datasets:

PRIVEE helps the data defenders select vulnerable datasets by showing a distribution of the values of the privacy-related attributes through a combination of histograms (for numerical attributes) and bar charts (for categorical attributes), as shown in Figure 3b. This combination is repeated for each dataset, ranked according to their degree of vulnerability. It is also responsive to the privacy-related attributes selected through the filter area. The vulnerable categories for these attributes and their labels are shown in bright red to help defenders efficiently select vulnerable datasets.

Comparing Joinability Risk:

PRIVEE automatically computes the possible pairs from the datasets selected from either Projection View or the Vulnerable Datasets View and ranks them according to their joinability risk. The visual cues, shown in Figure 3c, help the data defender compare different datasets and select the high-risk pairs on a priority basis. Overall information about the risk score distribution allows flexible selection of dataset pairs of varying risk.

Identifying disclosures:

The disclosure of sensitive information can depend on multiple factors, subject to evaluation by the data defender. In this *Disclosure Evaluation View*, as shown in Figure 3d, PRIVEE lets the data defender analyze the matching records generated for a specific dataset pair and a join key selected from the Risk Assessment View. PRIVEE also suggests other features to help the defenders select a better join key, helping them understand the relationship between different attributes and possible disclosures.

6 Triage Joinable Groups (G1)

Data defenders need to analyze the degree of joinability between datasets. Hence, the design requirements for addressing tasks T1 and T2 are to develop human-in-the-loop clustering methods responsive to multiple definitions of privacy relevance, along with transparency in analyzing cluster signatures. This enables defenders to develop a mental model of the context and the degree of the potential vulnerability of subsequent joins. In this section, we discuss the analytical methods and visualizations to find and triage the joinable groups.

6.1 Weighted clustering for finding joinable datasets

Converting Data Attributes to Word Embeddings: The joinability of two datasets is a function of *shared attributes*. Hence, the datasets with similar attributes should be more joinable. Attribute names in open datasets are often noisy and inconsistent, making it computationally difficult to perform a binary search for the presence or absence of certain attributes. We focus on the idea that similar attribute names can capture the semantic similarity among multiple datasets that might have a similar context. We use a word-embedding approach that simultaneously satisfies the need to capture datasets' joinability and their semantic similarity. *Word embeddings* can be defined as real-valued, fixed-length, dense, and distributed representations that can capture the lexical semantics of words [2, 4]. Hence, we converted the data attributes into their corresponding word embedding form using Python's spaCy library [55] and created a vector representation for the attribute space of

each dataset. The vectors with a smaller distance between themselves signify datasets with similar attributes, hence more joinable.

Adding Weights for Privacy-related attributes: At this stage, all the data attributes have equal importance in the vector representation of a dataset; hence, datasets with attributes like *version*, *version number*, etc. may be marked similar to each other. However, these attributes may not have much significance in the context of privacy. Hence, we decided to add weights to some of the privacy-related attributes identified from the seed dataset corpus. Attributes like *age*, *race*, *gender* and *age at arrest* were selected, and adding more weights to these attributes signifies that datasets having these attributes may be marked as more joinable. Any disclosure using these datasets can be considered a high risk, which will help further triage the datasets.

Cosine similarity is widely used to measure the similarity between words and documents [14, 56]. However, word embeddings are mere representations of the words, and multiplying them with numeric weights would not increase the cosine similarity between two datasets. Hence we introduced a *weight vector* where we assign a weight if the privacy-related attributes selected by the data defender are present in the dataset. If a data defender selects the privacy-related attributes [*age, gender, race*], then the corresponding weight vector for a dataset with only the age and gender attributes would be [x, 0, x], where x represents the weight assigned to the privacy-related attributes. We concatenate these weight vectors with the corresponding word embedding vectors to get the final vector representation of each dataset.

Projecting the datasets and finding Clusters: Each dataset is now represented by a vector with more than 300 elements/dimensions, and comparing these datasets using a 2-D or 3-D plot would be challenging if all the dimensions were used. Hence we used the t-SNE dimensionality reduction algorithm to reduce these into two-dimensional vectors [57]. A 2-D projection of the datasets might not readily reveal dataset groupings. Hence, we experimented with clustering algorithms like KMeans [58], DBSCAN [27, 50], Birch [64], and OPTICS [3, 49]. After a careful analysis of the clusters' quality and the cluster density scores, we selected the DBSCAN algorithm.

Evaluating the clusters: There can be multiple groups of similar/joinable datasets, which would lead to the creation of multiple clusters. A data defender may find it challenging to evaluate all of these clusters. Hence we have employed a few cluster evaluation techniques to triage these clusters **(T1)**.

One of such metrics is the *Calinski-Harabasz Index* which is defined as the ratio of the between-cluster dispersion and the intercluster dispersion, where dispersion means the sum squared distance between the samples and the barycenter [7]. A higher score signifies that the different clusters are far away, implying better cluster formation. We designed an experiment to evaluate the difference in the results from this metric along with other metrics like Silhouette Score [46] and Davies-Bouldin Index [18] and selected the Calinski-Harabasz Index since we observed that it could efficiently guide defenders in finding meaningful,

joinable datasets. Further details about this experiment can be found in the supplementary materials.

Finding vulnerable data distributions: A particular cluster can have multiple datasets with vulnerable data distributions, leading to the disclosure of sensitive information when joined with other individual record-level datasets. Hence, we found such data distributions and ranked these datasets according to their degree of vulnerability **(T2)**.

In order to evaluate the degree of vulnerability, we first analyzed all the datasets and created the record points for the privacy-related attributes present in them. Record points are the unique categories for a specific attribute, while *vulnerable record points* are those record points that have very few records for them, as shown in Table 1. These datasets are then sorted based on the number of such vulnerable record points present and the frequency of the most vulnerable record point. The intuition here is that a dataset with more vulnerable record points is more prone to disclosure risk using these privacy-related attributes.

6.2 Visualizing joinable group signatures

We designed the Projection View to provide an overview of the datasets and the joinable groups **(T1)** and perform an automatic evaluation of the vulnerable data distributions of the datasets in each joinable group **(T2)**. Data defenders can review the group signatures through the different components of the Projection View and update the parameters to see the details and the data distribution of the datasets that match their mental model of privacy relevance. The components of these views are described as follows:

Joinable groups: Given a set of datasets selected based on their metadata, defenders need to find groups of datasets that can be joined together. The analytical process is performed automatically by PRIVEE, leading to the formation of joinable clusters, which are represented using a multi-dimensional projection plot, as illustrated in Figure 4a. Here, a red dot represents an individual record-level dataset in a particular cluster, while the grey dots represent the datasets not in that cluster. During this design study, we realized that some of the datasets are highly joinable due to their similarity in the attribute space, which would cause overlapping of the dots in a cluster. Hence, the overlapping datasets are represented by a single dot with the number of overlapping datasets inscribed in it. For example, Figure 4a shows a cluster of seven highly similar datasets represented using a red dot. This view contains multiple projection plots, where each plot represents a group of joinable datasets. It helps the data defender quickly compare the different groups from a single view. The dual color encoding scheme (red-grey) helps visually differentiate between the datasets in a group and the other datasets. Initially, a scatterplot with different colors for the different clusters was also considered for this view. However, it was realized that it is challenging to assign perceptually different colors to each cluster when the number of clusters is large, due to the limits of perception. Hence, a multiple plot design approach was chosen with the two-color encoding scheme.

Transparent explanation of joinability and vulnerability: Understanding the cluster signatures is crucial in understanding the reason behind the genesis of a joinable group **(T1)**

and the presence of data vulnerabilities **(T2)**. Since we have construed these dataset groups based on the similarity in their attribute space, it is essential to understand the frequency of the attributes present in these groups. Hence, bar charts become the natural choice for displaying the most frequent attributes in a group and their frequency, as illustrated in Figure 4b. These bar charts are sorted according to the attribute frequency, yet the frequencies of the privacy-related attributes are shown first. The vulnerable datasets are also represented using bar charts (for categorical attributes) / histograms (for numerical attributes) for each of the privacy-related attributes present in them. However, bar charts can have the limitation of visual scalability where only a certain number of bars can be shown due to space constraints [26]. In order to overcome this limitation, we also introduce word clouds of the attributes, as shown in Figure 4c. All the attributes present in at least two datasets in a joinable group are represented in this word cloud, with the size channel representing their frequency.

The bar chart in Figure 4b explains the similarity of the datasets since all seven of these datasets have *gender* and *race* attributes, thus transparently explaining the group signatures. Besides overcoming the visual scalability limitation of the bar chart, the word cloud also helps the data defenders look for other attributes of interest that may have a lower frequency but have much larger relevance in the context of privacy. For example, attributes like *victim age* and *offender age* may not be significant for a general user; however, a data defender working with law enforcement may find them interesting since these attributes are used in the police datasets. PRIVEE enables the data defender to update the default selection of the privacy-related attributes, which triggers a re-rendering of the whole Projection View, thus automatically calculating new groups of joinable datasets with extra weightage to the newly added privacy-related attributes *victim age* and *offender age*. Together, these Projection View components enable human-in-the-loop dataset grouping that is adaptive to various definitions of privacy relevance by transparently displaying measures to evaluate cluster signatures.

7 Compare Joinability Risks (G2)

Dataset groups from the Projection View can lead to multiple pairwise combinations of datasets, where the data defenders need to analyze each pair for their joinability risk. Hence, the design requirement for addressing G2 is to facilitate efficient visual comparison of the risk profile of dataset pairs and guide defenders towards focusing on high-risk dataset pairs. In this section, we describe the metrics that can help a data defender quantify the risk of joinability between the candidate datasets and the subsequent use of visual cues to compare and prioritize the joinable pairs.

7.1 Metrics for Joinability risk comparison

Multiple metrics that can help the data defenders compare the joinability risks between different dataset pairs were explored during the design study process. In this subsection, we define the mathematical formulas for the different metrics that highlighted the joinability risks better and were selected as part of the PRIVEE workflow.

Metric based on attribute profile: Shannon's entropy is a measure of the uncertainty of a random variable [11]. It has been widely used as a privacy metric [1, 21, 42, 52],

as higher entropy signifies more unique values for that attribute, thus resulting in higher disclosure risk. Hence, we used this metric to help defenders find joinable attributes for a pair of datasets. For a pair of datasets (say A and B), we first calculated Shannon's entropy of each of their shared attributes according to equation 1 and kept their maximum as the entropy score for that attribute. The intuition here is that the attributes with higher entropy can be offered as suggestions to the defender for the join key.

$$H(X_J) = -\sum_{i=1}^{n} P(x_{J_i}) \ln P(x_{J_i}) \quad (1)$$

where X_J represents attribute X in dataset J ($J \in \{A, B\}$), $H(X_J)$ represents the entropy of an attribute present in dataset J while x_{J_i} represents each category of the attribute X_J in dataset J.

Metric based on dataset pairs in a join: Since the joinability of two datasets depends upon the number of shared features/attributes between them, the joinability risk score can be calculated as a function of the number of shared attributes and the number of privacy-related attributes between a pair of candidate datasets. The formulae for the *joinability risk score* can be defined as follows:

$$\text{risk} = \alpha * p + (c - p) \quad (2)$$

where α is the empirical risk ratio (a constant), p is the number of privacy-related attributes and c is the number of shared attributes.

The joinability risk score depends on the empirical risk ratio, and to determine its value, we designed an experiment to calculate the risk scores of all the possible combinations of joinable pairs from the seed datasets (${}^{426}C_2 = 90,525$ combinations). We observed that the value $\alpha = 50$ works well to separate the dataset pairs with privacy-related attributes and pairs without them; hence the empirical risk ratio was fixed at the value of 50. We have included further details about this experiment in the supplementary materials.

7.2 Visual risk assessment

PRIVEE uses multiple visual analytic components to encode the joinability risk metrics, and these components together form the Risk Assessment View. This subsection describes how we map these metrics with the components of this view so that data defenders can pro-actively analyze the risk between the candidate datasets.

Comparing shared attributes set: The shared attributes' entropy metric encodes the attribute profile information, potentially highlighting if an attribute should be included in the join key. In the Risk Assessment View, these attributes and the entropy are represented using a descending *sorted bar chart* between the dataset names, as illustrated in Figure 5c. The horizontal position shows the different attributes, while the vertical position encodes the entropy of these attributes. The bars for the privacy-related attributes are colored in violet (plum kingdom), while the other bars were colored in grey, thus following the

similar colorblind-safe two-color strategy used in the other views. During an initial design iteration, each shared attribute was represented using a small rectangular box, with each box containing the attribute name in it. However, we realized that this design leads to the loss of information about the difference in entropy between the different shared attributes. This led to the current design of the sorted bar charts where the data defender can analyze the entropy, select any number of the shared attributes as the join key for the dataset pair and evaluate them for disclosures.

Comparing risks: Each dataset pair (Figure 5b) is represented with a combination of the following components: dataset names, shared attributes, and the joinability risk bar. These pairs are sorted according to the risk score. Thus, a top-ranked dataset pair would imply higher chances of joinability. In order to highlight the joinability risk score between the dataset pairs, the Risk Assessment View has a *joinability risk bar* for each dataset pair **(T3)**, as shown in Figure 5d. This bar is filled with a linear gradient between the grey and red colors, representing low-risk and high-risk dataset pairs. The exact risk score is highlighted using a black vertical bar. The choice of the colors, following the two-color scheme used across the different views in PRIVEE, helps express the joinability risk score on a scale of low to high scores. This view also shows an overview of the shared privacy-related attributes and the risk score distribution between the dataset pairs using a horizontal bar chart and a histogram (Figure 5a and Figure 5e). PRIVEE also automatically selects the joining attributes based on their entropy and privacy relevance, which the data defender can further augment.

8 Identifying Disclosures (G3)

The design requirement for addressing tasks T4 and T5 is to let the defenders *judge the degree of sensitive information* that can ultimately be disclosed through the joins. Since an apriori definition of risky features is insufficient, PRIVEE also suggests additional features to defenders for diagnosing sensitive matches. In this section, we first discuss the methods used for evaluating the disclosures, followed by the design of the visual cues that can help evaluate them.

8.1 Methods for disclosure evaluation

During the red-teaming exercise, we realized that the join key could vastly influence the disclosure of sensitive information. In this subsection, we discuss two methods for disclosure evaluation:

Based on the low frequency of matching records: *Matching records* are the number of records present in the joined dataset. Hence, the presence of matching records can indicate the possible disclosures at the record level. However, the number of matching records may vary according to the choice of attributes in the join key and the type of records present in the datasets. For example, when joined on attributes x and y, dataset A and dataset B may have 200 matching records, but when joined on the attributes x, y, and z, they may have only 20 matching records. This implies that the attribute combination x, y, and z have a better chance of discovering an actual disclosure than the combination x and y. We have also

observed that matching records may contain duplicates if the original datasets have duplicate or blank entries.

Based on the mutual information between the joining attributes: The selection of the joining attributes is an iterative process in PRIVEE. Mutual information measures the amount of information one random variable contains about another [12] and quantifies the mutual dependence of the two attributes of a dataset. Hence, we use normalized mutual information to suggest other features that defenders can use for detecting disclosures. PRIVEE automatically calculates the normalized mutual information between the joining attributes and the other attributes of the joined dataset. Next, it finds the top-5 attributes with the highest mutual information score and lets defenders consider those features for detecting matches **(T5)**.

8.2 Visual cues for evaluating disclosures

The design of the Disclosure Evaluation View follows Shneiderman's mantra [53], where PRIVEE first provides an overview of the matched records, then allows the defender to explore them, and finally lets them view the record details on demand. Here we discuss the comparative visual cues [17] that aid in disclosure evaluation:

Exploration of matching records: Parallel Sets is a visualization method for the interactive exploration of categorical data, which shows the data frequencies instead of the individual data points [34]. PRIVEE shows the matching records using a modified parallel sets visualization, as illustrated in Figure 3d. Here, each attribute of the join key is represented using a stacked bar, where the height of the stacks represents the frequency of the different categories of that attribute. In the case of a numerical attribute, a histogram replaces the stacked bar and shows its data distribution. The numerical data is then divided into four equal bins to map them with the categories of the other join key attributes. The parallel sets for the privacy-related attributes are colored in violet, while that for the other attributes are colored in grey, following the similar color scheme used in the other views. The categories across the numerical and categorical attributes are connected using ribbons. Each ribbon represents the number of records in the joined dataset belonging to both categories. A simple click interaction on any of these ribbons opens a pop-up window showing the details of the records represented by the selected line.

This design helps detect both identity and attribute disclosures through the matching records **(T4)**. The thickness of the line may represent the identity disclosure, while the height of the stacked bar shows the attribute disclosure. For example, if there is only one record with a certain combination of all the join key attributes, this would be represented by a thin ribbon across the parallel sets visualization. This may potentially lead to identity disclosure if an individual is uniquely identified with this combination of the join key. Suppose if an attribute has only one category, then the corresponding stack height would cover all the height allocated to a certain attribute, revealing that all the individuals belonging to both the datasets have a particular feature and leading to attribute disclosure. This Disclosure Evaluation View helps the data defenders ascertain the degree of the sensitive information

disclosed by visualizing the overall relationship between the different attributes of the matching records yet retaining the granularity of the dataset at the record level.

Suggesting potential joining attributes: PRIVEE uses bar charts and histograms to encode the top-5 features with high mutual information with the join key attributes. These suggestions are positioned on the left and right-hand sides of the parallel sets, representing the feature suggestions from either of the datasets (Figure 3d). The privacy-related attributes are also highlighted in violet, while the others are colored in grey, following a color scheme similar to the interface's other views. Selecting any attributes from the feature suggestions would also update this visualization to include the newly selected attributes. These attributes can be used as suggestions for improving the initial set of joining attributes **(T5)**. The data distributions and the ranking of the attributes help boost defenders' understanding of the risky feature set that can be used as the join key.

9 Case Study: Risk Confidante

We report a case study that our data privacy collaborator and coauthor co-developed using the web interface of PRIVEE. He is a senior researcher with more than 15 years of experience in privacy-preserving data analysis and used PRIVEE as a privacy auditor. Specifically, he wanted to determine if there are any disclosure risks with the health-related datasets published in the open data portals and validate the role of PRIVEE as a risk confidante for data defenders.

Our collaborator selected the aggregated datasets in the interface PRIVEE along with the privacy-related attributes *age* and *race*; and then filtered them with the keyword "health" (see Figure 6a). He also enabled the Vulnerable Datasets switch to check if there are any vulnerabilities in the data distributions of these datasets. At this point, our collaborator observed that the first few clusters do not have such vulnerable datasets. However, the fourth cluster has the dataset *Whole Person Care Demographics 2* [60] from the open data portal of San Mateo county [54]. This dataset had only 1 record where the race was Hawaiian (Figure 6c) (**T2**). This was a significant cause of concern since if somebody knows a person in that county who identified as Hawaiian, then any dataset with a similar race category could potentially expose her health records. Thus, he started analyzing the risk of joining this dataset with all the individual-record level datasets available through PRIVEE, as shown in Figure 6c (**T3**). He decided to join these dataset pairs on the selected privacy-related attributes and the location attribute *geocodedcolumnn* since he wished to find datasets containing information relevant to this location. He observed that none of the top-4 dataset pairs yield any matching record when joined on these attributes (Figure 6d). Thus, our collaborator concluded that though this aggregated dataset has a meager count of a particular race, it does not lead to any disclosure (**T4**). He also analyzed a few other vulnerable datasets similarly but found no disclosures. Thus, PRIVEE acts as a risk confidante for the data defenders where they can analyze the disclosure risks for the vulnerable datasets in the presence of other open datasets. He also observed that he had not seen a tool with similar capabilities for interactive risk calibration and triage and commented: "*this is a great visual tool to explore privacy risks of open data, with the ability to visualize privacy risk across datasets in a dynamic manner*".

10 Case Study: Trusted Informer

We report a case study that a researcher developed using the PRIVEE web interface. He is a senior researcher and university professor with over 25 years of experience in the fields of big data, cyber security, and scientific visualization. He focused on validating the role of PRIVEE as a trusted informer for the data defenders.

The researcher started by choosing the New Orleans Open Data portal [39] and observed 7 datasets on the Projection View, which were so similar in their attribute space that they were displayed using an overlapping circle with the number of datasets inscribed. Using the attribute distribution bar chart, he observed that none of the default privacy-related attributes (age, race, gender) were present in this group of datasets. However, on analyzing the word cloud, he made an interesting observation that attributes like *victim age* and *offender age* were present in these datasets, as shown in Figure 7a **(T1)**. Since, from his background knowledge, he knew that these attributes are generally present in police datasets, he updated the list of privacy-related attributes to select some of the similar attributes like *victim age*, *victim gender*, *victim race*, and *offender age*. As PRIVEE helps to triage the joinable groups of datasets based on the data defender's definition of privacy relevance, the Projection View was updated to reflect the change in privacy-related attributes.

He selected all these seven datasets in order to compare the joinability risks of the 21 possible pairwise combinations in the Risk Assessment View **(T3)**. Since he wanted to focus only on the high-risk pairs, he filtered out the low-risk pairs using the Risk Score Distribution histogram. Joining the first pair of datasets, the researcher observed that there are no matching records between them.

Next, he selected a pair of datasets, namely *Electronic Police Report* 2016 and *Electronic Police Report* 2015, but augmented the PRIVEE-suggested join key attributes and made the following selection: *location*, *victim age*, *offender age*, *victim race*, *victim gender*, *offender gender*, as illustrated in Figure 7b **(T3)**. He joined these datasets and observed 14 matching records in the Disclosure Evaluation View. He inspected further details about a certain record and observed that a 22-year-old black male was charged with attempted robbery with a gun against a 27-year-old white male at 6*XX* Tchoupitoulas St on 13th July 2015 at 01 : 00 hrs and again on 30th April 2016 at 03: 00 hrs with attempted simple robbery **(T4)**. Next, from the feature suggestions offered by PRIVEE **(T5)**, he selected the attribute *disposition*, which shows the status of a particular incident. He observed that only one record was open in 2015 but closed in 2016 (Figure 7c). On inspecting further details, as shown in Figure 7d, he found out that an incident of a runaway female juvenile of age 17 was reported at 85*XX* Dinkins St on 26th February 2015, and the same incident was closed through a supplemental report one and half years later on 7th December 2016 **(T4)**.

The researcher concluded that this is an example of identity disclosure where individuals were identified using PRIVEE even when the addresses were partially masked in de-identified datasets. He was also shown an earlier version of the PRIVEE interface during the case study. He commented that the new changes "*improved the rich functionalities*"

of PRIVEE and added that this interface could "*help experienced data custodians analyze disclosure risks and potentially find examples of disclosures*".

11 Discussion

When plugged into the open data stream, PRIVEE can act as both a risk profiler and a trusted informer that oversees risks while providing an appropriate level of control to defenders for integrating their domain knowledge using an end-to-end workflow. One of the lessons learned during this design study is that an interface helping defenders evaluate disclosures should enable seamless communication across sources and implications of risks while responding to the myriad definitions of privacy relevance. PRIVEE is bootstrapped by a default view that quickly adapts to the data defenders' inputs, allowing them to leverage appropriate levels of control while automating parts of the analysis process.

In its current implementation, one of the limitations of PRIVEE is scalability, concerning the number of records of each processed dataset and the size of the seed input that is used for bootstrapping. We have limited the number of records to 100, 000 to avoid interaction latency.

There is also the need to incorporate greater automation in the selection of privacy-relevant, personal datasets without manual inter vention. During this design study process, we learned that automation of this workflow is inherently challenging as privacy-relevance is subjective and open data are noisy; hence, training a model to mimic human judgment is difficult. Our approach of specifying a seed set outside the PRIVEE workflow is an important methodological choice allowing us to focus on the most vulnerable datasets and anticipated attack scenarios. Currently, PRIVEE only assesses joinability risk between pairs of datasets. It is certainly possible that there could be other scenarios like when multiple datasets are joined progressively, the risks propagate through the links. However, based on the feedback of our data privacy collaborator, we consider the risk scenarios handled in PRIVEE to be the necessary first steps toward assessing more complex combinations and variants of disclosure risks.

12 Conclusion

PRIVEE, the visual risk inspection workflow described in this design study paper, is a first step towards allowing data defenders both the control and efficiency needed to minimize disclosure risks from the joinability of open datasets. Through our case studies with data privacy experts, we demonstrated a key takeaway that the visualizations and interactions were effective in end-to-end exploration and diagnosis of actual disclosure of sensitive information or identity of individuals. As an ongoing and future work, we will be exploring disclosure risks beyond joinable pairs. We will further augment our workflow with intelligent and scalable data processing capabilities in collaboration with big data experts. We also plan to conduct controlled studies for evaluating the usability of PRIVEE and its components with real-world cyber defenders.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The work reported in this publication was supported by the National Science Foundation (CNS-2027789) and the National Institutes of Health (R35GM134927). The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research.

REFERENCES

[1]. Alfalayleh M and Brankovic L. Quantifying privacy: A novel entropy-based measure of disclosure risk. In International Workshop on Combinatorial Algorithms, pp. 24–36. Springer, 2014.

[2]. Almeida F and Xexéo G. Word embeddings: A survey. arXiv preprint arXiv:1901.09069, 2019.

[3]. Ankerst M, Breunig MM, Kriegel H-P, and Sander J. Optics: Ordering points to identify the clustering structure. ACM Sigmod record, 28(2):49–60, 1999.

[4]. Bakarov A. A survey of word embeddings evaluation methods. arXiv preprint arXiv:1801.09536, 2018.

[5]. Bhattacharjee K, Chen M, and Dasgupta A. Privacy-preserving data visualization: Reflections on the state of the art and research opportunities. In Computer Graphics Forum, vol. 39, pp. 675–692. Wiley Online Library, Norrköping, Sweden, 2020.

[6]. Bhattacharjee K, Islam A, Vaidya J, and Dasgupta A. PRIVEE-NJIT dataset 10.7910/DVN/VHOR3V, 2022. doi: 10.7910/DVN/VHOR3V

[7]. Caliński T and Harabasz J. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1):1–27, 1974.

[8]. Charter OD. Our history - international open data charter https://opendatacharter.net/our-history/, 2020. (Accessed on 07/19/2021).

[9]. Chia PH, Desfontaines D, Perera IM, Simmons-Marengo D, Li C, Day W-Y, Wang Q, and Guevara M. Khyperloglog: Estimating reidentifiability and joinability of large data at scale. In 2019 IEEE Symposium on Security and Privacy (SP), pp. 350–364. IEEE, 2019.

[10]. City of Dallas Open Data https://www.dallasopendata.com/. (Accessed on 10/05/2021).

[11]. Cover TM. Elements of information theory John Wiley & Sons, 1999.

[12]. Cover TM, Thomas JA, et al. Entropy, relative entropy and mutual information. Elements of information theory, 2(1):12–13, 1991.

[13]. Culnane C, Rubinstein BI, and Teague V. Health data in an open world. arXiv preprint arXiv:1712.05627, 2017.

[14]. Dai AM, Olah C, and Le QV. Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998, 2015.

[15]. Dasgupta A, Kosara R, and Chen M. Guess me if you can: A visual uncertainty model for transparent evaluation of disclosure risks in privacy-preserving data visualization. VizSec, pp. 1–10, 2019.

[16]. Dasgupta A, Maguire E, Alfie A-R, and Chen M. Opportunities and challenges for privacy-preserving visualization of electronic health record data In Proceedings of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records, 2014.

[17]. Dasgupta A, Wang H, Nancy O, and Burrows S. Separating the wheat from the chaff: Comparative visual cues for transparent diagnostics of competing models. IEEE Transactions on Visualization and Computer Graphics, 2020.

[18]. Davies DL and Bouldin DW. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, pp. 224–227, 1979. [PubMed: 21868852]

[19]. De Montjoye Y-A, Hidalgo CA, Verleysen M, and Blondel VD. Unique in the crowd: The privacy bounds of human mobility. Scientific reports, 3(1):1–5, 2013.

[20]. De Montjoye Y-A, Radaelli L, Singh VK, and Pentland AS. Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221):536–539, 2015. [PubMed: 25635097]

[21]. Diaz C, Seys S, Claessens J, and Preneel B. Towards measuring anonymity In International Workshop on Privacy Enhancing Technologies, pp. 54–68. Springer, 2002.

[22]. Dong Y, Takeoka K, Xiao C, and Oyamada M. Efficient joinable table discovery in data lakes: A high-dimensional similarity-based approach In 2021 IEEE 37th International Conference on Data Engineering (ICDE), pp. 456–467. IEEE, 2021.

[23]. Douriez M, Doraiswamy H, Freire J, and Silva CT. Anonymizing nyc taxi data: Does it matter? In 2016 IEEE international conference on data science and advanced analytics (DSAA), pp. 140–148. IEEE, 2016.

[24]. Dwork C. Differential privacy and the us census In Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, pp. 1–1, 2019.

[25]. Dwork C, Roth A, et al. The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci, 9(3–4):211–407, 2014.

[26]. Eick SG and Karr AF. Visual scalability. Journal of Computational and Graphical Statistics, 11(1):22–43, 2002.

[27]. Ester M, Kriegel H-P, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd, vol. 96,34, pp. 226–231, 1996.

[28]. City of Fort Lauderdale Police Department Open Data https://fortlauderdale.data.socrata.com/. (Accessed on 10/05/2021).

[29]. Fung BC, Wang K, Chen R, and Yu PS. Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (Csur), 42(4):1–53, 2010.

[30]. Green B, Cunningham G, Ekblaw A, Kominers P, Linzer A, and Crawford SP. Open data privacy. Berkman Klein Center Research Publication, pp. 17–07, 2017.

[31]. O. K. O. D. Group. Open definition, 2015.

[32]. Hutchins EM, Cloppert MJ, Amin RM, et al. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. Leading Issues in Information Warfare & Security Research, 1(1):80, 2011.

[33]. Kao C-H, Hsieh C-H, Chu Y-F, Kuang Y-T, and Yang C-K. Using data visualization technique to detect sensitive information re-identification problem of real open dataset. Journal of Systems Architecture, 80:85–91, 2017.

[34]. Kosara R, Bendix F, and Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE transactions on visualization and computer graphics, 12(4):558–568, 2006. [PubMed: 16805264]

[35]. Kum H-C, Ragan ED, Ilangovan G, Ramezani M, Li Q, and Schmit C. Enhancing privacy through an interactive on-demand incremental information disclosure interface: Applying {Privacy-by-Design} to record linkage In Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019), pp. 175–189, 2019.

[36]. Lavrenovs A and Podins K. Privacy violations in riga open data public transport system In 2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE), pp. 1–6. IEEE, 2016.

[37]. Lloyd D and Dykes J. Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study. IEEE transactions on visualization and computer graphics, 17(12):2498–2507, 2011. [PubMed: 22034371]

[38]. Miller RJ, Nargesian F, Zhu E, Christodoulakis C, Pu KQ, and Andritsos P. Making open data transparent: Data discovery on open data. IEEE Data Eng. Bull, 41(2):59–70, 2018.

[39]. City of new orleans — open data https://datadriven.nola.gov/home/. (Accessed on 11/02/2021).

[40]. Nouwens M, Liccardi I, Veale M, Karger D, and Kagal L. Dark patterns after the gdpr: Scraping consent pop-ups and demonstrating their influence In Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–13, 2020.

[41]. NYC Open Data https://opendata.cityofnewyork.us/. (Accessed on 10/05/2021).

[42]. Oganian A and Domingo Ferrer J. A posteriori disclosure risk measure for tabular data based on conditional entropy. SORT 2003, Vol. 27, Num. 2 [July-December], 2003.

[43]. Ohm P. Broken promises of privacy: Responding to the surprising failure of anonymization. Ucla L. Rev, 57:1701, 2009.

[44]. Open Data Kansas City https://data.kcmo.org/. (Accessed on 10/05/2021).

[45]. Rocher L, Hendrickx JM, and De Montjoye Y-A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1):1–9, 2019.

[46]. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65, 1987.

[47]. Rubinstein IS and Hartzog W. Anonymization and risk. Wash. L. Rev, 91:703, 2016.

[48]. Ruggles S, Fitch C, Magnuson D, and Schroeder J. Differential privacy and census data: Implications for social and economic research. In AEA papers and proceedings, vol. 109, pp. 403–08, 2019.

[49]. Schubert E and Gertz M. Improving the cluster structure extracted from optics plots. In LWDA, 2018.

[50]. Schubert E, Sander J, Ester M, Kriegel HP, and Xu X. Dbscan revisited, revisited: why and how you should (still) use dbscan. ACM Transactions on Database Systems (TODS), 42(3):1–21, 2017.

[51]. Sekara V, Alessandretti L, Mones E, and Jonsson H. Temporal and cultural limits of privacy in smartphone app usage. Scientific reports, 11(1):1–9, 2021. [PubMed: 33414495]

[52]. Serjantov A and Danezis G. Towards an information theoretic metric for anonymity In International Workshop on Privacy Enhancing Technologies, pp. 41–53. Springer, 2002.

[53]. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. In The craft of information visualization, pp. 364–371. Elsevier, 2003.

[54]. SMC Datahub https://datahub.smcgov.org/. (Accessed on 10/07/2021).

[55]. spacy industrial-strength natural language processing in python https://spacy.io/. (Accessed on 11/14/2021).

[56]. Thongtan T and Phienthrakul T. Sentiment classification using document embeddings trained with cosine similarity In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 407–414, 2019.

[57]. Van der Maaten L and Hinton G. Visualizing data using t-sne. Journal of machine learning research, 9(11), 2008.

[58]. Vassilvitskii S and Arthur D. k-means++: The advantages of careful seeding In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027–1035, 2006.

[59]. Visualizing re-identification risk using data studio — data loss prevention documentation — google cloud https://cloud.google.com/dlp/docs/visualizing_re-id_risk. (Accessed on 06/28/2022).

[60]. Whole Person Care Demographics 2 — SMC Datahub https://datahub.smcgov.org/dataset/Whole-Person-Care-Demographics-2/qqdq-93h5. (Accessed on 10/07/2021).

[61]. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al. The fair guiding principles for scientific data management and stewardship. Scientific data, 3(1):1–9, 2016.

[62]. Zang H and Bolot J. Anonymization of location data does not work: A large-scale measurement study In Proceedings of the 17th annual international conference on Mobile computing and networking, pp. 145–156, 2011.

[63]. Zenko M. Red Team: How to succeed by thinking like the enemy Basic Books, 2015.

[64]. Zhang T, Ramakrishnan R, and Livny M. Birch: an efficient data clustering method for very large databases. ACM sigmod record, 25(2):103–114, 1996.

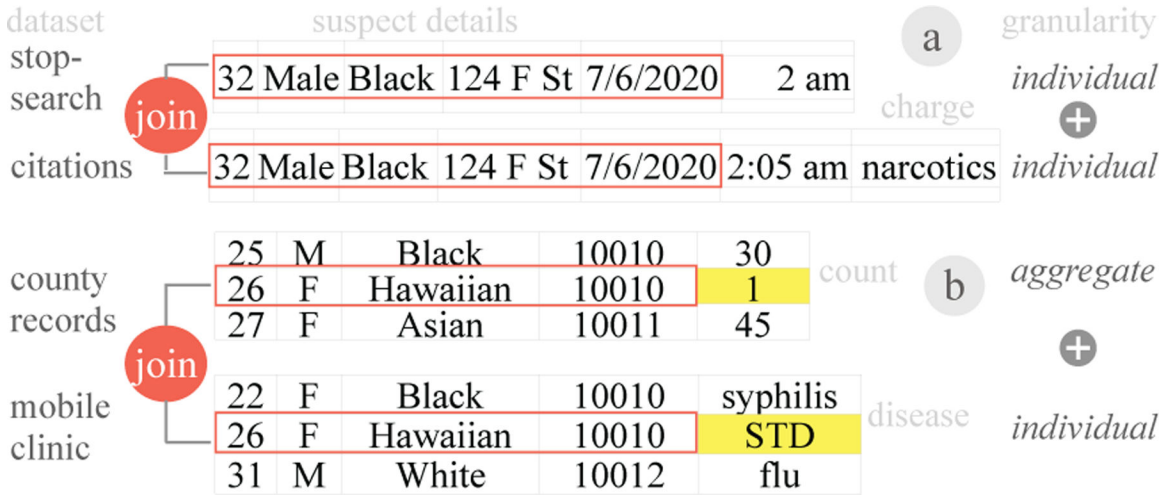

Figure 1: Open Data Disclosure risks using real datasets:

(a) Two *individual-level* de-identified open datasets can be joined using quasi-identifiers like age, sex, race, location, and date in order to identify an individual and reveal sensitive information about them like their citation charge. (b) *Aggregated-level* datasets can cause disclosure risk when a record with a meager value can be joined with another individual-level dataset, e.g., the only 26-year-old Hawaiian female living in a particular zip code has STD.

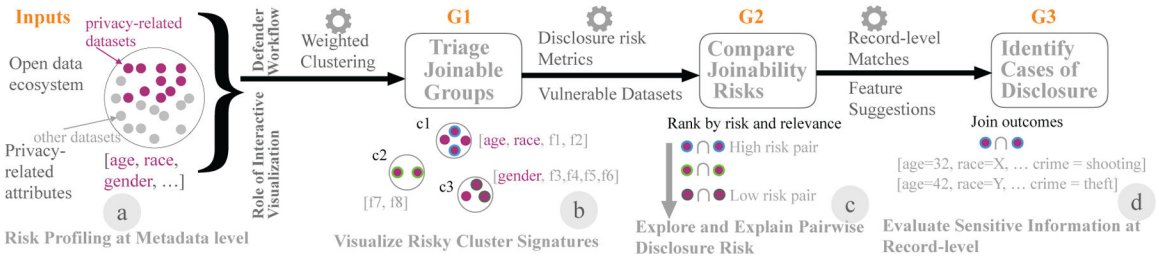

Figure 2: PRIVEE is an end-to-end risk inspection workflow for open datasets

that informs the defender in the analytical loop about potential disclosure risks in the presence of joinable datasets. Interactive visualization plays a crucial role in bootstrapping the risk inspection process via risk profiling, triaging and explaining risk signatures, and ultimately detecting instances of true disclosure at a record level. Colored borders track datasets across the goals.

Figure 3: Interface Design:

The design of PRIVEE comprises rich interaction among filters and multiple views: (a) Filter area helps select datasets based on metadata like tags, data granularity, and privacy-related attributes; (b) Projection View lets the defenders compare the signatures of different joinable groups of datasets and evaluate vulnerable data distributions; (c) Risk View helps compare the risk for dataset pairs and select the high-risk pairs; (d) Disclosure Evaluation View helps to analyze the matching records for potential disclosures.

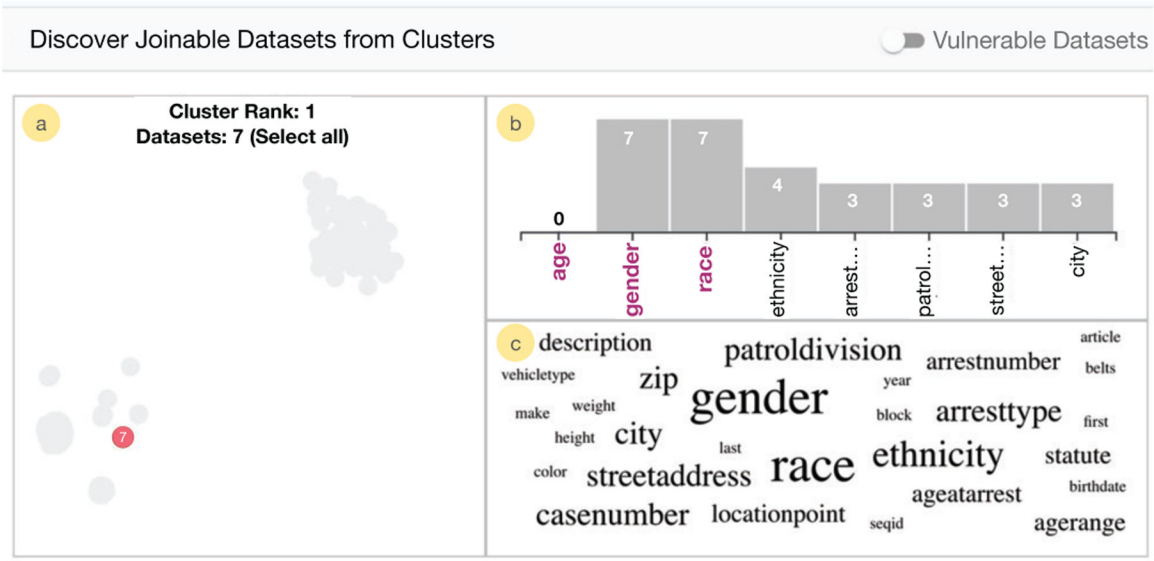

Figure 4: Projection View:

A group of joinable datasets is represented using (a) a projection plot. The (b) frequency distribution bar chart and (c) word cloud for the attributes of a group of joinable datasets help in the transparent explanation of the group signatures.

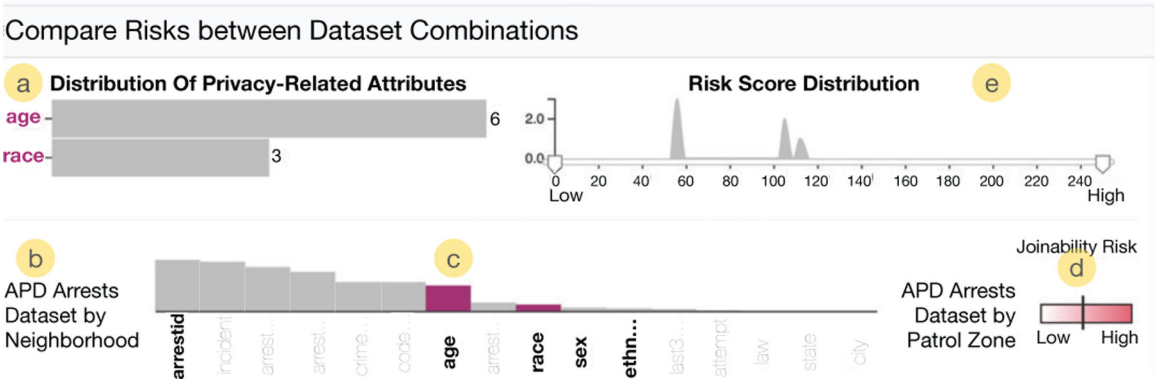

Figure 5: Risk Assessment View:

(a) The distribution of privacy-related attributes can affect the joinability risks between (b) dataset pairs. Data defenders can compare the risk between these pairs by analyzing the (c) sorted bar chart showing the shared attributes and the joinability risk score represented by the (d) risk score bar. They can use the (e) risk score distribution histogram to focus on the dataset pair of their interest.

Figure 6: PRIVEE as a risk confidante for defenders:

(a) selecting datasets based on their metadata like the popular tag “health” and their granularity of records, (b) finding and diagnosing the vulnerable data distributions and observing that there is only 1 record for the race “Hawaiian”, (c) comparing the joinability risk with the individual record-level datasets and (d) evaluating the disclosures with the top 4 individual-level datasets and observing that there is no disclosure.

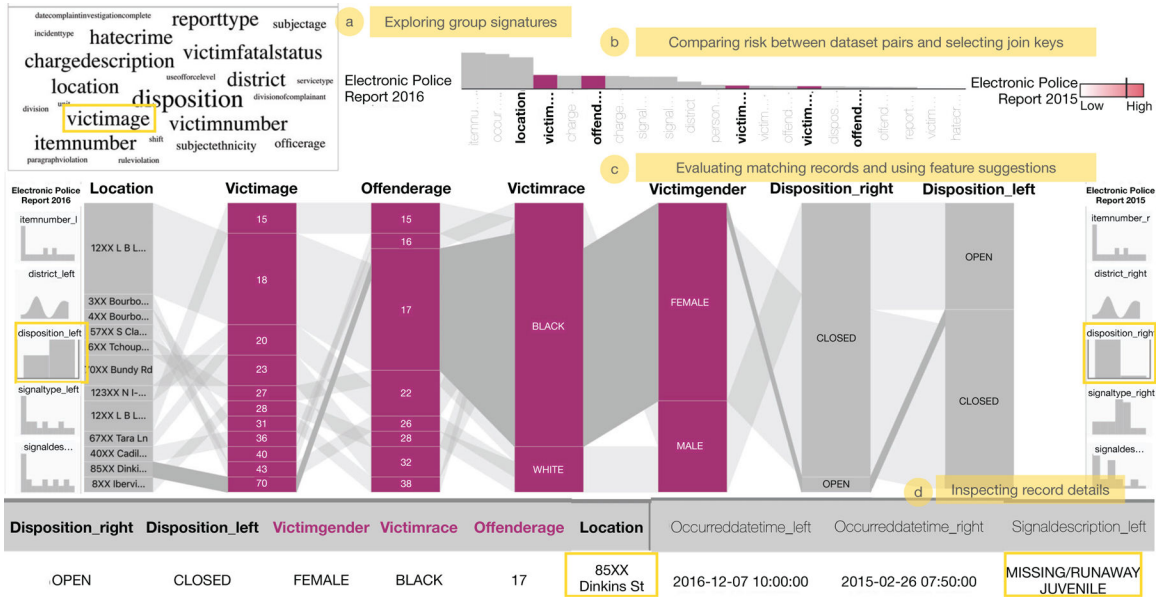

Figure 7: PRIVEE as a trusted informer for defenders:

(a) understanding group signatures and updating privacy-related attributes, (b)comparing the risk between dataset pairs, (c) evaluating the matching records using the feature suggestions shows that only one incident was open in 2015 but closed in 2016, (d) inspecting record details shows that a runaway juvenile can be identified despite the location being partially masked.

Table 1:

Sample record points

Record points	Description
[“age”, 11, 1]	For age=11, there is only 1 record
[“age”, 15, 5]	For age=15, there are 5 records
[“gender”, “F”, 2]	For gender=“F”, there are 2 records