Differential Inference Testing: A Practical Approach to Evaluate Sanitizations of Datasets

Ali Kassem, Gergely Acs, Claude Castelluccia, Catuscia Palamidessi

Differential Inference Testing: A Practical Approach to Evaluate Sanitizations of Datasets

Ali Kassem
Inria
Grenoble, France
ali.kassem@inria.fr

Gergely Ács
Crysys Lab, BME-HIT
Budapest, Hungary
acs@crysys.hu

Claude Castelluccia
Univ. Grenoble Alpes, Inria
F-38000 Grenoble, France
claude.castelluccia@inria.fr

Catuscia Palamidessi
Inria, École Polytechnique
Univ. of Paris Saclay, Paris, France
catuscia@lix.polytechnique.fr

***Abstract*—In order to protect individuals' privacy, data have to be "well-sanitized" before sharing them, i.e. one has to remove any personal information before sharing data. However, it is not always clear when data shall be deemed well-sanitized. In this paper, we argue that the evaluation of sanitized data should be based on whether the data allows the inference of sensitive information that is specific to an individual, instead of being centered around the concept of re-identification. We propose a framework to evaluate the effectiveness of different sanitization techniques on a given dataset by measuring how much an individual's record from the sanitized dataset influences the inference of his/her own sensitive attribute. Our intent is not to accurately predict any sensitive attribute but rather to measure the impact of a single record on the inference of sensitive information. We demonstrate our approach by sanitizing two real datasets in different privacy models and evaluate/compare each sanitized dataset in our framework.**

***Index Terms*—Sanitization, Inferences, Machine Learning, k-Anonymity, ℓ-Diversity, Differential Privacy**

I. Introduction

Nowadays, organizations own large volumes of data about individuals. Sharing those data provides several benefits for both organizations and individuals. But, at the same time, it puts individuals' privacy at high risk. A straightforward countermeasure to protect individuals' privacy, known as *pseudoanonymization*, is to exclude explicit identifiers such as name, address, and phone number. However, it has been shown that pseudoanonymization is not sufficient to protect individuals' privacy as the remaining information such as date of birth, gender, and zip code can be used to re-identify individuals [19], [20].

In order to provide more guarantees about individuals' privacy, more sophisticated techniques have been proposed to sanitize data from information that may lead to re-identification. Examples of such *sanitization techniques* are mechanisms that rely on data suppression and generalization (known as anonymization techniques) [12], [16], [18], and those that rely on noise addition like in differential privacy [6]. Nevertheless, there is neither well-defined scheme to evaluate the robustness of sanitization techniques, nor a clear understanding for "when data is regarded as well-sanitized". The European General Data Protection Regulation considers data as properly sanitized (anonymized) if "data subject is no longer identifiable". A more specific approach can be found in the Working Party 29 opinion on 05/2014 about "Anonymization Techniques", which considers the following three privacy risks: "re-identification", "linkability" and "inference".

In this paper, we argue that inferences should be the primary concern when it comes to individuals' privacy. In particular, we see identity disclosure as one way among others to infer information about individuals. Actually, mitigating "identity disclosure" is the primary goal of pseudoanonymization, however, it is not always relevant to data sanitization. Indeed, if a dataset is "completely-sanitized", then assigning an identity to a certain record is pointless as the records will be highly noised or aggregated. However, as far as the effectiveness of sanitization is concerned, we should be aware about the precise meaning of information inference as preventing any kind of inferences usually lead to useless data [7]. Indeed, the ultimate usefulness of a dataset is always to infer new information. So, as a trade off between privacy and utility, sanitized data should not allow the inferences of "private" information , but at the same time, they have to allow the inference of some "public" information about the population, i.e., the acquisition of any generalizable knowledge. The acceptability or unacceptability of an inference can be based on two criteria:

1) *The basis of the inference*: is the inference performed on the records of one (or a small group of) individual(s) or a large group of individuals. We will call these kinds of inferences *private* and *public*, respectively.
2) *The nature of the inference*: can the inference be used to discriminate users? Can it have a very negative (for example social or financial) impact?

The intuition behind the first criterion is that if an adversary cannot prove that the records of a user were used to generate the inference, then, by definition, these records are "protected". Note that, it might happen that a model which is exclusively built on population characteristics also accurately predicts some sensitive information of individuals who are member

Most of this work was achieved when Ali Kassem was a postdoc in the Inria team COMETE, in collaboration with the PRIVATICS team. Gergely Ács was supported by the Premium Post Doctorate Research Grant of the Hungarian Academy of Sciences (MTA) and the Higher Education Excellence Program of the Ministry of Human Capacities (BME FIKP-MI/FM). Claude Castelluccia was supported by the French National Research Agency in the framework of the "Investissements d'avenir" program (ANR-15-IDEX-0002). Catuscia Palamidessi was supported by the French National Research Agency in the framework of REPAS project (ANR-16-CE25-0011).

of this population [3]. However, we do not consider this to be a privacy breach as long as the population, which is used to build the model, is large enough. Instead, as in [15], we believe that there are acceptable and unacceptable disclosures: "learning statistics about a large population of individuals is acceptable, but learning how an individual differs from the population is a privacy breach". For instance, inferring an attribute value about the population of a large city, or a rule like "a man smoking between 1 and 4 cigarettes per day is 3 times more likely to die from lung cancer than a non-smoker" should be acceptable. But, deriving some information about the inhabitants of a building may or may not be acceptable depending on the number of people in the building.

As regards the second criterion, the inference nature is partly subjective and involves ethical and legal considerations [7]. In this paper, we focus on the first criterion and propose a framework called *Differential Inference Testing* to assess the inference basis. Namely, we make the following contributions:

- We propose an inference-based framework that can be used to evaluate the robustness of a given sanitized dataset against a specific adversary that is modeled by an inference algorithm (Section III). In particular, the adversary builds a machine learning model in order to infer an individual's sensitive attribute from his publicly known attributes in the sanitized dataset. We consider the attack successful (and the data not "well-sanitized") if the adversary obtains sufficiently different (but perhaps inaccurate) results depending on whether the target individual's record was used to train the model or not, i.e., the output of the inference potentially leaks some individual specific information aside from more general population characteristics. Our approach is reminiscent of Differential Privacy [6], however, it also differs from that in several aspects that we detail in Section II-B. A key feature of our testing procedure is that it needs to have access only to the sanitized data itself and requires no knowledge about the sanitization technique. Thus, it can be used to assess datasets that are sanitized by organizations which may prefer not to disclose their sanitization techniques. Even more, the verifier (e.g., Data Protection Authority) of a sanitization process does not need to understand or analytically analyze its privacy guarantees which can be very tedious and error-prone [13].
- We use our framework to evaluate two datasets after being sanitized in the k-anonymity, ℓ-diversity, and differential privacy (Section IV) models. In this paper, we consider microdata, but our solution is general and can be applied to any type of datasets, such as aggregated data.

II. Related Work

A. Testing Data Sanitization

To the best of our knowledge, our approach is the first one that proposes a *general* practical test to evaluate sanitized datasets by making distinction between acceptable and unacceptable inferences. Yet, there are some prior related works [1], [4], [5], [17].

The authors of [4] propose a framework to test whether a machine learning (ML) model can predict sensitive attribute values from a given sanitized dataset. But, they consider all types of inferences as privacy breaches. More precisely, their framework tests, for every record, whether the ML model can predict the true value of the sensitive attribute. If the ML model succeeds to predict the true value (what they call "empirical utility"), then the sanitization technique does not pass the test. Note that the framework does not consider whether the prediction was obtained from the record of the target individual (that was somehow poorly sanitized) or from the records of other users (that happen to be correlated with the target individual). By contrast, we propose a framework that does not consider data utility (i.e., ignores the accuracy of inferences in absolute sense), but instead tests whether an inference is private (depends on the target individual) or public. In our framework, a dataset is deemed "well-sanitized" if it can be shown that, for any user, the resulting inferences based on this dataset do not depend on the contribution of a single user but on the contribution of all users together: the inference accuracy should not change too much whether the user's record is included or not in the dataset. Such a dataset protects against "private" inferences while still allowing "public" inferences.

Recently, [1] and [5] have proposed statistical techniques to identify the violations of differential privacy. Unlike these approaches, our method considers the sanitization technique as a black-box and only requires access to the sanitized datasets. This can be a favorable feature if the sanitization schemes are proprietary and their exact operations are not published. Also, our testing procedure is more general as it can be applied beyond differential privacy.

Pyrgelis et al. [17] used machine learning for membership inference on aggregated location data. They build a single binary classifier to predict a given individual's presence in the sanitized data. By contrast, we follow a more general approach and measure how much the inference of a particular sensitive information/attribute is affected by a single individual's data using a specified distance measure. For this purpose, we build two classifiers; one which predicts the sensitive attribute using all the sanitized dataset and another one which uses the sanitized data excluding the individual's data, then report the difference between the output of these classifiers according to the chosen distance measure. Of course, we can easily turn our approach into membership inference by combining the output of the two models into a single binary classifier to infer membership. However, the choice of different distance measures allows to incorporate different privacy requirements into our framework which makes our approach more general. For instance, membership may be already publicly known, but not some sensitive attributes.

B. Differential Inference Testing vs. Differential Privacy

Our approach is inspired by differential privacy [6]. Indeed like differential privacy, it guarantees that the inferences one can derive from a sanitized dataset are similar, whether or

not the record of a certain individual is included. However it differs from differential privacy in several aspects:

- Our approach provides a method to measure the robustness of sanitized *datasets*, and to compare different sanitizations of the same dataset. Differential privacy, on the other hand, is a property of the sanitization scheme and not of the sanitized dataset.
- For differential privacy to hold, the sanitization must be done probabilistically (typically, by adding controlled noise to the answer to the query). Our approach, on the contrary, can also be applied to deterministic sanitization techniques, like k-anonymity [18] and ℓ-diversity [16].
- The possible inferences one can make in differential privacy are strictly related to the query for which the mechanism is defined without any further restrictions on how the inference model is built. In our case the inferences are produced by a machine learning algorithm, which constitutes a parameter of the framework.
- In differential privacy the metric used to compare the inferences in the dataset with and without a certain individual is fixed and based on the upper bound to the likelihood ration. In our setting, the comparison is based on a parametric notion of distance between distributions.
- Differential privacy relies on tedious and error-prone analytical analyses of the privacy guarantee, while our approach uses easy-to-implement empirical evaluation of a very similar (but weaker) guarantee [1], [14].

III. Differential Inference Testing

In this section, we introduce the notion of *indifferentiability* (Section III-A), then we propose a testing procedure in order to evaluate the indifferentiability of a given sanitized dataset against a certain inference model (Section III-B).

A. Model

Given a sanitized dataset, our approach tests whether the inference of some sensitive attribute(s) is influenced by the presence of any single individual in the dataset. If the "amount" of this influence is large, then the inference leaks some private information, i.e., any information that potentially differentiates the individual from the rest. In this case, the dataset is not sanitized properly. Conversely, smaller influence indicates stronger sanitization. In order to measure such an influence, we propose the notion of δ-indifferentiability defined in Definition 1. Without loss of generality, we express the sensitive attribute(s) to be inferred using a single attribute S, which can be any function of other attributes. Note that, the explicit distinction between quasi and sensitive attributes is only for demonstration purposes. Moreover, we assume that the contribution of every individual i to D is a single record (q^i, s^i) where q^i represents his quasi-identifiers and s^i represents his sensitive value.

Definition 1 (Indifferentiability): Let D be a dataset (Q, S) where Q is a tuple of quasi-identifiers and S is a sensitive attribute. Let D^{-i} denote the dataset obtained from D by removing the record (q^i, s^i) of individual i. Let $\mathcal{A}$ be a (possibly randomized) inference algorithm, and let f be a sanitization technique. Let $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$ denote the random variables describing the output of the models $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$ which are built according to $\mathcal{A}$ respectively using $f(D)$ and $f(D^{-i})$ to provide each, given a quasi-identifier tuple from Q, a prediction distribution over the domain of the attribute S. We say that $f(D)$ is δ-**indifferentiable** with respect to $\mathcal{A}$, if we have that

$$\forall (q^i, s^i) \in D,\ \mathsf{distance}\big(\mathcal{M}_{f(D)}, \mathcal{M}_{f(D^{-i})}\big) \leq \delta$$

where distance is a statistical distance measure.

Somewhat abusing the notation, $\mathcal{M}$ denotes both the model and the random variable describing its output henceforth.

The evaluation of the sanitized dataset $f(D)$ depends on the value of δ. If δ is small enough (depending on the case study) then, for every individual i in D, the inference about i's sensitive value does not strongly depend on the i's record (i.e., public inference). On the other hand, for larger δ's, there may exist i's in D such that the inference about the i's sensitive value depends on the i's record (private inference).

Definition 1 does not consider any external knowledge about i or D, however, it can be generalized in a straightforward way to capture any such possible knowledge. For instance, in case of Bayesian inference, auxiliary information can be used to compute the prior probabilities, and thus favorizing one value of the sensitive attribute over the others.

The inference algorithm $\mathcal{A}$ represents the adversarial strategy to predict/infer the value of the sensitive attribute. The choice of inference algorithm $\mathcal{A}$ depends on the case study and is a task of the analyst. Indeed, the differentiability δ of sanitized dataset $f(D)$ depends on the considered inference algorithm $\mathcal{A}$. In this paper, we use Bayesian inference as an inference algorithm for it is simplicity and popularity. Note that, the aim of our framework is not to accurately predict any sensitive attribute but rather to measure the impact of a single record on the inference of sensitive information.

We note that $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$ belong to the same model family since they are built using the same algorithm $\mathcal{A}$. For instance, if $\mathcal{M}_{f(D)}$ is a neural network then $\mathcal{M}_{f(D^{-i})}$ is also a neural network with the same architecture and with the same hyper parameters, but with potentially different model parameters as they are trained using two different training datasets $f(D)$ and $f(D^{-i})$. In the rest of the paper, we may use $\mathcal{M}$ and $\mathcal{M}^{-i}$ to refer to $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$, respectively. Definition 1 assumes that the output of a model $\mathcal{M}$ is a vector of probability values, i.e., a *prediction distribution* on the possible values of the sensitive attribute. Specifically, if there are n possible sensitive values $s_1, \ldots, s_n$, then for some record (q^i, s^i), $\mathcal{M}(q^i) = \{(s_1, p_1^i), \ldots, (s_n, p_n^i)\}$ where p_j^i denotes $\mathcal{M}$'s confidence that $s^i = s_j$. In the rest of the paper, we refer to the number of possible sensitive values by n, and we write $\mathcal{M}(q^i) = (p_1^i, \ldots, p_n^i)$ when the related sensitive values are clear from the context. Similarly for model $\mathcal{M}^{-i}$, we write $\mathcal{M}^{-i}(q^i) = (p_1^{-i}, \ldots, p_n^{-i})$.

Finally, distance denotes a distance measure (such as total variation distance, KL-divergence, etc.) chosen by the analyst.

The choice of distance should depend on the privacy requirements, and it fundamentally impacts the result of our approach.

B. Testing Procedure

We propose a procedure to find, given a sanitized dataset $f(D)$ and an inference algorithm $\mathcal{A}$, the minimal distance δ such that $f(D)$ is δ-indifferentiable with respect to $\mathcal{A}$ (that is the the minimal δ that satisfies Definition 1). In order to perform the test, one should also have access to $f(D^{-i})$ for every $(q^i, s^i) \in D$. Nevertheless, the sanitization technique f itself is not needed by our testing procedure. The testing procedure runs through the following steps:

1) Choose a record $(q^i, s^i) \in D$ for some individual i.
2) Use $\mathcal{A}$ to build two models $\mathcal{M}$ and $\mathcal{M}^{-i}$ respectively using datasets $f(D)$ and $f(D^{-i})$. For example, in the case where $\mathcal{A}$ is a machine learning algorithm, then $f(D)$ and $f(D^{-i})$ will act as training datasets.
3) Provide q^i as an input to the model $\mathcal{M}$. The output of $\mathcal{M}$ takes values from the set of all prediction distributions which corresponds to an n-dimensional simplex in $\mathbb{R}^n$ (i.e., $\mathcal{M}(q^i) = (p_1^i, \ldots, p_n^i)$ over the domain of the sensitive attribute S, where $\sum_{j=1}^n p_j^i = 1$).
4) Repeat the last step for $\mathcal{M}^{-i}$ whose output also takes values from the set of all prediction distributions (i.e., $\mathcal{M}^{-i}(q^i) = (p_1^{-i}, \ldots, p_n^{-i})$ over the domain of the sensitive attribute S, where $\sum_{j=1}^n p_j^{-i} = 1$).
5) Compute the distance $d^i = \mathsf{distance}(\mathcal{M}, \mathcal{M}^{-i})$.
6) Repeat Steps 1-5 for every individual i in D.
7) Return the maximal distance d^i for every $i \in D$, as δ.

The outputs of $\mathcal{M}$ and $\mathcal{M}^{-i}$ can be described by random variables whose output range is the n-dimensional simplex in $\mathbb{R}^n$. Indeed, the sanitization algorithm f is a possibly randomized black-box mechanism, which means that the output distributions of $\mathcal{M}^i$ and $\mathcal{M}^{-i}$ can only be approximated by sampling. However, sampling from the n-dimensional simplex is not scalable if n is large and/or there are many records in D. Hence, in this paper, we rely on the following simplification; we approximate the distribution of every coordinate of the prediction distribution independently, and compare the approximated distributions of the corresponding coordinates. More precisely, for every $1 \leq j \leq n$, let $\mathcal{P}_j^i$ and $\mathcal{P}_j^{-i}$ denote the random variables describing the values of p_j^i and p_j^{-i}, respectively. Then, $\mathsf{distance}(\mathcal{M}, \mathcal{M}^{-i}) = \sum_{j=1}^n \mathsf{div}(\mathcal{P}_j^i, \mathcal{P}_j^{-i})$, where div is a distance measure (or divergence) between distributions. In this paper, we use the 1^{st} Wasserstein distance (or Earth Mover's Distance, shortly EMD) as such a distance measure, that is, $\mathsf{div}(\mathcal{P}_j^i, \mathcal{P}_j^{-i}) = \mathsf{EMD}(\mathcal{P}_j^i, \mathcal{P}_j^{-i}) = \inf_\pi \int_{\mathbb{R}} \int_{\mathbb{R}} |x - y| d\pi(x, y)$, where the infimum is taken over all probability measures π on $\mathbb{R} \times \mathbb{R}$ with marginals $\mathcal{P}_j^i$ and $\mathcal{P}_j^{-i}$. Intuitively, EMD measures how far one has to move the probability mass of $\mathcal{P}_j^i$ to turn it into $\mathcal{P}_j^{-i}$, where "farness" between the values of $\mathcal{P}_j^i$ and $\mathcal{P}_j^{-i}$ is measured by their absolute distance[1]. We approximate the empirical measures of $\mathcal{P}_j^i$ and $\mathcal{P}_j^{-i}$ and compute the EMD between these empirical measures (see [2] for details). In particular, if $x_1^i, \ldots, x_N^i$ and $x_1^{-i}, \ldots, x_N^{-i}$ denote the samples taken from the distributions of $\mathcal{P}_j^i$ and $\mathcal{P}_j^{-i}$, respectively, then $\mathsf{EMD}(\mathcal{P}_j^i, \mathcal{P}_j^{-i})$ can be approximated by:

$$\mathsf{EMD}(\hat{\mathcal{P}}_j^i, \hat{\mathcal{P}}_j^{-i}) = \frac{1}{N} \sum_{k=1}^{N} |x_{(k)}^i - x_{(k)}^{-i}| \quad (1)$$

where $\hat{\mathcal{P}}_j^i$ and $\hat{\mathcal{P}}_j^{-i}$ denote the empirical measures of $\mathcal{P}_j^i$ and $\mathcal{P}_j^{-i}$, respectively, and $x_{(k)}^i$ denote the k^{th} order statistic (k^{th} smallest value) of samples $x_1^i, \ldots, x_N^i$ (analogously to $x_{(k)}^{-i}$).

We use EMD as it makes randomized and deterministic sanitizations comparable in our framework. In particular, the uncertainty of the adversary has two sources; one is measured deterministically by the inference algorithm and represented by the prediction confidences of each sensitive value in its output. The second source of uncertainty stems from the "artificially" introduced perturbation in the sanitization process (e.g., by the Laplace Mechanism in differential privacy) which induces a probability distribution on these (deterministic) confidences. Unlike traditional divergences like total variation distance or the max-ratio distance used in differential privacy, EMD also considers the value of the inference algorithm's output and not only the distribution of these values.

Example. Consider the dataset presented in Table Ia. It has two quasi-identifier attributes: "Age" (an integer) and "Gender" (M:Male or F:Female), and a sensitive attribute: "Disease" which can take two values (Flu and Cancer). Table II represents a 2-anonymous version of this dataset (as every record is syntactically indistinguishable from at least another record considering their quasi-identifiers)[2].

TABLE I: Original dataset, and related result.

(a) Dataset.

#	Age	Gender	Disease
1	28	M	Flu
2	36	M	Flu
3	47	F	Cancer
4	53	M	Flu
5	72	F	Flu

(b) $\mathcal{M}(q^i)$, $\mathcal{M}^{-i}(q^i)$, and d^i.

#	$\mathcal{M}(q^i)$	$\mathcal{M}^{-i}(q^i)$	d^i
1	(1, 0)	(1/2, 1/2)	1
2	(1, 0)	(1/2, 1/2)	1
3	(2/3, 1/3)	(1, 0)	2/3
4	(2/3, 1/3)	(1/2, 1/2)	1/3
5	(2/3, 1/3)	(1/2, 1/2)	1/3

Let i denote the individual that corresponds to record 4 from Table Ia. If we know that i is a 53 years old male, then we can infer from Table IIa the following prediction distribution about his disease: $Pr[\text{Flu} \mid (53, \text{M})] = \frac{2}{3}$ and $Pr[\text{Cancer} \mid (53, \text{M})] = \frac{1}{3}$, i.e., $\mathcal{M}(q^4) = \{(\text{Flu}, \frac{2}{3}), (\text{Cancer}, \frac{1}{3})\}$, as $(53, \text{M})$ belongs to the second equivalence class of Table IIa which is composed of the last three records (Step 3). Notice that we used a very simple inference algorithm $\mathcal{A}$ here for simplicity (i.e., computing the probability of a sensitive value conditioned on the values of all the quasi-identifiers), though one can use any sophisticated inference model in practice.

[1] EMD permits different "farness" measures other than the absolute difference $|x - y|$. We chose this metric due to its simplicity and fast computation.

[2] In our example, the sanitization technique f is k-anonymity [18].

Now, if we remove the 4^{th} record from the original dataset (Table Ia) then apply 2-anonymity we obtain Table IIb. It is important to remove the record from the original dataset before applying sanitization again. Hence, the new sanitized dataset $f(D^{-4})$ (after removing the record) can be different from the first sanitized dataset $f(D)$ (obtained by sanitizing the whole original dataset). The prediction distribution after removing the 4^{th} record is $\mathcal{M}^{-i}(q^4) = \{(\text{Flu}, \frac{1}{2}), (\text{Cancer}, \frac{1}{2})\}$ (Step 4).

Finally, considering EMD distance, then $d^i = \mathsf{distance}(\mathcal{M}^4, \mathcal{M}^{-4}) = \mathsf{distance}(\mathcal{M}(q^4), \mathcal{M}^{-i}(q^4)) = \sum_{j=1}^{n} \left|p_j^4 - p_j^{-4}\right| = \frac{1}{3}$ (Step 5) since $\mathcal{M}(q^4)$ and $\mathcal{M}^{-i}(q^4)$ are the only possible output of $\mathcal{M}$ and $\mathcal{M}^{-4}$, respectively (i.e., the sanitization scheme is deterministic). After repeating the previous steps for every record in the dataset (Step 6), the maximal distance $\delta = \max_{i \in D} d^i$ can be computed (that is the the minimal distance that satisfies Dentition 1), which is $\delta = 1 = \max\{1, 1, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}\}$ in our example. Table Ib depicts the distributions $\mathcal{M}(q^i), \mathcal{M}^{-i}(q^i)$ and the distance d^i for every record i in the dataset of Table Ia.

Notice that using a different distance metric the results can completely change. For example, if $\mathsf{distance}$ denotes the total variation distance (TVD), then $\mathsf{distance}(\mathcal{M}^4, \mathcal{M}^{-4}) = \mathsf{distance}(\mathcal{M}(q^4), \mathcal{M}^{-i}(q^4)) = 1 + 1 = 2$ which suggests that the data is blatantly non-private as $\mathsf{distance}(\mathcal{M}^i, \mathcal{M}^{-i}) \leq n$ for any i.

TABLE II: 2-anonymous versions of Table Ia.

(a) $f(D)$ (with $i = 4$).

#	Age	Gender	Disease
1	< 45	M	Flu
2	< 45	M	Flu
3	≥ 45	{M, F}	Cancer
4	≥ 45	{M, F}	Flu
5	≥ 45	{M, F}	Flu

(b) $f(D^{-4})$ (without $i = 4$).

#	Age	Gender	Disease
1	< 45	M	Flu
2	< 45	M	Flu
3	≥ 45	{M, F}	Cancer
5	≥ 45	{M, F}	Flu

IV. Evaluation

A. Datasets

We demonstrate our approach using two datasets: the UCI Adult (Census Income) dataset[3] and the "General Demographics" dataset from Internet Usage data[4]. Table III summarizes, for each dataset, its size ($|D|$), number of distinct record ($|D^\dagger|$), quasi-identifiers (QI), and sensitive attribute (SA) as well as the number of values that SA can take (n).

TABLE III: Datasets description.

Dataset	Adult	Internet Usage
$\|D\|$	10,000	9,799
$\|D^\dagger\|$	7,960	7,049
QI	"age", "education" "marital status" "hours per week" "native country"	"age", "race" "education attainment" "major occupation" "marital status"
SA	"occupation"	"household income"
n	14	9

[3]https://archive.ics.uci.edu/ml/datasets/Adult

[4]http://www.cc.gatech.edu/gvu/user_surveys/survey-1997-10

B. Sanitization

For sanitization techniques, we consider the basic Mondrian k-anonymity [9], Mondrian ℓ-diversity [10], and data perturbation required to satisfy differential privacy [6].

1) k-anonymity and ℓ-diversity: The Mondrian sanitization algorithm modifies the records by generalizing the quasi-identifiers until each record becomes syntactically indistinguishable from $k - 1$ other records (k-anonymity), or the correct sensitive value of any individual cannot be predicted with probability more than $1/\ell$ (ℓ-diversity)[5]. After being generalized, the data is then published with the related sensitive values.

2) Differential Privacy: As for differential privacy, noisy statistics of the microdata are computed which are then used to compute the prediction distribution of each record. In particular, contrary to k-anonymity and ℓ-diversity, differential privacy provides weak utility when it is directly applied on microdata. Instead of generating sanitized microdata, the differentially private prediction distributions $\mathcal{M}(q^i)$ and $\mathcal{M}^{-i}(q^i)$ are directly computed from the original dataset, and these distributions are compared in Step 5 of the Differential Inference Test described in Section III. In other words, our differential private sanitization technique f is coupled with a simple inference algorithm $\mathcal{A}$ that we describe below.

The sanitization technique f consists of releasing the perturbed conditional probabilities $Pr[s|q]$ for all sensitive attribute value s and value of quasi-identifier tuple q. These conditional probabilities are directly used as the prediction distributions $\mathcal{M}(q^i)$ and $\mathcal{M}^{-i}(q^i)$ in our Differential Inference Test (see Section III). Specifically, in order to obtain the differential private prediction distributions $\mathcal{M}(q^i)$ and $\mathcal{M}^{-i}(q^i)$ for a quasi-identifier tuple $q^i = (q_1^i, \ldots, q_m^i)$, we compute

$$Pr[s_k|q^i] = \frac{Pr[s_k, q^i]}{Pr[q^i]} \tag{2}$$

In Eq. (2), we calculate the joint probability $Pr[s_k, q^i]$ as

$$Pr[s_k, q^i] = \frac{C_{ik}}{\sum_i \sum_k C_{ik}} \tag{3}$$

and the marginal probability $Pr[q^i]$ as $Pr[q^i] = \sum_k Pr[s_k, q^i]$ where $C_{ik} = 1 + max(0,\ |\{(q, s) \in D : q = q^i \wedge s = s_k\}| + noise)$ and the noise is drawn from the Laplacian distribution $\mathcal{L}(0, 1/\epsilon)$ with zero mean and variance $2/\epsilon^2$. This perturbation technique is also referred to as the Laplace Mechanism in the literature of differential privacy[6]. Note that the addition of 1 to C_{ik} is the standard Laplacian correction in order to avoid zero value of the denominator in Eq. (3).

The privacy guarantee of differential privacy comes from the randomness of the Laplace Mechanism; if the variance of the added noise is larger, we have stronger guarantee (i.e., smaller ε), and the reverse direction holds for small variance. Differential privacy is formally defined in Definition 2.

[5]Using only $f(D)$ as a background knowledge for inference.

[6]The scale parameter of the Laplace noise is adjusted to the global sensitivity of the counts C_{ik} which is 1 in our application.

Definition 2 (ϵ-differential privacy [6]): A sanitization algorithm f guarantees ϵ-differential privacy if for any database D and D', differing on at most one record, and for any possible output $O \subseteq Range(f)$, $e^{-\epsilon} \leq \frac{Pr[f(D) \in O]}{Pr[f(D') \in O]} \leq e^{\epsilon}$.

In our case, the range of f is the space of all prediction distributions (i.e., vectors from an n-dimensional simplex), and an output O of f is a random vector from this space.

C. Pre-processing

Many sanitization techniques (such as Mondrian) generalize the attribute values according to a specific generalization hierarchy. In order to feed the learning algorithm with generalized data in our experiments, we use an encoding mechanism that is relative to the target record selected in Step 1 of our testing procedure (see Section III-B) and works as follows: a generalized quasi-attribute value q' (e.g., an interval or set) is represented by 1, if the corresponding quasi-attribute value q'' of the target record can also be generalized to q' (i.e., q'' is inside q' if q' is an interval, or q'' is a member of q' if it is a set). Otherwise, q' is represented by 0.

For example, consider the two generalized records: $r_1 = ([15, 25], \text{Female}, \{\text{France}, \text{Germany}\})$ and $r_2 = ([17, 20], \text{Male}, \{\text{Italy}, \text{Germany}\})$. Assuming that the target record, which is selected in the first step of our testing procedure, is $r_t = (16, \text{Male}, \text{France})$, then r_1 and r_2 will be encoded as follows: $encode(r_1, r_t) = (1, 0, 1)$ because $16 \in [15, 25]$, Male $\neq$ Female, and France $\in \{\text{France}, \text{Germany}\}$. Similarly, $encode(r_2, r_t) = (0, 1, 0)$ because $16 \notin [17, 20]$, Male = Male, and France $\notin \{\text{Italy}, \text{Germany}\}$.

An advantage of this encoding technique is that it depends on the target record which will be used as an input for the inference model. This may increase the sensitivity to the presence of the target record in the dataset, and thus help to better capture the difference between the two intended distributions. Another advantage of this approach is that it is very fast to compute and has to be done only once for each record (other approaches may require different encodings of the same record for the computation of $\mathcal{M}$ and $\mathcal{M}^{-i}$). Nevertheless, any encoding mechanism can be used in our framework as long as the encodings of each record is sufficiently different from that of the target record.

D. Differential Inference Test

1) Inference algorithm: For the purpose of inference $\mathcal{A}$ and the computation of the prediction distributions $\mathcal{M}(q^i)$ and $\mathcal{M}^{-i}(q^i)$, we use a Naive Bayes classifier[7] in the case of k-anonymity and ℓ-diversity, and the noised conditional probabilities in Eq. (3) in the case of differential privacy. In both cases, the inference algorithms use the encoded sanitized data to build the models $\mathcal{M}$ and $\mathcal{M}^{-i}$. The Naive Bayes classifier has been used by several prior works [3], [4], [11] to perform inference on sanitized data. Although Naive Bayes makes the simplistic assumption that the quasi-identifiers are independent, it usually performs remarkably well, especially when the size of the training dataset is not so large.

[7]We use the Bernoulli Naive Bayes from the sklearn python module.

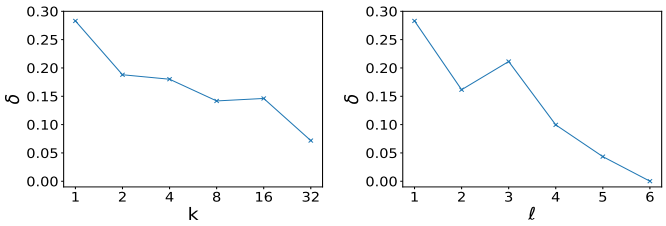
After choosing the encoding mechanism and the inference algorithm, we proceed according to the Differential Inference Test described in Section III-B: for every record r_i in the dataset, we train a model $\mathcal{M}$ where D includes r_i, and also train another model $\mathcal{M}^{-i}$ where D^{-i} excludes r_i. Then, the corresponding two prediction distributions of $\mathcal{M}$ and $\mathcal{M}^{-i}$ are approximated by sampling, and the distance $d^i = \mathsf{distance}(\mathcal{M}, \mathcal{M}^{-i})$ is computed for every i. Finally, we obtain the minimal distance δ that satisfies Dentition 1, i.e., $\delta = \max_{i \in D} d^i$. In what follows, "minimal δ" refers to the minimal distance δ that satisfies Dentition 1.

We emphasize that the (in)differentiability of a sanitized dataset depends on the inference algorithm $\mathcal{A}$, which represents the adversarial algorithm to infer sensitive information from the dataset. This is in stark contrast to differential privacy, which provides the same guarantee (i.e., the same ϵ value) against all inference algorithms. On the other hand, (in)differentiability (in Definition 1) can be empirically evaluated unlike differential privacy (in Definition 2) which relies on analytical evaluation that is often tedious and error-prone.

E. Results

1) k-anonymity and ℓ-diversity:

a) Adult Dataset: Figure 1 depicts the minimal δ (that satisfies Dentition 1 in the case of Adult dataset) depending on the privacy parameter (k or ℓ). The cases of $k = 1$ and $\ell = 1$ implies the absence of sanitization, i.e., the testing procedure is applied directly on the original data without any sanitization.



(a) k-anonymity. (b) ℓ-diversity.

Fig. 1: Minimal δ (Adult data).

From Figure 1, we can notice that:

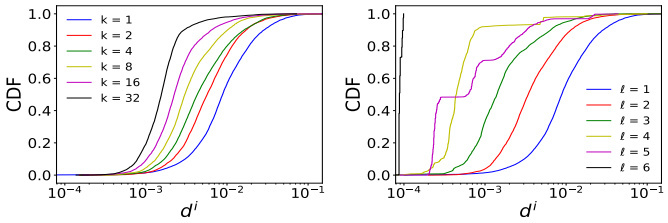
- δ is smaller for ℓ-diversity than for k-anonymity when k and ℓ have identical values. This is expected as, unlike k-anonymity, ℓ-diversity was designed to mitigate inference attacks, though not the same type of inference that we measure in our approach. Specifically, ℓ-diversity addresses the *absolute* accuracy of inferences. By contrast, we focus on the *relative* accuracy of inferences.
- δ decreases when the privacy parameter (k or ℓ) increases, however not monotonically. For instance, counterintuitively, the minimal δ increases when k increases from 8 to 16 and when ℓ increases from 2 to 3.

The second observation above shows that increasing the value of the privacy parameter may decrease the privacy guarantees against private inferences for some individuals (worst-case

privacy), even if the guarantees on average (average-case privacy) can be stronger. In particular, Figure 2 presents the Cumulative Distribution Function (CDF) of d^i. The CDF is the sum of the relative frequencies for all values that are less than or equal to the given value of d^i. Figure 2 shows that, for both k-anonymity and ℓ-diversity, the majority of d^i values are smaller for larger k or ℓ. The CDF illustrates the level of the average-case privacy, which increases if the value of the privacy parameter also increases (as one could expect for k-anonymity and ℓ-diversity). This emphasizes the fact that average-case privacy, which is usually adopted by companies and governments' regulations, does not always imply worst-case privacy, which is considered in our framework. Indeed, Dentition 1 has to be satisfied for every record in the dataset.

A closer investigation reveals that there are only few outlier records with large value of d^i when $\ell = 3$ or $k = 16$. For example, when $\ell = 3$, there is only one record r_i whose d^i value is greater than 10^{-1} (for this record, $d^i = 0.21$), which is the minimal δ in this case. Similarly, for 16-anonymity, there are only 3 outlier records which have much larger d^i values than others.

In order to achieve stronger sanitization (smaller δ), we remove the outlier records identified above from the original dataset, then repeat the entire testing procedure to compute a new value of minimal δ (remember that it is not sufficient to remove the outlier records only from the sanitized dataset). For $\ell = 2$, we obtain the following new values of minimal δ: 0.165 for $\ell = 2$ and 0.092 for $\ell = 3$, what one naturally expects when ℓ increases.



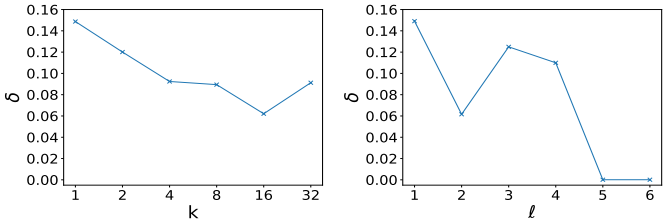
(a) k-anonymity. (b) ℓ-diversity.

Fig. 2: CDF of d^i (Adult data).

b) Internet Usage Dataset: Figures 3 and 4 present the minimal δ and the CDF of d^i, respectively, for the Internet Usage dataset. The results confirm the conclusion that average-case guarantees against private inferences often differs from the worst-case guarantees in practice due to the existence of a few outlier records with much worse privacy guarantee than the average. However, on average, increasing the privacy parameters k and ℓ results in stronger guarantee against private inferences using the Mondrian sanitization scheme.

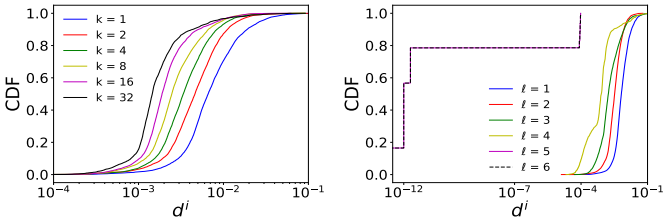
2) Differential Privacy:

a) Adult Dataset: Figure 5 presents the CDF and δ for the Adult dataset in the case of differential privacy (DP). We quantized the "age" and "hours per week" attributes, each, into 5 quantiles. This results into 2952 distinct records (instead of



(a) k-anonymity. (b) ℓ-diversity.

Fig. 3: Minimal δ (Internet Usage data).

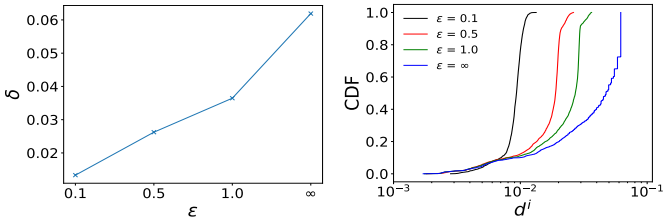


(a) k-anonymity. (b) ℓ-diversity.

Fig. 4: CDF of d^i (Internet Usage data).

7960) out of 10K records. $\epsilon = \infty$ corresponds to the case when no Laplace noise is added to the counts in Eq (2).

We performed $N = 25K$ samples per record in order to have an estimate of the prediction distribution for every individual. From these noisy predictions, we can compute the minimal δ (in Figure 5a) as it is described in Section III-B. The minimal δ curve shows that smaller ϵ indeed yields stronger protection, for every individual, against private inferences, as δ is monotonically decreasing with ϵ as one would expect.



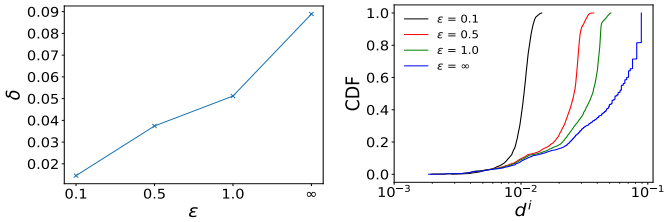
(a) Minimal δ. (b) CDF of d^i.

Fig. 5: CDF and δ (DP, Adult data, $N = 25K$).

b) Internet Usage Dataset: Figure 6 presents the CDF and δ for the Internet Usage dataset. Again, we have quantized the "age" attribute into 5 quantiles. This results into 2926 distinct records (instead of 7049) out of 9799 records.

The CDF and minimal δ curves (in Figure 6) show similar trends to the Adult dataset and confirm our observation that, for every individual, smaller ϵ yields stronger protection against private inferences.

Finally, we note that, for both datasets, the values of δ are smaller for k-anonymity and ℓ-diversity on average than for differential privacy (Figures 1, 3, 5, and 6). Hence, differential privacy even with $\epsilon = 0.1$ may provide weaker protection on



(a) Minimal δ. (b) CDF of d^i.

Fig. 6: CDF and δ (DP, Internet Usage data, $N = 25K$).

average against private inferences if we use the EMD distance measure defined in Eq. (1), but it is worst-case guarantee is superior to k-anonymity and even to ℓ-diversity if $\ell < 5$.

3) Computation time: All the experiments that are presented above were conducted on a machine with a 2.6 GHz Intel Core i7 Processor and 16 GB RAM. Table IV summarizes the average computation time of the differential inference test per privacy parameter (excluding the raw data case, i.e., where $k = \ell = 1$ and $\epsilon = \infty$). Note that the computation can be substantially improved since the sanitizations of the dataset (per user) are highly parallelizable.

TABLE IV: Average computation time.

Dataset	k-Anonymity	ℓ-Diversity	Differential Privacy
Adult data	60m 24s	76m 6s	5h 41m 51s
Internet data	47m 30s	50m 12s	4h 23m 22s

V. Conclusion

This paper presents an inference-based framework to evaluate the effectiveness of sanitization performed on a given dataset. In particular, we empirically measured how the sanitized dataset prevents the private inferences of sensitive attribute values. We demonstrated the usage of this framework on two datasets. Our framework allows to compare different sanitized datasets that might use different privacy models, such as k-anonymity, ℓ-diversity or differential privacy. It can potentially be employed by companies or DPAs (Data Protection Authorities) to test the robustness of sanitized datasets. It is important to note that our solution tests the robustness of sanitized datasets, not that of the underlying sanitization technique.

Our case study shows that ℓ-diversity and k-anonymity can provide stronger average protection against private inferences in our framework than differential privacy if ϵ is chosen to be too large. This result should be handled with caution, since these techniques have quite different adversary models and some attacks which are hard to be modeled using a machine learning algorithm in our framework can have devastating effect on ℓ-diversity and k-anonymity yet still difficult to launch against a differentially private dataset [8]. In particular, our model considers only a specific adversarial inference attack as well as some potentially defined extra background knowledge of the adversary. We also showed that increasing the value of k and ℓ results in stronger protection on average, but can also entail weaker worst-case guarantee when each individual is considered.

We believe that there is a need for a toolkit to test the robustness of sanitized datasets by implementing different re-identification or inference attacks. Our framework could be one component of such a toolkit. One benefit of the proposed testing tool is that the sanitized dataset is analyzed as a "black box", i.e. the sanitization algorithm does not need to be published. It is enough for the verifier to get access to an oracle that, given a dataset, outputs its sanitized version. We believe this is a desirable property for at least two reasons: (i) many companies are unwilling, for different reasons, to publish their sanitization algorithms, and (ii) the verifier does not need to go through the difficulty of understanding and analyzing the underlying algorithm.

In the proposed framework, the verifier can use his favorite inference models. This paper uses a Naive Bayes classifier, but other classifiers could be used. Evaluating our framework with other classifiers is part of our future work.

References

[1] B. Bichsel, T. Gehr, D. Drachsler-Cohen, P. Tsankov, and M. T. Vechev. Dp-finder: Finding differential privacy violations by sampling and optimization. In *ACM CCS*, pages 508–524, 2018.
[2] S. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics,and kantorovich transport distances. To appear in Memoirs of the AMS, 2016.
[3] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *ACC SIGKDD*, 2011.
[4] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Empirical privacy and empirical utility of anonymized data. In *Workshops Proceedings of ICDE*, 2013.
[5] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer. Detecting violations of differential privacy. In *ACM CCS*, pages 475–489, 2018.
[6] C. Dwork. Differential privacy. In *ICALP*, 2006.
[7] K. E. Emam and C. Alvarez. A critical appraisal of the article 29 working party opinion 05/2014 on data anonymization techniques. *Oxford Journals*, 5(1):73–87, 2015.
[8] S. R. Ganta, S. P. Kasiviswanathan, and A. D. Smith. Composition attacks and auxiliary information in data privacy. In *ACM SIGKDD*, pages 265–273, 2008.
[9] Q. Gong. Implementation of basic mondrian k-anonymity. goo.gl/kKpqZ6. Accessed: 01-02-2019.
[10] Q. Gong. Implementation of mondrian l-diversity. goo.gl/2ymJ14.
[11] D. Kifer. Attacks on privacy and definetti's theorem. In *ACM SIGMOD*, 2009.
[12] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, 2007.
[13] M. Lyu, D. Su, and N. Li. Understanding the sparse vector technique for differential privacy. *PVLDB*, 10(6):637–648, 2017.
[14] M. Lyu, D. Su, and N. Li. Understanding the sparse vector technique for differential privacy. *PVLDB*, 10(6):637–648, 2017.
[15] A. Machanavajjhala and D. Kifer. Designing statistical privacy for your data. *Commun. ACM*, 58(3), 2015.
[16] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1):3, 2007.
[17] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro. Knock knock, who's there? membership inference on aggregate location data. In *NDSS*, 2018.
[18] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI CS Laboratory, 1998.
[19] L. Sweeney. Uniqueness of simple demographics in the us population. *LIDAP-WP4. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA*, 2000.
[20] B. Woodward. The computer-based patient record and confidentiality. *New England Journal of Medicine*, 333(21):1419–1422, 1995.