Differential Inference Testing A Practical Approach to Evaluate Anonymized Data

Ali Kassem, Gergely Acs, Claude Castelluccia

Differential Inference Testing

A Practical Approach to Evaluate Anonymized Data

Ali Kassem
INRIA Grenoble
INRIA Saclay
Saclay, France
ali.kassem@inria.fr

Gergely Ács
Budapest University of
Technology and Economics
CrySyS Lab
Budapest, Hungary
acs@crysys.hu

Claude Castelluccia
INRIA Grenoble
Grenoble, France
claude.castelluccia@inria.fr

ABSTRACT

In order to protect individuals' privacy, governments and institutions impose some obligations on data sharing and publishing. Mainly, they require the data to be "anonymized". In this paper, we shortly discuss the criteria introduced by European General Data Protection Regulation to assess anonymized data. We argue that the evaluation of anonymized data should be based on whether the data allows individual based inferences, instead of being centered around the concept of re-identification as the regulation has proposed. We then propose a framework that allows us to evaluate a given (anonymized) dataset. Finally, we apply our framework to evaluate two real datasets after being anonymized using k-anonymity and l-diversity techniques.

Keywords

Anonymization, Inferences, Machine Learning, k-anonymity, l-diversity

1. INTRODUCTION

Nowadays, organizations own large volumes of data about individuals. Using and sharing these data may provide lot of benefits for both organizations and individuals. However, individuals' privacy must be preserved. To this end, several anonymization techniques [19, 17, 14, 15, 4] have been proposed. Nevertheless, there is neither well-defined scheme to evaluate the robustness of the anonymization techniques, nor a clear understanding for "when data is regarded as anonymous". Actually, defining effective criteria to answer the latter question is not an easy task. The aim of our work is to answer this question by proposing a framework that allows us to evaluate the robustness of an (anonymized) dataset. The evaluation is based on the type of the inferences that can be made from the dataset understudy.

Data Anonymization Assessment. The European Directive 95/46/EC considers a dataset as properly anonymized if "the data subject is no longer identifiable". To provide an interpretation for this Directive with regard to anonymization, the Working Party 29 (WP29) has published in April 2014 an "Opinion on Anonymization Techniques"[1]. This Opinion has put forward three criteria corresponding to three risks (namely, "singling out", "linkability" and "inference") that are claimed to be essential for anonymisation.

We argue that the WP29 criteria are neither necessary conditions nor effective means to assess anonymization techniques:

- They are not necessary because they do not take into account the type of information that can be derived. For instance, inferring or singling out noisy or useless data is not considered a real privacy threat. As an example, RAPPOR technology [6] does not jeopardize individual privacy (in particular, it provides strong deniability and differential privacy guarantees), however, it neither ensures "singling out" nor "linkability" criteria.

- As far as effectiveness is concerned, one should be aware about the precise meaning of "inference" criterion as preventing any kind of inferences usually lead to useless data. Indeed, the ultimate usefulness of a dataset is always to infer new information.

  The only way to make this criterion meaningful would be to qualify it and consider inferences about specific individuals with sufficient accuracy. For example, inferring an attribute about the population of a 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. However, deriving some information about the inhabitants of a building may or may not be acceptable depending on the size of the building.

Moreover, the concept of data re-identification is misleading: inference should be the primary concern. Actually, mitigating "identity disclosure" is the primary goal of pseudo-anonymization (where only identifiers are removed from the dataset), however, it is not relevant for data anonymization. Indeed, if a dataset is "correctly" anonymized, then assigning an identity to a certain record is pointless as the records will be highly noised or aggregated.

Public versus Private Inferences. We should avoid the fetichization of the notion of identification and rather see identity disclosure as one way among others to infer information about individuals. Accordingly, we have to deal with inferences as the main issue when individual's privacy is con-

[1]Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymization Techniques, 10 April 2014.

cerned. More specifically, anonymized datasets have to protect individuals against inferences of "private" information[2]? However at the same time, to provide some utility, they have to allow us to infer some "public" information about the population. 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 basis of the records of one (or a small group of) individual(s), that is *private inference*, or on the basis of the records of a large group of individuals (a population), that is *public inference*?

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 record(s) of a user was used to generate the anonymized dataset, then by definition this record is "protected". It might happen that the properties of the population can be used to build a model that can be applied to individuals with high accuracy [1]. However, we do not consider this to be a privacy breach as long as the group/population size is large enough. Instead, as in [16], 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".

The second criterion is partly subjective and involves ethical and legal considerations.

Toward Differential Inference Testing. In this paper, we focus on the first criterion and propose a scheme called *Differential Inference Testing* to assess the basis of the inference. However, we believe that the decisions to release a dataset should always be part of a rigorous privacy risk analysis, which systematically identifies the risks and the potential benefits of publishing the datasets [3]; especially that, certain public inferences can still be harmful.

In our scheme, a dataset is deemed "well-anonymized" if it can be shown that, for any user, the resulting inferences based on this dataset do not depend on the user's contribution (i.e., a single record) but on the contribution of other users (which may be correlated): the inference accuracy and certainty should be about the same whether the user's record is included or not in the dataset. Such a dataset protects against "private" inferences while still allowing "public" inferences.

Although our approach is inspired by differential privacy [5], it is not a privacy model *per se*. Instead, we propose a procedure that can be used to test the robustness of anonymized datasets or compare different anonymizations of the same dataset. Differential privacy is a property of the anonymization scheme and not that of the anonymized dataset. In particular, an anonymized dataset passing our test may still leak private information due to the following reasons. First, we only consider the posterior inference of *certain* attributes of each user *inside* the dataset. On the contrary, differential privacy bounds any information (posterior inference) which can differentiate *any* user inside or outside the dataset up to a threshold ε. That is, it may happen that an anonymized dataset, which is deemed robust by our approach, may still leak some private data of some individual who is either not in the dataset or has an attribute which is not tested by our scheme. However, we must note that differential privacy has at least three drawbacks. First, it usually provides very weak utility if micro-data should be released. Second, an anonymization scheme satisfying differential privacy must be randomized which often requires falsifying (perturbing) the data. This is often not tolerable in practice. Finally, the guarantee of differential privacy is hard to intuitively grasp. On the contrary, the interpretability of the robustness measured by our approach depends on the underlying inference algorithm. Hence, the data analyst has the option to choose a model where the inference results are relatively easy to interpret (such as decision trees).

Note that this paper focuses on microdata [3], but our solution is general and can be applied to any types of datasets, such as aggregated data.

Contributions. To the best of our knowledge, our approach is the first one that proposes a practical test to evaluate anonymized datasets by making distinction between acceptable and unacceptable inferences. Yet, there are some prior works, from which the one on empirical privacy [2] is the most related. This work proposes to test whether, e.g., a machine learning (ML) model can predict sensitive attribute values from a given (anonymized) dataset. But, it considers all types of inferences as privacy breaches. More precisely, the scheme tests, for every record, whether the ML model can predict the actual value of the sensitive attribute. If the ML model succeeds to predict the actual value (what they call "empirical utility"), then the anonymization technique does not pass the test. Note that the scheme does not consider whether the prediction was obtained from the record of the target individual (that was somehow poorly anonymized) or from the records of other users (that happen to be correlated with the target individual). By contrast, our framework does not consider data utility (i.e., does not concentrate whether the inferences are correct), but instead tests whether an inference is private (depends on the target individual) or public inference.

We make the following contributions:

- We propose an inference-based framework that can be used to evaluate the robustness of a given anonymized dataset against a specific inference model, e.g., a machine learning model (Section 2). Our approach evaluates the anonymized data itself, and deals with the related anonymization technique as a black-box. Thus, it can be used to assess datasets that are anonymized by organizations which may prefer not to provide ac-

[2]We regard private information as defined in [13], where a distinction between "personal information" and "private information" have been made. Private information is seen as "secrets that you can keep by withholding your data" whereas "personal data" could be derived by inference from datasets in which you are not necessarily involved.

[3]Microdata are datasets containing random samples of anonymous individual records, defined by a set of attributes (such as age, city, ...).

cess to their techniques.

- We use our framework to evaluate two datasets after being anonymized using k-anonymity and ℓ-diversity (Section 3).

In addition, the paper provides a conclusion and opens some perspectives (Section 4).

2. DIFFERENTIAL INFERENCE TESTING

The main idea is to test whether the prediction/inference of some sensitive attributes (such as salary, religion, occupation, etc.) is influenced by the presence of any single individual in the anonymized 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 anonymized properly. Conversely, smaller influence indicates stronger anonymization. To perform this test, the analyst selects (1) the attribute(s) whose private inference should be checked, and (2) the machine learning algorithm to build the corresponding inference model. In the rest, for simplicity, we assume that the analyst chooses a single sensitive attribute and a single learning algorithm for testing.

2.1 Model

Assume a dataset D with attributes (Q, S), where Q is a vector of quasi-identifier attributes and S is a sensitive attribute. Any attributes of an individual is a quasi-identifier which is easy to learn from other sources than D (e.g., social profiles, governmental registries which may contain ZIP codes, profession, birth date, etc.), whereas a sensitive attribute is only accessible from D and hence should be concealed (e.g., medical diagnosis, salary, etc.). Nevertheless, the distinction between quasi and sensitive attributes is context-dependent and should be analyzed in a systematic risk assessment.

Let $f(D)$ be an anonymized dataset computed from D, where f is an anonymization technique. Furthermore, let i be an individual with record $(q^i, s^i) \in D$, where q^i and s^i are respectively the quasi-identifiers and the sensitive value of the individual i. Consider an adversary that can perform an inference on s^i using $(q^i, f(D))$ and some external prior knowledge about i or D as evidences. For this purpose, the adversary builds a statistical inference model (e.g., machine learning model) using a learning algorithm $\mathcal{A}$ in order to infer the value of s^i given $(q^i, f(D))$.

In the following, we define the core concept of our paper called as the differentiability of an anonymized dataset.

Definition 1 (Differentiability). *Let D be a dataset with attributes (Q, S), and let D^{-i} denote a dataset obtained from D by removing the record of individual i. Let $\mathcal{A}$ be a (possibly) randomized inference algorithm, and f be an anonymization technique. Let $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$ denote the two models that are respectively built, using $\mathcal{A}$, from $f(D)$ and $f(D^{-i})$ to predict the attribute S. We say that $\mathcal{A}$ is an (ϵ, δ)-**differentiator** of $f(D)$ with respect to S, if for every record $(q^i, s^i) \in D$, we have that:*

$$\Pr[\texttt{distance}(\mathcal{M}_{f(D)}(q^i), \mathcal{M}_{f(D^{-i})}(q^i)) \geq \epsilon] \leq \delta$$

where `distance` *is a distance between the output distributions that is not necessarily a metric, and the probability is taken on the randomness of $\mathcal{M}_{f(D)}$, $\mathcal{M}_{f(D^{-i})}$, and f[4]. We say that $f(D)$ is (ϵ, δ)-**differentiable** with respect to $\mathcal{A}$ and S, if $\mathcal{A}$ is an (ϵ, δ)-differentiator of $f(D)$ with respect to S.*

A dataset $f(D)$ is not "well-anonymized" if there is an inference algorithm $\mathcal{A}$ which is (ϵ, δ)-differentiator with ϵ and δ that are both large enough. That is, $\mathcal{A}$ can produce models with quite different predictions or certainty of predictions depending on the presence of any single individual in D. In this case, $f(D)$ is likely to leak some private information of some users in D.

We note that $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$ belong to the same model family since they are built using 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 (but with potentially different parameters). In the rest of the paper, we use $\mathcal{M}$ and $\mathcal{M}^{-i}$ to refer to $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$, respectively.

We also note that, the output of 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 a 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 document, 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 $\mathcal{M}^{-i}(q^i) = (p_1^{-i}, \ldots, p_n^{-i})$.

2.2 Testing Procedure

Our approach is illustrated in Figure 1. The goal is to test whether an inference algorithm $\mathcal{A}$ is a differentiator of a sensitive attribute in a given (anonymized) dataset $f(D)$. The following steps are performed.

1. Given D, $f(D)$, and $f(D^{-i})$ for all $(q^i, s^i) \in D$. Choose an inference algorithm $\mathcal{A}$ and a sensitive attribute S for testing.
2. Choose an individual i from D.
3. Use $\mathcal{A}$ to build two models $\mathcal{M}$ and $\mathcal{M}^{-i}$ respectively from datasets $f(D)$ and $f(D^{-i})$, which are provided as training datasets to $\mathcal{A}$.
4. Use $\mathcal{M}$ and $\mathcal{M}^{-i}$ to predict the sensitive attribute S of the same individual i. This will result in two probability distributions $P^i = \mathcal{M}(q^i) = (p_1^i, \ldots, p_n^i)$ and

[4] The learning algorithm $\mathcal{A}$ is often a randomized algorithm (e.g., stochastic gradient descent used to compute the parameters of neural network), thus $\mathcal{M}_{f(D)}$ and $\mathcal{M}_{f(D^{-i})}$, which are built using $\mathcal{A}$, can be described by a random variable. Also, the anonymization algorithm f can be randomized (e.g., a differentially private sanitization algorithm), and thus it can be described by another random variable. The probability in Definition 1 is taken on the randomness of these variables.

$P^{-i} = \mathcal{M}^{-i}(q^i) = (p_1^{-i}, \ldots, p_n^{-i})$ over the domain of sensitive attribute values.

5. Compute the differentiability of $\mathcal{M}$ and $\mathcal{M}^{-i}$ based on Definition 1. To do so, the tail of the distribution of $\texttt{distance}(P^i, P^{-i})$ should be bounded by approximating the distribution of $\texttt{distance}(P^i, P^{-i})$ by sampling, i.e., executing Step 3 and 4 several times.

6. Repeat Steps 2-5 for all $(q^i, s^i) \in D$.

7. Evaluate the results using the set of all computed distances over all individuals. If $\texttt{distance}(P^i, P^{-i}) \geq \epsilon$ with probability at most δ for all $(q^i, s^i) \in D$, then $\mathcal{A}$ is an (ϵ, δ)-differentiator of S in $f(D)$.

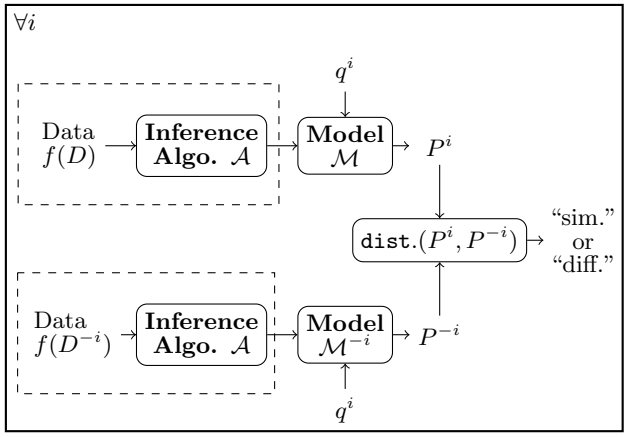


Figure 1: Differential inference testing.

As we do not test randomized anonymization techniques (i.e., f), and we use a deterministic learning algorithm $\mathcal{A}$ in this paper, in Step 5, we report only a single sample of $\texttt{distance}(P, P^{-i})$ *per* individual and ignore δ in the sequel. This single sample is used to verify whether the inference about attribute s^i leaks some private information about i (i.e., private inference) or not (i.e., public inference) using a certain threshold ϵ. If $\texttt{distance}(P^i, P^{-i}) < \epsilon$, then the inference about s^i does not strongly depend on the record of the individual i. That is, we would get almost the same prediction distribution even if i has not contributed to the original dataset. In this case, we say that P^i and P^{-i} are "similar". On the other hand, if $\texttt{distance}(P^i, P^{-i}) \geq \epsilon$, then the inference strongly depends on i's record, and we say that P^i and P^{-i} are "different". The choice of `distance` (and threshold ϵ) depends on the dataset and ultimately on the semantics of the sensitive attribute S. Examples of `distance` include the Earth Mover's Distance [18], the max divergence used by differential privacy [5], or any divergences between probability distributions.

We also note that the above testing approach implicitly assumes uniform prior on the prediction distributions. That is, we do not use any auxiliary information about any individuals in the above inference process since modeling a realistic prior is usually cumbersome in practice. Nevertheless, if provided, it should be straightfoward to incorporate such prior knowledge into the above inferences.

Finally, we stress that f is not needed by our testing procedure, but only $f(D)$ and $f(D^{-i})$ for each $(q^i, s^i) \in D$. This can be advantageous if f is confidential.

3. EVALUATION

In this section, we demonstrate our approach on two real datasets: UCI Adult (Census Income) data[5] with 48842 records and the "General Demographics" dataset from Internet Usage data[6] with 10108 records. For the Adult dataset, the attributes "age", "education", "marital status", "hours per week" and "native country" are used as quasi-identifiers, and "occupation" is used as sensitive attribute with 14 possible values. While for Internet dataset, quasi-identifier attributes are "age", "country", "education attainment", "major occupation", "marital status" and "race", and sensitive attribute is "household income" which takes 9 possible values. We note that, for Adult dataset, only 10,000 records (with no missing attributes) were randomly selected; and that, for Internet dataset, all the records where "age" takes the value "not-say" were removed (only 9799 records remains).

For anonymization techniques, we consider the basic Mondrian k-anonymity [8] and Mondrian ℓ-diversity [10]. Both approaches modify 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)[7].

3.1 Feature Encoding

Anonymization techniques such as k-anonymity and ℓ-diversity aim at preserving individuals' privacy by generalizing the original data according to a specific hierarchy. Continuous (numerical) attribute values are usually generalized into intervals (ranges), e.g., age 25 may be generalized into [20,30]. While categorical attribute values are usually generalized into sets or generalizations, e.g., country *France* may be generalized into {*France*, *Germany*} or *Europe*.

In order to use a generalized dataset in data mining, e.g., to train a machine learning model, it has to be encoded first. In our study, we use a novel encoding scheme of quasi-identifiers in anonymized datasets $f(D)$ and $f(D^{-i})$, called as TF-TF encoding. The TF-TF encoding depends on the current record of the individual i, and works (for both continuous and categorical attributes) as follows: an interval or set (*resp.* generalization) is represented by 1 (or *True*), if the corresponding value in the target record is inside the interval or the set (*resp.* child of the generalization). Otherwise, it is represented by 0 (or *False*).

Example. Consider a dataset that contains the two records: $r_1 =$ ([15,25], *Female*, {*France*, *Germany*}) and $r_2 =$ ([17,20], *Male*, {*Italy*, *Germany*}). Assume that the target record is $r_t =$ (16, *Male*, *France*). Then, the records r_1 and r_2 would be encoded as follows: $encode(r_1, r_t) = (1, 0, 1)$ because $16 \in [15, 25]$, *Male* $\neq$ *Female*, and *France* $\in$ {*France*, *Germany*}, and $encode(r_2, r_t) = (0, 1, 0)$ because $16 \notin [17, 20]$, *Male* =

[5]https://archive.ics.uci.edu/ml/datasets/Adult
[6]goo.gl/oXXok5
[7]Using only $f(D)$ as a background knowledge for inference.

Male, and *France* $\notin$ {*Italy*, *Germany*}.

The TF-TF encoding requires fewer number of features than other known encodings. Exactly, one feature is needed to represent each attribute. It also results in smaller computation time. Moreover, when used, TF-TF encoding results in a greater prediction accuracy. In Appendix A, we present a comparison between various encoding techniques, including TF-TF encoding.

3.2 Inference Algorithm

We use a Naive Bayes classifier for the purpose of inference. An implementation of Naive Bayes is available as BernoulliNB (BNB) from sklearn python module[8] . BNB is suitable when the training data is composed of binary features, which is the case for our encoding scheme presented above. Although Naive Bayes makes a simplistic assumption that the quasi-identifiers are independent, it usually performs remarkably well, especially when the size of the training dataset is small. Several prior works used Naive Bayes classifier to perform inference on anonymized data [2, 1, 12].

3.3 Distance Measure

We measure the distance between P^i and P^{-i} (see Section 2.2) by the Earth Mover's Distance (EMD). EMD represents the "amount of energy" (or cost) needed to transform one distribution to another, and is a metric for probability distributions. Formally, $\text{EMD}(P^i, P^{-i}) = \min_{\{g_{uv}\}} \sum_{u,v} g_{uv} d_{uv}$ such that $g_{uv} \geq 0$, $\sum_v g_{uv} \leq p^i_u$, $\sum_u g_{uv} \leq p^i_v$, where $\{g_{uv}\}$ denotes the set of all possible flows (each g_{uv} represents the amount of probability mass transported from s_u to s_v), and d_{uv} is the ground distance. The ground distance can be chosen according to the semantics of the sensitive attribute S. For example, if S has generalized values according to a generalization hierarchy then a suitable ground metric can be the path length between two nodes (two different values of S) in the corresponding generalization hierarchy. This is illustrated by Figure 2, where S takes a value from {*Engineer*, *Lawyer*, *Professional*, *Dancer*, *Writer*, *Artist*, *ANY*}. For example, the ground distance between *Engineer* and *Artist* is 3, and between *Dancer* and *Writer* is 2. Another example for ground distance is the geographical distance if S represents geographical locations (e.g., a visited location).

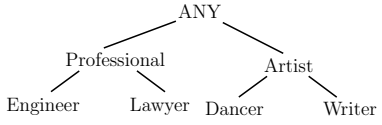


Figure 2: Example for a generalization hierarchy.

As the sensitive attribute S is never generalized in our experiments, we choose a simple ground distance for all sensitive attributes in this paper: we consider equal ground distance between any two different sensitive values, i.e., $d_{uv} = 1$ if $u \neq v$ and 0 otherwise. In that case,

$$\begin{aligned}\texttt{distance}(P^i, P^{-i}) &= \text{EMD}(P^i, P^{-i}) \\ &= \frac{1}{4} \sum_{j=1}^{n} \left| p^i_j - p^{-i}_j \right|\end{aligned}$$

which follows from [7].

This provides an intuitive interpretation of threshold ϵ in Definition 1. If $\mathcal{A}$ allows to infer a specific sensitive value with certainty c using $f(D)$, then c can change with at most 4ϵ if a single individual is removed from the anonymized dataset[9]. For example, if the profession of an individual is *Lawyer* with probability 0.5, then this probability becomes 0.5 ± 0.02 if $\epsilon = 0.02$ after the removal of the individual's record from the training data $f(D)$.

3.4 Results

In what follows, we first report the number of records for which $\mathcal{M}$ and $\mathcal{M}^{-i}$ have different predictions (Section 3.4.1), and argue that this provides an inadequate evaluation of anonymization. Then, we measure the differentiability of $f(D)$ depending on the anonymization parameters k and ℓ (Section 3.4.2).

3.4.1 Comparing the actual predictions

We compare the predictions of the two models $\mathcal{M}$ and $\mathcal{M}^{-i}$. The predictions $\mathcal{M}$ and $\mathcal{M}^{-i}$ are defined as the mode of the prediction distributions P^i and P^{-i}, respectively, i.e., the sensitive value with the largest prediction probability[10].

Figures 3 and 4 depict for Adult and Internet datasets, respectively, the percentage of records for which $\mathcal{M}$ and $\mathcal{M}^{-i}$ provide different predictions depending on the anonymization parameters k and ℓ. For both k-anonymity and ℓ-diversity, in general, we have that the percentage of different predictions is greater for small privacy parameter values than for larger parameter values. Note that, in case of ℓ-diversity, ℓ changes between 1 and 7. The Adult and Internet datasets become fully generalized when $\ell > 7$ and $\ell > 5$, respectively. Hence, we evaluate ℓ-diversity only up to $\ell = 7$.

Although prior works often used the number of different predictions as an empirical measure of anonymization [2], it is not adequate if the adversary can observe the complete prediction distribution produced by $\mathcal{A}$. Indeed, the private information leaked by $f(D)$ can be substantial even if the mode of the prediction distributions are identical. For illustration, consider the case when $P^i = (0.49, 0.01, 0.5)$ and $P^{-i} = (0.01, 0.49, 0.5)$. Although both P^i and P^{-i} have the same predictions, $\text{EMD}(P^i, P^{-i}) = 0.24$. This means that, by observing $f(D)$, an adversary can learn a lot about what sensitive value individual i is (un)likely to have just because

[8]http://scikit-learn.org/stable/index.html

[9]In theory, this does not provide an upper bound on the leakage of private information. Indeed, Pinsker's inequality says that the relative entropy of P^i with respect to P^{-i} (i.e., their Kullback-Leibler divergence) is at least $8\epsilon^2$ if $\text{EMD}(P^i, P^{-i}) = \epsilon$, but the maximum of the relative entropy can be unbounded.

[10]This corresponds to the maximum likelihood decision.

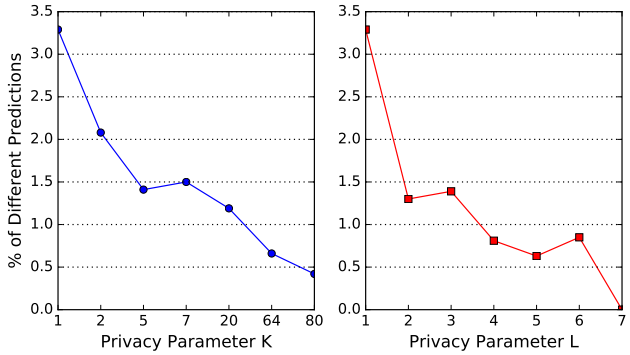


Figure 3: Percentage of different predictions for Adult data

i's record is included in D. Therefore, we study the differentiability of $f(D)$ next, which measure potential private data leakage more faithfully.

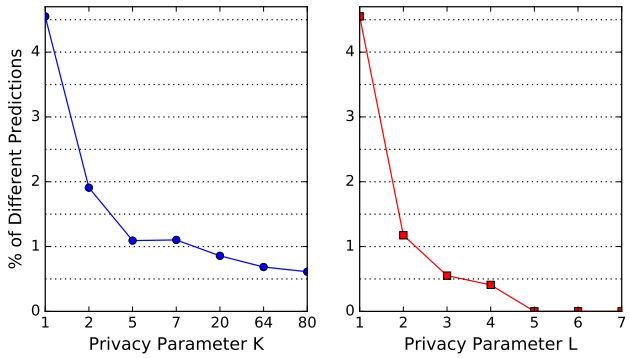


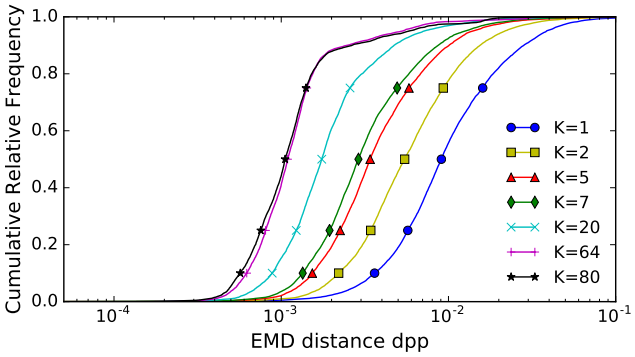
Figure 4: Percentage of different predictions for Internet data

3.4.2 *Comparing the prediction distributions*

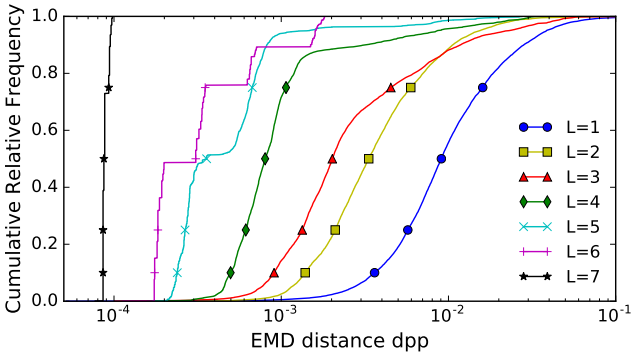
In this section, we evaluate the anonymized datasets by their differentiability with respect to BNB. The computation of differentiability is described in Section 2.2. Recall that the EMD distance between P^i and P^{-i} is computed for each individual i that has a record in D, and only this single value, denoted as d_{pp}, is reported per individual.

Figures 5 and 6 depict the cumulative relative frequency of d_{pp} with respect to privacy parameters k and ℓ. These figures show that:

- d_{pp} decreases when the privacy parameters (k or ℓ) increase. This means that, as expected, there are fewer private inferences and thus more robustness when the privacy guarantee is stronger.
- d_{pp} is smaller for ℓ-diversity than for k-anonymity when k and ℓ have identical values. This is also 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, where privacy breach occurs if the sensitive attribute of an individual can be inferred with too large accuracy. By contrast, we focus on the *relative* accuracy of inferences, i.e., the difference in inference accuracy caused by the inclusion of an individual's record.



(a) k-anonymity



(b) ℓ-diversity

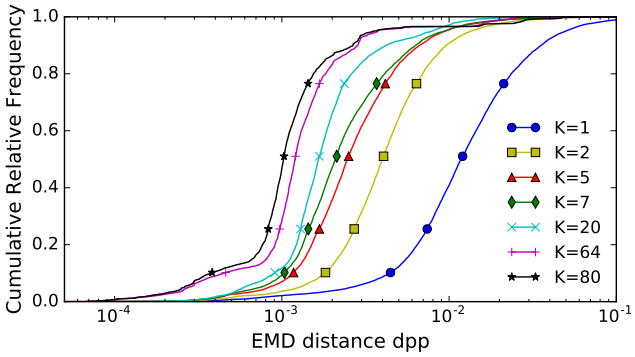
Figure 5: CDF for d_{pp} on Adult data

For example, 54% of all individuals in the Adult dataset have $d_{pp} > 10^{-2.5} \approx 0.003$ for 5-anonymity (red curve), and 3% for 5-diversity (cyan curve, see Fig. 5). The percentage of individuals for the same threshold of $10^{-2.5}$ are 35% for 5-anonymity, and 0% for 5-diversity in case of the Internet dataset (see Fig. 6). This means that only the 5-diversified Internet dataset satisfies $(10^{-2.5}, 0)$-differentiability with respect to BNB (see Definition 1)[11]. Indeed, in this case, there are no records for which $d_{pp} > 10^{-2.5} \approx 0.003$. For the other cases, in order to have a stronger robustness guarantee in our differentiability model[12], the datasets need to be anonymized with larger values of k or ℓ.

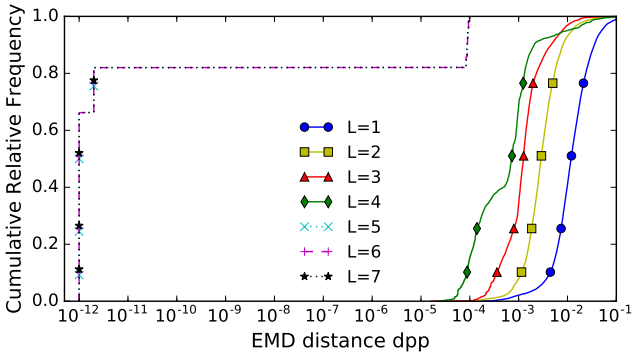
Finally, we compute the differentiability of the anonymized datasets depending on the privacy parameters k and ℓ. Ta-

[11]Notice that the 5-diversified Adult dataset does not satisfy $(10^{-2.5}, 0.03)$-differentiability either, because the probability δ is taken on the randomness of the BNB (see Definition 1) and not on dataset D.

[12]Removing records whose differentiability is larger than ϵ may not be sufficient, as anonymization has to be re-applied on the filtered data. There is no guarantee that the newly anonymized data has smaller differentiability.



(a) k-anonymity



(b) ℓ-diversity

Figure 6: CDF for d_{pp} on Internet data

k	1	2	5	7	20	64	80
Max d_{pp}	0.32	0.197	0.137	0.126	0.178	0.234	0.235

Table 1: Max. d_{pp} (i.e., ϵ) for Adult with k-anonymity

ℓ	1	2	3	4	5	6	7
Max d_{pp}	0.32	0.21	0.213	0.04	0.028	0.0018	9.8e-05

Table 2: Max. d_{pp} (i.e., ϵ) for Adult with ℓ-diversity

ble 1, 2 show the maximum value of distance d_{pp} over all individuals for the Adult dataset, and Table 3 and 4 for the Internet dataset. The maximum value of d_{pp} corresponds to ϵ in our model to achieve $(\epsilon, 0)$-differentiability (see Definition 1). In general, stronger anonymization (in terms of ℓ) entails smaller value of ϵ, i.e., the differentiability decreases by increasing ℓ. However, this is not the case for k-anonymity, where larger k implies larger differentiability. On the other hand, the average value of d_{pp} over all individuals decreases by increasing k, as expected, which is shown in Fig. 7. This also demonstrates that worst-case measures, such as differentiability, do not always follow the same trend as average-case measures.

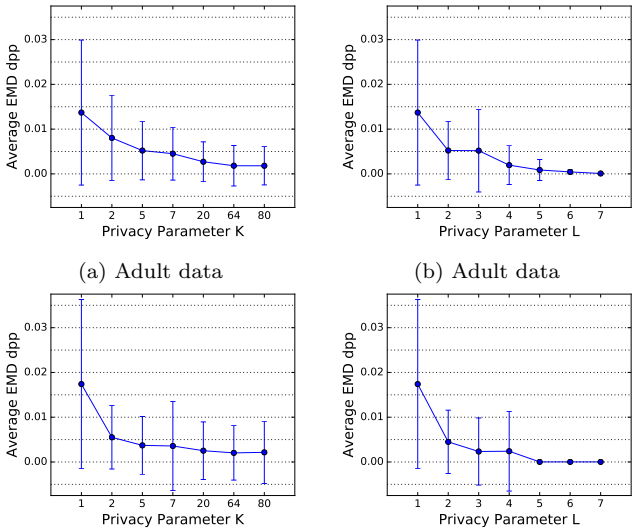
Table 1, 2, 3 and 4 also show that, for identical values of k and ℓ (if $k, \ell > 2$), ℓ-diversity implies strictly smaller values of ϵ than k-anonymity, i.e., it provides more robust anonymized datasets. In particular, ϵ remains in the order of 10^{-1} for k-anonymity even for large values of k, but it decreases to $\approx 10^{-4}$ for ℓ-diversity when $\ell > 7$ (Adult dataset), and $\ell > 5$ (Internet dataset). The more robust anonymization guarantee of ℓ-diversity is also confirmed by the average d_{pp} computed over all individuals, which is shown in Fig. 7 for both datasets.

k	1	2	5	7	20	64	80
Max d_{pp}	0.2	0.257	0.305	0.398	0.512	0.459	0.424

Table 3: Max. d_{pp} (i.e., ϵ) for Internet with k-anonymity

ℓ	1	2	3	4	5	6	7
Max d_{pp}	0.2	0.15	0.43	0.17	9.6e-05	9.6e-05	9.6e-05

Table 4: Max. d_{pp} (i.e., ϵ) for Internet with ℓ-diversity



(a) Adult data (b) Adult data

(c) Internet data (d) Internet data

Figure 7: Average d_{pp} with standard deviation

4. CONCLUSION

This paper presents an inference-based scheme aiming at evaluating the robustness of a given anonymized dataset. We illustrate how this scheme operates by analyzing two real datasets.

Our proposed scheme allows to compare different anonymized datasets that might use different anonymization models, such as k-anonymity, ℓ-diversity or Differential Privacy. It can be used by companies or DPAs (Data Protection Authorities) to test the robustness of anonymized datasets. It is important to note that our solution tests the robustness of anonymized datasets, not of the underlying anonymization.

Similarly to pen-testing tools that test the security of a system by performing a set of security attacks, we believe that there is a need for a toolkit to test the robustness of anonymized 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 anonymized dataset is analyzed as a "black box", i.e. the anonymization 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 anonymized version. We believe this is a desirable property for at least two reasons: (1) many companies are unwilling, for different reasons, to publish their anonymization algorithms. (2) The verifier does not need to go through the difficulty of understanding and analysing 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, and actually should, be used. Evaluating our scheme with other classifiers is part of our future work.

5. REFERENCES

[1] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA*, 2011.

[2] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Empirical privacy and empirical utility of anonymized data. In *Workshops Proceedings of the 29th IEEE International Conference on Data Engineering, ICDE, Brisbane, Australia*, 2013.

[3] S. J. De and D. L. Métayer. *Privacy Risk Analysis*. Synthesis Lectures on Information Security, Privacy, & Trust. Morgan & Claypool Publishers, 2016.

[4] C. Dwork. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP, Venice, Italy*, 2006.

[5] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC, NY, USA*, 2006.

[6] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA*, 2014.

[7] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[8] Q. Gong. Implementation of basic mondrian k-anonymity. goo.gl/kKpqZ6. Accessed: 16-09-2017.

[9] Q. Gong. Implementation of mondrian k-anonymity. goo.gl/eFYXBv. Accessed: 16-09-2017.

[10] Q. Gong. Implementation of mondrian l-diversity. goo.gl/2ymJ14. Accessed: 16-09-2017.

[11] A. Inan, M. Kantarcioglu, and E. Bertino. Using anonymized data for classification. In *Proceedings of the 25th International Conference on Data Engineering, ICDE, Shanghai, China*, 2009.

[12] D. Kifer. Attacks on privacy and definetti's theorem. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD, Providence, Rhode Island, USA*, 2009.

[13] J. Klucar and F. McSherry. Lunchtime for data privacy. goo.gl/Lfq8iw, 2016. Accessed: 16-09-2017.

[14] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE, Atlanta, GA, USA*, 2006.

[15] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the 23rd International Conference on Data Engineering, ICDE, Istanbul, Turkey*, 2007.

[16] A. Machanavajjhala and D. Kifer. Designing statistical privacy for your data. *Commun. ACM*, 58(3), 2015.

[17] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. *L*-diversity: Privacy beyond *k*-anonymity. *TKDD*, 1(1):3, 2007.

[18] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 2000.

[19] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. technical report. Technical report, SRI Computer Science Laboratory, 1998.

APPENDIX

A. ENCODING TECHNIQUES

In this appendix, we illustrate about the usefulness of TF encoding by making a comparison between various encodings with respect to accuracy, computational time and number of features required.

The experiment was performed on the Adult and Internet Usage datasets, which are presented in Section 3. The experiment runs as follows: (i) anonymize the dataset. Then, for every record: (ii) encode the anonymized data; (iii) use the encoded data to train a machine learning model. (iv) use the model to make a prediction on the current record. For anonymization, we used the python implementation of Mondrian k-anonymity [14] by Qiyuan Gong [9]; and for the privacy parameter K, we considered the values: 1, 3, 5, 8, 16, 32, 64, 128 where K=1 corresponds for no anonymization. For learning algorithm, we used Naive Bayes BernoulliNB (BNB).

In appendix A.1, we briefly present several possible encoding techniques, which are usually used to encode generalized data (cf. [11]). Then in appendix A.2, we illustrate the usefulness of the TF-TF encoding for our framework with respect to prediction accuracy and efficiency by comparing it to the other encodings.

A.1 Attribute Encodings

For continuous attributes, one can use one of the following

encodings:

1. Scaling: encode each interval by its midpoint, then rescale it into [0,1] or [-1,1]. One can use the following formula to rescale a value x into $x_0 \in [0,1]$: $x' = \frac{x-min}{max-min}$ where min and max are respectively the minimum and the maximum values. For example, let 15 and 90 are respectively the min and the max values of the attribute "age". Then, the value 30 is rescaled into 0.2, and the interval [35,55] is replaced by 45 then rescaled into 0.4.
2. Bounds: represent each interval by its bounds after rescaling them. Considering the previous example, the interval [35, 55] is represented by the two features vector (0.267 0.534), and the value 30 is represented by the vector (0.2 0.2).
3. Binary: represent attribute values as binary vectors. For a value x, apply *floor* function (returns the greatest integer less than or equal to x), then convert *floor*(x) into binary representation. An interval is represented by its midpoint. Note that, one can shift attribute values into [0, *max-min*] to minimize the number of bits required to represent all attribute values. The "age" attributes requires at least 7 bits (even after shifting from [15,90] into [0,75]). To encode the value 30, we first shift it into 30-15=15, then obtain the binary representation of 15 over 7 bits, that is (0 0 0 1 1 0 1). The interval [35, 55] is replaced by 45, shifted to 28, and then represented by (0 0 1 1 1 0 0). Note also that, one may use the *ceiling* function (*ceiling*(x) returns the least integer greater than or equal to x) instead of the *floor* function.
4. QuantilesB (before anonymization): Discretize attribute values into quantiles before anonymization, and then deal with it as categorical attribute.
5. QuantilesA (after anonymization): decompose attribute domain into quantiles. Then, replace each interval (or value) by a binary vector having a length equal to the quantiles number. The vector is composed in a way such that each entry that corresponds to a quantile which intersects with the considered interval is set to 1. For example, the "age" attribute with range [15,90], is decomposed into five quantiles: [15,29], [30,44], [45,59], [60,74], [75,90]. Then, the interval [35, 55] is represented by the vector (0 1 1 0 0).

For categorical attributes, one can use one of the following encodings:

1. Hot encoding: let n be the number of possible categories before anonymization. Then, each set of categories (generalization) is represented by a binary vector of length n where all and only the entries that correspond to its elements (children) are set to 1. For example, the attribute "gender", with only two categories *Male* and *Female*, is represented by a binary vector of length 2 where *Male* and *Female* can be replaced with (1 0) and (0 1) respectively. For a generalization (e.g., a combination of categories) that appears after anonymization, all the entries corresponds to its children are set to 1. For instance, in our example, the combination {*Male*, *Female*} is replaced by (1 1).
2. Binary: deal with each generalization that appears after anonymization as a new category (and forget about initial categories). Then, map the obtained categories into ordinal numbers, and finally replace the ordinal numbers by their binary representation. For the "gender" attribute, originally there are two categories (*Male* and *Female*). Assuming that these two categories still exist after anonymization and additionally the generalization {*Male*, *Female*} appears, then in total we will have three categories. These three categories can be mapped into 1, 2, and 3, then represented as (0 1), (1 0) and (1 1) respectively.
3. HotA (after anonymization): similar to the Binary case, deal with each generalization as a new category. Then, apply the Hot encoding to the new obtained categories, i.e., replace each category by a binary vector where only the corresponding entry is set to 1. For the "gender" attribute, assuming that we will have the three categories: *Male*, *Female*, and {*Male*, *Female*}, then we can represent them respectively by the vectors (1 0 0), (0 1 0), and (0 0 1).

A.2 Results and Discussion

We make a comparison between all the possible combinations of encodings for continuous and categorical attributes. This results in 24 different encodings. We refer to these encoding combinations by e_{ij}, where $i \in \{1,2,3,4,5,6\}$ and $j \in \{1,2,3,4\}$ respectively refer to continuous and categorical attributes encoding. Note that, we refer to TF encoding by 6 for continuous attributes, and by 4 for categorical attributes.

We note that, this experiment only considers the full anonymized dataset $f(D)$ aiming at evaluating the effect of various encodings on the prediction accuracy and efficiency. In the following, we present the results that have been obtained. All the experiments were conducted using an Intel machine (no TSX) 1VCPU 2.3 GHz, 2 GB RAM.

Accuracy (η). Figure 8 presents the average accuracies, obtained for different K values on Adult and Internet Usage dataset, with respect to various encodings. We can see that, for both datasets, the maximum average accuracy is achieved when TF-TF encoding (e_{64}) is used: 22.80% for Adult dataset, and 25.30% for Internet Usage dataset. Note that, for Adult dataset, same value is also achieved in case of QuantilesB-TF (e_{44}) where TF-TF encoding is also used (at the end), but continuous attributes are converted into categorical ones before anonymization. For Internet Usage dataset, the case of QuantilesB-TF comes second in descending order with an accuracy 24.17%.

Number of Features. TF encoding requires a small number of features (which is usually desirable): one feature per attribute. So, TF encoding is a best choice to minimize the number of features, especially when it comes for categorical attributes. In our case, the minimum number of features are respectively 5 and 6 for Adult and Internet Usage datasets, which are the numbers required when Scaling-TF, QuantilesB-TF, and TF-TF (e_{14}, e_{44}, and e_{64} respectively) encodings are used. The other encoding combinations require a number of features which range from 7

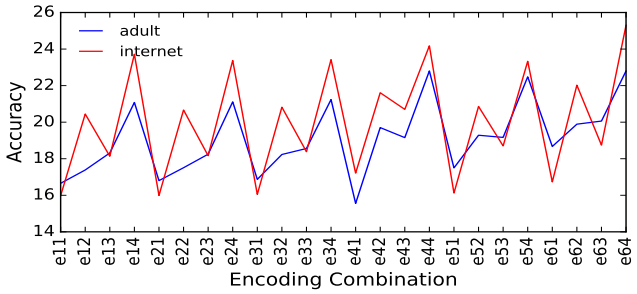


Figure 8: Average accuracy.

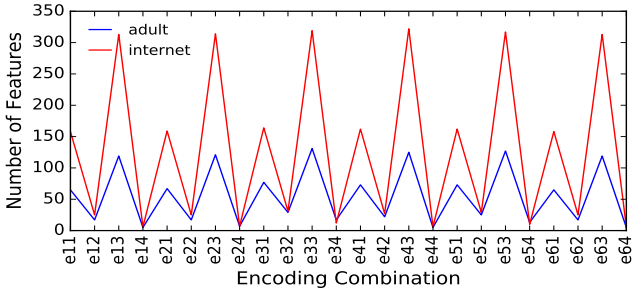


Figure 9: Average number of features.

(for Bounds-TF e_{24}) to 131 (for Binary-HotA e_{33}) in case of Adult dataset, and from 7 (for Bounds-TF e_{24}) to 322 (for Binary-HotA e_3) in case of Internet Usage dataset.

Time. The minimum time, for both datasets, is achieved

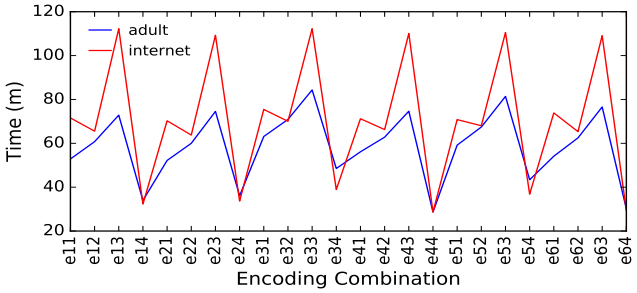


Figure 10: Average computation time.

in case of QuantilesB-TF (e_{44}) encoding: 28.67 minutes for Adult dataset and 28.64 minutes for Internet Usage dataset. The TF-TF encoding comes after with 30.17 minutes for Adult dataset and 29.37 for Internet Usage dataset. This means that converting continuous attributes into categorical ones before anonymization makes the process a bit faster. We note that, the measured time for Adult (*resp.* Internet) dataset is the time required to train 10K (*resp.* 9799) models and to make a prediction by each. The obtained results show that TF encoding improve the computation time when it comes to train a single model. However, a disadvantage of TF encoding is that to make multi-predictions a new model have to be trained for each input record. This is not the case for other encodings, at least when the training algorithm is deterministic (which is in our case). But, since in our framework a new model $\mathcal{M}^{-i}$ is required for each record regardless of the encoding used. Then in worst case, the number of models to be trained when TF encoding is used is about two times that number when TF encoding is not used. Counting that, the expected time that will be required to run our framework for:

- Adult dataset is: 57.34 minutes (28.67×2) in case of QuantilesB-TF encoding and 60.34 minutes (30.17×2) in case of TF-TF encoding. These time values are still smaller or not much greater than the time required by any encoding combination that does not consider TF encoding (neither for continuous nor for categorical attributes) which ranges from 52.19 (for Bounds-Hot e_{21}) to 84.34 minutes (for Binary-HotA e_{33}).
- Internet dataset is: 57,28 minutes (28.64×2) in case of QuantilesB-TF encoding and 58,74 minutes (29.37×2) in case of TF-TF encoding. These time values are smaller than the time required by any encoding combination that does not consider TF encoding. The time required by the these encodings ranges from 65.53 (for Scaling-Binary e_{12}) to 112.26 minutes (for Binary-HotA e_{33}).

To sum up, the obtained results show that TF-TF (or QuantilesB-TF) encoding provides a greater accuracy and requires a smaller number of features than the other encodings. Moreover, using TF-TF (or QuantilesB-TF) encoding results in a model training computation time that is about half the time required when most of the other encodings (that do not consider TF encoding) are used. Therefore, the TF-TF (or QuantilesB-TF) is suitable encoding to consider in order to run our framework.