TRANSACTIONS ON DATA PRIVACY 16 (2023) 153–189

Optimizing Privacy and Data Utility: Metrics and Strategies

Clémence Mauger, Gaël Le Mahec, Gilles Dequen

Université de Picardie Jules Verne - MIS Laboratory, 33 rue Saint-Leu, Amiens, 80000, France

E-mail: {clemence.mauger,gael.le.mahec,gilles.dequen}@u-picardie.fr

Received 5 April 2022; received in revised form 15 February 2023; accepted 25 February 2023

Abstract. k-anonymity is a PPDP anonymization model preventing identity disclosure by making each record of the table indistinguishable from $k-1$ others. To obtain a k-anonymous version of a table, a common technique is to generalize the quasi-identifier attributes values until records are grouped in equivalence classes of size at least k. The choice of records to be grouped will influence the amount of generalization to be performed and therefore the quality of the anonymized data (the more a value is generalized, the more precision it loses). The different k-anonymous versions of a table are therefore more or less interesting in terms of data utility. To assess the quality of a k-anonymized table, information loss metrics are often used. They can also be used within the k-anonymization process itself to choose the groupings of records resulting in the least data alteration. In this article, we propose a unified modeling of such metrics, faciliting their implementation and their use. We then analyze the behaviors of seven metrics when they are used in the k-anonymization process to guide the equivalence classes mergings. Our analyzes compare these seven metrics on two public tables for 14 values of k. After that, we turned to the limits of k-anonymity. In a k-anonymous table, the distribution of sensitive values in equivalence classes can lead to the disclosure of sensitive information about an individual. l-diversity and t-closeness anonymization models impose constraints that keep control over the distribution of sensitive values and therefore limit attribute disclosure. We continue our study on k-anonymization by proposing strategies aimed at optimizing the data alteration, the l-diversity and the t-closeness of the k-anonymous tables produced. Using two information loss metrics, we evaluate the seven optimization strategies on the two public tables first on real sensitive values distributions and then on 21 simulated sensitive values distributions. With this large study, we would like to understand how to choose a metric and an optimization strategy to provide k-anonymous database with strong guarantees on the data privacy and preserving as much as possible the data utility.

1 Introduction

The volume of collected data on Internet continues to grow. Among all the data collected on each of us, some have very great potential in many areas. For example, health data can be crucial information for the study of diseases. Consumers' data are of interest for marketing. Data on the behaviors and habits of citizens can help decision-making on public policies. However, publishing such data is also a significant risk for the people it concerns. To take these risks into account, the legislation of most countries has evolved towards ever greater protection. In order to continue to exploit data, it is therefore necessary to give guarantees

of privacy to individuals. Before publishing data, the identity of individuals and their sensitive data must be masked. This process is called anonymization [7].

Generally, the data sets considered are tables. Lines, or records, of these tables represent individuals and columns represent attributes. We distinguish three categories of attributes. The identifier attribute is a direct and unique link between an individual and a record of the table (e.g. social security number, name). In Table 1a, T is a table of height lines and four attributes. It has four equivalence classes that contain two records.

The quasi-identifier attributes could induce information disclosure if they are considered totally or partially and/or crossed with external data source (e.g. gender, location, age). In a table, we group records according to their quasi-identifying values: an *equivalence class* of the table contains all the records with exactly the same quasi-identifiyng values. Finally, the sensitive attributes contain the most useful information to protect and are the reason for publishing the table (e.g. disease, salary). The data editor should be very careful about qualifying an attribute as sensitive, especially when multiple sensitive attributes are present. If this attribute turns out to be quasi-identifying in the context of the publication, the protection offered by k-anonymity can be compromised. The problematic contexts can be the presence of several sensitive attributes or even the publication of the same attribute in two different databases, even if, considered independently, they are k-anonymous ([2]). In this article, we do not consider such a situation, focusing on optimizing the usefulness of a database with well-defined attributes.

The Privacy-Preserving Data Publishing [12, 11, 13], abbreviated as PPDP, is a research field whose objective is to ensure that the publication of the data does not permit to associate an individual with a record in the database or to learn more about the sensitive data of the individual. Among the anonymization models proposed in PPDP, k-anonymity was presented in [28] to struggle the identity disclosure [15] (ie the ability to associate an unique record of the table with an individual). In a k-anonymous table, each record has to be indistinguishable from at least $k-1$ other records with respect to the set of quasi-identifier attributes. In other words, each equivalence class of the table contains at least k records. By k-anonymizing a table, we guarantee that even if an adversary knows in which class of the k-anonymous table an individual is, he will not be able to associate a record with him except with a probability of $\frac{1}{k}$. k-anonymity is so a protection against identity disclosure. In Table 1b, T_{ano} is a 4-anonymous version of T. We first replace the identifier values by pseudonymes to achieve *pseudonymization*. Then, we construct two equivalence classes containing four records.

To achieve k-anonymity, techniques as generalization [28], bucketisation [33], suppression [29] or micro-agregation [8] have been proposed. In this article, we focus on the generalization technique for categorical data. In this approach, it is considered that the values are replaced by less specific (more general) values, in order to group them into a set of at least k entries. It implies that we dispose of generalization hierarchies [28] that lead how the values can be replaced to be grouped maintaining the semantic of the records. Such generalizations must be defined prior to apply the different algorithms we propose. These hierarchies, for those that are categorical, have been chosen considering the finest grain allowing semantic preservation. For each quasi-identifier attribute of the table, we so associate a generalization hierarchy that is a tree that represents the possible generalizations of each original value of the attribute. The original values in the table are the leaves of the hierarchy at level 0 and correspond to values without generalization. The upper levels increasingly generalize the values until the complete loss of information represented by the root of the hierarchy.

There are several ways to use the generalization technique to anonymize a table. When

performing a global recoding, the identical records of the original table are found in the same equivalence class of the anonymized table. On the other hand, in the case of local recoding, each record is considered independently and identical records can be generalized differently. Moreover, we talk about single dimensional recoding when we allow only one level of generalization for all the values of a quasi-identifier attribute. On the contrary, in a multidimensional recoding, several levels of generalization can be applied to the values of a quasi-identifier attribute. Incognito [16] is an example of framework using a single dimensional global recoding to k-anonymize a table. Mondrian [17] is based on a multidimensional global recoding and improves Incognito performance [1]. Examples of frameworks using a multidimensional local recoding are KACA [18], k-member [5], OKA [21] or GCCG [26]. Although multidimensional local recoding have shown good performance in data utility conservation, we will use multidimensional global recoding in our work.

Using generalization technique, construct a k-anonymous version of a table consists in partitioning the table into equivalence classes of size at least k. However, the number of such partitions is the k-associated Stirling number of the second kind [6] which is greater than 2^n, with n the number of records. Among these k-anonymous tables, some have undergone more generalizations of their values than others and are thus less interesting in terms of data utility. In order to classify tables according to the data utility, *information loss metrics* are frequently used. These metrics estimate the amount of information that has been lost between the original table and the k-anonymized table with generalization technique. Many metrics have been proposed like [14, 3, 18, 5]. The first contribution of this article will be to propose a model unifying the writing of metrics and simplifying their use. We will then compare the performance of information loss metrics when used in a k-anonymization algorithm to construct equivalence classes of size greater than k (cf. Section 3). The model and preliminary results limited to the adult data set was published in [23].

Although k-anonymity provides privacy guarantees, it does not totally protect against all attacks. As showned in [28] or in [30], k-anonymity is sometimes not effective against attribute disclosure. Indeed, if k-anonymity protects against the re-identification of individuals in a database, it does not respond to linking attacks, which can be posed by the lack of diversity of sensitive data (as all the de-identification techniques). In [22] and [19], authors propose l-diversity and t-closeness, two anonymization models to strengthen the privacy guarantees of k-anonymity. These models take a control on sensitive values distribution in each equivalence class. In a l-diverse table, each equivalence class contains at least l sensitive values that are fairly represented. In a table that has a t-closeness, sensitive values distribution is quite the same as the sensitive values distribution in the whole table. Our second contribution will be to propose strategies allowing to build the best k-anonymous tables possible in terms of limitation of data alteration and optimization of l-diversity and t-closeness (cf. Section 4). First results limited to adult data set and one optimisation metric was published in [24].

The rest of the article is organized as follows. In Section 2, we give notations to use the generalization technique and we present the anonymization algorithm *GAA*. In Section 3, we present a new modeling unifying the information loss metrics writing and simplying their use. We compare the performances of seven metrics when they are used in a k-anonymization algorithm to guide the mergings of equivalence classes. We conduct experiments on two public tables. In Section 4, we study the l-diversity and t-closeness models. We present seven new strategies to use in the anonymization algorithm *GAA* and that permit to optimize data alteration and l-diversity or t-closeness of the k-anonymous tables produced. We experiment on real data and on simulated data which represents 45

Identifier	Quasi-identifier		Sensitive
Name	*Gender*	*Race*	*Disease*
Ana	F	Lion	Cold
Bea	F	Dog	Bronchitis
Carole	F	Lion	Cold
Daphne	F	Dog	Conjunctivitis
Eric	M	Cat	Broken paw
Fred	M	Cat	Broken paw
Gui	M	Lion	Angina
Herve	M	Lion	Bronchitis

(a) A table T

Identifier	Quasi-identifier		Sensitive
Name	*Gender*	*Race*	*Disease*
P_1	*	Lion	Cold
P_2	*	Mammal	Bronchitis
P_3	*	Lion	Cold
P_4	*	Mammal	Conjonctivitis
P_5	*	Mammal	Broken paw
P_6	*	Mammal	Broken paw
P_7	*	Lion	Angina
P_8	*	Lion	Bronchitis

(b) A 4-anonymous version of a table T

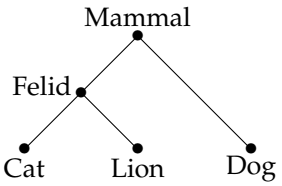
Table 1: A table and a 4-anonymous version of it

tables to study. Section 5 concludes the article and gives perspectives.

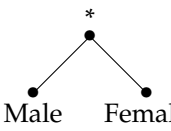
2 Generalization Technique and Anonymization Algorithm

2.1 Generalization Technique

As said in Section 1, we use generalization technique to k-anonymize tables. For each quasi-identifier attribute of the table, we construct a generalization hierarchy. Figure 1 presents an example of such generalization hierarchies for the attributes *Gender* and *Race* of the table 1a.



(a) Hierarchy of *Race*



(b) Hierarchy of *Gender*

Figure 1: Two generalization hierarchies

For the following definitions, we consider a m-tuple $\mathcal{H} = (H_1, \ldots, H_m)$ of hierarchies of $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$, with Q the set of quasi-identifier of the table. We denote by $\mathcal{R}$ the set of generalized records on $(\mathcal{A}, \mathcal{H})$.

Definition 1 (Generalization of a record and generalized table on $(T, \mathcal{H})$)**.** Let v and v' be two nodes of the generalization hierarchy H_j for $j \in [\![1, m]\!]$. v' is a generalization of v (or v can be generalized in v') if v' is on the path from v to the root of H_j.

Let $F = (f_1, \ldots, f_m, s)$ and $F' = (f'_1, \ldots, f'_m, s')$ be two records in $\mathcal{R}$. F' is a generalization of F if for all $j \in [\![1, m]\!]$, f'_j is a generalization of f_j and $s' = s$ (s is the sensitive attribute).

Let $T = \{E^1, \ldots, E^n\}$ be a table on $(\mathcal{A}, \mathcal{H})$ of cardinal $n \in \mathbb{N}^*$. A table T^{gen} on $(\mathcal{A}, \mathcal{H})$ is a *generalized table on* $(T, \mathcal{H})$ if $|T^{gen}| = n$ and for all $i \in [\![1, n]\!]$, the record F^i of T^{gen} is a generalization of E^i.

We denote by $\mathcal{T}^{gen}_{(T,H)}$ the set of generalized tables on $(T, \mathcal{H})$. For the sake of clarity, we will simply use $\mathcal{T}^{gen}$ when there is no ambiguity.

To generalize subsets of records in a table, we use the notion of *Lowest Commun Ancestor*, abbreviated as LCA, in a tree defined in [4]. The LCA of a set of nodes of a generalization hierarchy is the ancestor of the nodes that is located farthest from the root.

Definition 2 (Generalization of a subset of records)**.** Let T be a table on $(\mathcal{A}, \mathcal{H})$ of cardinal $n \in \mathbb{N}^*$. Set $\mathcal{E} = \{E^{s_1}, \ldots, E^{s_t}\}$ a subset of records to generalize in T with $t \in [\![1, n]\!]$. For each $x \in [\![1, t]\!]$, set $E^{s_x} = (e_1^{s_x}, \ldots, e_m^{s_x})$. We construct a table $gen_T(\mathcal{E}) = \{F^1, \ldots, F^n\}$ in which, for each $i \in [\![1, n]\!]$:

$$\begin{cases} F^i = (\mathrm{LCA}(e_1^{s_1}, \ldots, e_1^{s_t}), \ldots, \mathrm{LCA}(e_m^{s_1}, \ldots, e_m^{s_t})) & \text{if } E^i \in \mathcal{E} \\ F^i = E^i & \text{else} \end{cases}$$

For example, to obtain T_{ano} of Table 1b, we put generalizations (*, Lion) in records 1, 3 ,7 and 8 of T and generalizations (*, Mammal) in records 2,4, 5 and 6.

2.2 Anonymization Algorithm

In order to k-anonymize tables considering several anonymization models and measures, we use the *Greedy Anonymization Algorithm*, abbreviated as *GAA*. Its pseudo-code is in Algorithm 1. k-anonymity, l-diversity and t-closeness being NP-hard (cf. [25, 10], [34, 9], [20] respectively), *GAA* uses an heuristic method to guide production of an anonymous table. *GAA* aims to produce a version of the table which respects an anonymization model by optimizing equivalence classes mergings according to a predeterminate strategy.

Algorithm 1 Greedy Anonymization Algorithm

Require: $\mathcal{A} = \{Q_1, \ldots, Q_m, S\}$ a set of $m \in \mathbb{N}^*$ quasi-identifier attributes and one sensitive attribute, $\mathcal{H} = (H_1, \ldots, H_m)$ a m-tuple of hierarchies of the quasi-identifier attributes of $\mathcal{A}$, T a table on $(\mathcal{A}, \mathcal{H})$, Φ an anonymization model and *Strat* a merging selection strategy
Ensure: A generalized table on $(T, \mathcal{H})$ that respects the Φ model
1: **procedure** *GAA*($\mathcal{A}$, $\mathcal{H}$, T, Φ, *Strat*)
2: **while** T does not respect the Φ model **do**
3: Choose arbitrarly an equivalence class C_s of T of minimal size that does not respect the Φ model
4: Search an equivalence class C of T different from C_s that respects the conditions of *Strat* strategy
5: $T \leftarrow gen_T(C_s \cup C)$
6: **end while**
7: Return T
8: **end procedure**

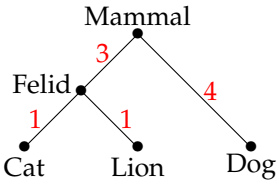
At each step, *GAA* does the merging of two equivalence classes. The first class to merge, denoted by C_s, is arbitrarly chosen among the equivalence classes that do not respect the anonymization model Φ (e.g. among the equivalence classes of size less than k for a k-anonymization). The choice of the second class to merge, denoted by C, is determined by the condition of the *Strat* strategy: it is one whose merging with C_s optimizes the conditions of the strategy.

3 Comparison of Information Loss Metrics

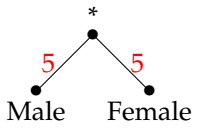
For a given table, there potentially exist several k-anonymous versions of it. Considering a table of n records, the number of potential k-anonymous versions of it, corresponding to the number of partitionings of n elements in subsets of size greater than k, is greater than $\sum_{i=1}^{k} \binom{n}{k}$ which is of the order of 2^n. More precisely, it is the k-associated Stirling number of the second kind $\left\{ {n \atop m} \right\}_{\geq k}$ representing the number of ways to partition n labeled elements in m unlabeled subsets containing at least k elements. It is therefore important to be able to classify these k-anonymous tables according to their quality in terms of data utility. The notion of information loss metrics is often used to assess the quality of anonymous tables. An information loss metric is a map from a table to $\mathbb{R}$ that estimates the amount of information that is lost during an anonymization process. Although the general principle remains the same, several information loss metrics have been proposed [35, 5, 18, 23]. All these metrics are defined to measure the loss of information by considering the generalization hierarchies of the different attributes. They assume that the "shapes" of the hierarchies (i.e. the degrees and connections of its nodes) describe the amount of information of generalized attribute values. Other techniques like micro-aggregation assume that attribute values can be aggregated to reflect more or less multiple values grouped together in the same set (e.g. using the average of multiple values, ...). For both approaches, determining the "cost" of such a grouping is not insignificant and depends on the approach itself. A comparison between these different approaches is hard and depends on the context in which the data are used. For instance, consider an age attribute for three records with 20, 25, and 30 years old. It can be aggregated over 25 years or generalized into a the set [20,30]. 25 can be considered better because it is at most 5 years from the original values, but 25 says nothing on the range of the values and can be unusable for certain uses requiring this information (age range of patients, ...). On the contrary, for some applications, the aggregate values may be more interesting than the range of values. One way to evaluate the final k-anonymous database is to use it as input to an ML algorithm or to measure the result of a data mining query. But such comparisons must define specific tasks on the data and are highly dependent on the data or queries on it. Thus, such a comparison is valid for a particular task or set of tasks but may lead to very different results when applied to another query on the data. In this paper, we focus on metric-based techniques and for each metric we use its own criteria to evaluate it. Indeed, the results we present are computed as a ratio of the worst case and calculated for and with the metric itself.

Most of the time, the notations used to define these information loss metrics are not the same from one article to another. Moreover, few justifications are given for the choice of one metric over another.

In this section, we will conduct a comparative study of information loss metrics. The objective is, firstly, to propose a unified and easily usable writing of information loss metrics in Section 3.1. A new definition and a matrix-based modeling will be presented. In Section 3.2, we will present three information loss metrics from the literature and four metrics from our work. Finally, in Section 3.3, we will try to evaluate the performance of several information loss metrics when used in a k-anonymization algorithm with experiments on two public datasets.



(a) Hierarchy of *Race*



(b) Hierarchy of *Gender*

Figure 2: Two generalization hierarchies with weighted edges

3.1 A Modeling of Information Loss Metrics

To unify the writing of the information loss metrics and to simplify their use, we propose a new modeling. First, we propose a definition of an information loss metric based on weights to be placed on the edges of the generalization hierarchies of the quasi-identifier attributes of the table. Second, for a given metric, we define a matrix for each quasi-identifier attribute in the table. This matrix is labeled by the nodes of the generalization hierarchy of the quasi-identifier in row and in column. A value of the matrix will correspond to the cost for the chosen information loss metric of generalizing the node labeling the row into the common ancestor of the nodes labeling the row and the column.

In Definitions 3 to 5, we consider a set $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ of m quasi-identifier attributes and a m-tuple $\mathcal{H} = (H_1, \ldots, H_m)$ of generalization hierarchies of $\mathcal{Q}$.

Definition 3 (Information loss metric)**.** For each $j \in [\![1, m]\!]$, we weight the edges of H_j with values in $\mathbb{R}^+$. For two nodes x and x' of H_j, if there exists an edge (x, x') in H_j, we denote by $\omega(x, x')$ the weight of (x, x'). A *metric on* $\mathcal{H}$ is a m-tuple of sets of weights on the hierarchies of $\mathcal{H}$:

$$\mu = (\mu_1, \ldots, \mu_m),$$

with $\mu_j = \{\omega(x, x') \ : \ (x, x') \text{ edge of } H_j\}$ for $j \in [\![1, m]\!]$.

Thus, to define a metric, we define a set of weights on the edges of the generalization hierarchies. For example, the weights in the edges of the hierarchies of *Gender* and *Race* in Figure 2 define an information loss metric. For this metric, the cost of generalizing Dog in Mammal is 4 (see Figure 2a) and the cost of generalizing Male or Female in * is 5 (see Figure 2b).

Before presenting the costs matrices associated with an information loss metric, we define the generalization cost for a metric of a node of a hierarchy in one of its generalizations. Recall that a node v' is a generalization of a node v in a hierarchy if v' is on the path from v to the root of the hierarchy.

Definition 4 (Generalization cost of a node v in a node v')**.** Let $\mu = (\mu_1, \ldots, \mu_m)$ be a metric on $\mathcal{H}$. For all $j \in [\![1, m]\!]$, the application $cost_{H_j} : H_j \times H_j \to \mathbb{R}$ associates with each couple of nodes of H_j:

$$\begin{array}{rccl} cost_{H_j} : & H_j \times H_j & \longrightarrow & \mathbb{R} \\ & (v, v') & \longmapsto & \begin{cases} \displaystyle\sum_{(x,x') \in \mathcal{E}(v,v')} \omega(x, x') & \text{if } v' \text{ is a generalization of } v \\ 0 & \text{else} \end{cases} \end{array}$$

with $\mathcal{E}(v, v')$ the set of edges in the path from v to v' and $\omega(x, x') \in \mu_j$ for all $(x, x') \in \mathcal{E}(v, v')$.

If v' is a generalization of v in H_j, $cost_{H_j}(v, v')$ represents the cost for μ of the path from v to v' i.e. the generalization cost of v to v' for μ. For the sake of clarity, we will simply use $cost(v, v')$ in the rest of the paper.

The costs matrix defines the cost of generalizing any pair of nodes in a hierarchy. The rows and columns of the matrix being labeled by the nodes of the hierarchy, any coefficient represents the cost of generalizing the node labeling the row into the lowest common ancestor of the node labeling the row and the node labeling the column.

Definition 5 (Costs matrix)**.** For all $j \in [\![1, m]\!]$, the *costs matrix* of H_j for μ, denoted by M_{μ,H_j}, is defined as:

- rows and columns of the matrix are labeled by the nodes of H_j
- for all couple of nodes (v, v') of H_j, $M_{\mu,H_j}(v, v') = cost(v, \mathrm{LCA}(v, v'))$

Example 1. Consider the hierarchy H_{Race} of *Race* attribute in Figure 2a and the metric defined by the following weights: $\omega(\text{Cat}, \text{Felid}) = 1, \omega(\text{Lion}, \text{Felid}) = 1, \omega(\text{Dog}, \text{Mammal}) = 4$ and $\omega(\text{Felid}, \text{Mammal}) = 3$.

The costs matrix of H_{Race} for μ is:

$$M_{\mu,H_{Race}} = \begin{array}{c} \\ \text{Cat} \\ \text{Lion} \\ \text{Dog} \\ \text{Felid} \\ \text{Mammal} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{Cat} & \text{Lion} & \text{Dog} & \text{Felid} & \text{Mammal} \end{array} \\ \begin{pmatrix} 0 & 1 & 4 & 1 & 4 \\ 1 & 0 & 4 & 1 & 4 \\ 4 & 4 & 0 & 4 & 4 \\ 0 & 0 & 3 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

For instance, the coefficient $M_{\mu,H_{Race}}(\text{Cat}, \text{Dog})$ is computed as:

$$\begin{aligned} M_{\mu,H_{Race}}(\text{Cat}, \text{Dog}) &= cost(\text{Cat}, \mathrm{LCA}(\text{Cat}, \text{Dog})) \\ &= cost(\text{Cat}, \text{Mammal}) \\ &= \omega(\text{Cat}, \text{Felid}) + \omega(\text{Felid}, \text{Mammal}) \\ &= 1 + 3 = 4 \end{aligned}$$

For Definitions 6 and 7, we consider a set $\mathcal{A} = \{Q_1, \ldots, Q_m, S\}$ of $m \in \mathbb{N}^*$ quasi-identifier attributes and one sensitive attribute, a m-tuple $\mathcal{H} = (H_1, \ldots, H_m)$ of hierarchies of $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$, a table T on $(\mathcal{A}, \mathcal{H})$ of cardinal $n \in \mathbb{N}^*$ such that $T = \{E^1, \ldots, E^n\}$ and a metric μ on $\mathcal{H}$.

Thanks to the costs matrices, we compute the generalization cost of a generalized table.

Definition 6 (Generalization cost of a generalized table)**.** Let M_{μ,H_j} be the costs matrix of H_j for all $j \in [\![1, m]\!]$.

We define the application $\overline{\mu} : \mathcal{R} \times \mathcal{R} \to \mathbb{R}^+$ which associates with two records on $\mathcal{R}$ the cost of generalizing these tuples for μ:

$$\begin{array}{rccl} \overline{\mu}: & \mathcal{R} \times \mathcal{R} & \longrightarrow & \mathbb{R}^+ \\ & (F, F') & \longmapsto & \displaystyle\sum_{j=1}^{m} M_{\mu,H_j}(f_j, f'_j) + M_{\mu,H_j}(f'_j, f_j) \end{array},$$

with $F = (f_1, \ldots, f_m, s)$ and $F' = (f'_1, \ldots, f'_m, s')$. In other words, $\overline{\mu}(F, F')$ is the cost to generalize F and F′ so they are in the same equivalence class.

We define the application $\mu_T : \mathcal{T}^{gen} \to \mathbb{R}$ which associates with each generalized table on $(T, \mathcal{H})$ its generalization cost according to T:

$$\mu_T(T^{gen}) = \sum_{i=1}^{n} \overline{\mu}(F^i, E^i),$$

with $T^{gen} = \{F^1, \ldots, F^n\}$.
$\mu_T(T^{gen})$ is then the total cost of the generalization from T to T^{gen}.

To obtain an estimation of the gap between the generalization cost for μ of a generalized table and the cost for μ of the table in which all information is lost, we define the *alteration* of a generalized table for μ. Alteration is a percentage so it is easily understandable whatever the metric chosen.

Definition 7 (Alteration)**.** Let T^* be the generalized table on $(T, \mathcal{H})$ such that for all $F^i \in T^*$ with $i \in [\![1, n]\!]$, $F^i = (r_1, \ldots, r_m, s_i)$ with s_i the sensitive value of E^i and r_j the root of H_j for all $j \in [\![1, m]\!]$.
We define the application $alt_\mu(T) : \mathcal{T}^{gen} \to \mathbb{R}$ which associates with each generalized table on $(T, \mathcal{H})$ its *alteration* for μ:

$$\begin{array}{rccc} alt_\mu(T): & \mathcal{T}^{gen} & \longrightarrow & \mathbb{R} \\ & T^{gen} & \longmapsto & \frac{\mu_T(T^{gen})}{\mu_T(T^*)} \times 100 \end{array} .$$

3.2 Information Loss Metrics

In the following, we will present seven information loss metrics in the light of the modeling defined in Section 3.1. Three come from the literature: *Distortion* [18], *NCP* [35] and *Total* [5]. Four have been exposed in a previous work [23]: *Lost Leaves Metric* (*LLM*), *Normalized Lost Leaves Metric* (*NLLM*), *Wid Lost Leaves Metric* (*WLLM*) and *Wid Normalized Lost Leaves Metric* (*WNLLM*).
First of all, we introduce notations and definitions. The *height* of a generalization hierarchy is the number of nodes in the longest path. We denote by lvl(v) the level of v in the generalization hierarchy and by nl(v) the number of leaves in the subtree rooted in v.

Definition 8 (Weighting on a set of quasi-identifier attributes)**.** Let $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ be a set of $m \in \mathbb{N}^*$ quasi-identifier attributes and $\mathcal{H} = (H_1, \ldots, H_m)$ be a m-tuple of hierarchies of $\mathcal{Q}$. Set h_{H_j} the height of H_j for $j \in [\![1, m]\!]$ and $h_{\max} = \max\limits_{1 \leq j \leq m} h_{H_j}$.
We define the application $w_1 : \mathcal{Q} \to \mathbb{R}$ which associates with each quasi-identifier attribute of $\mathcal{Q}$ its weight for w_1:

$$\begin{array}{rccc} w_1: & \mathcal{Q} & \longrightarrow & \mathbb{R} \\ & Q_j & \longmapsto & 1 - \frac{(h_{H_j} - 1)^m}{\sum_{i=1}^{m} (h_{H_i} - 1)^m} \end{array} ,$$

for $j \in [\![1, m]\!]$. It is the wid presented in [27].
We define the application $w_2 : \mathcal{Q} \to \mathbb{R}$ which associates with each quasi-identifier attribute of $\mathcal{Q}$ its weight for w_2:

$$\begin{array}{rccc} w_2: & \mathcal{Q} & \longrightarrow & \mathbb{R} \\ & Q_j & \longmapsto & \frac{h_{\max}}{h_{H_j}} \end{array} ,$$

for $j \in [\![1, m]\!]$.

Let $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ be a set of $m \in \mathbb{N}^*$ quasi-identifier attributes and $\mathcal{H} = (H_1, \ldots, H_m)$ be a m-tuple of hierarchies of $\mathcal{Q}$. Let $j \in [\![1, m]\!]$. Let x and x' be two nodes of H_j such that there is an edge from x to x' and $\text{lvl}(x) = \text{lvl}(x') - 1$. Let r_j be the root of H_j. For each studied metric, we give the weight $\omega(x, x')$ to put on the edge (x, x').

3.2.1 Metrics of the Literature

Distortion *Distortion* is a metric exposed in [18]. We use the weighting w_1 as in [27] to take into account the heights of the hierarchies: these multipliers put a penalty on hierarchies of small height (e.g. $Gender$, an attribute with only one possible generalization, has a bigger w_1 than Age, an attribute that could have several levels of generalization).
The weight to put on the edge (x, x') for *Distortion* is:

$$\omega(x, x') = \frac{\frac{1}{h_{H_j} - \text{lvl}(x')}}{\sum_{i=1}^{h_{H_j}-1} \frac{1}{h_{H_j} - i}} \times w_1(Q_j).$$

NCP *Normalized Certainty Penalty*, or *NCP*, is a metric that deals with the number of lost leaves when a generalization is applied to a value. A normalization step is done by dividing by the number of leaves in the hierarchy. It comes from [35]. The weight to put on the edge (x, x') for *NCP* is:

$$\omega(x, x') = \frac{\text{nl}(x') - \text{nl}(x)}{\text{nl}(r_j)}.$$

Total Exposed in [5], *Total* focuses on the level of the nodes in the hierarchy. The closer a node is to the root, the higher its generalization cost for *Total*. A normalization step is done by dividing by the height of the hierarchy minus 1. The weight to put on the edge (x, x') for *Total* is:

$$\omega(x, x') = \frac{\text{lvl}(x') - \text{lvl}(x)}{h_{H_j} - 1}.$$

3.2.2 *LLM* and Three Variants

We now present *LLM* and three of its variants as they are defined in our previous work [23].

LLM The *Lost Leaves Metric*, or *LLM*, is a metric based on the number of lost leaves when a generalization is done. It is the same idea as in *NCP* but we add the weighting w_2 to take into account the heights of the hierarchies. The weight to put on the edge (x, x') for *LLM* is:

$$\omega(x, x') = (\text{nl}(x') - \text{nl}(x)) \times w_2(Q_j).$$

NLLM The first variant of *LLM* is the *Normalized Lost Leaves Metric* or *NLLM*. A normalization step is added by dividing by the number of leaves in the hierarchy. It is the *NCP* metric in which we add the weighting w_2. The purpose of this variant is to study the effects of the normalization on the node generalization cost when the weighting used is w_2. The weight to put on the edge (x, x') for *NLLM* is:

$$\omega(x, x') = \frac{\text{nl}(x') - \text{nl}(x)}{\text{nl}(r_j)} \times w_2(Q_j).$$

WLLM The second variant of *LLM* is the *Wid Lost Leaves Metric* or *WLLM*. The weighting used is w_1. The purpose is to compare the performances of the two weightings. The weight to put on the edge (x, x') for *WLLM* is:

$$\omega(x, x') = (\text{nl}(x') - \text{nl}(x)) \times w_1(Q_j).$$

WNLLM The last variant of *LLM* is the *Wid Normalized Lost Leaves Metric* or *WNLLM*. In this metric, a normalization step on the node generalization cost is done and the weighting used is w_1. The purpose of this variant is to study the effects of the normalization on the node generalization cost when the weighting used in w_1. The weight to put on the edge (x, x') for *WNLLM* is:

$$\omega(x, x') = \frac{\text{nl}(x') - \text{nl}(x)}{\text{nl}(r_j)} \times w_1(Q_j).$$

To better understand what the metrics are calculated on, we decompose the compute of the value's cost in two phases. The first one consists in assigning to the value an intermediate cost and the second one is a multiplication of this intermediate cost by a multiplier depending on the attribute which the value belongs. We note μ_{inter} to represent the first phase, for μ a metric, and μ_{multi} for the second phase.

For instance, for *Distortion*, $\textit{Distortion}_{inter} = \frac{\frac{1}{h_{H_j} - \text{lvl}(x')}}{\sum_{i=1}^{h_{H_j}-1} \frac{1}{h_{H_j}-i}}$ and $\textit{Distortion}_{multi} = w_1(Q_j)$.

For *NLLM*, we have $\textit{NLLM}_{inter} = \frac{\text{nl}(x') - \text{nl}(x)}{\text{nl}(r_j)}$ and $\textit{NLLM}_{multi} = w_2(Q_j)$. Finally, $\textit{NCP}_{inter} = \textit{NLLM}_{inter}$ and $\textit{NCP}_{multi} = 1$. Table 2 lists characteristics of the two phases for the seven studied metrics.

	Distortion	*NCP*	*Total*	*LLM*	*NLLM*	*WLLM*	*WNLLM*
μ_{inter} depends on the level of the node in the hierachy	✓	×	✓	×	×	×	×
μ_{inter} depends on the number of lost leaves	×	✓	×	✓	✓	✓	✓
Normalization on μ_{inter}	✓	✓	✓	×	✓	×	✓
μ_{multi} penalizes the small hierarchies	w_1	×	×	w_2	w_2	w_1	w_1

Table 2: Characteristics of the seven studied metrics

3.3 Experiments

In this section, we will study the performances of information loss metrics when used in a k-anonymization process. We present the experimental protocol and then we analyze the results obtained.

3.3.1 Experimental Protocol

We conduct experiments on two public and available online tables: the *Adult data set*[1] [31] and an extract of records from the voter list of Florida state[2] [32]. In *Adult data set*, we conserve nine attributes (*Age*, *Gender*, *Race*, *Marital status*, *Education*, *Native country*, *Work class*,

[1][Online; accessed on June 2019] `https://archive.ics.uci.edu/ml/index.php`
[2][Online; accessed on May 2020] `http://flvoters.com/`

Occupation and *Salary*). *Adult data set* contains 30162 complete records (we removed the lines with missing values). In order to have two tables with the same number of records and so to simplify comparisons between the two tables, we randomly extract 30162 records from the Florida voter list. The extract is denoted *florida_30162*. In *florida_30162*, we conserve five attributes (*Zipcode*, *Year of birth*, *Gender*, *Race* and *Affiliation*). All table attributes will be considered quasi-identifiers.

Let $\mathcal{M} = \{$*Distortion*, *NCP*, *Total*, *LLM*, *NLLM*, *WLLM*, *WNLLM*$\}$ be the set of information loss metrics training that we are going to study.

For each table, for each metric of $\mathcal{M}$, we will use *GAA* (cf. Section 2.2) to k-anonymize the table for 14 values of k between 3 and 15000. The range of k values is large and the greatest of them are not realistic for a table of 30162 records. We tested these high values to analyze the behavior of the algorithms and optimization strategies. The choice of a suitable value for k is not easy and depends on the acceptance of the individuals whose data is made public.

The strategy to give to *GAA* as *Strat* parameter has for condition:

$$C \in \{C' \in \mathcal{C}(T) - C_s \,:\, \overline{\mu}(C_s, C') = \min_{C'' \in \mathcal{C}(T) - C_s} \overline{\mu}(C_s, C'')\}.$$

This condition reflects the fact that C is chosen so that its generalization cost with C_s for the metric μ is minimal.

In other words, for a table T , for a metric $\mu \in \mathcal{M}$, for an integer k, we will build a k-anonymous version of T with *GAA* by guiding the mergings of equivalence classes to be performed thanks to the metric μ.

To compare k-anonymous tables obtained by running *GAA* with the seven metrics, we use three quality criteria.

Definition 9 (Quality criteria)**.** Let $\mathcal{A} = \{Q_1, \ldots, Q_m, S\}$ be a set of $m \in \mathbb{N}^*$ quasi-identifier-identifiers attributes and one sensitive attribute. Let $\mathcal{H} = (H_1, \ldots, H_m)$ be a m-tuple of hierarchies of $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ with r_j the root of H_j for all $j \in [\![1, m]\!]$. Let T be a table on $(\mathcal{A}, \mathcal{H})$ of cardinal $n \in \mathbb{N}^*$. Set $T = \{E^1, \ldots, E^n\}$ with $E^i = (e^i_1, \ldots, e^i_m, s^i)$ for all $i \in [\![1, n]\!]$. Let $T^{gen} = \{F^1, \ldots, F^n\}$ be a generalized table on $(T, \mathcal{H})$ with $F^i = (f^i_1, \ldots, f^i_m, s^i)$ for all $i \in [\![1, n]\!]$. Let $\mathcal{M}$ be a set of information loss metrics.

Mean alteration on $\mathcal{M}$ The mean alteration of T^{gen} on $\mathcal{M}$ is:

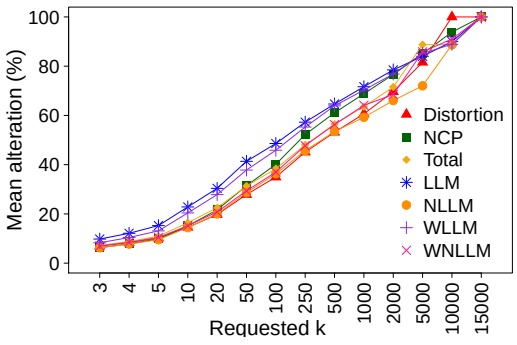
$$alt_{\mathcal{M},T}(T^{gen}) = \frac{1}{|\mathcal{M}|} \times \sum_{\mu \in \mathcal{M}} alt_{\mu,T}(T^{gen}).$$

Percentage of generalized values The percentage of generalized values of T^{gen} according to T is:

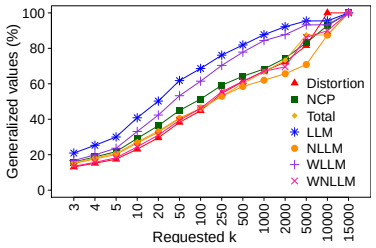
$$p_{\text{gen},T}(T^{gen}) = \frac{|\{(i,j) \in [\![1,n]\!] \times [\![1,m]\!] \,:\, \text{lvl}(f^i_j) > \text{lvl}(e^i_j)\}|}{n \times m} \times 100.$$

Percentage of generalized values at the root (i.e. at the maximum level of the hierarchy) The percentage of generalized values at the root of T^{gen} according to T is:

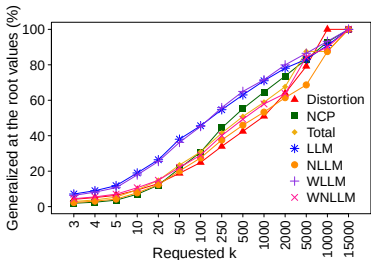
$$p_{\text{root},T}(T^{gen}) = \frac{|\{(i,j) \in [\![1,n]\!] \times [\![1,m]\!] \,:\, f^i_j = r_j\}|}{n \times m} \times 100.$$



(a) Mean alteration on $\mathcal{M}$

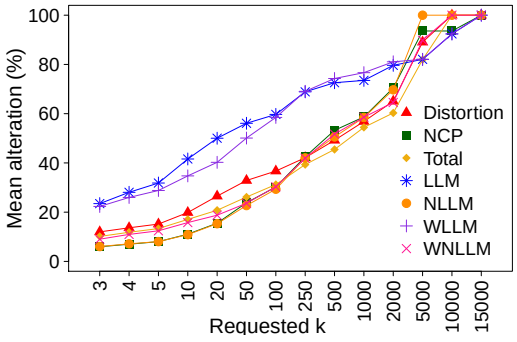


(b) Percentage of generalized values

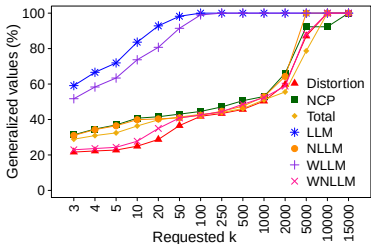


(c) Percentage of generalized at the root values

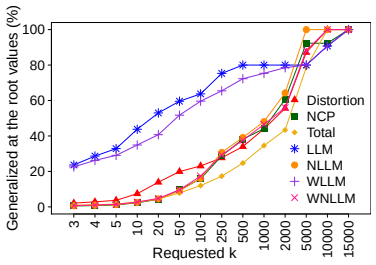
Figure 3: Experimental results on *Adult data set*



(a) Mean alteration on $\mathcal{M}$



(b) Percentage of generalized values



(c) Percentage of generalized at the root values

Figure 4: Experimental results on *florida_30162*

	Mean alteration		Percentage of generalized values		Percentage of generalized at the root values	
	Metric	*NAUC*	Metric	*NAUC*	Metric	*NAUC*
Adult data set	*NLLM*	56.07	*NLLM*	59.63	*Distortion*	48.35
	Distortion	57.3	*WNLLM*	63.24	*NLLM*	49.74
	WNLLM	59.29	*Total*	63.59	*WNLLM*	53.1
	Total	60.06	*Distortion*	63.66	*Total*	54.96
	NCP	64.65	*NCP*	66.28	*NCP*	59.64
	WLLM	66.6	*WLLM*	80.06	*LLM*	66.93
	LLM	68.12	*LLM*	84.43	*WLLM*	68.03
florida_30162	*Total*	50.4	*Total*	49.75	*Total*	31.03
	Distortion	53.96	*Distortion*	50.07	*Distortion*	41.32
	WNLLM	54.33	*WNLLM*	51.43	*WNLLM*	42.59
	NLLM	55.25	*NLLM*	52.92	*NCP*	42.85
	NCP	56.13	*NCP*	54.4	*NLLM*	45.64
	LLM	72.91	*WLLM*	99.4	*WLLM*	72.71
	WLLM	74.48	*LLM*	99.76	*LLM*	77.79

Table 3: *NAUC* for the seven metrics for mean alteration, percentage of generalized values and percentage of generalized at the root values computed on $[3, 2000]$ on both tables

Thus, for each k-anonymous table, we compute its mean alteration on $\mathcal{M}$, its percentage of generalized values and its percentage of generalized values at the root. Results are presented in Figures 3 and 4.

For each table, to have a representative value of the results, we compute for each metric a mean value for each quality criterion. To do so, we use the following formula: for a continuous function f on $[a, b] \subset \mathbb{R}$, the mean value of f on $[a, b]$ is $\frac{1}{b-a}\int_a^b f(x)dx$. Thus, for a metric and a quality criterion, we call *Normalized Area Under Curve*, abbreviated in *NAUC*, of the metric for the criterion the mean value of the results for the criterion obtained by the k-anonymous tables produced with *GAA* by using the metric for $k \in [3, 15\,000]$. As our three quality criteria are percentages to be minimized, the *NAUC* are also percentages and if a metric has a low *NAUC* for a criterion, we say that this metric has good performance for this criterion. Because metrics have a chaotic behavior for high values of k, we compute the *NAUC* on $[3, 2000]$. Results are presented in Table 3.

3.3.2 Results Analysis

Results analysis will be done in two parts. The first part consists in comparing metrics performance according to the three quality criteria. In the second part, we analyze the effects of three characteristics of the metrics definitions on the performance of the metrics.

For mean alteration on $\mathcal{M}$, we notice that *NLLM* has the best *NAUC* on *Adult data set* (cf. line *Adult data set* and column Mean alteration of Table 3). For *florida_30162*, *Total* is the best metric with its *NAUC* around 50% (cf. line *florida_30162* and column Mean alteration of Table 3). *LLM* and *WLLM* are the worst metrics for both tables for this criterion. On *florida_30162*, *LLM* has a *NAUC* for mean alteration around 73% whereas *NCP* has a *NAUC* around 56%.

For the percentage of generalized values, we notice in line *Adult data set* and column Percentage of generalized values of Table 3 that *NLLM* has the best *NAUC* on *Adult data set*. The percentage of generalized values of the k-anonymous versions of *Adult data set* for $k \in [3, 2000]$ produced with *NLLM* is on average around 60%. *Distortion*, *Total* and *WNLLM* have *NAUC* around 63% for the percentage of generalized values on *Adult data set*. *LLM* and

WLLM are the worst for this criterion on *Adult data set*. It is the same report on *florida_30162* (cf. line *florida_30162* and column Percentage of generalized values of Table 3). The percentage of generalized values of the k-anonymous versions of *florida_30162* for $k \in [3, 2000]$ produced with *LLM* and *WLLM* is more than 99% on average. This observation is visible on Figure 4b: the curves of *LLM* and *WLLM* reach 100% from $k = 250$.

For the percentage of generalized values at the root, we observe the same results. *LLM* and *WLLM* have the worst *NAUC* on both tables. On *Adult data set* (cf. line *Adult data set* and column Percentage of generalized at the root values of Table 3), *NLLM* is among the best metrics with its *NAUC* around 50%. On *florida_30162* (cf. line *florida_30162* and column Percentage of generalized at the root values of Table 3), *Total* obtains a *NAUC* of 10 points lower than the *NAUC* of the second best metric.

To conclude on this first point, *LLM* and *WLLM* obtained the worst results for the three quality criteria in terms of *NAUC* computed in $[3, 2000]$. Regarding metrics with the best results, *NLLM* seems to be the more interesting to use on *Adult data set* and *Total* achieves to produce good k-anonymous versions of *florida_30162* with regard to the three quality criteria.

After the metrics presentation in Section 3.2, we listed some characteristics of the metrics definitions. To define a metric, we gave a formula computing the weights to put in the hierarchies edges. This formula can be break down in two phases: the node generalization cost and weighting on the set of quasi-identifier attributes. We have identified three main characteristics for the different metrics (see Table 2):

1. the node generalization cost depends on hierarchy height or hierarchy width: Increasing the cost based on the height of the hierarchy assumes that the more time a data is generalized, the more its utility decreases (e.g. generalizing a "cat" to "felid" is more precise than generalize it to "mammal"). But this assumption ignores the details of original data and the distribution of possible values. Indeed, it can be very different to generalize a data in a set of two possibilities compared to generalizing it in a set of several tens (for instance a hierarchy which contains the dog and the wolf as only "canids" but the 40 species of felids grouped in a single generalization "felidae"). Considering the width of the hierarchy attempts to take into account such a situation.

2. weighting on the set of quasi-identifier attributes is w_1, w_2 or none: This weighting is set to take into account the height differences between the hierarchies of the quasi-identifier attributes in the calculation of generalization cost of a record. For example, generalizing a cat to a felid, in a hierarchy of height 4 (cat $\rightarrow$ felid $\rightarrow$ mammal $\rightarrow$ animal) represents 25% of the maximum level of generalization compared to generalizing "Male" to "*", i.e. a 100% generalization (total loss of information).

3. there is or not a normalization step in the node generalization cost computation: The normalization step reduces the generalization costs of the different attributes to common values between 0 and 1. A proportion of the maximum score is considered rather than a "raw" score. This avoids favoring the generalization of one attribute over another.

For each previous point, each metric of $\mathcal{M}$ satisfies exactly one affirmation. For instance, for *NCP*, the node generalization cost depends on hierarchy width (point 1) and has a normalization step (point 3). Any weighting on the set of quasi-identifier attributes is used for *NCP* (point 2).

For each point, we will study if the differences in the metrics definitions have consequences on the results obtained for the three quality criteria and for both tables. We focus on the *NAUC* computed in $[3, 2000]$ (cf. Table 3).

For the point 1, we seek to know if the way to compute the node generalization cost influences the metrics performance. We distinguish two groups of metrics in $\mathcal{M}$: those for which node generalization cost depends on the hierarchy height (*Distortion* and *Total*) and those for which the node generalization cost depends on the hierarchy width (*NCP*, *LLM*, *NLLM*, *WLLM* and *WNLLM*). For *Adult data set*, for the three quality criteria (cf. line *Adult data set* of Table 3), this characteristic seems to have no impact on metrics performance. *NLLM* (computation on the hierarchy width) has the lowest *NAUC* on $[3, 2000]$ for two criteria and *Distortion* (computation on hierarchy height) has the lowest *NAUC* on $[3, 2000]$ for the last criterion. In contrast, for *florida_30162*, for the three criteria (cf. line *florida_30162* of Table 3), *Distortion* and *Total* have the lowest *NAUC* on $[3, 2000]$. These are the only two metrics in which the node generalization cost depends on the hierarchy height.

For the point 2, we study the impact of the application or not of a weighting on the set of quasi-identifier attributes. We presented two weightings in Definition 8. The weighting w_1 is used in *Distortion*, *WLLM* and *WNLLM* (group w_1). The weighting w_2 is used in *LLM* and *NLLM* (group w_2). For *NCP* and *Total*, no weighting is used (group "no weighting"). For both tables, for the three quality criteria, no group of metrics, based on the weighting, obtains significantly better results than the others. For instance, for *Adult data set*, for mean alteration (cf. line *Adult data set* and column Mean alteration of Table 3) and percentage of generalized values (cf. line *Adult data set* and column Percentage of generalized values of Table 3), *NLLM* and *LLM* have respectively the best and the worst *NAUC* on $[3, 2000]$ whereas the weighting w_2 is used in both metrics. The application of a weighting on the quasi-identifier attributes alone does not appear to be a determining factor in obtaining a metric with good performance.

For the point 3, we study the effects of a normalization step in the computation of the node generalization cost in metrics definitions on the metrics performance. For instance, in the *NCP* definition, the number of leaves in the subtree rooted in the node is divided by the total number of leaves in the hierarchy. In metrics definitions of $\mathcal{M}$, either we do a normalization step in the computation of the node generalization cost (*Distortion*, *Total*, *NCP*, *NLLM* and *WNLLM*) or we do not (*LLM* and *WLLM*). For the three quality criteria (cf. Table 3), *LLM* and *WLLM* have the highest *NAUC* on $[3, 2000]$ for both tables. Gaps with *NAUC* on $[3, 2000]$ of other metrics are important (excepted for mean alteration on *Adult data set*). For instance, we denote a gap of around 27 points between the *NAUC* of *WLLM* and *NLLM* for the percentage of values generalized at the root. It seems so that a normalization step in the computation of the node generalization cost greatly influences the metric performance. This characteristic seems essential for the metric to produce good quality k-anonymous tables according to the three quality criteria studied.

3.4 Summary

In this section, our purpose was to compare performance of information loss metrics when they are used in a k-anonymization algorithm.

To do so, we defined in Section 3.1 an information loss metric on a set of generalization hierarchies as a set of weights put on the edges of the hierarchies. With this definition, we presented a modeling that permits to simplify the use of information loss metrics: we defined costs matrices associated with a metric. Thanks to costs matrices, we explained the alteration of a generalized table used to compare the quality of k-anonymous tables.

Then, we got interested in several information metrics in Section 3.2. *Distortion*, *NCP* and *Total* are metrics already defined in articles and *LLM* and three of its variants come from our research.

Finally, in Section 3.3, we experimented on two public datasets to compare the performance of the chosen metrics during a k-anonymization process. We used three quality criteria to compare the quality of the obtained k-anonymous tables. We note that the metrics with best performance are not the same in the two tables studied. On *Adult data set*, *NLLM* is among the best metrics for the three criteria. On *florida_30162*, *Total* produces the best k-anonymous tables in view of the three quality criteria. Regarding the worst metrics, our experiments show that metrics without a normalization step on the computation of the node generalization cost do not achieve to produce good quality k-anonymous tables. Indeed, metrics that do not do this normalization step, *LLM* and *WLLM*, obtain the worst results for the three quality criteria on both tables.

We have seen that the metric used to produce the best k-anonymous tables is not the same on the two tables. To continue this work, it would be interesting to look for criteria based on the characteristics of the table or on the shape of the hierarchies and allowing to choose the best metric to use to k-anonymize a table.

4 Optimization Strategies

In Section 3, we k-anonymize tables containing exclusively quasi-identifier attributes. But, a weakness of k-anonymity is that a lack of diversity could appear in the sensitive values in equivalence classes of k-anonymous table and potentially disclose information [22]. As k-anonymity does not take into account sensitive attributes, nothing guarantees that equivalence classes of a k-anonymous table will not present very unbalanced sensitive values distributions. Intituively, the greater the requested k value is, the more important the probablity of a better sensitive values distribution in the equivalence classes is. It is therefore necessary to define what better distribution of sensitive values means. We study here l-diversity [22] and t-closeness [19], two anonymization models controlling sensitive values distribution in equivalence classes of a table.

Our purpose is still to produce k-anonymous tables. In this section, we develop new strategies, each corresponding to a way to guide equivalence classes mergings in *GAA* (cf. Section 2.2). Thanks to these strategies mixing generalization cost for a metric, l-diversity and t-closeness, we hope that produced k-anonymous tables have other privacy guarantees than the one brought by k-anonymity.

In the rest of this section, we will present the definitions of l-diversity and t-closeness and two measures to evaluate levels of l-diversity and t-closeness of a table in Section 4.1. Then, we will detail seven strategies to use in the anonymization algorithm *GAA* to produce good quality k-anonymous tables in terms of l-diversity and t-closeness in Section 4.2. Finally, we will present our experiments and the results obtained in Section 4.3.

4.1 l-diversity, t-closeness and Measures

In this section, we consider a set $\mathcal{A} = \{Q_1, \ldots, Q_m, S\}$ of $m \in \mathbb{N}^*$ quasi-identifier attributes and one sensitive attribute, a m-tuple $\mathcal{H} = (H_1, \ldots, H_m)$ of generalization hierarchies of $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$ and a table T on $(\mathcal{A}, \mathcal{H})$ of cardinal $n \in \mathbb{N}^*$.

Definition of l-diversity was first given by Machanavajjhala et al. in [22]. It is an anonymization model that guarantees that, in each equivalence class of a table, a set of sensitive

values is well represented. An equivalence class C is well represented by l sensitive values if there exists at least $l \geq 2$ sensitive values in C such that the l most frequent values in C have mostly the same frequency of occurrence.

In other words, the idea of l-diversity is to guarantee that, in each equivalence class, there are at least l distinct sensivite values present in sufficient quantity. Thus, although an adversary knows the equivalence class of the anonymous table in which is an individual, probability to associate a sensitive value to the individual is low. It avoids situations where belonging to an equivalence class gives strong clues about the sensitive values.

In our work, we use the *entropy l-diversity* [22].

Definition 10 (Entropy l-diversity)**.** Let $l \in \mathbb{N}^*$. T is *entropic l-diverse* if, for each equivalence class C of T, we have:

$$-\sum_{s \in S_C} p_T(s, C) \ln(p_T(s, C)) \geq \ln(l),$$

with $p_T(s, C) := \frac{1}{|C|}|\{E = (e_1, \ldots, e_r, s_E) \in C \,:\, s_E = s\}|$ the proportion of sensitive value s in records of C and $S_C = \{s \in S \,:\, \exists E = (e_1, \ldots, e_m, s_E) \in C \,:\, s_E = s\}$ the set of distinct sensitive values in C.

We use the previous definition to introduce a measure of l-diversity of a table.

Definition 11 (l-diversity measure)**.** We define the application $l_{div,T} : \mathcal{C}(T) \to \mathbb{R}$ which associates with each equivalence class of T:

$$\begin{array}{rccl} l_{div,T} : & \mathcal{C}(T) & \longrightarrow & \mathbb{R} \\ & C & \longmapsto & \exp\left(-\sum\limits_{s \in S_C} p_T(s, C) \ln\left(p_T(s, C)\right)\right) \end{array}$$

with $S_C = \{s \in S \,:\, \exists E = (e_1, \ldots, e_m, s_E) \in C \,:\, s_E = s\}$.

We define the application $l_{div} : \mathcal{T}^n_{(\mathcal{A},\mathcal{H})} \to \mathbb{R}$ which associates with each table on $(\mathcal{A}, \mathcal{H})$ its *l-diversity value*:

$$\begin{array}{rccl} l_{div} : & \mathcal{T}^n_{(\mathcal{A},\mathcal{H})} & \longrightarrow & \mathbb{R} \\ & T & \longmapsto & \min\limits_{C \in \mathcal{C}(T)} l_{div,T}\left(C\right) \end{array} .$$

Defined by Li et al. in [19], t-closeness requests sensitive values distribution in each equivalence class to be not further than a threshold t from sensitive values distribution in the whole table. We say that the equivalence class has a t-closeness. The knowledge of sensitive values distribution in the whole table is a prerequisite to achieve t-closeness.

In other words, the idea of t-closeness is to guarantee that, in each equivalence class, sensitive values distribution is approximatively similar to sensitive values distribution in the whole table. Thus, although an adversary knows which equivalence class an individual belongs to, probability of finding the sensitive value of the individual is the same as in the whole table.

We use here the definition of t-closeness with norm 1.

Definition 12 (Associated distance to norm 1 on $\mathbb{R}^n$)**.** Let $n \in \mathbb{N}^*$.

Norm 1 on $\mathbb{R}^n$ is:

$$\begin{array}{rccl} \|.\|_1 : & \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & (x_1, \ldots, x_n) & \longmapsto & \sum_{i=1}^n |x_i| \end{array} .$$

Associated distance to norm 1 on $\mathbb{R}^n$ is:

$$\begin{array}{rccl} d_{\|.\|_1} : & \mathbb{R}^n \times \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & (x, y) & \longmapsto & \|x - y\|_1 \end{array} .$$

Definition 13 (t-closeness of a table)**.** Set $S = \{s_1, \ldots, s_q\}$ for $q \in \mathbb{N}^*$. Let C be an equivalence class of T. Let $t \in [0, 1]$.

Set $P(S,T) = (p(s_1,T), \ldots, p(s_q,T))$ the vector of sensitive values distribution of S in T and $P_T(S,C) = (p_T(s_1,C), \ldots, p_T(s_q,C))$ the vector of sensitive values distribution of S in C.

C has a t-closeness for $d_{\|.\|_1}$ if:

$$d_{\|.\|_1}(P_T(S,C), P(S,T)) \leq t.$$

T has a t-closeness for $d_{\|.\|_1}$ if each equivalence class of T has a t-closeness for $d_{\|.\|_1}$.

We use the previous definition to introduce a measure of t-closeness of a table.

Definition 14 (t-closeness measure)**.** We define the application $t_{clo,T} : \mathcal{C}(T) \to \mathbb{R}$ which associates with each equivalence class of T:

$$\begin{array}{rccl} t_{clo,T}: & \mathcal{C}(T) & \longrightarrow & \mathbb{R} \\ & C & \longmapsto & d_{\|.\|_1}(P_T(S,C), P(S,T)) \end{array},$$

with $P(S,T) = (p(s_1,T), \ldots, p(s_q,T))$ and $P_T(S,C) = (p_T(s_1,C), \ldots, p_T(s_q,C))$.

We define the application $t_{clo} : \mathcal{T}^n_{(\mathcal{A},\mathcal{H})} \to \mathbb{R}$ which associates with each table on $(\mathcal{A},\mathcal{H})$ its *t-closeness value* for $d_{\|.\|_1}$:

$$\begin{array}{rccl} t_{clo}: & \mathcal{T}^n_{(\mathcal{A},\mathcal{H})} & \longrightarrow & \mathbb{R} \\ & T & \longmapsto & \max\limits_{C \in \mathcal{C}(T)} t_{clo,T}(C) \end{array}.$$

4.2 Optimization Strategies

In this section, we present seven optimization strategies mixing generalization cost for an information loss metric, value of l-diversity and value of t-closeness to use in *GAA* as *Strat* parameter. These strategies give conditions of selection of the equivalence class C to merge with C_s at each round of *GAA*.

Each strategy aims to optimize one or several of the following measures:

- generalization cost represented by $\overline{\mu}$ with μ an information loss metric defined in Section 3.2
- l-diversity value represented by l_{div} defined in Section 4.1
- t-closeness value represented by t_{clo} defined in Section 4.1

Two approaches are possible:

1. two measures are successively considered. The selected merging is chosen among the mergings optimizing the second measure among those optimizing the first measure (e.g. selection of the best mergings minimizing the generalization cost and then, among these mergings, selection of the one which maximizes the l-diversity)
2. the selected merging is chosen among those optimizing two measures in the same time

In the first approach, we evaluate the merging of C_s with all other equivalence classes according to a first measure then, among classes that obtain the best result for the first measure, we choose a class such that the merging with C_s gives the best result for the second measure.

In the second approach, the aim is to optimize equivalence classes mergings according to two measures simultaneously. To do so, we define two applications: one mixing generalization cost and l-diversity value, the other mixing generalization cost and t-closeness value (cf. Definition 15). Recall that generalization cost and t-closeness value are to be minimized in a table and l-diversity value is to be maximized.

For Definitions 15 to 17, we consider a set $\mathcal{A} = \{Q_1, \ldots, Q_m, S\}$ of $m \in \mathbb{N}^*$ quasi-identifier attributes and one sensitive attribute, a m-tuple $\mathcal{H} = (H_1, \ldots, H_m)$ of hierarchies of $\mathcal{Q} = \{Q_1, \ldots, Q_m\}$, a table T on $(\mathcal{A}, \mathcal{H})$ of cardinal $n \in \mathbb{N}^*$ with $\mathcal{C}(T)$ its set of equivalence classes and a metric μ on $\mathcal{H}$.

Definition 15 (Generalization cost, l-diversity value and t-closeness value)**.** We define the application $\overline{\mu}_{l_{div}} : \mathcal{C}(T) \times \mathcal{C}(T) \to \mathbb{R}$ which associates with each couple of equivalence classes of T their generalization cost for μ divided by the l-diversity value of the generalized table in which both classes are merged:

$$\begin{array}{rccc} \overline{\mu}_{l_{div}} : & \mathcal{C}(T) \times \mathcal{C}(T) & \longrightarrow & \mathbb{R} \\ & (C, C') & \longmapsto & \frac{\overline{\mu}(C,C')}{l_{div}(gen_T(C \cup C'))} \end{array} .$$

We define the application $\overline{\mu}_{t_{clo}} : \mathcal{C}(T) \times \mathcal{C}(T) \to \mathbb{R}$ which associates with each couple of equivalence classes of T their generalization cost for μ multiplied by the t-closeness value of the generalized table in which both classes are merged:

$$\begin{array}{rccc} \overline{\mu}_{t_{clo}} : & \mathcal{C}(T) \times \mathcal{C}(T) & \longrightarrow & \mathbb{R} \\ & (C, C') & \longmapsto & \overline{\mu}(C, C') \times t_{clo}\left(gen_T(C \cup C')\right) \end{array} .$$

To minimize $\overline{\mu}_{l_{div}}$, we can minimize $\overline{\mu}$ or maximize l-diversity value. The idea is to propose a tradeoff between generalization cost for μ and l-diversity value. In other words, we allow a generalization cost for μ x times more important if the merging has a l-diversity value of x.

To minimize $\overline{\mu}_{t_{clo}}$, we can minimize $\overline{\mu}$ or minimize t-closeness value. The idea is to propose a tradeoff between generalization cost for μ and t-closeness value. In other words, we allow a generalization cost of x if the merging has a t-closeness value of $\frac{1}{x}$.

To lighten the descriptions of optimization strategies, we define subsets of equivalence classes respecting certain properties.

Definition 16 (Subsets of equivalence classes)**.** Let $C \in \mathcal{C}(T)$. We define five subsets of $\mathcal{C}(T)$ depending on C.

The set of equivalence classes such that their merging with C minimizes the generalization cost for μ is:

$$\Delta_{\overline{\mu}}(C, \mathcal{C}(T)) = \{C' \in \mathcal{C}(T) - C \,:\, \overline{\mu}(C, C') = \min_{C'' \in \mathcal{C}(T) - C} \overline{\mu}(C, C'')\}.$$

The set of equivalence classes such that their merging with C maximizes the l-diversity value is:

$$\Delta_{l_{div}}(C, \mathcal{C}(T)) = \{C' \in \mathcal{C}(T) - C \,:\, l_{div}\left(gen_T(C \cup C')\right) = \max_{C'' \in \mathcal{C}(T) - C} l_{div}\left(gen_T(C \cup C'')\right)\}.$$

The set of equivalence classes such that their merging with C minimizes the t-closeness value is:

$$\Delta_{t_{clo}}(C, \mathcal{C}(T)) = \{C' \in \mathcal{C}(T) - C \,:\, t_{clo}(gen_T(C \cup C')) = \min_{C'' \in \mathcal{C}(T)-C} t_{clo}(gen_T(C \cup C''))\}.$$

The set of equivalence classes such that their merging with C minimizes $\overline{\mu}_{l_{div}}$ is:

$$\Delta_{\overline{\mu}_{l_{div}}}(C, \mathcal{C}(T)) = \{C' \in \mathcal{C}(T) - C \,:\, \overline{\mu}_{l_{div}}(C, C') = \min_{C'' \in \mathcal{C}(T)-C} \overline{\mu}_{l_{div}}(C, C'')\}.$$

The set of equivalence classes such that their merging with C minimizes $\overline{\mu}_{t_{clo}}$ is:

$$\Delta_{\overline{\mu}_{t_{clo}}}(C, \mathcal{C}(T)) = \{C' \in \mathcal{C}(T) - C \,:\, \overline{\mu}_{t_{clo}}(C, C') = \min_{C'' \in \mathcal{C}(T)-C} \overline{\mu}_{t_{clo}}(C, C'')\}.$$

We define the seven following strategies by the conditions of selection of the equivalence class C to merge with C_s in *GAA*.

Definition 17 (Optimization strategies)**.** Let Φ be an anonymization model. Let C_s be the chosen equivalence class of minimum size in a round of $GAA(\mathcal{A}, \mathcal{H}, T, \Phi, Strat)$.

Strategy 1 (S1) C is chosen among the classes such that their merging with C_s minimizes the generalization cost for μ: $C \in \Delta_{\overline{\mu}}(C_s, \mathcal{C}(T))$.

Strategy 2 (S2) C is chosen among the classes such that their merging with C_s maximizes the l-diversity value among those such that their merging with C_s minimizes the generalization cost for μ: $C \in \Delta_{l_{div}}(C_s, \Delta_{\overline{\mu}}(C_s, \mathcal{C}(T)))$.

Strategy 3 (S3) C is chosen among the classes such that their merging with C_s minimizes the generalization cost for μ among those such that their merging with C_s maximizes the l-diversity value: $C \in \Delta_{\overline{\mu}}(C_s, \Delta_{l_{div}}(C_s, \mathcal{C}(T)))$.

Strategy 4 (S4) C is chosen among the classes such that their merging with C_s minimizes the generalization cost for μ divided by the l-diversity value: $C \in \Delta_{\overline{\mu}_{l_{div}}}(C_s, \mathcal{C}(T))$.

Strategy 5 (S5) C is chosen among the classes such that their merging with C_s minimizes the t-closeness value among those such that their merging with C_s minimizes the generalization cost for μ: $C \in \Delta_{t_{clo}}(C_s, \Delta_{\overline{\mu}}(C_s, \mathcal{C}(T)))$.

Strategy 6 (S6) C is chosen among the classes such that their merging with C_s minimizes the generalization cost for μ among those such that their merging with C_s minimizes the t-closeness value: $C \in \Delta_{\overline{\mu}}(C_s, \Delta_{t_{clo}}(C_s, \mathcal{C}(T)))$.

Strategy 7 (S7) C is chosen among the classes such that their merging with C_s minimizes the generalization cost for μ multiplied by the t-closeness value: $C \in \Delta_{\overline{\mu}_{t_{clo}}}(C_s, \mathcal{C}(T))$.

Strategy 1, in which generalization cost only is taken into account, will be a repository for the measures to optimize. Indeed, for l-diversity and t-closeness values, the mergings choices made by this strategy can be considered random. For the generalization cost for μ, we hope that k-anonymous tables produced using Strategy 1 have good results. Strategies 2 to 4 optimize l-diversity and generalization cost for μ. Strategies 5 to 7 optimize t-closeness and generalization cost for μ.

4.3 Experiments

In this section, we will study the performance of optimization strategies when used in a k-anonymization process. For this, we will return to the experimental protocol set up and we will analyze the results obtained.

4.3.1 Experimental Protocol

As in Section 3.3, we choose *Adult data set* and *florida_30162* as experimental tables.
We do two kind of experiments depending on the sensitive attribute considered in the table:

Real data the sensitive attribute is an attribute of the original table

Simulated data the sensitive attribute is a new column added to the table and following a predeterminated distribution

For experiments on real data, we consider two configurations of *Adult data set* in which *Age* and *Marital status* are the sensitive attributes and one configuration of *florida_30162* in which *Affiliation* is the sensitive attribute. We denote these tables by $Adult_{Age}$, $Adult_{Mar}$ and $florida\_30162_{Aff}$.
For experiments on simulated data, we create seven sensitive attributes with 5, 10, 20, 50, 100, 200 or 500 possible values. Then, for each attribute, we generate sets of 30162 values whose distribution is Equivalent, Geometric or Standard Normal. Finally, we add these sets as new columns to *Adult data set* and *florida_30162*. We obtain 21 configurations for each table in which the sensitive attribute is one of the new attribute and the original attributes are quasi-identifier.
In both cases, we compare k-anonymous tables with three measures:

- alteration (to be minimized)
- l-diversity value (to be maximized)
- t-closeness value (to be minimized)

For certain strategies and for the alteration measure, we have to specify an information loss metric. As we saw in Section 3, the metrics with the best performance on *Adult data set* and *florida_30162* are *NLLM* and *Total* respectively. We therefore carried out our two kind of experiments using *NLLM* and *Total*. We note for both types of experiments that the behaviors of the strategies are slightly equivalent when we use *NLLM* and when we use *Total*. Thus, the choice of metric does not seem to influence the performance of optimization strategies when they are used in a k-anonymization process. In the following, we will only comment the results obtained using the *NLLM* metric.

4.3.2 On real Data

Firstly, we study two strategies that are not in our seven strategies: a strategy in which only l-diversity value is to be optimized and a strategy in which only t-closeness value is to be optimized. For both strategies, generalization cost is not taken into account. The aim is to justify the introduction of strategies mixing l-diversity and t-closeness values and generalization cost.

Denote by Strategy l_{div} (Strategy t_{clo}) the strategy only optimizing on l-diversity value (t-closeness value).

For both strategies, we produce with *GAA* k-anonymous versions of the tables $Adult_{Age}$, $Adult_{Mar}$ and $florida\_30162_{Aff}$ for $k \in \{3, 4, 5, 10, 20, 100, 250, 500, 1000, 2000, 5000, 10000, 15000\}$.

For Strategy l_{div}, we compute alteration for *NLLM* and l-diversity value of each k-anonymous table. Figures 5a, 6a and 7a present the results on the three tables. Each graph has two curves: the purple triangle curve represents alteration for *NLLM* of the k-anonymous table according to k and the red square curve represents l-diversity value of the k-anonymous table according to k. Recall that l-diversity value is to be maximized.

In order to analyze these graphs, we compute maximum l-diversity value for each sensitive attribute studied. This value corresponds to the case in which all the records of the table are in the same equivalence class. We obtain $l_{div}\left(Adult^*_{Age}\right) \simeq 50.03$, $l_{div}\left(Adult^*_{Mar}\right) \simeq 3.53$ and $l_{div}\left(florida\_30162^*_{Aff}\right) \simeq 3.16$.

For $Adult_{Mar}$ and $florida\_30162_{Aff}$, we note in Figures 6a and 7a that l-diversity value of k-anonymous tables is quickly equal to the maximum l-diversity value: for $Adult_{Mar}$, the maximum l-diversity value is reached from the 50-anonymous table and for $florida\_30162_{Aff}$, it is reached from the 10-anonymous table. For $Adult_{Age}$, the curve grows slower but the maximum l-diversity value is reached from the 500-anonymous table.

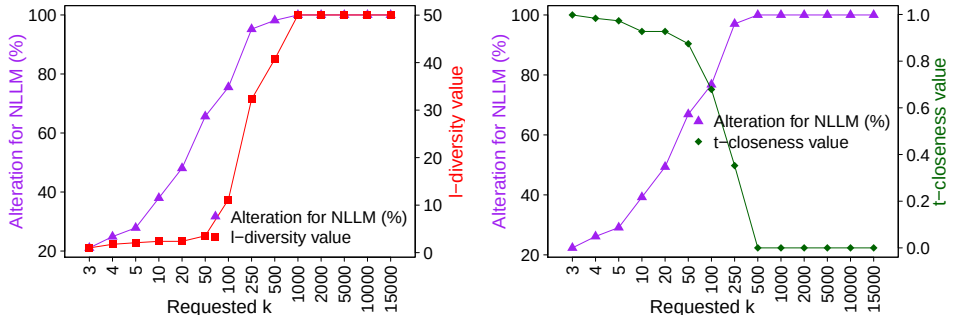
Nevertheless, alteration for *NLLM* of k-anonymous tables is high from the first k values. For $Adult_{Mar}$, alteration of the 3-anonymous table is close to 60% and for $florida\_30162_{Aff}$, alteration of the 3-anonymous table is higher than 97%. To compare, if we use in *GAA* the Strategy 1 guiding mergings only with generalization cost, the 3-anonymous version of $Adult_{Mar}$ produced has an alteration around 2.77% and the 3-anonymous version of $florida\_30162_{Aff}$ produced has an alteration around 1.23%.

Let's study results obtained with Strategy t_{clo}. For this strategy, we compute the alteration for *NLLM* and t-closeness value of each k-anonymous table. Figures 5b, 6b and 7b present the results on the three tables. Each graph has two curves: the purple triangle curve represents alteration for *NLLM* of the k-anonymous table according to k and the green diamond curve represents t-closeness value of the k-anonymous table according to k. Recall that t-closeness value is between 0 and 1 and is to be minimized in a table. Purple triangle curve is the alteration depending on the requested value of k (read on left y axis); red square curve is the l-diversity value depending on the requested value of k (read on right y axis); green diamonds are t-closeness values depending on the requested value of k (read on right y axis).

For $Adult_{Mar}$ and $florida\_30162_{Aff}$, we note in Figures 6b and 7b that t-closeness value is quickly close to 0: for $Adult_{Mar}$, t-closeness value is close to 0 from the 50-anonymous table and for $florida\_30162_{Aff}$, t-closeness value is 0 from the 10-anonymous table. For $Adult_{Age}$, the curve decreases slower but t-closeness value is 0 from the 500-anonymous table.

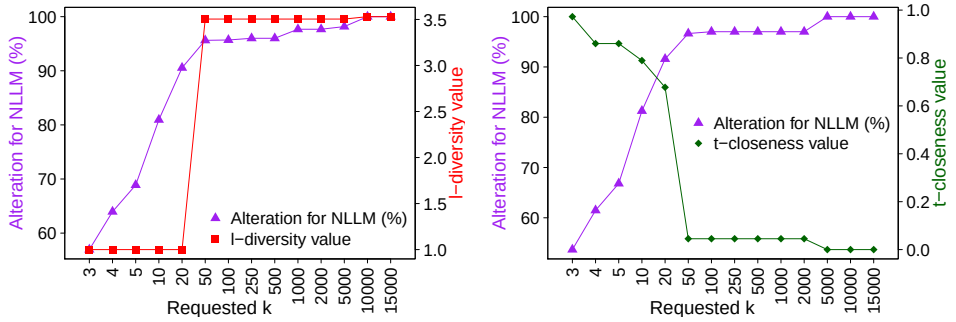
Nevertheless, alteration for *NLLM* of k-anonymous tables is high from the first k values. For $Adult_{Mar}$, alteration of the 3-anonymous table is close to 60% and for $florida\_30162_{Aff}$, alteration of the 3-anonymous table is close to 97%.

To conclude on this first experiment, although l-diversity and t-closeness values are quickly optimized in the k-anonymous tables when we use Strategies l_{div} and t_{clo}, this implies a high alteration of the k-anonymous tables from the first k values. It is 100% when l-diversity and t-closeness values are optimal in the k-anonymous table.



(a) Optimization of l-diversity only (b) Optimization of t-closeness only

Figure 5: Optimization on sensitive attribute *Age* of *Adult data set*- read the right y axis for l-diversity (squares) or t-closeness (diamonds) values, read the left y axis for alteration values (triangles)



(a) Optimization of l-diversity only (b) Optimization of t-closeness only

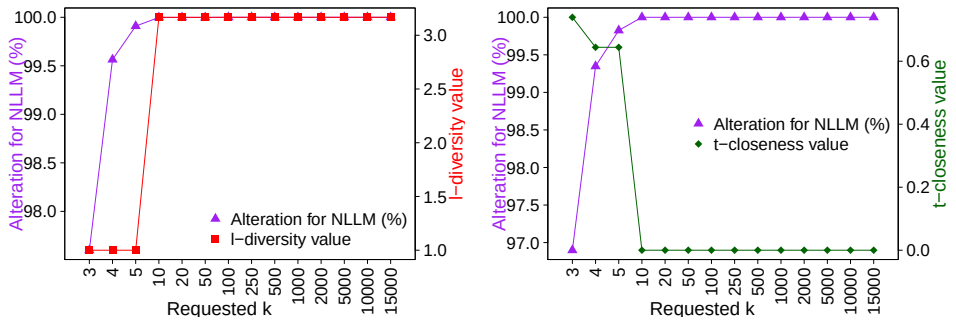
Figure 6: Optimization on sensitive attribute *Marital status* of *Adult data set*- read the right y axis for l-diversity (squares) or t-closeness (diamonds) values, read the left y axis for alteration values (triangles)

Thus, results obtained for Strategies l_{div} and t_{clo} suggest that an optimization taking into account only sensitive values distribution do not permit to maintain a reasonable alteration in k-anonymous tables. In general case, it is necessary to guide equivalence classes mergings in *GAA* considering both l-diversity value or t-closeness value and generalization cost.

Secondly, we study the performances of the seven strategies introduced in Section 4.2 when sensitive attribute considererd is an attribute of the table. Consider again the tables $Adult_{Age}$, $Adult_{Mar}$ and $florida\_30162_{Aff}$.

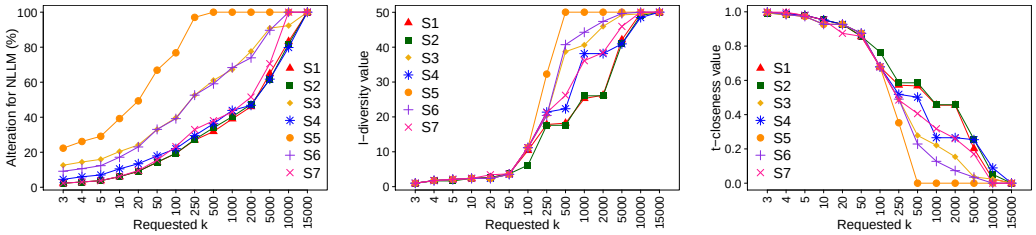
For each table, for each strategy, we apply *GAA* using the strategy on the table for each $k \in \{3, 4, 5, 10, 20, 100, 250, 500, 1000, 2000, 5000, 10\,000, 15\,000\}$. For each k-anonymous table produced, we compute its alteration for *NLLM*, its l-diversity value and its t-closeness value.

Figures 8, 9 and 10 present results for the three measures on the three tables (see Figures 11, 12 and 13 for results obtained with the *Total* metric). Each graph has seven curves corresponding to the seven strategies studied.



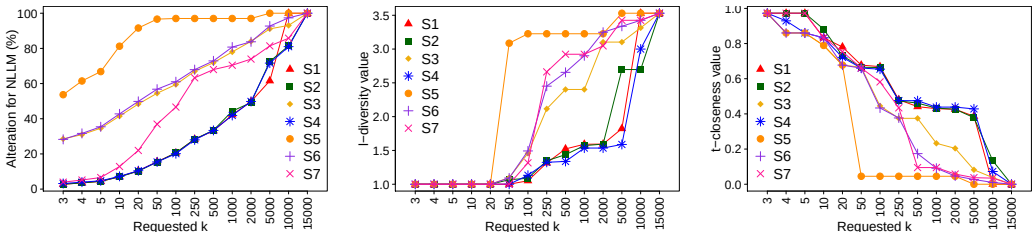
(a) Optimization of l-diversity only (b) Optimization of t-closeness only

Figure 7: Optimization on sensitive attribute *Affiliation* of *florida_30162*- read the right y axis for l-diversity (squares) or t-closeness (diamonds) values, read the left y axis for alteration values (triangles)



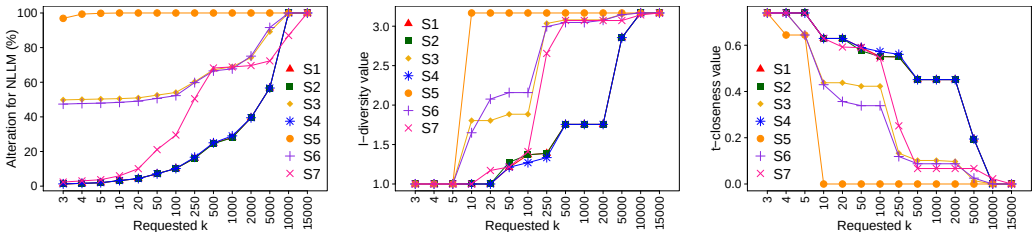
(a) Alteration for *NLLM* (b) l-diversity value (l_{div}) (c) t-closeness value (t_{clo})

Figure 8: Alteration for *NLLM*, l-diversity value and t-closeness value of k-anonymous versions of *Adult data set* with *Age* as sensitive attribute produced using the seven optimization strategies



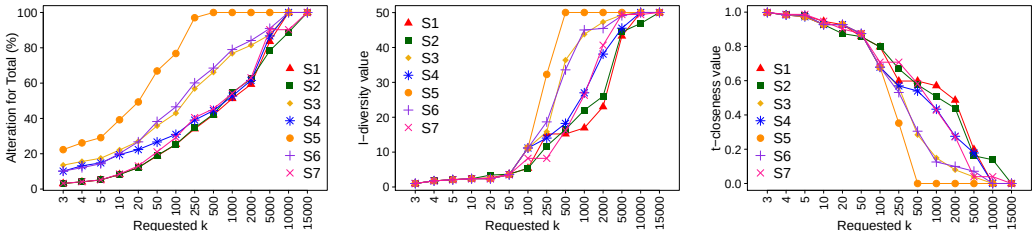
(a) Alteration for *NLLM* (b) l-diversity value (l_{div}) (c) t-closeness value (t_{clo})

Figure 9: Alteration for *NLLM*, l-diversity value and t-closeness value of k-anonymous versions of *Adult data set* with *Marital status* as sensitive attribute produced using the seven optimization strategies



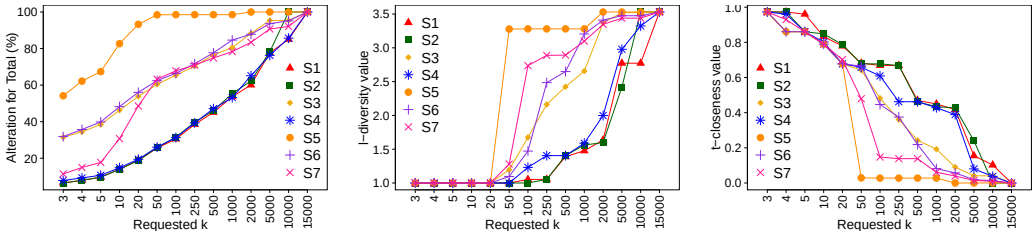
(a) Alteration for *NLLM* (b) l-diversity value (l_{div}) (c) t-closeness value (t_{clo})

Figure 10: Alteration for *NLLM*, l-diversity value and t-closeness value of k-anonymous versions of *florida_30162* with *Affiliation* as sensitive attribute produced using the seven optimization strategies



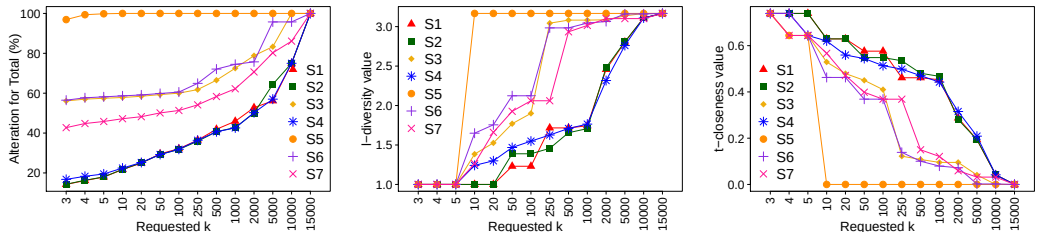
(a) Alteration for *Total* (b) l-diversity value (l_{div}) (c) t-closeness value (t_{clo})

Figure 11: Alteration for *Total*, l-diversity value and t-closeness value of k-anonymous versions of *Adult data set* with *Age* as sensitive attribute produced using the seven optimization strategies



(a) Alteration for *Total* (b) l-diversity value (l_{div}) (c) t-closeness value (t_{clo})

Figure 12: Alteration for *Total*, l-diversity value and t-closeness value of k-anonymous versions of *Adult data set* with *Marital status* as sensitive attribute produced using the seven optimization strategies



(a) Alteration for *Total* (b) l-diversity value (l_{div}) (c) t-closeness value (t_{clo})

Figure 13: Alteration for *Total*, l-diversity value and t-closeness value of k-anonymous versions of *florida_30162* with *Affiliation* as sensitive attribute produced using the seven optimization strategies

For each table, to have a representative value of the results, we compute for each strategy a mean value for each measure. To do so, we use the following formula: for a continuous function f on $[a, b] \subset \mathbb{R}$, the mean value of f on $[a, b]$ is $\frac{1}{b-a} \int_a^b f(x)dx$. For each table, for each strategy, for each measure, we compute a mean value of the results for this measure obtained by the k-anonymous versions of this table produced with *GAA* using this strategy.

Nevertheless, by computing a mean value of the results for a measure, it is not easy to compare the results obtained by the strategies on two different tables. Instead of a mean value, we associate to each strategy a percentage for each measure and for each table. For each table, for each strategy, for each measure, the percentage obtained is called *NAUC* of the strategy for the measure on the table.

Alteration is already expressed in percentage. So, for each table, for each strategy and for each measure, *NAUC* of the strategy for the measure on the table is the mean value of the curve. For t-closeness value, we multiply the mean value by 100 to obtain a percentage. For l-diversity value, we take the percentage of the mean value over the maximum l-diversity value.

Table 4 presents the *NAUC* computed on $[3, 15\,000]$ of the seven strategies for the three measures and the three tables (see Table 5 for results obtained with the *Total* metric).

We now study the results for the three measures.

For alteration for *NLLM* (cf. column Alteration for *NLLM* of Table 4), we observe similar results for the three tables. Strategies 1 (optimization on generalization cost), 2 (optimization on generalization cost then on l-diversity value) and 4 (optimization on generalization cost divided by l-diversity value) have the best *NAUC* for alteration ; Strategies 3 (optimization on l-diversity value then on generalization cost), 5 (optimization on generalization cost then on t-closeness value) and 6 (optimization on t-closeness value then on generalization cost) have the worst *NAUC* for alteration ; Strategy 7 (optimization on generalization cost multiplied by t-closeness value) has an intermediate *NAUC* for alteration. For instance, on $florida\_30162_{Aff}$, we observe in line $florida\_30162_{Aff}$ and column Alteration for *NLLM* of Table 4 that Strategies 1, 2 and 4 have a *NAUC* for alteration around 72%, Strategies 3, 5 and 6 have a *NAUC* for alteration greater than 90% and Strategy 7 has a *NAUC* for alteration around 80%.

We conclude that k-anonymous tables produced using Strategies 3, 5 and 6 are more altered on average than k-anonymous tables produced using Strategies 1, 2 and 4. This observation is confirmed by the shape of the curves in graphs of Figures 8a, 9a and 10a. Moreover, Strategy 5 has a *NAUC* for alteration greater than 99% for the three tables. This

	Alteration for *NLLM*		*l*-diversity value		*t*-closeness value	
	Strategy	*NAUC*	Strategy	*NAUC*	Strategy	*NAUC*
$Adult_{Age}$	4	69.71	5	98.56	5	1.36
	2	70.07	6	96.29	6	4.65
	1	71.19	3	94.9	3	7.15
	7	79.46	7	90.41	7	12.14
	3	88.02	4	86.65	4	17.44
	6	89.77	1	83.35	1	17.63
	5	99.6	2	82.22	2	20.13
$Adult_{Mar}$	4	72.92	5	97.85	5	1.2
	2	73.37	7	94.01	6	3.7
	1	76.58	6	93.85	7	4.95
	7	83.35	3	89.46	3	9.4
	3	90.43	1	73.97	1	20.69
	6	92.45	2	72.64	4	24.48
	5	99.28	4	66.77	2	24.99
$florida\_30162_{Aff}$	2	72.71	5	99.98	5	0.02
	1	72.72	3	98.74	6	3.06
	4	72.79	6	98.68	3	3.1
	7	80.34	7	97.43	7	4.93
	3	90.24	2	86.65	2	15.98
	6	90.89	1	86.64	1	15.98
	5	100.0	4	86.6	4	16.01

Table 4: *NAUC* for the seven strategies for alteration for *NLLM*, *l*-diversity value and *t*-closeness value computed on $[3, 15000]$ on the three tables

	Alteration for *Total*		*l*-diversity value		*t*-closeness value	
	Strategy	*NAUC*	Strategy	*NAUC*	Strategy	*NAUC*
$Adult_{Age}$	2	79.97	5	98.56	5	1.36
	7	83.9	3	95.71	3	5.18
	1	84.51	6	95.17	6	6.22
	4	85.93	7	90.61	7	11.47
	3	90.87	4	88.6	4	13.59
	6	92.42	1	82.16	1	17.88
	5	99.6	2	81.99	2	20.6
$Adult_{Mar}$	1	78.39	5	99.16	5	0.43
	4	78.96	6	96.4	7	2.45
	2	83.96	3	96.01	6	3.8
	7	90.13	7	95.98	3	6.37
	6	92.93	4	82.25	4	13.24
	3	93.26	2	78.25	2	17.12
	5	99.83	1	73.78	1	18.24
$florida\_30162_{Aff}$	4	67.55	5	99.98	5	0.02
	1	67.86	3	98.64	6	2.37
	2	69.51	6	98.62	3	3.77
	7	82.07	7	97.14	7	4.63
	3	89.54	1	88.63	1	15.09
	6	91.3	2	88.59	2	15.2
	5	100.0	4	87.76	4	15.82

Table 5: *NAUC* for the seven strategies for alteration for *Total*, *l*-diversity value and *t*-closeness value computed on $[3, 15000]$ on the three tables

strategy does not seem suitable for maintaining reasonable alteration in k-anonymous tables.

For l-diversity value (cf. column l-diversity value of Table 4) and t-closeness value (cf. column t-closeness value of Table 4), we also observe similar results on the three tables. Recall that l-diversity value is to be maximized and t-closeness value is to be minimized.

Strategy 5 has much better *NAUC* for l-diversity and t-closeness values than the other strategies. For instance, on *florida_30162*$_{Aff}$, the *NAUC* of Strategy 5 for l-diversity value is 99.98% and its *NAUC* for t-closeness value is around 0.02%. It means that k-anonymous tables produced using Strategy 5 almost all have optimal l-diversity and t-closeness values. This observation is confirmed by the shape of the curves in graphs of Figures 10b and 10c: we observe that optimal l-diversity and t-closeness values are reached from the 10-anonymous table.

Strategies 1, 2 and 4 have the worst *NAUC* for l-diversity and t-closeness values on the three tables. However, their *NAUC* for l-diversity value are greater than 66% and their *NAUC* for t-closeness value are lower than 25%.

To conlude on experiments on real data, we first studied two strategies which only take into account l-diversity value (Strategy l_{div}) or t-closeness value (Strategy t_{clo}). Results obtained on three tables showed that k-anonymous tables produced using these strategies in *GAA* have optimal l-diversity and t-closeness values for a large range of k values. However, alteration of k-anonymous tables is high even for small k values. If data were published in this way, they would be of no use. Using strategies mixing l-diversity value or t-closeness value and generalization cost is so justified.

Then, we evaluated the seven strategies presented in Section 4.2 on three tables. Results, similar for the three tables, showed that Strategies 1, 2 and 4 are the best strategies to limit alteration in k-anonymous tables. However, these strategies have the worst results for l-diversity and t-closeness values. Strategy 5 is equivalent to Strategies l_{div} and t_{clo}: l-diversity and t-closeness values are quickly optimized to the detriment of very high alteration from the first k values.

Strategies 3 and 6 do not achieve to limit alteration in k-anonymous tables but maintain good levels of l-diversity and t-closeness in the k-anonymous tables.

Strategy 7 has no significant advantage over other strategies: it obtains intermediate results for the three measures.

4.3.3 On simulated Data

In this section, we study the performances of the seven strategies when sensitive attributes are simulated values added to the table following a predeterminated distribution. We consider 42 tables as explained in Section 4.3: 21 configurations of *Adult data set* and 21 configurations of *florida_30162*.

For each configuration, for each strategy, we apply *GAA* using the strategy on the configuration for each $k \in \{3, 4, 5, 10, 20, 100, 250, 500, 1000, 2000, 5000\}$. For each k-anonymous table produced, we compute alteration for *NLLM*, l-diversity value and t-closeness value. As in experiments on real data, for each configuration, for each measure, for each strategy, we compute the *NAUC* on $[3, 5000]$ of the strategy for the measure on the configuration.

In order to study *NAUC*, for each configuration, for each measure, we make a ranking of the strategies according to their *NAUC* for the measure on the configuration. The strategy that obtained the best *NAUC* for the measure for the configuration is ranked 1 and the color associated with it is green. The strategy that obtained the worst *NAUC* for the measure for the configuration is ranked 7 and the color associated to it is red.

		Equivalent distribution							Geometric distribution							Standard Normal distribution							Mean
		5	10	20	50	100	200	500	5	10	20	50	100	200	500	5	10	20	50	100	200	500	
Strategy 1	Alteration for *NLLM*	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	51.974
	l-diversity value	6	6	6	7	4	7	6	6	6	4	5	3	4	6	6	7	7	7	5	4	6	93.829
	t-closeness value	6	7	6	6	4	7	4	7	6	7	4	7	7	6	7	7	7	7	6	6	6	9.628
Strategy 2	Alteration for *NLLM*	3	2	2	3	3	2	3	2	3	3	4	3	3	1	1	2	3	4	3	3	2	55.282
	l-diversity value	7	7	5	6	7	5	5	5	5	5	7	7	6	5	5	6	6	6	7	6	7	93.775
	t-closeness value	7	6	5	7	6	5	6	6	5	6	6	5	5	5	6	6	5	5	7	7	7	9.619
Strategy 3	Alteration for *NLLM*	5	5	5	5	5	6	5	6	6	6	6	6	6	6	6	5	6	6	5	5	5	77.286
	l-diversity value	4	2	2	2	3	2	4	4	2	2	2	2	2	3	3	2	2	2	6	2	4	94.86
	t-closeness value	3	2	2	2	3	2	5	3	3	2	2	2	2	2	3	3	3	2	3	2	4	8.031
Strategy 4	Alteration for *NLLM*	2	3	3	2	2	3	6	3	2	2	2	4	4	4	3	3	2	2	2	2	4	56.619
	l-diversity value	5	5	4	5	5	4	2	7	7	7	6	4	3	4	7	5	5	5	4	3	2	94.189
	t-closeness value	5	5	7	5	7	4	2	5	7	5	7	4	3	7	5	5	6	6	4	3	3	9.191
Strategy 5	Alteration for *NLLM*	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	99.938
	l-diversity value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	99.528
	t-closeness value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.455
Strategy 6	Alteration for *NLLM*	4	4	6	6	6	5	4	5	5	5	5	5	5	5	5	6	5	5	6	6	6	75.462
	l-diversity value	3	3	7	3	2	3	3	2	3	6	3	6	7	7	2	3	3	3	2	7	3	94.186
	t-closeness value	4	3	4	3	2	3	3	2	2	3	3	3	6	4	2	2	2	3	5	4	2	8.29
Strategy 7	Alteration for *NLLM*	6	6	4	4	4	4	2	4	4	4	3	2	2	3	4	4	4	3	4	4	3	63.345
	l-diversity value	2	4	3	4	6	6	7	3	4	3	4	5	5	2	4	4	4	4	3	5	5	94.168
	t-closeness value	2	4	3	4	5	6	7	4	4	4	5	6	4	3	4	4	4	4	2	5	5	8.887

(a) For *Adult data set*

		Equivalent distribution							Geometric distribution							Standard Normal distribution							Mean
		5	10	20	50	100	200	500	5	10	20	50	100	200	500	5	10	20	50	100	200	500	
Strategy 1	Alteration for *NLLM*	3	3	4	4	4	5	6	4	4	4	4	4	5	5	4	4	4	4	4	4	6	61.146
	l-diversity value	6	6	6	4	5	3	4	6	5	3	4	5	4	2	6	5	6	3	4	3	2	94.317
	t-closeness value	6	4	4	4	2	3	2	4	4	4	3	4	3	2	6	4	4	3	2	2	2	8.001
Strategy 2	Alteration for *NLLM*	2	1	1	3	2	1	2	2	3	2	2	1	1	1	2	2	1	1	2	2	1	50.17
	l-diversity value	7	7	7	7	6	7	7	5	6	6	6	4	6	6	5	6	7	7	7	7	7	93.7
	t-closeness value	7	7	7	7	6	7	6	7	6	5	7	5	5	7	4	6	7	7	7	7	7	9.635
Strategy 3	Alteration for *NLLM*	6	4	5	5	5	4	5	5	6	6	6	6	6	4	6	6	6	6	5	6	3	73.326
	l-diversity value	2	2	2	2	3	5	2	2	2	2	2	3	2	3	2	2	2	2	2	2	4	94.822
	t-closeness value	3	2	2	2	5	5	3	5	2	3	2	2	2	4	3	2	2	2	3	4	5	8.058
Strategy 4	Alteration for *NLLM*	1	2	2	2	1	2	3	1	2	3	3	2	2	3	1	3	2	3	1	1	4	51.005
	l-diversity value	5	5	5	5	7	4	5	7	4	5	5	6	5	5	7	7	5	6	5	6	6	93.897
	t-closeness value	5	6	6	6	7	4	5	6	5	6	5	7	6	5	7	7	6	6	5	6	6	9.349
Strategy 5	Alteration for *NLLM*	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	99.999
	l-diversity value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	99.927
	t-closeness value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.073
Strategy 6	Alteration for *NLLM*	5	5	6	6	6	6	4	6	5	5	5	5	4	6	5	5	5	5	6	5	5	71.669
	l-diversity value	4	4	3	3	2	2	3	3	7	4	7	7	3	4	3	3	3	4	3	4	3	94.224
	t-closeness value	4	5	3	3	3	2	4	2	7	2	4	6	4	3	2	3	5	4	4	5	3	8.419
Strategy 7	Alteration for *NLLM*	4	6	3	1	3	3	1	3	1	1	1	3	3	2	3	1	3	2	3	3	2	55.592
	l-diversity value	3	3	4	6	4	6	6	4	3	7	3	2	7	7	4	4	4	5	6	5	5	93.976
	t-closeness value	2	3	5	5	4	6	7	3	3	7	6	3	7	6	5	5	3	5	6	3	4	9.031

(b) For *florida_30162*

Figure 14: Strategies ranking according to their *NAUC* for alteration for *NLLM*, l-diversity value and t-closeness value on 42 tables

		Equivalent distribution							Geometric distribution							Standard Normal distribution							Mean
		5	10	20	50	100	200	500	5	10	20	50	100	200	500	5	10	20	50	100	200	500	
Strategy 1	Alteration for *Total*	3	2	2	2	2	1	3	2	2	3	1	2	1	3	2	2	2	2	2	2	2	71.088
	l-diversity value	7	7	5	7	5	7	6	5	5	5	6	3	6	6	6	6	5	6	6	5	5	94.089
	t-closeness value	5	7	6	7	4	6	5	6	6	7	6	4	7	3	6	6	6	6	4	6	5	9.143
Strategy 2	Alteration for *Total*	2	3	1	1	1	2	1	3	3	2	3	4	2	2	1	3	3	1	1	1	1	71.028
	l-diversity value	5	6	7	6	7	6	7	6	6	4	3	4	3	5	7	5	6	7	7	6	6	94.004
	t-closeness value	7	6	7	6	7	7	7	5	5	6	7	6	5	5	7	5	5	7	7	7	7	9.356
Strategy 3	Alteration for *Total*	6	5	5	5	4	3	4	6	6	6	6	6	4	4	4	5	6	5	5	3	4	83.542
	l-diversity value	4	4	2	2	2	3	3	3	4	2	2	2	4	4	4	3	2	2	2	4	3	95.011
	t-closeness value	2	4	2	3	2	3	3	3	3	3	3	2	4	4	4	4	3	2	3	5	3	7.961
Strategy 4	Alteration for *Total*	1	1	3	3	6	6	6	1	1	1	4	3	5	6	3	1	1	4	4	6	6	79.478
	l-diversity value	6	5	6	5	4	2	2	7	7	6	4	5	5	2	5	7	7	5	4	2	2	95.095
	t-closeness value	6	5	5	5	6	2	2	7	7	5	5	5	2	2	5	7	7	4	5	2	2	7.283
Strategy 5	Alteration for *Total*	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	99.939
	l-diversity value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	99.528
	t-closeness value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.455
Strategy 6	Alteration for *Total*	5	4	6	6	5	5	5	5	5	5	5	5	3	5	5	4	5	6	6	5	3	84.474
	l-diversity value	3	2	4	3	3	4	5	2	2	3	5	6	7	7	3	4	3	3	3	7	4	94.39
	t-closeness value	3	3	4	2	3	4	6	2	2	2	2	3	6	7	3	3	2	3	2	4	4	8.04
Strategy 7	Alteration for *Total*	4	6	4	4	3	4	2	4	4	4	2	1	6	1	6	6	4	3	3	4	5	79.973
	l-diversity value	2	3	3	4	6	5	4	4	3	7	7	7	2	3	2	2	4	4	5	3	7	94.421
	t-closeness value	4	2	3	4	5	5	4	4	4	4	4	7	3	6	2	2	4	5	6	3	6	8.376

(a) For *Adult data set*

		Equivalent distribution							Geometric distribution							Standard Normal distribution							Mean
		5	10	20	50	100	200	500	5	10	20	50	100	200	500	5	10	20	50	100	200	500	
Strategy 1	Alteration for *Total*	1	2	2	1	2	1	1	1	2	3	2	3	2	1	1	2	2	2	2	2	1	56.386
	l-diversity value	7	7	7	7	6	7	7	6	7	6	7	6	7	6	7	5	7	6	7	6	7	93.698
	t-closeness value	6	7	7	6	6	7	7	6	6	7	7	7	7	7	7	7	7	6	7	6	7	9.524
Strategy 2	Alteration for *Total*	2	3	4	3	1	3	2	3	3	4	3	4	4	2	2	3	4	3	4	1	3	58.893
	l-diversity value	6	6	6	6	7	6	6	7	5	7	6	7	6	5	6	6	6	7	5	7	6	93.725
	t-closeness value	4	5	5	7	7	6	6	7	5	6	5	6	5	5	6	6	6	7	6	7	4	9.358
Strategy 3	Alteration for *Total*	5	6	5	5	5	5	4	6	6	5	6	6	6	4	6	5	6	6	6	6	5	76.422
	l-diversity value	3	2	2	3	2	2	4	3	3	3	2	2	3	7	3	2	2	2	4	2	3	94.719
	t-closeness value	3	2	2	3	2	3	4	4	4	3	2	3	3	6	3	2	2	2	4	2	5	8.391
Strategy 4	Alteration for *Total*	3	1	3	4	4	4	6	2	1	2	4	2	1	6	3	4	3	1	1	4	6	61.757
	l-diversity value	5	5	3	4	4	4	2	5	6	5	5	4	4	2	4	4	3	3	2	3	2	95.745
	t-closeness value	7	6	3	5	4	5	2	5	7	5	4	4	4	2	5	5	4	4	2	4	2	7.343
Strategy 5	Alteration for *Total*	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	99.999
	l-diversity value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	99.927
	t-closeness value	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.073
Strategy 6	Alteration for *Total*	4	5	6	6	6	6	5	5	5	6	5	5	5	5	5	6	5	5	5	5	4	75.471
	l-diversity value	4	3	4	2	3	3	3	4	2	2	3	3	2	3	5	7	5	4	3	4	5	94.612
	t-closeness value	5	3	4	2	3	2	3	2	2	2	3	2	2	3	4	3	3	3	3	5	6	8.446
Strategy 7	Alteration for *Total*	6	4	1	2	3	2	3	4	4	1	1	1	3	3	4	1	1	4	3	3	2	61.431
	l-diversity value	2	4	5	5	5	5	5	2	4	4	4	5	5	4	2	3	4	5	6	5	4	94.227
	t-closeness value	2	4	6	4	5	4	5	3	3	4	6	5	6	4	2	4	5	5	5	3	3	8.775

(b) For *florida_30162*

Figure 15: Strategies ranking according to their *NAUC* for alteration for *Total*, *l*-diversity value and *t*-closeness value on 42 tables

Figure 14 presents the rankings obtained for each measure on each table: Figure 14a for *Adult data set* and Figure 14b for *florida_30162* (see Figure 15 for results obtained with the *Total* metric). Lines of each table correspond to the three measures for the seven strategies and columns correspond to number of possible values of the seven created attributes for the three distributions.

We first study the results obtained for *Adult data set* and *florida_30162* separately.

For *Adult data set*, in Figure 14a, we observe that Strategies 1, 2 and 4 are ranked first for alteration for *NLLM* on a majority of configurations of *Adult data set* (lines (Strategy i, Alteration for *NLLM*) for $i \in \{1, 2, 4\}$). However, they are less well ranked than the other strategies for the l-diversity and t-closeness values on a majority of configurations: we notice that they are often ranked 5, 6 and 7 for these two measures.

Strategies 3 and 6 are ranked last for alteration for *NLLM* on configurations of *Adult data set*. However, they are well ranked for l-diversity ant t-closeness values.

Strategy 5 is last ranked for alteration rankings on all the configurations of *Adult data set*. However, it is ranked first for l-diversity and t-closeness values rankings on all the configurations of *Adult data set* (lines (Strategy 5, l-diversity) and (Strategy 5, t-closeness) of Figure 14a).

Always in Figure 14a, we observe that Strategy 7 has intermediate results but heterogeneous for the three measures on configurations of *Adult data set*. It is often ranked 4 (in 30 rankings out of 63) but, for the rest, it is ranked as much in the first strategies as in the last.

For *florida_30162*, in Figure 14b, we observe that Strategies 2, 4 and 7 are the top-ranked strategies for alteration for *NLLM* on a majority of configurations of *florida_30162*.
Strategies 2 and 4 are the less well ranked for l-diversity and t-closeness values on a majority of configurations of *florida_30162*. Strategy 7 obtains heterogeneous results for these two measures.

As on *Adult data set*, Strategies 3 and 6 are among the best ranked strategies for l-diversity and t-closeness values but are among the worst ranked strategies for alteration for *NLLM* on a majority of configurations of *florida_30162*.

Strategy 5 is ranked first for l-diversity and t-closeness values but is ranked 7 for alteration for all the configurations of *florida_30162*.

As Strategy 7 on *Adult data set*, Strategy 1 obtains intermediate results for the three measures: it is ranked 4 for 28 rankings out of 63 on the configurations of *florida_30162*. However, with Strategy 1, generalization cost only is to be optimized during the choice of the equivalence classes mergings in *GAA*. This may be due to the fact that method used in *GAA* is heuristic and does not guarantee global optimality of the produced k-anonymous table.

To conclude on these first observations, we notice that Strategy 5 has a extreme behavior: it optimizes l-diversity and t-closeness values leading to major alteration on all the configurations of *Adult data set* and *florida_30162*. On both tables, Strategies 2 and 4 manage to limit alteration for a majority of configurations but they do not rank well in the rankings of l-diversity and t-closeness values. On the contrary, Strategies 3 and 6 favor l-diversity and t-closeness values but k-anonymous tables produced are more altered for a majority of configurations. Strategies 1 and 7 obtain different and heterogeneous results on both tables.

Secondly, for each table, for each strategy, for each measure, we compute the mean of the *NAUC* obtained by the strategies for the measure on the configurations of the table. Thus we have a representative value of the measure for the strategy on all the configurations of the table.

We add the Mean column to tables of Figures 14a and 14b. Values are rounded at 10^{-3}.

Recall that for alteration for *NLLM* and t-closeness value, *NAUC* is to be minimized: the closer a strategy has a *NAUC* for one of these two measures to 0% the better its performance for the measure is. However, for l-diversity value, *NAUC* is to be maximized: the closer a strategy has a *NAUC* for l-diversity value to 100% the better its performance for l-diversity value is.

We notice that, for both tables, mean *NAUC* for l-diversity and t-closeness values are good for all the strategies. Mean *NAUC* for l-diversity value are greater than 93% for the seven strategies: the lower mean *NAUC* for l-diversity value is 93.775% on *Adult data set* and 93.7% on *florida_30162*, both of them obtained by Strategy 2. Mean *NAUC* for t-closeness value are lower than 10% for the seven strategies: the higher mean *NAUC* for t-closeness value is 9.628% on *Adult data set* and 9.635% on *florida_30162*, obtained by Strategies 1 and 2 respectively. On an another side, mean *NAUC* for alteration are greater than 50% for all the strategies and for both tables.

These observations suggest that an increase of mean *NAUC* of a strategy for alteration brings an optimization of mean *NAUC* of the strategy for l-diversity and t-closeness values. In other words, the more the k-anonymous tables produced using this strategy in *GAA* are altered, the more the l-diversity and t-closeness values of these k-anonymous tables are optimized. Indeed, a k-anonymous table with a high alteration has usually few equivalence classes but they contain a lot of records. As l-diversity and t-closeness models deal with sensitive values distribution in each equivalence class of the table, it is easier to respect these models when sensitive values are numerous in each equivalence class.

4.4 Summary

In this section, our purpose was to propose strategies that permit to produce k-anonymous tables with interesting l-diversity and t-closeness values. Indeed, we noticed that k-anonymous tables sometimes suffer from a lack of diversity of the sensitive values in their equivalence classes. We therefore take into account the distribution of sensitive values in equivalence classes during a k-anonymization process.

In Section 4.1, we first came back to l-diversity and t-closeness, two anonymization models whose requirements relate on the distribution of sensitive values in the equivalence classes of a table. We introduced two measures that permit to evaluate the quality of a table in terms of l-diversity and t-closeness: l-diversity value and t-closeness value.

In Section 4.2, we presented seven optimization strategies to be used in *GAA* to guide equivalence classes mergings. The proposed strategies permit to optimize, at each round of the algorithm, alteration and l-diversity or t-closeness value of the table. Strategy 1 only optimizes generalization cost. Strategies 2 to 4 optimize l-diversity and generalization cost for μ. Strategies 5 to 7 optimize t-closeness and generalization cost for μ.

In Section 4.3, we experimented to compare the performance of the seven strategies on the production of k-anonymous tables. We considered two tables, *Adult data set* and *florida_30162*, and three measures to evaluate the quality of the k-anonymous tables produced, alteration for *NLLM*, l-diversity value and t-closeness value. We conducted two types of experiments according to the choice of the sensitive attribute in the table: either the sensitive attribute is present in the table (it contains real data) or the sensitive attribute is a new generated column (it contains simulated data following a predetermined distribution).

Experiments on real data showed that Strategies 1, 2 and 4 are better at limiting alteration than optimizing l-diversity and t-closeness values. On contrary, Strategies 3, 6 and 7 are better at favoring l-diversity and t-closeness values than limiting alteration. Strategy 5

behaves like a strategy optimizing only l-diversity value or t-closeness value. On an another side, experiments on simulated data suggest that the performance of the strategies are equivalent in terms of optimizing l-diversity and t-closeness values and that only the alteration results are at consider when determining the strategy for producing the best k-anonymous versions of a table.

To continue this work, we could adopt an opposite approach to that proposed in this section. Instead of trying to build a k-anonymous table by optimizing the values of l-diversity and t-closeness, we could build a table respecting the models of l-diversity or t-closeness by optimizing the value of k at each merging of equivalence classes in *GAA*.

5 Conclusion

In this article, we have sought to optimize the data utility in k-anonymous tables. We first presented the *GAA* algorithm in Section 2.2 for producing k-anonymous tables. With this algorithm, equivalence classes mergings are performed until a table respecting the desired anonymization model is obtained. The choice of mergings is made thanks to a strategy given as a parameter of the algorithm.

At first, we were interested in Section 3 in information loss metrics. These metrics permit to evaluate the quantity of information lost during the anonymization of a table by the generalization technique. As many k-anonymous versions of a table exist, it is necessary to have a way to compare them: information loss metrics fulfill this role. Our first contribution consisted in proposing a model unifying the writing of information loss metrics and simplifying their use. A metric is defined as a set of weights to put on the edges of generalization hierarchies of quasi-identifier attributes. From these weights, we constructed a costs matrix for each quasi-identifier attribute containing the costs of generalizing the nodes of the hierarchy of this attribute. Then, we compared the performance of seven information loss metrics when they are used in a k-anonymization algorithm to guide the equivalence classes mergings to be performed. For it, we conducted experiments on the tables *Adult data set* and *florida_30162*. Results showed that the metrics with best performances are not the same in the two tables studied according to three quality criteria. On *Adult data set*, *NLLM* is among the best metrics for the three criteria. On *florida_30162*, *Total* produces the best k-anonymous tables in view of the three quality criteria. Regarding the worst metrics, our experiments showed that metrics without a normalization step on the computation of the node generalization cost do not achieve to produce good quality k-anonymous tables. Indeed, metrics that do not do this normalization step, *LLM* and *WLLM*, obtain the worst results for the three quality criteria on both tables. In conclusion of this study, we nevertheless believe that the choice of the metric to use to k-anonymize a table depends in particular on the table to be k-anonymized, the requested value of k and the generalization hierarchies chosen for the quasi-identifier attributes.

Secondly, we sought to limit one of the weaknesses of k-anonymity in Section 4. In some k-anonymous tables, it may happen that a lack of diversity appears in the sensitive values of the equivalence classes. To remedy this problem, l-diversity and t-closeness models have been proposed in [22] and [19] respectively. These two anonymization models give constraints on the sensitive values distribution in the equivalence classes of the table. By relying on l-diversity and t-closeness, we therefore sought to produce k-anonymous tables preserving a good data utility while keeping control over the senstive values distribution in equivalence classes. Then, we have proposed seven strategies to be used in *GAA* with the objective of optimizing the alteration and the values of l-diversity or t-closeness of the k-

anonymous tables produced. To compare the performance of these seven strategies during a k-anonymization process, we conducted experiments on real data and on simulated data. We used three measures to evaluate the quality of the k-anonymous tables produced: alteration for *NLLM*, l-diversity value and t-closeness value. Experiments on real data showed that Strategies 1, 2 and 4 are better at limiting alteration than optimizing l-diversity and t-closeness values. On the contrary, Strategies 3, 6 and 7 are better at favoring l-diversity and t-closeness values than limiting alteration. Strategy 5 behaves like a strategy optimizing only l-diversity value or t-closeness value. Experiments on simulated data suggest that the performance of the strategies are equivalent in terms of optimizing l-diversity and t-closeness values and that only the alteration results are at consider when determining the strategy for producing the best k-anonymous versions of a table.

Whether working with information loss metrics as in Section 3 or with optimization strategies as in Section 4, we believe that choosing the best metric or strategy to use to k-anonymize a table depends on many parameters. We could in particular be interested in the characteristics of the table and of the generalization hierarchies provided. The objective would then be to propose a procedure based on this information and making it possible to determine the most appropriate metric or strategy for k-anonymizing the table while keeping the best data utility and privacy.

References

[1] Vanessa Ayala-Rivera, Patrick McDonagh, Thomas Cerqueus, Liam Murphy, et al. A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Transactions on data privacy*, 7(3):337–370, 2014.

[2] Muzammil M Baig, Jiuyong Li, Jixue Liu, and Hua Wang. Cloning for privacy protection in multiple independent data publications. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 885–894, 2011.

[3] Roberto J. Bayardo and Rakesh Agrawal. Data privacy through optimal k-anonymization. In *21st International Conference on Data Engineering (ICDE'05)*, pages 217–228, 4 2005.

[4] Michael A. Bender, Martín Farach-Colton, Giridhar Pemmasani, Steven Skiena, and Pavel Sumazin. Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms*, 57(2):75–94, 2005.

[5] Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. Efficient k-anonymization using clustering techniques. In Ramamohanarao Kotagiri, P. Radha Krishna, Mukesh Mohania, and Ekawit Nantajeewarawat, editors, *Advances in Databases: Concepts, Systems and Applications*, pages 188–200, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[6] Harold Connamacher and Julia Dobrosotskaya. On the uniformity of the approximation for r-associated stirling numbers of the second kind. *Contributions to Discrete Mathematics*, 15(3):25–42, 2020.

[7] Tore Dalenius. Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics*, 2(3):329, 1986.

[8] Josep Domingo-Ferrer and Vicenç Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195–212, 9 2005.

[9] Riccardo Dondi, Giancarlo Mauri, and Italo Zoppis. On the complexity of the l-diversity problem. In *International Symposium on Mathematical Foundations of Computer Science*, pages 266–277. Springer, 2011.

[10] Yang Du, Tian Xia, Yufei Tao, Donghui Zhang, and Feng Zhu. On multidimensional k-anonymity with local recoding generalization. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 1422–1424. IEEE, 2007.

[11] Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. Privacy in trajectory micro-data publishing: a survey. *Transactions on Data Privacy*, 13:91–149, 2020.

[12] Benjamin C.M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 1st edition, 2010.

[13] Joao M. Gonçalves, Diogo Gomes, and Rui L Aguiar. Privacy in data publishing for tailored recommendation scenarios. *Trans. Data Priv.*, 8(3):245–271, 2015.

[14] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 279–288, New York, NY, USA, 2002. ACM.

[15] Diane Lambert. Measures of disclosure risk and harm. *Journal of official statistics*, 9:313–313, 1993.

[16] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 49–60, New York, NY, USA, 2005. ACM.

[17] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25, 4 2006.

[18] Jiuyong Li, Raymond Chi-Wing Wong, Ada Wai-Chee Fu, and Jian Pei. Achieving k-anonymity by clustering in attribute hierarchical structures. *Data Warehousing and Knowledge Discovery*, pages 405–416, 2006.

[19] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[20] Hongyu Liang and Hao Yuan. On the complexity of t-closeness anonymization and related problems. In *International Conference on Database Systems for Advanced Applications*, pages 331–345. Springer, 2013.

[21] Jun-Lin Lin and Meng-Cheng Wei. An efficient clustering method for k-anonymization. In *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society*, PAIS '08, pages 46–50, New York, NY, USA, 2008. ACM.

[22] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3–es, 2007.

[23] Clémence Mauger, Gaël Le Mahec, and Gilles Dequen. Modeling and evaluation of k-anonymization metrics. In *Privacy Preserving Artificial Intelligence Workshop of AAAI 2020*, 2020.

[24] Clémence Mauger, Gaël Le Mahec, and Gilles Dequen. Multi-criteria optimization using l-diversity and t-closeness for k-anonymization. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, pages 73–88. Springer, 2020.

[25] Adam Meyerson and Ryan Williams. On the complexity of optimal k-anonymity. In *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 223–228, New York, NY, USA, 2004. ACM.

[26] Sang Ni, Mengbo Xie, and Quan Qian. Clustering based k-anonymity algorithm for privacy preservation. *Int. J. Netw. Secur.*, 19(6):1062–1071, 2017.

[27] Md Ileas Pramanik, Raymond Y.K. Lau, and Wenping Zhang. K-anonymity through the enhanced clustering method. In *2016 IEEE 13th Int. Conf. on e-Business Engineering (ICEBE)*, pages 85–91, 11 2016.

[28] Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[29] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[30] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[31] UCIrvine. Machine Learning Repository, 1987. [Online; accessed on June 2019] `https://archive.ics.uci.edu/ml/index.php`.

[32] US Government. Registered voters in the State of Florida, U.S.A. [Online; accessed on May 2020] `http://flvoters.com/`.

[33] Xiaokui Xiao and Yufei Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 139–150, 2006.

[34] Xiaokui Xiao, Ke Yi, and Yufei Tao. The hardness and approximation algorithms for l-diversity. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 135–146, 2010.

[35] Jian Xu, Wei Wang, Jian Pei, Xiaoyuan Wang, Baile Shi, and Ada Wai-Chee Fu. Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 785–790, New York, NY, USA, 2006. ACM.