Using t-closeness anonymity to control for non-discrimination

Salvatore Ruggieri

Dipartimento di Informatica, Università di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy

E-mail: ruggieri@di.unipi.it

Abstract. We investigate the relation between t-closeness, a well-known model of data anonymization against attribute disclosure, and α-protection, a model of the social discrimination hidden in data. We show that t-closeness implies $bd_f(t)$-protection, for a bound function $bd_f()$ depending on the discrimination measure $f()$ at hand. This allows us to adapt inference control methods, such as the *Mondrian* multidimensional generalization technique and the *Sabre* bucketization and redistribution framework, to the purpose of non-discrimination data protection. The parallel between the two analytical models raises intriguing issues on the interplay between data anonymization and non-discrimination research in data protection.

Keywords. Data anonymity, non-discrimination, t-closeness, α-protection.

1 Introduction

Several inference control methods have been proposed in privacy-preserving data mining for protecting micro-data from the risk of revealing confidential information, such as identities and sensitive attribute values [6, 7, 9, 15]. Ultimately, private data protection consists of data transformations, such as perturbations, generalizations, or suppressions, that achieve a measurable level of privacy, according to some formal model, such as k-anonymity [33, 34] or t-closeness [20]. The challenge here is to trade-off the achieved level of privacy with an unavoidable data utility loss.

Conceptually, a similar problem occurs in the blooming field of discrimination-aware data mining [28]. Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category. Some groups, traditionally subject to discrimination, are explicitly listed as "protected" groups by human rights laws, including women, black people, immigrant workers, minority ethnic groups, and so on. As for data privacy, the release of micro-data can be subject to discrimination threats. Consider the following examples:

- An employer may notice from public census data that the race or sex of workers act as proxy of the workers' productivity in his specific industry segment and geographical region. The employer may then use those visible traits of individuals, rather than their unobservable productivity, for driving (discriminatory) decisions in job interviews. This behavior is known as *statistical* discrimination [28].

- A data mining model used to profile applicants to a bank loan may learn from past application records some patterns of traditional prejudices that led to negative decisions for applicants belonging to a minority group. The profiles assigned by the model to new applicants may then be biased against that minority group.

Privacy preserving data publishing techniques based on group anonymization tackle similar problems, where individuals are partitioned into groups and each group must ensure some property such as k-anonymity or t-closeness. In this paper, we investigate whether data anonymization techniques for privacy protection can be adapted to sanitize a dataset of historical decisions with regard to discrimination threats before releasing the data publicly (*non-discrimination data publishing*) or before using them for training a classifier (*discrimination-free classification*). In the first example above, the employer should not be able to derive (for his industry sector and region) any signal stronger than some maximal threshold of different productivity among groups of workers of different race or sex. In the second example, the learning algorithm should not find out[1] any condition denoting past discriminatory practices against minority groups of applicants.

Using a common notation based on itemset mining, in this paper we investigate the relation between the model of α-protection for discrimination (which is parametric in a discrimination measure) [30] and the model of t-closeness for data anonymization [20]. Both approaches are based on the key idea of contrasting proportions on subsets of data. However, t-closeness considers the distribution of a sensitive attribute (e.g., proportions of diseases), while α-protection considers the joint distribution of a discriminatory attribute and a decision (e.g., proportions of denied loans for women and men). We formally prove that *t-closeness implies $bd_f(t)$-protection*, for an appropriate bound function $bd_f()$ depending on the reference discrimination measure $f()$. The converse does not hold, due to a form of the Simpson's paradox on proportions that prevents α-protection from having the generalization property of t-closeness. We exploit the implication result above to devise a generalization-based algorithm, called *dMondrian*, that is a variant of a well-known generalization approach for k-anonymity [19]; and a bucketization and redistribution algorithm, called *dSabre*, that is a variant of the state-of-the-art framework for t-closeness [5]. These data transformation techniques provide a formal guarantee on the maximum level of discrimination present in a sanitized dataset before it is released. This is a major technical advancement over existing discrimination data sanitization approaches [11, 13, 16]. We provide both theoretical and experimental analyses. On the theoretical side, by means of a polynomial reduction of t-closeness to α-protection, we conclude that the problem of finding an optimal generalization for α-protection is NP-hard. On the experimental side, we illustrate the performances of the proposed algorithms in terms of discrimination data protection, of information loss and utility, and of running time. Summarizing, the parallel highlighted between the analytical models of t-closeness and α-protection raises intriguing issues on the interplay between data anonymization and non-discrimination research.

This paper is organized as follows. Sect. 2 recalls basic notions using a common notation based on itemset mining. Sect. 3 revises and extends the model of α-protection. The implication between t-closeness and α-protection is formalized in Sect. 4. The data sanitization algorithms *dMondrian* and *dSabre* are presented in Sect. 5. Experiments are reported in Sect. 6, and related work is discussed in Sect. 7. Conclusions summarize the contributions of the paper and some open research directions. Appendix A provides a polynomial time reduction of t-closeness to α-protection.

[1] Discriminatory predictions can still be produced, however, due to a possible bias of the learning algorithm [4]. The design of *discrimination-free* algorithms is out of the scope of the paper.

2 Preliminaries

We recall notation and concepts from itemset mining [14]. They allow us to express in a common framework basic definitions of data anonymity and discrimination analysis.

Let $\mathcal{R}$ be a relational table (or, simply, a table or a dataset) with attributes $A_1, \ldots, A_N$. Tuples in the table denote individuals, and attribute values denote information about individuals. An *A-item* is a term $A = v$, where A is an attribute and $v \in dom(A)$, the domain of A. We assume that $dom(A)$ is categorial for every attribute A. Continuous domains can be accounted for by discretizing values into ranges. An item is any A-item. An *itemset* $\mathbf{X}$ is a set of items. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$. A tuple σ from $\mathcal{R}$ *supports* $\mathbf{X}$ if for every $A = v$ in $\mathbf{X}$, we have $\sigma[A] = v$, where $\sigma[A]$ is the value of the attribute A in the tuple σ. The *cover* of $\mathbf{X}$ is the set of tuples that support $\mathbf{X}$: $cover_{\mathcal{R}}(\mathbf{X}) = \{\sigma \in \mathcal{R} \mid \sigma \text{ supports } \mathbf{X}\}$. We omit the subscript $\mathcal{R}$ if it is clear from the context. The *support* of $\mathbf{X}$ is the size $|cover(\mathbf{X})|$ of its cover. The relative support of $\mathbf{X}$ is $supp(\mathbf{X}) = |cover(\mathbf{X})|/|\mathcal{R}|$. $\mathbf{X}$ is a *frequent* itemset if $supp(\mathbf{X}) \geq minsupp$, where $minsupp$ is a given threshold. $\mathbf{X}$ is *closed* if there is no $\mathbf{Y} \supset \mathbf{X}$ with $cover(\mathbf{Y}) = cover(\mathbf{X})$. A closed itemset is a representative member of the class of equivalence of itemsets with a same cover [2].

In privacy-aware data mining, attributes of a dataset are partitioned into *quasi-identifiers* (QIs) and *sensitive* attributes. Quasi-identifiers, such as *ZIP code*, *gender*, and *birth-date*, can potentially identify an individual when joined with some external knowledge. We restrict here to the case that $A_1, \ldots, A_{N-1}$ are the QIs and A_N is the only sensitive attribute. Let us introduce the notion of QI itemset.

Definition 1. A *QI itemset* $\mathbf{Q}$ is an itemset containing *one and only one* A-item for every QI attribute A, and no A-item for sensitive attributes A.

With our restriction, $\mathbf{Q}$ has the form $A_1 = v_1, \ldots, A_{N-1} = v_{N-1}$. The *q-block* (also known as the *equivalence class*) of $\mathbf{Q}$ is the cover of $\mathbf{Q}$. A q-block denotes a set of individuals sharing the same QI characteristics and possibly differing only in the value of the sensitive attribute.

In discrimination-aware data mining, attributes of a table of historical decisions are partitioned into *potentially discriminatory* (PD) attributes, such as *sex* and *race*; *potentially*[2] *non-discriminatory* (PND) attributes, such as *education* and *skills*; and *decision attributes*, such as *hired*. We restrict here to the case of only one PD attribute, say A_{N-1}, with binary values "*protected*" and "*unprotected*", and of only one decision attribute, say A_N, with binary values "+" (positive decision) and "-" (negative decision). Thus, $A_1, \ldots, A_{N-2}$ is the set of PND attributes.

Definition 2. A *PND itemset* $\mathbf{B}$ is an itemset containing *at most one* A-item for every PND attribute A, and no A-item for PD or decision attributes A.

With our restriction, $\mathbf{B}$ has the form $A_{\pi_1} = v_1, \ldots, A_{\pi_k} = v_k$ where $\pi_1, \ldots, \pi_k$ are distinct numbers from $\{1, \ldots, N-2\}$. The *context of possible discrimination* denoted by $\mathbf{B}$ is the cover of $\mathbf{B}$. A context of possible discrimination regards a set of individuals sharing some PND characteristics and possibly differing in the value of the PD and sensitive attributes. Notice that QI itemsets contain exactly *one* item for every QI attribute, whilst PND itemsets contain *at most one* item for every PND attribute.

[2] The use of PD (resp., PND) attributes in decision making does not necessarily lead to (or prevent from) discriminatory decisions [4, 30]. This motivates the adjective "potentially".

	decision		
group	-	+	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

$$RD = p_1 - p_2 \quad ED = p_1 - p$$

Figure 1: Difference-based discrimination measures.

3 Discrimination analysis

3.1 Discrimination measures

Consider a dataset of historical decisions about granting a benefit (e.g., a loan, a job, a wage increase, a school admission). A common tool for statistical analysis is provided by a 2×2, or 4-fold, contingency table, as shown in Fig. 1. Different outcomes between two groups are measured in terms of the proportion of people in each group with a specific outcome. The proportions of negative decisions for the protected-by-law group (p_1), the unprotected-by-law group (p_2) and the overall dataset (p) are considered. A general legal principle is then to consider *group proportional representation* [28] in decision outcomes as a quantitative measure of discrimination against a protected-by-law (briefly, protected) group. Group proportional representation can be measured as differences or rates of these proportions. Measures defined as differences of proportions include:

- *risk difference* (RD $= p_1 - p_2$), also known as *absolute risk reduction*, measures the difference in the proportion of negative decisions between the protected and the unprotected group;
- and *extended difference* (ED $= p_1 - p$), measures the difference in the proportion of negative decisions between the protected group and the whole population.

The terminology is borrowed from bio-statistics and epidemiological comparative studies between two dichotomous groups. Statistical tests and confidence intervals have been considered both in the legal and in the data mining literature [26, 32].

The degree of observed disproportionate burden suffered by the protected group is monotonically increasing for the two discrimination measures. Since one is interested in contexts with a larger proportion of negative decisions, or equivalently for a smaller proportion of positive decisions, for the protected group compared to the unprotected group or to the average, the values of interest for difference measures are those greater than 0. Values lower than 0 are also worth to be investigated since they measure phenomena of affirmative actions or of favoritism (see [30]). However, such cases must be analysed separately, and this is why the discrimination measures are not defined as absolute values of differences of proportions. This is a minor difference with what happens in measures of data anonimity. Finally, observe that RD $= 0$ iff ED $= 0$ iff $ad = bc$. The last condition describes a situation of *statistical parity*, with equal chances of obtaining a negative decision for the protected group, the unprotected group, and the whole population. The rankings imposed by the various existing measures on contingency tables can be dramatically different [25]. Since national legislations have adopted different measures, discrimination analysis has then to be parametric in the discrimination measure at hand.

3.2 Discrimination protection

The actual discovery of discriminatory situations and practices may be an extremely difficult task. A huge number of possible contexts may, or may not, be the theater for discrimination. To see this point, consider the case of gender discrimination in credit approval. Although an analyst may observe that no discrimination occurs in general, i.e., when considering the whole available decision records, it may turn out that it is extremely difficult for women to obtain credit for a particular purpose, e.g., in the case of car loans. Using the itemset notation, we would then be interested in checking the value of discrimination measures over the context of possible discrimination denoted by the cover of the PND itemset *purpose=car*, where the protected group is *sex=female*. However, many small or large such contexts may exist that conceal discrimination, and therefore all possible specific situations should be considered as candidates, consisting of all possible combinations of variables and variable values, i.e., in our words, of all PND itemsets. This problem has been tackled first in [30] by extracting and ranking classification rules on the basis of a discrimination measure. We restate here the analysis framework using[3] PND itemsets and measures defined over the 4-fold contingency table of their cover. Let us start introducing some notation.

Definition 3. For a PND itemset $\mathbf{B}$, we denote by $f(\mathbf{B})$ the value of a measure $f()$ over the contingency table of the set $cover(\mathbf{B})$.

The proportions and measures in Fig. 1 extend then to a generic PND itemset $\mathbf{B}$ by restricting to only the tuples in the cover of $\mathbf{B}$. Once provided with a quantitative measure of the degree of discrimination and with a threshold between "legal" and "illegal" degree, we can isolate contexts of possible discrimination where the measure is above such a threshold. Let us extend the definition of α-protection, originally introduced in [30].

Definition 4. Let $f()$ be a measure defined over a contingency table, and α a fixed threshold. A PND itemset $\mathbf{B}$ is α-protective w.r.t. $f()$ if $cover(\mathbf{B}) = \emptyset$ or $f(\mathbf{B}) \leq \alpha$. Otherwise, c is α-discriminatory.

The problem of *discrimination discovery* consists then of extracting and ranking PND itemsets that are α-discriminatory, i.e., having a non-empty cover and a measure value greater than the threshold α. Such itemsets denote contexts that need further consideration by a legal expert, e.g., for determining a legitimate justification, such as a genuine occupational requirement [28] or explainable discrimination [17]. The problem we tackle in this paper is the one of *non-discrimination data publishing*, which consists of transforming a dataset so that it contains no α-discriminatory PND itemset. Throughout the paper, we assume that PD attributes are available in the dataset under analysis. This is not the case, for instance, in the problem of *indirect discrimination discovery* [11, 30], where the analysis has to rely on background knowledge relating attributes in the dataset to unavailable PD attributes.

The original work [30] and its actual implementation [31], resort to frequent itemset mining in defining α-protection, thus restricting to PND and PD itemsets with a minimum support. Technically this means that condition $cover(\mathbf{B}) = \emptyset$ in Def. 4 is relaxed to $a < minsupp$, where a is from the contingency table in Fig. 1. Since in this paper we will show sufficient conditions for establishing α-protection, our results extend to the relaxed version as well.

[3] [30] reasons over 4-fold contingency tables of classification rules $\mathbf{A}, \mathbf{B} \rightarrow \mathbf{C}$, where $\mathbf{A}$ is a PD itemset, $\mathbf{B}$ is a PND itemset, and $\mathbf{C}$ is a decision item. We can restrict to consider only the PND itemset $\mathbf{B}$ because we assume only one PD attribute and only one decision attribute. Hence, there is only one 4-fold contingency table for a given $\mathbf{B}$. Our approach extends to the case of multiple PD and decision attributes by considering in Def. 3 the maximum value of $f()$ over all possible 4-fold contingency tables for $cover(\mathbf{B})$.

From a legal perspective, however, the relaxed version is unable to capture the illegal practice of tokenism, which we will consider in the next subsection.

The notion of α-protection extends to a whole dataset.

Definition 5. A relational table is α-protective w.r.t. a measure $f()$ if every closed PND itemset is α-protective w.r.t. $f()$.

Since the context of possible discrimination of a PND itemset corresponds to the individuals sharing the characteristics stated by the itemset, this definition amounts at checking the proportional representation principle, as measured by $f()$, for all such contexts. We can restrict to consider *closed* PND itemsets, since they are representative itemsets. The cover of a non-closed PND itemset is checked when considering the closed itemset in its class of equivalence. Moreover, since there is a unique closed PND itemset for every class of equivalence, there is no duplicate analysis of a same context of possible discrimination.

3.3 Tokenism and discrimination measures

Due to a division by zero in the discrimination measure, α-protection may be undefined. This occurs for ED when $n_1 = 0$ and for RD when $n_1 = 0$ or $n_2 = 0$. Such cases are unlikely for the whole dataset, but they can readily occur for (small) covers of PND itemsets Let us discuss here the legal interpretation of those conditions. First, consider the ED measure. When $n_1 = 0$, all individuals in the cover belong to the unprotected group. Is that a discriminatory practice? It may be so. If in the context there are $d = 999$ individuals with positive decision and $c = 1$ individuals with negative decision, then the unprotected group is favored by the absence of competitors from the protected group. The single individual with negative decision, called a *token*, may be there to create the false appearance of equality and to prevent charges of discrimination. This illegal practice is known as *reverse tokenism* (whereas tokenism [18] consists of granting a benefit to a few members of a minority group to create the false appearance of inclusiveness). A measure of the disproportion between negative decisions for the unprotected group and for the whole population is the difference $p_- - c/n_2$ between the proportion p_- of negative decision in the the overall dataset and the proportion of the unprotected group with negative decision (c/n_2).

Definition 6. We denote by p_- the fraction of tuples with negative decision in a relational table, i.e., $p_- = supp(A_N = -)$.

The ratio c/n_2 coincides with m_1/n, since $n_1 = 0$. Thus, ED can be extended as follows:

$$ED = \begin{cases} p_- - p & \text{if } n_1 = 0 \\ p_1 - p & \text{otherwise} \end{cases}$$

This looks intuitive: when the proportion a/n_1 of individuals from the protected group with negative decision is undefined (because $n_1 = 0$), we simply consider the expected proportion p_-. The above extended definition of ED, the extension of RD, and the extensions of ratio-based measures that we will introduce in Sect. 4.3 can be obtained, in a general way, by revising the definitions of p_1 and p_2 in Fig. 1 as follows:

$$p_1 = \begin{cases} p_- & \text{if } n_1 = 0 \\ a/n_1 & \text{otherwise} \end{cases} \qquad p_2 = \begin{cases} p_- & \text{if } n_2 = 0 \\ c/n_2 & \text{otherwise} \end{cases}$$

These extensions cover relevant legal contexts and will not interfere with linking α-protection to t-closeness.

4 Data anonymization and non-discrimination

Several partition-based schemes of privacy in data disclosure are defined by proof conditions over the q-blocks of a released dataset. k-anonymity [33, 34] requires that the support of any non-empty q-block is at least k. l-diversity [24] requires that the number of distinct values of the sensitive attribute in a non-empty q-block is at least l. t-closeness [20] requires that the distribution of the sensitive values in a non-empty q-block is close to the distribution in the overall dataset, according to a distance function between distributions. We concentrate on t-closeness, making the further assumption that the *sensitive attribute is binary*, with values $\star$ and $\bullet$. This assumption will be needed later on, when mapping the decision attribute (which is binary) to a sensitive attribute. Moreover, under this assumption, known distance measures between distributions collapse to variational distance. Let us recall the notion of t-closeness from [20].

Definition 7. Let $p_\star$ be the fraction of tuples in a relational table with sensitive value $\star$. A q-block is t-close if it is empty or, called p the fraction of tuples in it with sensitive value $\star$, if $|p - p_\star| \leq t$. A relational table is t-close if all of its q-blocks are t-close.

Variational distance is the average distance, in an equal-distance space, between the proportion of a sensitive value in the q-block and its proportion in the whole dataset. For a binary attribute, variational distance is: $\frac{1}{2}(|p - p_\star| + |(1-p) - (1-p_\star)|)$, which boils down to $|p - p_\star|$. The parameter t in t-closeness takes values from 0 to $max\{p_\star, 1 - p_\star\}$, which is the largest possible variational distance.

The proof conditions required by t-closeness closely resemble those of α-protection. QI itemsets and PND itemsets play similar roles, partitioning individuals/tuples into groups (q-blocks and contexts of possible discrimination) for which some bounds on the distributions of values must be satisfied. However, QI itemsets fix *all* of the values of QI attributes, whilst PND itemsets fix *some* of the values of PND attributes. This occurs because a generalization property holds for t-closeness but not for α-protection – as it will be shown in Sect. 4.1. Another analogy is that both t-closeness and α-protection impose a maximum difference between two proportions computed over a group of individuals. However, the proportions compared in t-closeness regard the distribution of the sensitive attribute only, whilst the proportions in α-protection regard the joint distribution of the PD and the decision attributes.

The proof conditions of t-closeness are stronger than those of α-protection. Because they impose that the proportion of a (sensitive) value is bounded in each q-block, one can derive bounds on the relative proportions of the value in any two given q-blocks (in particular, two q-blocks which differ only in the value of the PD attribute). This is precisely the idea exploited in Sect. 4.2 to show that t-closeness imply $bd_f(t)$-protection, for some bounding function $bd_f()$ depending on the discrimination measure $f()$.

4.1 On the generalization property

We have observed that t-closeness proof conditions are restricted to QI itemsets and need not to consider explicitly their subsets. This is a consequence of the generalization property of t-closeness [20, 21], for which generalizing two or more values of a QI attribute to a common value (or even removing the attribute from the dataset) leads to a dataset that is t'-close for some $t' \leq t$. We prove this statement using the itemset-based notation.

Lemma 8. *Consider a t-close relational table. The cover of any subset* $\mathbf{B}$ *of a QI itemset is t-close.*

dept	*sex*	*admitted*
A	female	no
A	female	no
A	female	no
A	female	no
A	female	no
A	female	yes
A	female	yes
A	male	no
A	male	yes
A	male	yes

PND itemset *dept=A*
$ED = 5/7 - 6/10 = 0.114$

dept	*sex*	*admitted*
B	female	no
B	female	yes
B	male	no
B	male	no
B	male	no
B	male	yes
B	male	yes
B	male	yes
B	male	yes
B	male	yes

PND itemset *dept=B*
$ED = 1/2 - 4/10 = 0.10$

PND itemset empty
(both departments)
$ED = 6/9 - 10/20 = 0.167$

Figure 2: Example of the Simpson's paradox.

Proof. By sensitive group, we mean the set of tuples whose sensitive attribute takes the value $\star$. Let $p_\star$ be the proportion of the sensitive group in the overall dataset (as in Def. 7). It is immediate to observe that:

$$cover(\mathbf{B}) = \bigcup_{\mathbf{B}' \in base(\mathbf{B})} cover(\mathbf{B}')$$

where $base(\mathbf{B})$ is the set of QI itemsets $\mathbf{B}'$ such that $\mathbf{B}' \supseteq \mathbf{B}$. Notice that the covers in the union are disjoint. This implies that the size s of the sensitive group in $cover(\mathbf{B})$ can be expressed as: $s = \sum_{\mathbf{B}' \in base(\mathbf{B})} s_{\mathbf{B}'}$ where $s_{\mathbf{B}'}$ is the size of the sensitive group in $cover(\mathbf{B}')$. The fraction $p(\mathbf{B})$ of the sensitive group in the cover of $\mathbf{B}$ is $(\sum_{\mathbf{B}' \in base(\mathbf{B})} s_{\mathbf{B}'})/|cover(\mathbf{B})|$, which can be written as a weighted average of proportions:

$$p(\mathbf{B}) = \frac{\sum_{\mathbf{B}' \in base(\mathbf{B})} |cover(\mathbf{B}')|\ p(\mathbf{B}')}{|cover(\mathbf{B})|}$$

where $p(\mathbf{B}')$ is the fraction of the sensitive group in the cover of $\mathbf{B}'$. This implies that $max\{p(\mathbf{B}') \mid \mathbf{B}' \in base(\mathbf{B})\} \geq p(\mathbf{B}) \geq min\{p(\mathbf{B}') \mid \mathbf{B}' \in base(\mathbf{B})\}$. Since the dataset is t-close, we have that $p(\mathbf{B}') \in [p_\star - t, p_\star + t]$ for every $\mathbf{B}' \in base(\mathbf{B})$. Therefore, $p(\mathbf{B}) \in [p_\star - t, p_\star + t]$ as well. □

The generalization property does not hold instead for α-protection, due to a form of the Simpson's paradox when comparing differences or ratios of two proportions. This motivates requiring the proof conditions of α-protection for all (closed) PND-itemsets.

Example 9. A real life example of the Simpson's paradox occurred in a legal case [3] regarding bias against women in a university admission exam. Fig. 2 shows a table with the university department, sex of applicant, and the exam outcome for a fictious set of individuas. Here, *dept* is a PND attribute, *sex* is the PD attribute, and *admitted* is the decision attribute. There are 4 applicants admitted to department A out of a total of 10: there are 2 women out of 7, and 2 men out of 3. The extended difference ED is then $5/7 - 6/10 = 0.114$.

When considering applicants to department B, ED is $1/2 - 4/10 = 0.10$. Then, for all PND itemsets with exactly one item, i.e., *dept=A* and *dept=B*, the ED measure is bounded by 0.114. However, for the empty PND itemset (denoting applicants to any department), the extended difference is greater, since it amounts at $6/9 - 10/20 = 0.167$.

Another subtle consequence of the lack of the generalization property is that α-protection has to be defined for *discrete* PND attributes only. In fact, by the same arguments of the above example, α-protection for contexts such as *age=25* and *age=26* does not necessarily generalize to α-protection for the context *age* $\in$ *[25,26]*. Thus, in order to extend the definition to continuous attributes, all PND itemsets with ranges should be explicitly checked. This is computationally expensive. Moreover, frequent itemset mining tools rarely support the extraction of itemsets with ranges. Our results will allow us to draw conclusions about α-protection of PND itemsets with ranges without having to explicitly check them.

4.2 From t-closeness to α-protection

As shown in the previous subsection, t-closeness and α-protection impose distinct proof conditions on datasets. In particular, t-closeness proof conditions are stronger than the ones of α-protection, which motivates searching for some implication result. Assume a given relational table, with fixed PND, PD and decision attributes. The problem we will investigate is: *does there exist a partition of the attributes into QIs and sensitive attributes, such that t-closeness of the table implies its α-protection for some α?* We answer affirmatively by showing that a t-close table is $bd_f(t)$-protective for an appropriate bound function $bd_f()$ that depends on the discrimination measure $f()$. An alternative way of stating such a result, is that a sufficient condition for α-protection is to show $bd_f^{-1}(\alpha)$-closeness. Recall that p_- is the fraction of the negative decision in a relational table (see Def. 6).

Theorem 10. *Fix as QIs the set of PND attributes plus the PD attribute, and as sensitive attribute the decision attribute. If the table is t-close then it is $bd_f(t)$-protective w.r.t. $f \in \{ED, RD\}$, where $bd_{RD}(t) = bd_{ED}(t) = min\{2t, t + \hat{p}_-, 1\}$ and $\hat{p}_- = min\{p_-, 1 - p_-\}$.*

Proof. Consider a PND itemset $\mathbf{B}$ with non-empty cover, and with contingency table as in Fig. 1. We will show the stronger result $|f(\mathbf{B})| \leq bd_f(t)$.

Let us start with RD. The bound $|\text{RD}(\mathbf{B})| \leq 1$ is trivial. First, we observe that $p_1, p_2 \in [p_- - t, p_- + t]$. This is immediate when $n_1 = 0$ (resp., $n_2 = 0$) due to the extended definition of p_1 (resp., p_2) provided in Sect. 3.3. When $n_1 > 0$ (resp., $n_2 > 0$), this is stated by Lemma 8 since the table is t-close. By triangle inequality, the bounds for p_1 and p_2 imply $|\text{RD}| = |p_1 - p_2| \leq 2t$. Moreover, we claim that $|\text{RD}| \leq t + p_-$. We distinguish two cases. If $|\text{RD}| = \text{RD}$ then $|\text{RD}| = p_1 - p_- + p_- - p_2 \leq p_1 - p_- + p_- \leq t + p_-$ since $p_1 \in [p_- - t, p_- + t]$. If $|\text{RD}| = -\text{RD}$ then $|\text{RD}| = p_2 - p_- + p_- - p_1 \leq p_2 - p_- + p_- \leq t + p_-$ since $p_2 \in [p_- - t, p_- + t]$. We also claim that $|\text{RD}| = |(1 - p_1) - (1 - p_2)| \leq t + (1 - p_-)$. We distinguish two cases. If $|\text{RD}| = (1 - p_1) - (1 - p_2)$ then $|\text{RD}| = (1 - p_1) - p_- + p_- - (1 - p_2) \leq p_- - p_1 + (1 - p_-) \leq t + (1 - p_-)$ since $p_1 \in [p_- - t, p_- + t]$. If $|\text{RD}| = (1 - p_2) - (1 - p_1)$ then $|\text{RD}| = (1 - p_2) - p_- + p_- - (1 - p_1) \leq p_- - p_2 + (1 - p_-) \leq t + (1 - p_-)$ since $p_2 \in [p_- - t, p_- + t]$. Summarizing, $|\text{RD}| \leq t + \hat{p}_-$, and then $|\text{RD}| \leq bd_{RD}(t)$.

Consider now ED. Since $p = m_1/n = (p_1 n_1 + p_2 n_2)/(n_1 + n_2)$ is the weighted average of p_1 and p_2, we have $min\{p_1, p_2\} \leq p \leq max\{p_1, p_2\}$, and then $|\text{ED}| = |p_1 - p| \leq |p_1 - p_2| = |\text{RD}| \leq bd_{RD}(t) = bd_{ED}(t)$. □

Basically, the bound $2t$ originates from triangle inequality. If the negative decision rates for the protected and unprotected groups cannot deviate from the average negative decision rate more than t, then they cannot deviate each other more than $2t$. The bound $t+\hat{p}_-$ is also intuitive. It derives from the observation that the extreme value of the negative decision rate for the protected (resp., unprotected) group in a q-block can be distant from p_- (resp., $1-p_-$) at most $\hat{p}_- = min\{p_-, 1-p_-\}$. Notice the symmetry: the bounds are independent from the choice of which group is protected and which one is unprotected. This is due to the fact that $p_2 - p_1 = -(p_1 - p_2)$, namely the risk difference considering one group as protected is the opposite of the risk difference considering the other group as protected. Their absolute value, which is the one bounded in the proof of the theorem, is the same.

Intuitively, Thm. 10 states that a dataset does not contain discrimination (more than a threshold $bd_f(t)$) if it is not possible to be confident (more than a threshold t) on the decision assigned to an individual by exploiting the differences in the fraction of positive and negative decisions between the protected and the unprotected groups in a privacy attack assuming as QIs the set of PND attributes plus the PD attribute. Notice that the role of an "attacker" here is played by the anti-discrimination analyst, whose objective is to unveil from data a context where negative decisions are biased against the protected group.

The upper bound $bd_f(t)$ of Thm. 10 is sharp.

Example 11. Consider a dataset with uniform distribution of the decision attribute, i.e., $p_- = \hat{p}_- = 0.5$. Following Thm. 10, we fix such an attribute as sensitive. The dataset is clearly t-close for $t = 0.5$. Consider now two PND itemsets with the following contingency tables:

	decision		
group	-	+	
protected	1	0	1
unprotected	0	1	1
	1	1	2

	decision		
group	-	+	
protected	1	0	1
unprotected	0	d	d
	1	d	$d+1$

For the left table, RD $= 1 - 0 = 1 = bd_{RD}(t) = min\{2 \cdot 0.5, 0.5 + 0.5, 1\}$. For the right table, ED $= 1 - 1/(d+1)$ is arbitrarily close to the bound $bd_{ED}(t) = bd_{RD}(t) = 1$ for a sufficiently large d. Later on, in Sect. 6 we will experiment on real datasets, and we will show that bounds are typically reached also in practice.

The converse of Thm. 10 does not hold in general. A counter-example is provided by the Simpson's paradox table from Ex. 9.

Example 12. Reconsider the table in Fig. 2. The distribution of the decision attribute in the overall dataset is uniform: 10 yes and 10 no, hence $p_- = \hat{p}_- = 0.5$. From Ex. 9, we know that the dataset is 0.167-protective w.r.t. the ED measure, since the maximal value of ED over PND itemsets is 0.167. Fix now as QIs the attributes *dept* and *sex*, and as sensitive the attribute *decision*. The dataset is not 0.0835-close, where $0.0835 = bd^{-1}_{ED}(0.167) = max\{0.167/2, 0.167 - 0.5\}$. In fact, the q-block of the QI itemset *dept=A, sex=female* includes 5 no and 2 yes, with a variational distance of $|5/7 - 1/2| = 0.214$. As shown by the following calculations, the dataset is t-close only for $t \geq 0.214$:

dept=A, sex=female	*dept=A, sex=male*	*dept=B, sex=female*	*dept=B, sex=male*
$\vert 5/7 - 1/2 \vert = 0.214$	$\vert 1/3 - 1/2 \vert = 0.167$	$\vert 1/2 - 1/2 \vert = 0.0$	$\vert 3/8 - 1/2 \vert = 0.125$

Let us now consider range items or, more generally, disjunctive items.

Definition 13. A *disjunctive item* is a term $A = v_1 \vee \ldots \vee A = v_n$, where A is an attribute and $\{v_1, \ldots, v_n\} \subseteq dom(A)$. A tuple σ *supports* that disjunctive-item if $\sigma[A] \in \{v_1, \ldots, v_n\}$.

Range items $A \in [v_1, v_2]$ for a continuous attribute A are a special case of disjunctive items of the form $A = v_1 \vee A = v_1 + 1 \vee \ldots \vee A = v_2$. We say that an itemset includes disjuntive items if it consists of a subset of items and disjunctive items. The notions of cover and support of an itemset clearly extend in presence of disjunctive items. As observed after Ex. 9, α-protection of a dataset does not imply that a PND itemset with disjunctive items is α-protective. Thm. 10 allows us to draw conclusions on α-protection of PND itemsets with disjunctive items at no additional cost.

Corollary 14. *Fix as QIs the set of PND attributes plus the PD attribute, and as sensitive attribute the decision attribute. If the table is t-close then every PND itemset, possibly with disjunctive items, is $bd_f(t)$-protective w.r.t. $f \in \{ED, RD\}$ and $bf_f()$ as in Thm. 10.*

Proof. In the proof of Thm. 10, it is exploited the property that a PND itemset $\mathbf{B}$ is t-close. Such a property can be shown also for $\mathbf{B}$ including disjunctive items by a simple generalization of the proof of Lemma 8, where $base(\mathbf{B})$ is now defined as the set of QI itemsets $\mathbf{B}'$ such that every item $A = v$ in it appears in $\mathbf{B}$ as a single item or in a disjunctive item. □

As an example, if a dataset including the *age* attribute is 0.1-close, then a PND itemset with disjunctive items, such as *age* ∈ *[25, 26]*, is 0.2-protective w.r.t. RD, where $0.2 \geq bd_{RD}(0.1)$.

4.3 Ratio-based discrimination measures

Several ratio-based discrimination measures have been defined over the 4-fold contingency table of Fig. 1 (see [28]), including:

$$RR = \frac{p_1}{p_2} \quad RC = \frac{1-p_1}{1-p_2} \quad OR = \frac{p_1(1-p_2)}{(1-p_1)p_2} = \frac{a/b}{c/d} \quad ER = \frac{p_1}{p} \quad EC = \frac{1-p_1}{1-p}$$

risk ratio or *relative risk* (RR), defined as the ratio of the proportions of negative decisions; *relative chance* or *selection rate* (RC), defined as the ratio of the proportions of positive decisions; and *odds ratio* (OR), defined as the ratio of the odd of the protected group ($p_1/(1-p_1) = a/b$) over the odd of the unprotected group ($p_2/(1-p_2) = c/d$). The variants of RR and RC when the protected group is compared to the average proportion p of negative decisions, rather than to the proportion for the unprotected group, include the *extended ratio* or *extended lift* (ER), and the *extended chance* (EC). All ratio measures extend to deal with (reverse) tokenism by the revised definitions of p_1 and p_2 provided in Sect. 3.3.

The degree of observed disproportionate burden suffered by the protected group is monotonically increasing for RR, OR, and ER, and monotonically decreasing for RC and EC (for these measures, the inequality in Def. 4 of α-protection is replaced by $f(\mathbf{B}) \geq \alpha$). All ratio measures boil down to 1 in the case of statistical parity, namely when $ad = bc$. However, they rank contexts of possible discrimination in a different order (see [25]).

By exploiting the lower and upper bounds imposed on proportions p_1 and p_2 by t-closeness, upper bounds on α-protection are readily derived for ratio-based measures as well.

Theorem 15. *Fix as QIs the set of PND attributes plus the PD attribute, and as sensitive attribute the decision attribute. If the table is t-close then it is $bd_f(t)$-protective w.r.t. f, where:*

- $bd_{RR}(t) = bd_{ER}(t) = min\{p_- + t, 1\}/max\{p_- - t, 0\}$,
- $bd_{RC}(t) = bd_{EC}(t) = max\{1 - p_- - t, 0\}/min\{1 - p_- + t, 1\}$,
- $bd_{OR}(t) = min\{p_- + t, 1\}min\{1 - p_- + t, 1\}/(max\{p_- - t, 0\}max\{1 - p_- - t, 0\})$.

Proof. Following the reasoning of Thm. 10, the bounds of p_1, p_2 can be further refined to $p_1, p_2 \in [max\{p_- - t, 0\}, min\{p_- + t, 1\}]$ by observing that $p_1, p_2 \in [0, 1]$. This implies $(1 - p_1), (1 - p_2) \in [max\{1 - p_- - t, 0\}, min\{1 - p_- + t, 1\}]$. The upper bounds on RR, ER and OR, and the lower bounds on RC and EC follow immediately. □

Notice the symmetry again: the bound for RC is equal to the inverse of the bound for RR but considering "+" as the negative decision. This is due to the fact that risk chance amounts at the risk ratio when considering "+" as the negative decision, but in the former case one is interested in lower bounds and in the latter case in upper bounds. Also, notice that $bd_{OR}(t) = bd_{RR}(t)/bd_{RC}(t)$ reflects the fact that OR = RR / RC.

5 Discrimination data protection algorithms

As an application of Thms. 10 and 15, we can resort to inference control methods for t-closeness to the purpose of controlling the degree of α-protection in a dataset. This provides us with a means to prevent discrimination inference attacks, such as in the examples from the introduction. In the α-protection model, these attacks consists of inferring contexts of possible discrimination with a discrimination measure greater than a given threshold α.

Observe that t-closeness is a single sufficient condition for establishing α-protection w.r.t. a wide range of discrimination measures. That is, discrimination protection does not need specialized methods for sanitizing a dataset w.r.t. different measures. Rather, we only have to select the value $t_\alpha = bd_f^{-1}(\alpha)$ such that a dataset sanitized for t_α-closeness is guarranteed to be α-protective.

In this section, we devise two algorithms for discrimination data protection inspired by the observation above. We consider non-perturbation methods which rely on partial reductions in details of data by partitioning a dataset into classes of equivalence. In particular, *generalization* (also called *recoding*) maps domain values to less specific values, according to a user defined hierarchy for categorial domains, or by grouping values into ranges in an ordered or continuous domain. We adapt two well-known generalization algorithms: *Mondrian*, originally designed for k-anonymity; and *Sabre*, which is the state-of-the-art method for t-closeness. The overall approach consists of *mimicking the algorithms of data anonimization except in the part where values of the PD attribute may be generalized*. The motivation for this is that collapsing protected and unprotected individuals into a single group, would make 4-fold contingency tables and, *a fortiori*, discrimination measures undefined.

A natural question arises about whether optimal generalization methods exists in data sanitization for discrimination protection, where "optimal" means that the number of generalizations introduced by the method is minimal. The answer is negative. In Appendix A, we deal with the opposite problem tackled so far, namely with showing t-closeness using α-protection. We present a polynomial time reduction under the general assumption that the sensitive attribute is multi-valued. This is an interesting result on its own. α-protection is at least as difficult as t-closeness even in the (apparently) simple framework of binary PD and decision attributes. NP-harness of generalization methods for discrimination protection follow then from analogous results for t-closeness.

Theorem 16. *The problem of finding an α-protective partition of a dataset with the minimum number of generalizations is NP-hard.*

Proof. See Appendix A. □

We are then forced to look at heuristics algorithms for discrimination protection.

Algorithm 1 dMondrian.Anonymize($\mathcal{P}$, t)

```
if no d-allowable cut for $\mathcal{P}$ then
    return PND_ranges($\mathcal{P}$)
else
    $A \leftarrow$ choose_PND_dimension($\mathcal{P}$)
    $v \leftarrow$ find_median($\mathcal{P}$, $A$)
    $\mathcal{P}_1 \leftarrow \{\sigma \in \mathcal{P} \mid \sigma[A] \leq v\}$
    $\mathcal{P}_2 \leftarrow \{\sigma \in \mathcal{P} \mid \sigma[A] > v\}$
    return Anonymize($\mathcal{P}_1$, $t$) $\cup$ Anonymize($\mathcal{P}_2$, $t$)
end if
```

5.1 From Mondrian to dMondrian

A well-known multidimensional recoding model for k-anonymity was proposed in [19], together with an intuitive greedy algorithm called *Mondrian*. Alg. 1 reports an adaption of the algorithm, called *dMondrian*, for data protection against discrimination inference. *dMondrian* follows a divide & conquer pattern common to space partition algorithms, such as in *kd*-tree construction and in decision tree induction. Starting from a set of tuples $\mathcal{P}$ (initially the whole table to be sanitized), the procedure computes, if it exists, an axis-parallel *cut* along a PND attribute that partitions $\mathcal{P}$ into subsets $\mathcal{P}_1$ and $\mathcal{P}_2$ on which the procedure is recursively applied. If no such cut exists for $\mathcal{P}$, then the values of every PND attribute are replaced by the range "$[min, max]$" of such an attribute in $\mathcal{P}$. This substitution is performed by the *PND_ranges()* function in Alg. 1. Differently from what would occur by a direct application of the *Mondrian* algorithm, we prevent cutting on the PD attribute, which in Thms. 10 and 15 plays the role of a QI. Moreover, notice that the *PND_ranges()* function changes only the values of PND attributes. The combined effect is that PD values are always left unchanged by *dMondrian*.

Let us now discuss how cuts are defined, starting from their definition in the original *Mondrian* algorithm. We assume that the domain $dom(A)$ of any PND attribute A is ordered. This is immediate for continuous attributes, while it requires an additional input from the user for categorial attributes. A (multidimensional) cut $A \leq v$ is *allowable* [19] if it partitions a k-anonymous set of tuples into two sets (respectively, tuples σ such that $\sigma[A] \leq v$, and tuples σ such that $\sigma[A] > v$) that are both k-anonymous. Notice that a cut $A \leq v$ is allowable iff the cut $A \leq v_m$ is allowable, where v_m is the median value of A in the set $\mathcal{P}$. Since the median value leads to the most balanced partitions, the *Mondrian* algorithm adopts median-partitioning. Its extension to t-closeness, called *tMondrian* [20], simply requires that each partition resulting from a cut is t-close, instead of (or in addition to being) k-anonymous. We extend the notion of allowable cuts to non-discrimination protection by introducing d-allowable (for "discrimination allowable") cuts, which check the t-closeness proof condition (where the sensitive attribute is now the decision attribute) for the subsets of the protected and unprotected groups in the resulting partitions. The need for checking *two* subsets is motivated by the fact that generalizations over the PD attribute are not permitted, hence cuts must explicitly check t-closeness over the two PD attribute values. Let us first introduce a useful notation.

Definition 17. The *maximum variational distance* of a contingency table (see Fig. 1) is:

$$\tau = max\{|p_1 - p_-|, |p_2 - p_-|\}$$

$\tau(\mathcal{R})$ denotes the maximum variational distance of the contingency table of the dataset $\mathcal{R}$.

Sample dataset

ID	purpose	emp	sex	decision
1	housing	no	female	-
2	housing	no	female	-
3	housing	no	female	+
4	housing	no	male	-
5	housing	no	male	+
6	housing	yes	female	-
7	housing	yes	female	+
8	housing	yes	female	+
9	housing	yes	male	-
10	housing	yes	male	-
11	housing	yes	male	+
12	housing	yes	male	+
13	car	no	female	+
14	car	no	male	-
15	car	no	male	+
16	car	yes	female	-
17	car	yes	male	+

Output of dMondrian

ID	purpose	emp	sex	decision
1	housing-car	no	female	-
2	housing-car	no	female	-
3	housing-car	no	female	+
13	housing-car	no	female	+
4	housing-car	no	male	-
14	housing-car	no	male	-
5	housing-car	no	male	+
15	housing-car	no	male	+
6	housing-car	yes	female	-
16	housing-car	yes	female	-
7	housing-car	yes	female	+
8	housing-car	yes	female	+
9	housing-car	yes	male	-
10	housing-car	yes	male	-
11	housing-car	yes	male	+
12	housing-car	yes	male	+
17	housing-car	yes	male	+

Figure 3: Sample dataset and *dMondrian* output for $t = 0.25$.

We are now in the position to introduce d-allowable cuts.

Definition 18. Let $\mathcal{P}$ be a set of tuples. A cut $A \leq v$ is *d-allowable* for $\mathcal{P}$, where A is a PND attribute and $v \in dom(V)$, if for $\mathcal{P}_1 = \{\sigma \in \mathcal{P} \mid \sigma[A] \leq v\} \neq \emptyset$ and $\mathcal{P}_2 = \{\sigma \in \mathcal{P} \mid \sigma[A] > v\} \neq \emptyset$, it turns out $\tau(\mathcal{P}_1) \leq t$ and $\tau(\mathcal{P}_2) \leq t$.

By replacing conditions $\mathcal{P}_1, \mathcal{P}_2 \neq \emptyset$ with $|\mathcal{P}_1|, |\mathcal{P}_2| \geq k$, we add the requirement of k-anonymity for the dataset in output. Differently from k-anonymity, however, if $A \leq v$ is d-allowable, then $A \leq v_m$ is not necessarily d-allowable. However, we keep using the heuristics of *Mondrian* of testing cuts only at median values (see *find_median()* in Alg. 1) because it has two main advantages. First, *dMondrian* can be used to control both α-protection and k-anonymity at the same time. Second, the search space of the algorithm is kept reasonably low. In fact, it consists of a search tree space with $O(D)$ levels, where $D = \sum_{i=1,\ldots,N-2} log\,|dom(A_i)|$ is the sum of the logarithm of domain sizes of PND attributes. Notice that the actual number of levels depends on the parameter t, which prunes the search space. At each level, a scan of the tuples in the input dataset $\mathcal{R}$ is performed to compute d-allowable cuts, which requires $O(ND|\mathcal{R}|)$, where N is the number of attributes. Summarizing, the computational complexity of *dMondrian* is $O(ND^2|\mathcal{R}|)$.

Finally, when more than one PND attribute has a d-allowable cut, the function *choose_PND_dimension()* implements the heuristics of choosing the cut that produces the lowest maximal variational distance, i.e., such that $max\{\tau(\mathcal{P}_1), \tau(\mathcal{P}_2)\}$ in Def. 18 is minimal.

Example 19. The sample dataset in Fig. 3 regards past decisions on granting a loan to applicants based on the purpose of the loan, on whether the applicant is employed, and on the sex of the applicant. The proportion of negative decisions is $p_- = 8/17 = 0.47$. Fix $t = 0.25$. The following contingency tables:

purpose=housing	decision		
sex	-	+	
female	3	3	6
male	3	3	6
	6	6	12

purpose=car	decision		
sex	-	+	
female	1	1	2
male	1	2	3
	2	3	5

emp=no	decision		
sex	-	+	
female	2	2	4
male	2	2	4
	4	4	8

emp=yes	decision		
sex	-	+	
female	2	2	4
male	2	3	5
	4	5	9

allow for computing the maximum variational distances for the cuts w.r.t. the *purpose* and *emp* attributes. It turns out: $\tau(\textit{purpose=housing}) = max\{|0.5 - 0.47|, |0.5 - 0.47|\} = 0.03$; $\tau(\textit{purpose=car}) = max\{|0.5-0.47|, |0.33-0.47|\} = 0.14$; $\tau(\textit{emp=no}) = max\{|0.5-0.47|, |0.5-0.47|\} = 0.03$; $\tau(\textit{emp=yes}) = max\{|0.5 - 0.47|, |0.4 - 0.47|\} = 0.07$. Both cuts on *purpose* and on *emp* are d-allowable. *dMondrian* selects the cut on *emp* because:

$$0.07 = max\{\tau(\textit{emp=no}), \tau(\textit{emp=yes})\} < max\{\tau(\textit{purpose=housing}), \tau(\textit{purpose=car})\} = 0.14$$

The search tree space of *dMondrian* consists of this single cut. In fact, for both branches *emp=no* and *emp=yes*, there is no d-allowable cut on *purpose*. E.g., the contingency table:

emp=no			
purpose=car	decision		
sex	-	+	
female	0	1	1
male	1	1	2
	1	2	3

shows that $\tau(\textit{emp=no,purpose=car}) = max\{|0 - 0.47|, |0.5 - 0.47|\} = 0.47 > t = 0.25$. In both branches, the *purpose* attribute takes values *housing* and *car*. They are generalized to the range *housing-car*. The dataset in output by *dMondrian* is shown in Fig. 3. The horizontal line separates the outputs of the two branches of the search tree.

Summarizing, for an input table $\mathcal{R}$, *dMondrian.Anonymize(*$\mathcal{R}, t$*)* returns a t-close dataset, hence $bd_f(t)$-protective, obtained by generalizing PND attributes. Actually, in the limit case that there is no d-allowable cut for the whole $\mathcal{R}$, the procedure returns *PND_ranges(*$\mathcal{R}$*)*, which is $\tau(\mathcal{R})$-close, because cuts involving the PD attribute are not permitted. Therefore, for $t \leq \tau(\mathcal{R})$ we can only conclude that *dMondrian.Anonymize(*$\mathcal{R}, t$*)* is $bd_f(\tau(\mathcal{R}))$-protective.

Theorem 20. *dMondrian.Anonymize(*$\mathcal{R}, t$*) is* $bd_f(max\{t, \tau(\mathcal{R})\})$*-protective w.r.t.* $f()$.

Example 21. In Sect. 6, we will experiment on the *German credit* dataset. It consists of 1000 tuples of applicants to a loan, with the sex of applicants as the PD attribute. There are 300 negative decisions, hence $p_- = 0.3$. Restricting to female applicants, there are 109 negative decisions out of 310 applicants, hence $p_1 = 0.352$. For males, there are 191 negative decisions out of 690 applicants, hence $p_2 = 0.277$. Any call to the *Anonymize()* procedure leaves unchanged the PD attribute. Hence, the itemset *sex=female* is 0.052-close (where $0.052 = |p_1 - p_-|$), and the itemset *sex=male* is 0.023-close (where $0.023 = |p_2 - p_-|$). For $t < 0.052 = \tau(\mathcal{R})$, the only way to obtain a t-close output would be to generalize all attributes, including the PD one. While this is reasonable in the case of t-closeness, and it can occur in *Mondrian*, it is not for α-protection, because it would make discrimination measures undefined. As a consequence, the empty PND itemset has a RD value of $p_1 - p_2 = 0.075$. There is no way of generalizing the input dataset to obtain a value of the RD measure lower than 0.075. This is an intrinsic limitation of non-perturbation methods.

Algorithm 2 dSabre.Anonymize($\mathcal{P}$, t)

```
$\mathcal{A} \leftarrow cover(A_{N-1} = protected, A_N = -)$
$\mathcal{B} \leftarrow cover(A_{N-1} = protected, A_N = +)$
$\mathcal{C} \leftarrow cover(A_{N-1} = unprotected, A_N = -)$
$\mathcal{D} \leftarrow cover(A_{N-1} = unprotected, A_N = +)$
$\phi \leftarrow (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$
return Dichotomize(($|\mathcal{A}|$, $|\mathcal{B}|$, $|\mathcal{C}|$, $|\mathcal{D}|$), $t$)
```

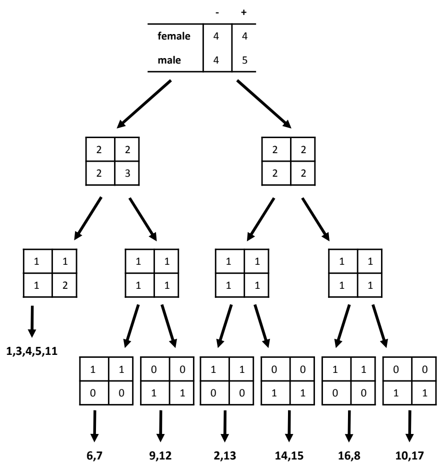
Algorithm 3 dSabre.Dichotomize((a, b, c, d), t)

```
1:  if $a + b + c + d < 2$ then
2:      $\mathcal{G} \leftarrow$ takeOut($\phi$, $(a, b, c, d)$)
3:      return PND_ranges($\mathcal{G}$)
4:  else
5:      $a_1 \leftarrow$ floor($a/2$), $a_2 \leftarrow a - a_1$
6:      $b_1 \leftarrow$ floor($b/2$), $b_2 \leftarrow b - b_1$
7:      $c_1 \leftarrow$ ceil($c/2$), $c_2 \leftarrow c - c_1$
8:      $d_1 \leftarrow$ ceil($d/2$), $d_2 \leftarrow d - d_1$
9:      if $\tau(a_1, b_1, c_1, d_1) \leq t$ and $\tau(a_2, b_2, c_2, d_2) \leq t$ then
10:         return Dichotomize($(a_1, b_1, c_1, d_1)$, $t$) $\cup$ Dichotomize($(a_2, b_2, c_2, d_2)$, $t$)
11:     else
12:         $\mathcal{G} \leftarrow$ takeOut($\phi$, $(a, b, c, d)$)
13:         return PND_ranges($\mathcal{G}$)
14:     end if
15: end if
```

5.2 From Sabre to dSabre

Sabre [5] (Sensitive Attribute Bucketization and REdistribution framework) is the state-of-the-art algorithm for data sanitization with regard to the t-closeness model. We adapt the algorithm to deal with discrimination data protection by proposing *dSabre* in Algs. 2–3. As in *dMondrian*, the core algorithm of *dSabre* is a recursive procedure *Dichotomize*, shown as Alg. 3, which generates equivalence classes. In *dMondrian*, the split of a node in the search tree is "guided" by the dataset to be sanitized. Specifically, d-allowable cuts depend on the distribution of values in the set of tuples at current node. However, tuples may be unevenly distributed to child nodes, even though the median partitioning heuristics tries and do so.

In *dSabre*, each node is associated with a contingency table, with the root having the contingency table of the whole dataset. Then a node is split by evenly distributing the entries of its contingency table between the left and the right child (lines 5-8 in Alg. 3). The split is simply based on numbers, not on attribute values. Nevertheless, it can be checked whether it is allowable because the proof conditions of Def. 18 are actually stated on numbers from contingency tables. The notations $\tau(a_1, b_1, c_1, d_1)$ and $\tau(a_2, b_2, c_2, d_2)$ at line 9 denote the maximum variational distance of the contingency tables of the two child nodes. An allowable split ensures that there are tuples in the dataset that can be selected according to the entries of the contingency tables of the child nodes. If the split is allowable, the procedure is recursively called on its child nodes (line 10). If the split is not allowable (lines 12-13) or the input contingency table cannot be split into non-empty ones (line 1-3), then an equivalence class has to be generated. How? A contingency table (a, b, c, d) states that a tuples have to be chosen from the protected group with negative decision, b from the protected group with positive decision, c from the unprotected group with negative decision, and d from



Output of dSabre

ID	*purpose*	*emp*	*sex*	*decision*
1	housing	no-yes	female	-
3	housing	no-yes	female	+
4	housing	no-yes	male	-
5	housing	no-yes	male	+
11	housing	no-yes	male	+
6	housing	yes	female	-
7	housing	yes	female	+
9	housing	yes	male	-
12	housing	yes	male	+
2	housing-car	no	female	-
13	housing-car	no	female	+
14	car	no	male	-
15	car	no	male	+
16	housing-car	yes	female	-
8	housing-car	yes	female	+
10	housing-car	yes	male	-
17	housing-car	yes	male	+

Figure 4: *dSabre* search tree space and output for the dataset in Fig. 3 and $t = 0.25$.

the unprotected group with positive decision. Such four buckets are pre-selected at lines 1-4 of Alg. 2, and stored in the variable ϕ. The generation of an equivalence class then takes out from each bucket an appropriate number of tuples. The resulting set of tuples $\mathcal{G}$ (see lines 2 and 12 of Alg. 3) is generalized to an equivalence class through the *PND_ranges()* procedure, which replaces PND values by their range of variation. In order to have such ranges as small as possible, a nearest-neighbor approach is adopted. The first element to be inserted in $\mathcal{G}$ is chosen randomly from the bucket[4] with the lowest ratio of its current size divided by the number of elements to be selected from it. The other elements are selected from each bucket by looking at the nearest-neighbors of the first element in such a bucket. We adopt Euclidian distance between tuples based on the normalized[5] rank order of the values of attributes. This is motivated by the fact that ranges in output are intended to minimize the number of distinct values generalized. Finally, by adding to the test condition at line 1 of Alg. 3 "$a_1 + b_1 + c_1 + d_1 \geq k$ and $a_2 + b_2 + c_2 + d_2 \geq k$", we achieve, in addition, k-anonymity of the dataset in output.

Example 22. Reconsider Ex. 19. Recall that $p_- = 0.47$ and $t = 0.25$. The search tree space of *dSabre* is shown in Fig. 4 (left). The top contingency table is the one for the whole dataset. It is split between the left and the right branches by halving numbers, as described in Alg. 3, as far as the maximum variational distance is lower than or equal to t. E.g., for the leftmost leaf in Fig. 4, we have $\tau(1,1,1,2) = max\{|0.5 - 0.47|, |0.33 - 0.47|\} = 0.14 \leq t$. Its contingency table cannot be further split, since by halving the numbers, we get $\tau(1,1,0,1) = max\{|0.5 - 0.47|, |0 - 0.47|\} = 0.47 > t$. For the other leaves, we have $\tau(1,1,0,0) = \tau(0,0,1,1) = max\{|0.5 - 0.47|, |0.47 - 0.47|\} = 0.03 \leq t$ by recalling that p_1 is set to p_- (resp., $p_2 = p_-$) when $n_1 = 0$ (resp., $n_2 = 0$) – see Sect. 3.3. For contingency tables in the leaves of the search tree space, *dSabre* selects tuples in the buckets in ϕ. E.g., for the leftmost leaf in Fig. 4, the first tuple is randomly selected in the bucket of females with negative decision, say the one with ID$= 1$. The other four tuples are its nearest-neighbors,

[4] This heuristics has better experimental performances than the original *Sabre* approach of a random selection of the bucket from which to pick up the first element.

[5] A value with rank $r \in [1, dom(A)]$ in an ordered attribute A, has normalized rank $(r-1)/(|dom(A)|-1)$. E.g., the normalized ranks of values 1, 5 and 100, are $(1-1)/2 = 0$, $(2-1)/2 = 0.5$ and $(3-1)/2 = 1$.

		PND atts			t min-max	
name	tuples	discrete	continuous	p_-	$\tau(\mathcal{R})$	$\hat{p}_-$
German credit	1000	6	1	0.3	0.052	0.7
Adult	48842	6	1	0.761	0.087	0.761
Census-Income	299285	9	3	0.938	0.028	0.938

Table 1: Summary information on experimental dataset. $\tau(\mathcal{R})$ is the maximum variational distance of the whole dataset. $\hat{p}_- = max\{p_-, 1 - p_-\}$.

with respect to the PND attributes, from the other bucktes. Tuples with IDs 3, 4, and 5 have its same PND attribute values. There is no further tuple with the same PND values in the bucket of male with positive decisions. Thus, the fifth tuple is the one with ID$=$ 11, which differs in the value of *emp*. All the selected tuples assume instead the value *purpose=housing*. The equivalence class of that contingency table and of the others in the search tree space are outputted by *dSabre* as shown in Fig. 4 (right), separated by horizontal lines. Contrasting them with the output of *dMondrian* (see Fig. 3), we observe that six less generalizations of values have been performed by *dSabre*. Tuples with ID$=$ 6, 7, 9, 12, 14, and 15 are left unchanged by *dSabre*.

Let us consider now the computational complexity of *dSabre*. The search tree space of *dSabre* consists of at most $log|\mathcal{R}|$ levels, with a total of $L \leq |\mathcal{R}|$ leaves. At each leaf, one tuple is chosen from some bucket and the others are searched among its nearest-neighbors. In the worst case, we have a total complexity of $O(L + NL(|\mathcal{R}| - L))$, where N is the number of attributes. This is maximum when $L = |\mathcal{R}|/2$, and amounts at $O(N|\mathcal{R}|^2)$; and it is minimum when $L = |\mathcal{R}|$, and amounts at $O(|\mathcal{R}|)$. Interestingly, when $L = 1$, it amounts at $O(N|\mathcal{R}|)$. Thus, complexity starts increasing as the search space increase ($L = 1$), reach a maximum ($L = |\mathcal{R}|/2$), and then goes down to a minimum ($L = |\mathcal{R}|$). Since the number L of leaves is a (monotonically increasing) function of the parameter t, which prunes the search space, we expect that *dSabre* has a bell shape running time as a function of t.

We conclude this subsection with the same protection guarantee of Thm. 20.

Theorem 23. *dSabre.Anonymize($\mathcal{R}, t$) is $bd_f(max\{t, \tau(\mathcal{R})\})$-protective w.r.t. $f()$.*

6 Experiments

We have implemented the *dMondrian* and *dSabre* algorithms in Java, adopting some optimizations well-suited for divide & conquer algorithms [27]. The input relational table is stored in main memory by columns; each column stores integer indexes to actual values; and integer indexes respect the ordered of values in the domain of the column attribute. This structure save space, while not impacting on the performances. In fact, *dMondrian* compares column values by their order rank, and *dSabre* computes distance based on normalized order rank. *dSabre* uses *kd*-trees to store buckets in ϕ. The additional space required is linear in the size of the input dataset.

The experiments reported in this section consider classical datasets available from the UCI Machine Learning repository.[6] The *German credit* dataset consists of 1000 records over bank account holders. We set 7 PND attributes: credit_history, purpose, credit_amount, employment, other_payment_plans, housing, and existing_credits. The PD attribute is personal_status with "not-single women" as the protected group. The decision attribute is the

[6] `http://archive.ics.uci.edu/ml`

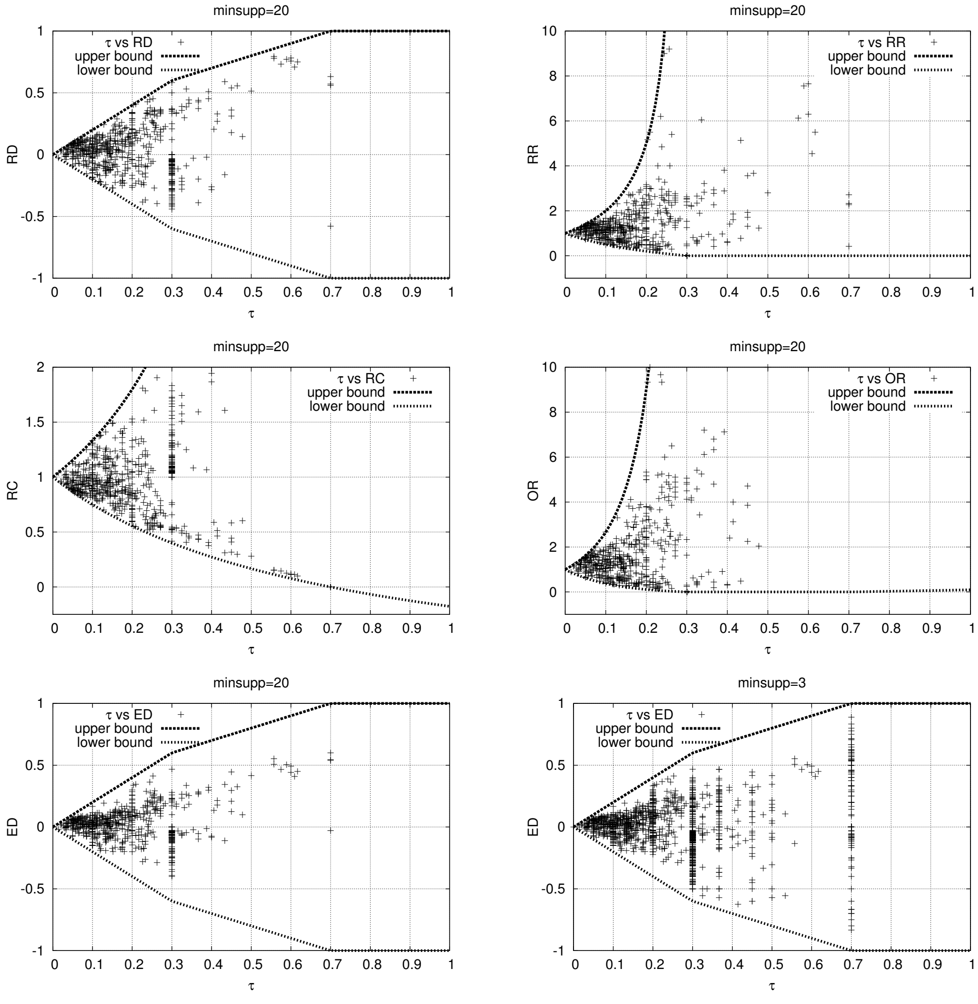


Figure 5: *German credit* dataset. Scatter plots of maximum variational distance (τ) vs discrimination measures.

bad/good credit rating assigned to the bank account holder, with "bad credit" as the negative decision. The *Adult* dataset contains census information on 48848 individuals. We set 6 PND attributes: age, workclass, education, marital-status, occupation, and relationship. The PD attribute is race, with "non-white" individuals as the protected group. The decision attribute is income, which can be <50K or ≥50K dollars, with "<50K" as negative decision. Finally, *Census-Income* is another census dataset, which contains 299285 tuples with 12 PND attributes: age, class of worker, education, wage per hour, marital status, industry, occupation, sex, member of a union, region of residence, household summary, worked weeks. PD and decision attributes are the same as in *Adult*. Summary information on the datasets is reported in Table 1. In particular, t min-max is the meaningful range for the parameter t in input to *dMondrian* and *dSabre* (see Ex. 21 for the calculation of $\tau(\mathcal{R})$ for *German credit*).

The datasets fit well the motivations of the introduction. In the *German credit* dataset, we want to protected non-single women from being associated to failure in re-paying loans (in

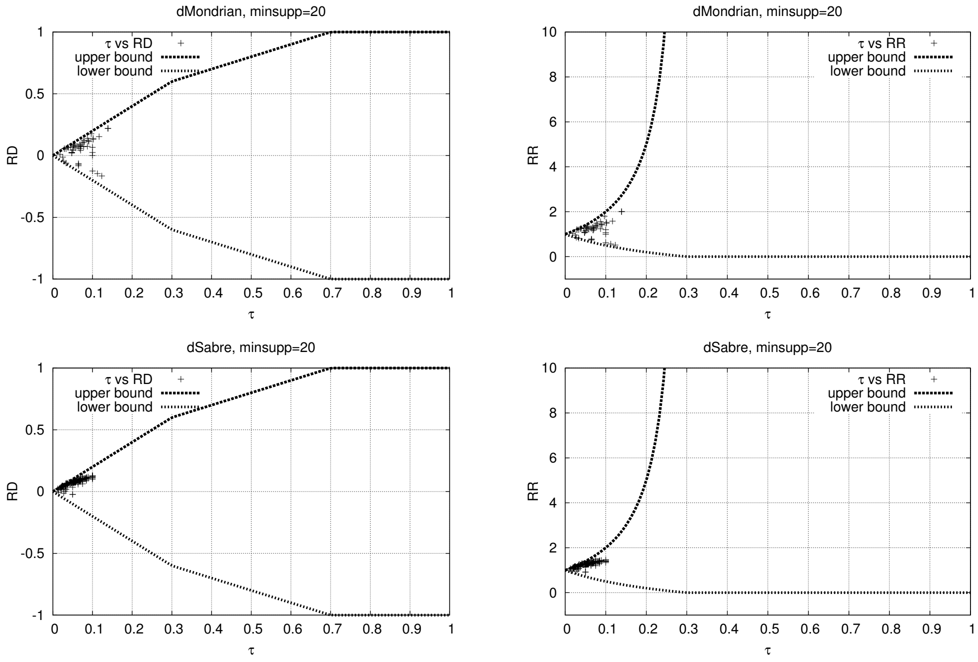


Figure 6: *German credit* dataset. Scatter plots after processing the dataset by *dMondrian* (top) and *dSabre* (bottom), with parameter $t = 0.15$.

some context not known *a-priori*), hence with the risk of being discriminated in subsequent credit evaluation. In the *Adult* and *Census-Income* datasets, we want to protect non-whites from being associated with low income, hence with the risk of being discriminated, e.g., in home loans, or segregated, e.g., in suburban neighborhoods or in specific job positions.

Fig. 5 reports scatter plots for the *German credit* dataset computed as follows. A point refers to a closed PND itemset with minimum support of 20 (or 2%), i.e., such that $n \geq 20$ in the 4-fold contingency table of its cover. The y-axis is the value of a discrimination measure for that cover. The x-axis is the maximum variational distance τ (see Def. 17). The bounds on discrimination measures imposed by variational distance as stated in Thms. 10 and 15 are also shown in Fig. 5 – now including both lower and upper bounds. *The scatter plots highlight that those bounds are not of theoretical interest only, but they can be reached in practice.* When comparing proportions of the protected group vs the unprotected group (as in RD, RR, RC, and OR) there are contexts reaching the bounds even for high values of τ. When comparing proportions of the protected group vs the general population (as in ED, ER, and EC) the bounds are reached for low values of τ, or when focusing on contexts with very low minimum support – as it can be noticed by contrasting the two scatter plots for the ED measure, one with minimum support of 20 and the other of 3. Intuitively, this is due to the fact that variation of the protected group from the average is always lower or equal than variation of the protected group from the unprotected group.

Fig. 6 shows the scatter plots for the RD and RR measures after the *German credit* dataset has been sanitized by *dMondrian* and *dSabre* for input parameter $t = 0.15$. As expected, there is no closed PND itemset with maximum variational distance $\tau > 0.3$, with RD $>$ $bf_{RD}(0.15) = max\{2 \cdot 0.15, 0.15 + 0.3, 1\} = 0.3$, nor with RR $>$ $bf_{RR}(0.15) = min\{0.3 +$

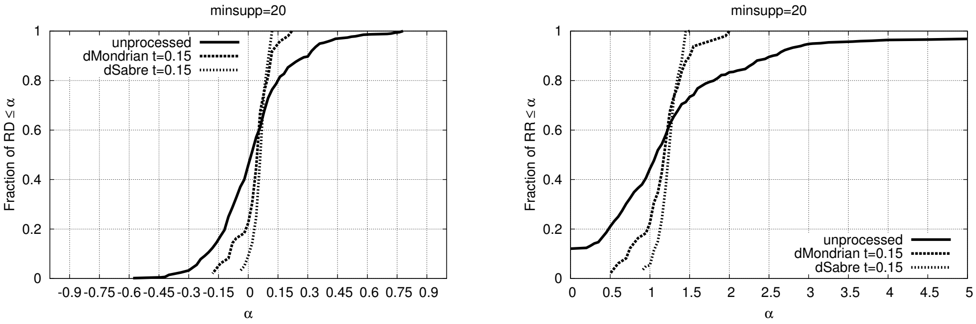


Figure 7: *German credit* dataset. Distributions of RD and RR values.

$0.15, 1\}/max\{0.3 - 0.15, 0\} = 3$. *Since our results hold irrespectively of the minimum support, such a statement is true for any (closed) itemset* – even though the figure plots only closed PND itemsets with a minimum support of 20. The dataset sanitized by *dMondrian* has a wider range of the discrimination measure values, whilst *dSabre* yields narrow ranges. This is made clearer by Fig. 7, showing the distributions of the RD and RR measures for the original and the processed datasets. Distributions flattened around 0 for RD and around 1 for RR denote less discriminatory datasets. Both *dMondrian* and *dSabre* improve the degree of discrimination of the sanitized dataset over the original one, with values below the upper bounds $bd_{RD}(0.15)$ and $bd_{RR}(0.15)$. However, *dSabre* produces less discriminatory datasets.

The actual maximum and minimum RD and ED values in a processed dataset are shown in Fig. 8 for the *German credit* and *Adult* datasets. Such extreme values depend on the minimum support used to extract closed PND itemsets to be plotted. Formally, the function:

$$bd_f(\mathcal{R}, k) = max\{f(\mathbf{B}) \mid \mathbf{B} \text{ closed PND itemset s.t. } |cover_{\mathcal{R}}(\mathbf{B})| \geq k\} \qquad (1)$$

returns the maximum value of a discrimination measure $f()$ over PND closed itemsets of a dataset $\mathcal{R}$ with minimum support k. The maximum value of a dataset processed by *dSabre* with input parameter t and for a minimum support k is then $bd_f(\mathit{dSabre.Anonymize}(\mathcal{R}, t), k)$. Contrast the bottom plots in Fig. 8, which report the extreme values for different minimum support thresholds. Those values approach the $bd_f(t)$ limit for lower and lower minimum support thresholds of closed PND itemsets. Summarizing, *if no requirement is stated on a minimum support threshold of the contexts of possible discrimination, then the bounds stated in Thms. 10 and 15 can be considered sharp in practice.*

Discrimination data protection has to be traded-off with an unavoidable information loss. Since our approach has adapted existing algorithms for data anonymity, we inherit the information loss performances of such algorithms. According to [5], we consider the standard information loss metric (LM) defined as:

$$\text{LM} = \sum_{\text{QI itemset } \mathbf{Q}} supp(\mathbf{Q})\, L(\mathbf{Q}) \qquad L(\mathbf{Q}) = \sum_{i=1,\ldots,N-1} \frac{range(v_i) - 1}{|dom(A_i)| - 1}$$

where a QI attribute in our context is any PND or PD attribute, and, for $\mathbf{Q}$ being $A_1 = v_1, \ldots, A_{N-1} = v_{N-1}$, the loss $L(\mathbf{Q})$ is the sum of losses of its attribute values, with $range(v_i)$ denoting the number of values of $dom(A_i)$ generalized by v_i. LM ranges from 0 (no value has been generalized) to 1 (all values collapsed).

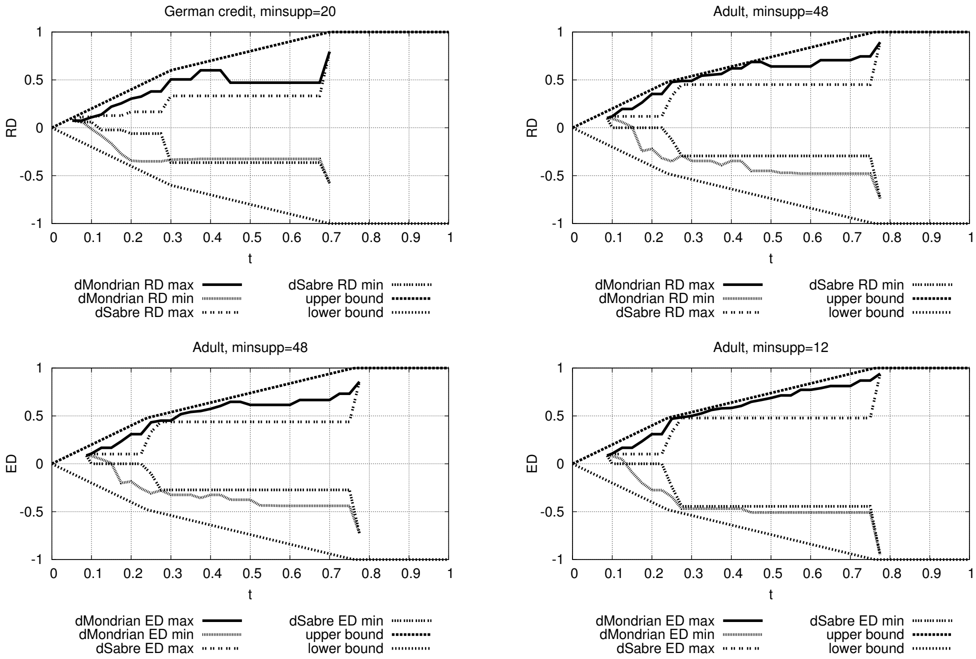


Figure 8: Extreme values of scatter plots after processing *German credit* and *Adult* by *dMondrian* and *dSabre*. Ranges of input parameter t are shown in Table 1.

Fig. 9 shows the information loss due to data sanitization by *dMondrian* and *dSabre* at the variation of the input parameter t. The superiority of *dSabre* over *dMondrian* for lower values of t is visible, and it is inherited from the better performances of *Sabre* over *tMondrian* for the t-closeness data anonymity model [5]. Plots of *dSabre* have abrupt lines because a whole new level at a time appears in the search tree as t grows (see Fig. 4). In summary, an analyst has to trade-off the benefits of a formal bound on discrimination measures achieved by data sanitization (see Fig. 8) with the loss of data utility that is introduced (see Fig. 9). This is analogous to what occurs in the context of data anonymization for privacy protection. *Our approach provides the necessary tools for applying the trade-off analysis in the scenario of discrimination data sanitization.* The bottom right plot in Fig. 9 instantiates the trade-off analysis for the RD and ED discrimination measures, by showing the loss metric for the output of *dSabre* at the variation of the bound $bd_f(t)$ imposed by Thm. 10.

Discrimination data protection has to be traded-off also with utility of the sanitized dataset. According to [5] again, we consider as utility metric the median relative error defined as follows. Consider a count query:

```
SELECT COUNT(*)
FROM dataset
WHERE $A_{\pi_1}$ in [$v_1$, $w_1$] AND ... AND $A_{\pi_n}$ in [$v_n$, $w_n$] AND $A_N$ in [$v_{n+1}$, $w_{n+1}$]
```

where $A_{\pi_1}, \ldots, A_{\pi_n}$ are PND or PD attributes, A_N is the decision attribute, and ranges $[v_i, w_i]$ are drawn uniformly from the domain of the attributes. Such count queries occur frequently in typical data analysis reports or data mining algorithms. The relative error of a query is defined as $|est - prec|/prec$ where $prec$ is the count result over the original dataset,

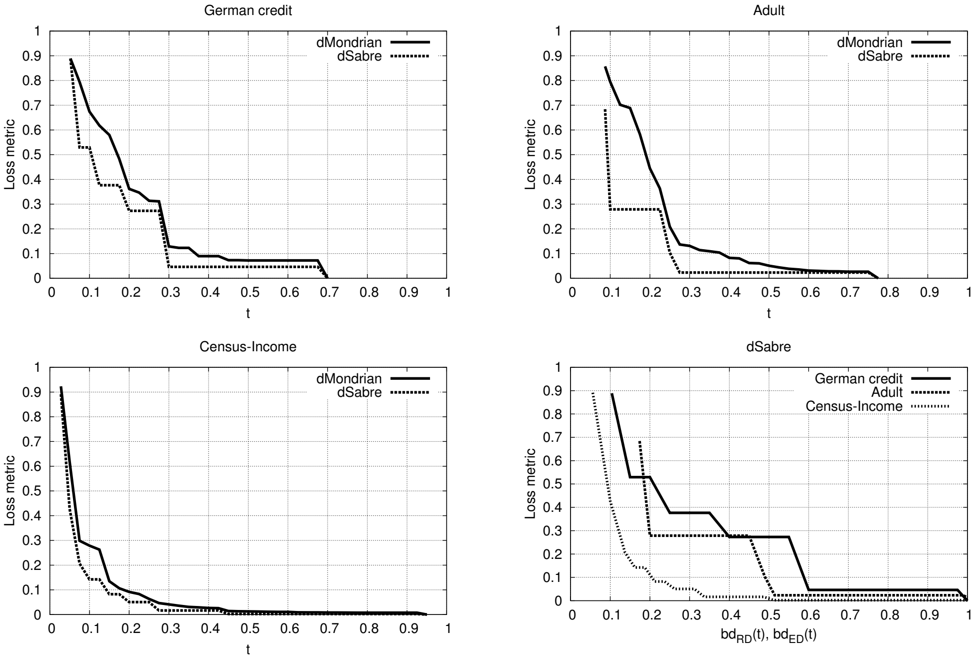


Figure 9: Information loss metric for *dMondrian* and *dSabre*.

and est is the count result estimated on the sanitized dataset. The estimation assumes a uniform distribution of tuples in the domains of values of an attribute. The median relative error is the median error over 10K randomly generated count queries with n uniformly distributed in $[1, 5]$ and with non-zero $prec$ value (needed to prevent division by zero).

Fig. 10 shows the utility metric of datasets sanitized by *dMondrian* and *dSabre* at the variation of the input parameter t. *dSabre* definitely outperforms *dMondrian* on all values of t for *German credit* and *Adult*, but not for *Census-Income*. The reason can be attributed to the "curse of dimensionality", a well-known phenomenon in k-anonymity [1]. *dSabre* is more sensitive than *dMondrian* to high dimensionality datasets or high cardinality attributes due to the direct use of nearest-neighbor search. The bottom right plot in Fig. 10 shows the utility metric plots for *Census-Income* where the two highest cardinality attributes (wage per hour and age) have been removed. *dSabre* is now comparable to *dMondrian*.

Let us consider now efficiency issues. Fig. 11 reports the elapsed execution times of the sanitization procedures on the *Adult* and *Census-Income* datasets for a commodity PC with Intel Core i5-2410@2.30 GHz with 4Gb of RAM and Windows 7 OS. Elapsed times of *dMondrian* increase with t due to the fact that the search tree grows with t, and each level of the search tree adds a complexity linear in the size of the dataset. Elapsed time plots follow a typical pattern of space partitioning algorithms, such as decision tree builders [29], at the variation of the stopping parameter. *dSabre* exhibits a different pattern, which is consistent with its computational complexity discussed in Sect. 5.2. For low t values the search space is small, for high t values the need for costly nearest-neighbor searches decreases, and for middle t values the cost reaches a maximum. Fig. 11 highlights such a bell shape.

In summary, *dSabre* is preferable to *dMondrian* for datasets with low dimensionality and low attribute cardinality because it has similar running time performances (Fig. 11), lower

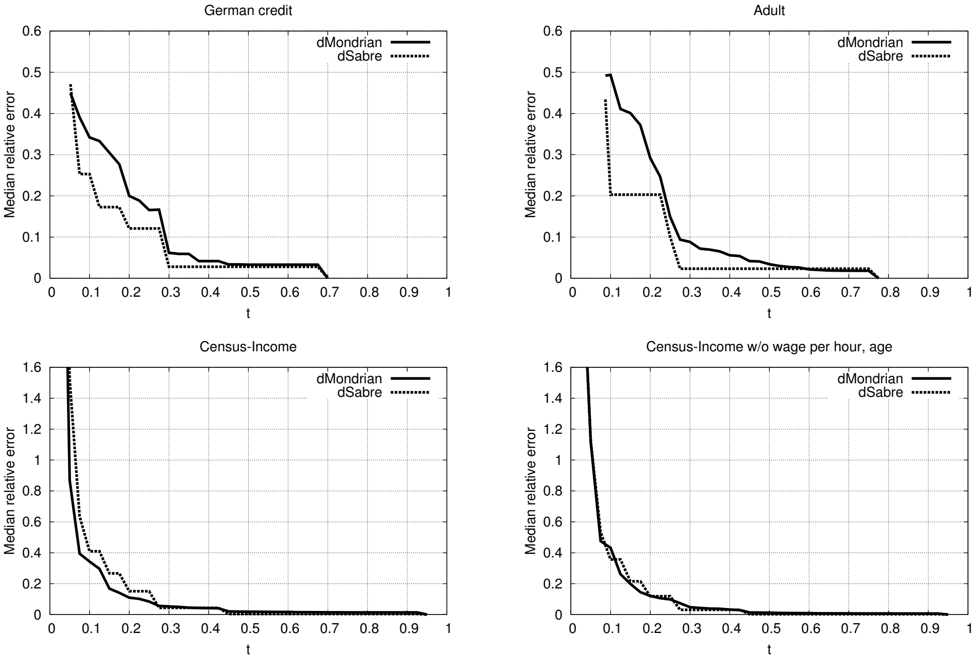


Figure 10: Median relative error utility metric for *dMondrian* and *dSabre*.

ranges of the discrimination measures in the sanitized output (Fig. 7 and 8), better information loss (Fig. 9), and better utility (Fig. 10). *dMondrian* is instead preferrable to *dSabre* for medium to large dimensionality datasets or for datasets with large attribute cardinalities.

7 Related work

Data sanitization for discrimination prevention. Approaches for building classifiers that do not make discriminatory decisions may rely on data sanitization of the training set [28]. Existing techniques for data sanitization adopt perturbation approaches by changing values of the PD attribute or of the decision attribute. The approaches in [11, 16, 23] massage the dataset by promoting (from - to + decision value) some individuals of the protected group and/or demoting (from + to -) individuals of the unprotected group using some heuristics. [11] ranks individuals on the basis of the number of contexts of possible discrimination they appear in. [16] adopts the prediction confidence of a classifier for ranking individuals in the protected and in the unprotected groups. [23] adopts a measure of the bias observable by an individual in the decisions between the top nearest neighbors of the protected group and the top nearest neighbors of the unprotected group. None of the approaches provides a formal guarantee on the level of α-protection of the sanitized dataset, as we did with the bounds of Thms. 10 and 15. As an additional limitation, [11] deals with nominal attributes only, because it heavily relies on association rule mining. We can instead cope with continuous attributes by letting the sanitization algorithms perform a discrimination-oriented discretization. Moreover, [16] considers the RD measure only at the grain of the whole dataset, i.e., only a single context of possible discrimina-

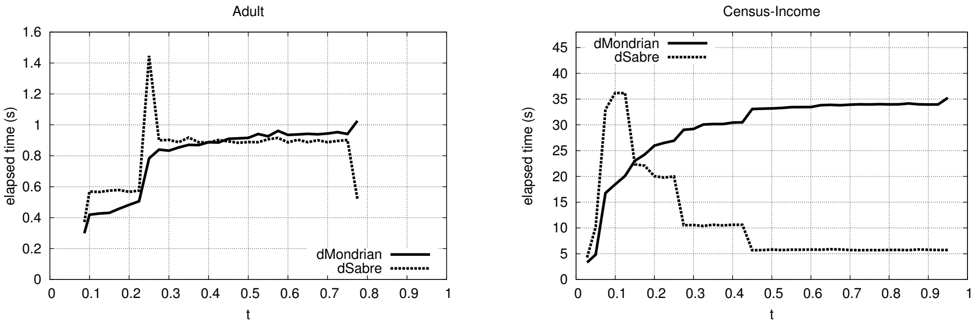


Figure 11: Elapsed times of *dMondrian* and *dSabre* on *Adult* and *Census-Income*.

tion is guaranteed to be sanitized. Nevertheless, since there are intrinsic limitations of non-perturbation methods (see Ex. 21), an hybrid approach trading off massaging with generalization must necessarily be pursued to achieve the best from the two approaches.

Impact of anonymization on non-discrimination. [10] discusses how data anonymization techniques that improve k-anonymity affect the degree of α-protection w.r.t. the RR (called *slift*) and ER (called *elift*) measures. The techniques of global and local recoding generalizations, and of cell, record, and value suppressions are considered. For instance, it is found that generalization may lead an α-protective dataset to be non α-protective anymore. This result may seem in contrast with our findings. Two observations clarify this. First, if the original dataset is α-protective, and the transformed dataset is only α'-protective for $\alpha < \alpha'$ and for no lower α', then this is still consistent with the fact that the original dataset is t-close and $\alpha < \alpha' \leq bd_f(t)$, where $bd_f(t)$ is the bound from Thms. 10 and 15. Second, [10] assumes a relaxed definition of α-protection, where the (equivalent of the) PND-itemset $\mathbf{B}$ in Def. 4 has a minimum support k, rather than a non-empty cover. As a consequence of generalization, some infrequent PND itemsets may become frequent, and then their discrimination measure value becomes relevant, whilst in the original dataset it was not accounted for. Using the notation (1), this can be formally stated as:

$$bd_f(Anonymize(\mathcal{R}), k) \not\leq bd_f(\mathcal{R}, k).$$

Achieving both anonymization and non-discrimination. The recent paper [12] considers the problem of sanitizing a dataset both w.r.t. k-anonymity for a given set of QIs, and w.r.t. α-protection (w.r.t. RR and ER) for PD attributes that are assumed to be at also QIs. It is the first paper that applies a non-perturbation transformation methods for discrimination prevention. The paper considers k-anonymity, a model for preventing identity disclosure attacks, whilst we reasoned about t-closeness, a privacy requirement for preventing attribute disclosure attacks. The paper adopts a full domain generalization approach, which generalizes all instances of a value in the domain of an attribute to a same ascendent of that value in the hierarchy of the domain. The authors adapt the *Incognito* algorithm for searching the space of all possible k-anonymous full domain generalizations that are also α-protective. Full domain generalization is one of the earliest approaches for k-anonimity, and it has been superseded, both in utility of the sanitized dataset and in efficiency of sanitization algorithms, by the multidimensional generalization approaches considered in this paper. However, *Incognito* is optimal with regard to full domain generalizations, whilst *Mondrian* and *Sabre* are heuristical algorithms (since the multidimensional generalization problem is NP-hard). Moreover, multidimensional generalization suffers from the problem of data exploration, caused by the co-existence of specific and generalized values in the sanitized dataset, which makes data exploration and interpretation more difficult than in the case of full domain generalization. Finally, compared to this paper, [12] considers other alternative privacy models in addition to k-anonymity (attribute disclosure and differential privacy) and alternative discrimination analyses (indirect discrimination and genuine occupational requirement), but it does not cope with tokenism nor with disjunctive items.

[13] proposes a methodology for achieving both k-anonymity and α-protection (w.r.t. RR) in knowledge disclosure, specifically in the disclosure of frequent itemsets. The approach consists of first applying additive sanitization to control for k-anonymity, and then a form of anti-discrimination additive sanitization to control for α-protection. The second step does not affect k-anonymity since it *adds* tuples to the dataset. In contrast, our approach can tackle (the stronger model of) t-closeness and α-protection in a single step. Since we establish t-closeness to ensure $bd_f(t)$-protection, we have that the dataset sanitized is *at the same time* t-close and $bd_f(t)$-protective. Moreover, we have highlighted simple variants of *dMondrian* and *dSabre* imposing also k-anonymity on the sanitized datataset.

Differential privacy and non-discrimination. Throughout this paper, we have considered the legal principle of "group proportional representation": fixed a *subset* of PND attribute values, the protected and unprotected groups must be proportionally represented in a context of possible discrimination. This is sometimes known as *group fairness*. Another legal principle is to "treat like cases alike", leading to a notion of *individual fairness*. Here, similarity of individuals is measured by some (agreed and public) distance function, defined over the space of individuals' characteristics, i.e., over *all* PND attributes. [8] formalizes the principle of individual fairness by expressing similarity of decisions as a distance between the probability distributions of decisions assigned to individuals. The approach is then related theoretically to ϵ-differential privacy where, rephrasing the legal principle of individual fairness, "tables that are alike should answer alike" to a given query. Our work complements [8]: together they relate concepts of group/individual fairness in discrimination analysis to techniques of t-closeness/ϵ-differential privacy in data anonymization.

(n, t)**-closeness.** t-closeness has been relaxed in [21] to (n, t)-closeness by requiring the distribution in a q-block to be close to the distribution of some *natural* superset of the q-block whose size is at least n. Natural supersets are those obtained by generalizing (over a hierarchy of values) one of the values in the QI itemset of the q-block. For instance, natural supersets of *age=20-25* and *zip=561** are *age=20-30* and *zip=56*** respectively, assuming that *20-30* is the father of *20-25* and *56*** is the father of *561** in the hierarchy of attribute values. Our results can be extended to (n, t)-closeness by replacing any usage of p_- by the proportion p_-^{nats} of negative decisions in the natural superset of the PND-itemset under consideration.

Utility metrics for discrimination analysis. We have adopted an utility measure of the sanitized dataset based on (median relative error of) count queries. Some literature on discrimination prevention [4, 12, 16, 23] adopt instead the classification accuracy of the predictions made by a classifier trained from a sanitized dataset. Such an approach, however, would require a "ground truth" test set having no discriminatory decision. Otherwise, we would be comparing a possibly fair prediction (yet inaccurate) with a possibly discriminatory one (yet accurate).

[11] considers utility metrics for discrimination sanitization of *knowledge*, specifically of association rules. The authors exploit approaches from the privacy literature on association rule hiding (see [35]). They consider a *misses cost* of sanitization, as the percentage of patterns that are not anymore extractable from the sanitized dataset, and a *ghost cost*, as the percentage of new patterns that were not extractable from the original dataset.

8 Conclusions

The contribution of this paper was twofold.

First, we have related the analytical tools of t-closeness in privacy data anonymization and of α-protection in non-discrimination data analysis by showing that t-closeness implies $bd_f(t)$-protection for a discrimination measure $f()$, and by showing that t-closeness reduces to α-protection in polynomial time. The former result covers contexts of possible discrimination expressed as conjunctions of, possibly disjunctive, items. The latter result allowed us to conclude NP-harness of multidimensional generalization-based data sanitization for discrimination data protection.

Second, by exploiting the discovered implication, we have *systematically* obtained *dMondrian*, a multidimensional generalization algorithm, and *dSabre*, a bucketization and redistribution algorithm, as adaptions of well-known algorithms for k-anonymity and t-closeness. This is is a methodological contribution linking data anonymization research to non-discrimination research. Our results guarantee a formal bound on the discrimination protection of the dataset sanitized by such algorithms. No previous approach on discrimination-aware data mining can provide such a guarantee.

Experiments showed that *dSabre* performs better than *dMondrian*, for low-dimensionality datasets, but it suffers from a curse of dimensionality problem. As one would expect, an anti-discrimination analyst has to trade-off discrimination protection with information loss and utility of the sanitized dataset. The proposed framework provides the necessary tools for conducting such trade-off anayses.

Several challenging issues for future research are raised by the results of this paper, regarding the interplay of privacy-preserving data mining and discrimination-aware data mining. A first research direction is to study how the relation between t-closeness and α-protection extends to the case of multi-valued decisions, such as interest rate range, and PD attributes, such as *race*, without resorting to binarization. The problem here is that, while several distance functions between probability distributions have been considered in the data privacy literature, the discrimination literature restricts to statistical test (such as the χ^2-test) of the significance of the uneven distribution of values in a $n \times m$ contingency table. It would be interesting to import distribution distances from the privacy literature to the purpose of defining discrimination measures in the general case. Second, the idea of relating models of privacy to models of discrimination can be at the basis of the adaption of approaches for *privacy-by-design*, where some guarantee is provided on data privacy while performing a certain task, to design approaches for *anti-discrimination-by-design*, where now the guarantee is on non-discrimination while taking a certain decision. Finally, an intriguing research line follows from the parallel between the role of an anti-discrimination analyst and the one of an attacker. It is then natural to investigate whether attack models considered in the privacy literature can be translated into helpful, legally-grounded, methodologies for discrimination analysis in the hands of anti-discrimination authorities.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *Proc. of Int. Conf. on Very Large Data Bases (VLDB 2005)*, pages 901–909. ACM, 2005.

[2] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhal. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2(2):66–75, 2000.

[3] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975.

[4] T. Calders and I. Žliobaitytė. Why unbiased computational processes can lead to discriminative decision procedures. In B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky, editors, *Discrimination and Privacy in the Information Society*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 43–57. Springer, 2012.

[5] J. Cao, P. Karras, P. Kalnis, and K.-L. Tan. SABRE: a Sensitive Attribute Bucketization and REdistribution framework for *t*-closeness. *VLDB Journal*, 20(1):59–81, 2011.

[6] B.-C. Chen, D. Kifer, K. LeFevre, and A. Machanavajjhala. Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2(1-2):1–167, 2009.

[7] J. Domingo-Ferrer. A survey of inference control methods for privacy-preserving data mining. In C. C. Aggarwal and P. S. Yu, editors, *Privacy-Preserving Data Mining - Models and Algorithms*, volume 34 of *Advances in Database Systems*, pages 53–80. Springer, 2008.

[8] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In S. Goldwasser, editor, *Proc. of the 3rd Int. Conf. on Innovations in Theoretical Computer Science (ITCS 2012)*, pages 214–226. ACM, 2012.

[9] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4), 2010.

[10] S. Hajian and J. Domingo-Ferrer. A study on the impact of data anonymization on anti-discrimination. In J. Vreeken et al., editor, *Proc. of the IEEE ICDM 2012 Int. Workshop on Discrimination and Privacy-Aware Data Mining (DPADM)*, pages 352–359. IEEE Computer Society, 2012.

[11] S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. on Knowledge and Data Engineering*, 25(7):1445–1459, 2013.

[12] S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, pages 1–31, 2014.

[13] S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer, and F. Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In J. Vreeken et al., editor, *Proc. of the IEEE ICDM 2012 Int. Workshop on Discrimination and Privacy-Aware Data Mining (DPADM)*, pages 360–367. IEEE Computer Society, 2012.

[14] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[15] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[16] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2012.

[17] F. Kamiran and I. Žliobaitytė. Explainable and non-explainable discrimination in classification. In B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky, editors, *Discrimination and Privacy in the Information Society*, volume 3 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, pages 155–170. Springer, 2013.

[18] R. M. Kanter. Some effects of proportions on group life: Skewed sex ratios and responses to token women. *American Journal of Sociology*, 82(5):965–990, 1977.

[19] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proc. of the Int. Conference on Data Engineering (ICDE 2006)*, page 25. IEEE Computer Society, 2006.

[20] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proc. of the Int. Conference on Data Engineering (ICDE 2007)*, pages 106–115. IEEE

Computer Society, 2007.

[21] N. Li, T. Li, and S. Venkatasubramanian. Closeness: A new privacy measure for data publishing. *IEEE Trans. on Knowledge and Data Engineering*, 22(7):943–956, 2010.

[22] H. Liang and H. Yuan. On the complexity of t-closeness anonymization and related problems. In W. Meng, L. Feng, S. Bressan, W. Winiwarter, and W. Song, editors, *Database Systems for Advanced Applications*, volume 7825 of *LNCS*, pages 331–345. Springer, 2013.

[23] B. T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In C. Apté, J. Ghosh, and P. Smyth, editors, *Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2011)*, pages 502–510. ACM, 2011.

[24] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. on Knowledge Discovery from Data*, 1(1):Article 3, 2007.

[25] D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In S. Ossowski and P. Lecca, editors, *Proc. of ACM Int. Symposium on Applied Computing (SAC 2012)*, pages 126–131. ACM, 2012.

[26] M. J. Piette and P. F. White. Approaches for dealing with small sample sizes in employment discrimination litigation. *Journal of Forensic Economics*, 12(1):43–56, 1999.

[27] F. J. Provost and V. Kolluri. A survey of methods for scaling up inductive algorithms. *Data Min. Knowl. Discov.*, 3(2):131–169, 1999.

[28] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 2014. To appear, `doi:10.1017/S0269888913000039`.

[29] S. Ruggieri. YaDT: Yet another Decision tree Builder. In *Proc. of Int. Conf. on Tools with Artificial Intelligence (ICTAI 2004)*, pages 260–265. IEEE, 2004.

[30] S. Ruggieri, D. Pedreschi, and F. Turini. Data mining for discrimination discovery. *ACM Trans. on Knowledge Discovery from Data*, 4(2):Article 9, 2010.

[31] S. Ruggieri, D. Pedreschi, and F. Turini. DCUBE: Discrimination discovery in databases. In A. K. Elmagarmid and D. Agrawal, editors, *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010)*, pages 1127–1130. ACM, 2010.

[32] S. Ruggieri, D. Pedreschi, and F. Turini. Integrating induction and deduction for finding evidence of discrimination. *Artificial Intelligence and Law*, 18(1):1–43, 2010.

[33] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 1998)*, page 188. ACM, 1998.

[34] L. Sweeney. k-anonymity: A model for protecting privacy. *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[35] V. Verykios and A. Gkoulalas-Divanis. A survey of association rule hiding methods for privacy. In C. C. Aggarwal and P. S. Yu, editors, *Privacy- Preserving Data Mining: Models and Algorithms*, pages 267–289. Springer, 2008.

	decision		
group	-	+	
protected	a	$rs-a$	rs
unprotected	c	$rs-c$	rs
	$a+c$	$2rs-(a+c)$	$2rs$

$$a = rs \sum_{\substack{i=1,\ldots,M \\ p_i \geq p_{s_i}}} p_i \qquad c = rs \sum_{\substack{i=1,\ldots,M \\ p_i \geq p_{s_i}}} p_{s_i} \qquad s = |cover_{\mathcal{R}}(\mathbf{Q})| \qquad r = |\mathcal{R}|$$

Figure 12: Contingency table of $cover_{red(\mathcal{R})}(\mathbf{Q})$.

A Appendix

Throughout the paper, we have assumed that sensitive attributes are binary. Let us now relax this assumption. Let $dom(A_N) = \{s_1, \ldots, s_M\}$ be the set of sensitive attribute values, and p_{s_i} be fraction of tuples having sensitive value s_i in the whole dataset, i.e., $p_{s_i} = supp(A_N = s_i)$. Variational distance for a QI itemset $\mathbf{Q}$ can be rewritten as follows [20]:

$$VD(\mathbf{Q}) = \frac{1}{2} \sum_{i=1,\ldots,M} |p_i - p_{s_i}| = \sum_{\substack{i=1,\ldots,M \\ p_i \geq p_{s_i}}} p_i - p_{s_i}$$

where p_i is the fraction of tuples in the cover of $\mathbf{Q}$ with sensitive value s_i.

Let us now introduce a transformation that will be useful to reduce (in polynomial time) the notion of t-closeness to the one of α-protection. Let $\mathcal{R}$ be a relational table with attributes $A_1, \ldots, A_N$, where A_N is the sensitive, possibly multi-valued, attribute and A_1, $\ldots$, A_{N-1} are QIs. Let $r = |\mathcal{R}|$ be its size. We define a table $red(\mathcal{R})$ with attributes $A_1, \ldots, A_{N-1}$, *group*, *decision* where *group* has values *protected* and *unprotected*, and *decision* has values "-" and "+". Attributes $A_1, \ldots, A_{N-1}$ are considered as PND attributes, *group* as the PD attribute, and *decision* as the decision attribute. For every QI itemset $\mathbf{Q}$ defined as $a_1 = v_1, \ldots, a_{N-1} = v_{N-1}$ and having a non-empty cover $\mathcal{S} = cover(\mathbf{Q})$ of size $s = |\mathcal{S}|$, $red(\mathcal{R})$ includes $2rs$ tuples defined as follows. All tuples σ are such that $\sigma[A_i] = v_i$ for $i = 1, \ldots, N-1$, i.e., they support $\mathbf{Q}$ as well. The values of PD and decision attributes are defined according to the contingency table in Fig. 12. a/r is the number of tuples in $\mathcal{S}$ whose sensitive value occurs with frequency greater than or equal to its frequency in the whole dataset. $(a-c)/r$ is the number of tuples in $\mathcal{S}$ that should have changed their sensitive values to make the distribution of $\mathcal{S}$ equal to the one of the whole dataset. Both a and c are integer values, since p_i and p_{s_i} are fractions over s and r respectively. The multiplication by a factor of r in Fig. 12 is purposely designed for this. The following is immediate by definition of a and c.

Lemma 24. $a/(rs) - c/(rs) = VD(\mathbf{Q})$.

Finally, notice that $|red(\mathcal{R})| = 2|\mathcal{R}|^2$. We have the following polynomial time reduction.

Theorem 25. *A relational table $\mathcal{R}$ is t-close iff $red(\mathcal{R})$ is t-protective w.r.t. RD. Moreover, $red(\mathcal{R})$ can be computed in polynomial time from $\mathcal{R}$.*

Proof. $\mathit{red}(\mathcal{R})$ can be computed in polynomial time by first sorting $\mathcal{R}$ lexicographically, and then scanning covers of QI itemsets. For each cover of size s, the generation of the $2rs$ tuples of $\mathit{red}(\mathcal{R})$ takes linear time. Summarizing, $\mathit{red}(\mathcal{R})$ can be computed in $O(|\mathcal{R}|^2)$ time.

Suppose now that $\mathit{red}(\mathcal{R})$ is t-protective w.r.t. RD. Let $\mathbf{Q}$ be a QI itemset of $\mathcal{R}$ with non-empty cover. By definition of $\mathit{red}(\mathcal{R})$, $\mathbf{Q}$ can be seen as a PND itemset of $\mathit{red}(\mathcal{R})$. Then, by t-protection $\mathrm{RD}(\mathbf{Q}) \leq t$. From Fig. 12, it is immediate to observe that $\mathrm{RD}(\mathbf{Q}) = a/(rs) - c/(rs)$ and then, by Lemma 24, we conclude $\mathit{VD}(\mathbf{Q}) = \mathrm{RD}(\mathbf{Q}) \leq t$. Summarizing, $\mathcal{R}$ is t-close.

Conversely, suppose that $\mathcal{R}$ is t-close. Consider a PND itemset $\mathbf{B}$ with non-empty cover w.r.t. $\mathit{red}(\mathcal{R})$. We have:

$$\mathit{cover}_{\mathit{red}(\mathcal{R})}(\mathbf{B}) = \bigcup_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} \mathit{cover}_{\mathit{red}(\mathcal{R})}(\mathbf{B}') \tag{2}$$

where $\mathit{base}(\mathbf{B})$ is the set of QI itemsets $\mathbf{B}'$ such that $\mathbf{B}' \supseteq \mathbf{B}$. Notice that the covers in the union are disjoint. Let $a^{\mathbf{B}'}$, $c^{\mathbf{B}'}$ and $s^{\mathbf{B}'}$ be the values a, c and s in Fig. 12 for the cover of $\mathbf{B}'$. For the cover of $\mathbf{B}$, it turns out from (2) that:

$$a = \sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} a^{\mathbf{B}'} \qquad c = \sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} c^{\mathbf{B}'} \qquad s = \sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} s^{\mathbf{B}'} \tag{3}$$

which implies:

$$\mathrm{RD}(\mathbf{B}) = \frac{\sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} a^{\mathbf{B}'}}{rs} - \frac{\sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} c^{\mathbf{B}'}}{rs} = \frac{1}{s} \sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} \frac{(a^{\mathbf{B}'} - c^{\mathbf{B}'})}{r}$$

By Lemma 24, $(a^{\mathbf{B}'} - c^{\mathbf{B}'})/r = s^{\mathbf{B}'} \mathit{VD}(\mathbf{B}')$. Since $\mathcal{R}$ is t-close, we have $\mathit{VD}(\mathbf{B}') \leq t$ and then $(a^{\mathbf{B}'} - c^{\mathbf{B}'})/r \leq s^{\mathbf{B}'} t$. This implies:

$$\mathrm{RD}(\mathbf{B}) \leq \frac{t}{s} \sum_{\mathbf{B}' \in \mathit{base}(\mathbf{B})} s^{\mathbf{B}'} = t$$

where the last equivalence follows from the equation on s in (3). Summarizing, $\mathit{red}(\mathcal{R})$ is t-protective w.r.t. RD. □

Notice that in the proof of the only-if part of the theorem, we do not incur in the Simpson's paradox because the protected and the unprotected groups have been generated purposely with an equal number of tuples in each context of possible discrimination.

Problems that are hard for t-closeness translate to hard problems for α-protection. Thm. 16 is about NP-harness of the optimal generalization problem for discrimination protection.

Theorem 16.

Proof. By definition of $\mathit{red}(\mathcal{R})$ any generalization of attribute values for a dataset $\mathcal{R}$ with cost c corresponds to a generalization for $\mathit{red}(\mathcal{R})$ with cost $2cr$ (where the factor 2 is due to doubling tuples for protected and unprotected groups, and the factor r is due to multiplication of tuples), and vice-versa. By Thm. 25, we have that the optimal generalization problem for t-closeness of $\mathcal{R}$ is reducible in polynomial time to the optimal generalization problem for α-protection of $\mathit{red}(\mathcal{R})$. The conclusion of the theorem follows from NP-harness of optimal generalization problem for t-closeness [22]. □