Data Anonymity Meets Non-Discrimination

Salvatore Ruggieri
Dipartimento di Informatica, Università di Pisa
Largo B. Pontecorvo 3, 56127 Pisa, Italy
ruggieri@di.unipi.it

***Abstract*—We investigate the relation between t-closeness, a well-known model of data anonymization, and α-protection, a model of data discrimination. We show that t-closeness implies $bd(t)$-protection, for a bound function $bd()$ depending on the discrimination measure at hand. This allows us to adapt an inference control method, the *Mondrian* multidimensional generalization technique, 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 mining.**

I. 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 [1]. 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 [2] or t-closeness [3]. 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 [4]. 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.
- 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 on applications of members of 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 any condition based on an applicant's membership to a minority group denoting past discriminatory practices. Using a common notation based on itemset mining, we investigate the relation between the model of α-protection for discrimination (which is parametric in a discrimination measure) [5] and the model of t-closeness for data anonymization [3]. 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 $bd(t)$-protection*, for an appropriate bound function $bd()$ depending on the reference discrimination measure. 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 [6]. This data transformation technique can provide a formal guarantee on the maximum level of discrimination present in a sanitized dataset before it is released.

II. Preliminaries

We recall notation and concepts from itemset mining. 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 $V_1, \ldots, V_N$. Tuples in the table denote individuals, and attribute values denote information about individuals. A *V-item* is a term $V = v$, where V is an attribute and $v \in dom(V)$, the domain of V. We assume that $dom(V)$ is categorial (hence finite) for every attribute V. An item is any V-item. We denote by $\mathcal{I}$ the set of all items. An *itemset* $\mathbf{X}$ is a subset of $\mathcal{I}$. As usual in the literature, we write $\mathbf{X}, \mathbf{Y}$ for $\mathbf{X} \cup \mathbf{Y}$. A tuple t from $\mathcal{R}$ *supports* $\mathbf{X}$ if for every $V = v$ in $\mathbf{X}$, we have $t[V] = v$, where $t[V]$ is the value of the attribute V in the tuple t. The *cover* of $\mathbf{X}$ is the set of tuples that support $\mathbf{X}$: $cover(\mathbf{X}) = \{t \in \mathcal{R} \mid t \text{ supports } \mathbf{X}\}$. The *support* of $\mathbf{X}$ is the number of tuples $|cover(\mathbf{X})|$ in its cover. The relative support of $\mathbf{X}$ is $supp(\mathbf{X}) = |cover(\mathbf{X})|/|\mathcal{R}|$. $\mathbf{X}$ is a *frequent* itemset

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

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

$$RD = p_1 - p_2 \quad ED = p_1 - p \quad SD = n_1/n - b/m_2$$

Fig. 1. Discrimination measures.

if $supp(\mathbf{X}) \geq minsupp$, where $minsupp$ is a given threshold. $\mathbf{X}$ is *closed* if there is no $\mathbf{Y} \supset \mathbf{X}$ such that $cover(\mathbf{Y}) = cover(\mathbf{X})$. A closed itemset is a representative member of the class of equivalence of itemsets with a same cover.

In privacy-aware data mining, attributes of a disclosed table 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 $V_1, \ldots, V_{N-1}$ are the QIs and V_N is the only sensitive attribute. Let us introduce the notion of QI itemset.

Definition 2.1: A *QI itemset* $\mathbf{Q}$ is an itemset containing *one and only one* V-item for every QI attribute V, and no V-item for sensitive attributes V.

With our restriction, $\mathbf{Q}$ has the form $V_1 = v_1, \ldots, V_{N-1} = v_{N-1}$. The *q-block* (also known as the equivalence class) of $\mathbf{Q}$ is the cover of $\mathbf{Q}$.

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[1] 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 V_N, with binary values "*protected*" and "*unprotected*", and of only one decision attribute, say V_{N-1}, with binary values "+" (positive decsion) and "-" (negative decision). Thus, $V_1, \ldots, V_{N-2}$ is the set of PND attributes.

Definition 2.2: A *PND itemset* $\mathbf{B}$ as an itemset containing *at most one* V-item for every PND attribute V, and no V-item for PD or decision attributes V.

With our restriction, $\mathbf{B}$ has the form $V_{\pi_1} = v_1, \ldots, V_{\pi_k} = v_k$ 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 $\mathbf{B}$. Notice that QI itemsets contain exactly *one* item for every QI attribute, whilst PND itemsets contain *at most one* item for every PND attribute.

III. Discrimination analysis

A. 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* 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. In this paper, we mainly consider measures defined as differences of proportions, including:

- *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, yet other terms are also used in data mining [5], [8]. Let us formalize a further measure, which has not been considered so far in data mining. The *selection difference* SD is the difference between the fraction of people of the protected group in the overall dataset (n_1/n) and in the subset of positive decisions (b/m_2). This measure is well-known in the legal domain[2]. The degree of observed disproportionate burden suffered by the protected group is monotonic increasing for RD, ED, and SD. Since one is interested in contexts with a larger proportion of negative decisions for the protected group compared to the unprotected group or to the average, the values of interest for such measures are those greater than 0. Finally, observe that RD $= 0$ iff ED $= 0$ iff SD $= 0$ iff $ad = bc$. The last condition describes a situation of *statistical parity*, with equal chances of obtaining a negative (resp., positive) decision for the protected group, the unprotected group, and the whole population.

B. 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*. Many small or large such contexts may exist that conceal discrimination, and therefore all possible specific situations

[1]The use of PD (resp., PND) attributes in decision making does not necessarily lead to (or exclude) discriminatory decisions [5], [7]. This motivates the adjective "potentially".

[2]The *Castaneda rule* (named after the *Castadena vs. Partida* U.S. law case, 1977, concerning jury selection in a Texas county) states that the selected fraction of the protected group cannot exceed 3 standard deviations (w.r.t. a random selection) the fraction of the group in the overall population.

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 [5] by extracting and ranking classification rules on the basis of a discrimination measure. We restate here the analysis framework using PND itemsets and measures defined over the 4-fold contingency table of their cover. Let us start introducing some notation.

Definition 3.1: For a PND itemset $\mathbf{B}$, we denote by $f(\mathbf{B})$ the value of a measure $f()$ over the 4-fold contingency table of the set $cover(\mathbf{B})$.

The discrimination 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 discrimination and a threshold between "legal" and "illegal" degree, we are in the position to isolate contexts of possible discrimination where the measure is above such a threshold.

Definition 3.2: [5] Let $f()$ be a measure defined over a contingency table, and $\alpha \in \mathbb{R}$ 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.

With this approach, the problem of *discrimination discovery* consists of extracting 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 and evaluation by a legal expert. In actual implementation, [9] resorts to frequent itemset mining, thus restricting to PND itemsets with a minimum support rather than with a non-empty cover. From a legal perspective, this is reasonable, since it leaves out cases where numbers do not provide sufficient statistical confidence. Let us now extend the notion of α-protection to a whole dataset.

Definition 3.3: 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. This observation, which follows from the adoption of the itemset mining notation, has two main advantages over the original definition of [5], which considers every PND itemset. First, there is no duplicate analysis of the same context of possible discrimination, since the covers of two distinct closed itemsets denote different groups of individuals. Second, discrimination discovery is sped up, since the search for α-discriminatory PND itemsets is restricted to closed PND itemsets only.

C. Tokenism and discrimination measures

Due to a division by zero in the discrimination measure, α-protection may be undefined: this occurs for RD when $n_1 = 0$ or $n_2 = 0$; for ED when $n_1 = 0$; and for SD when $m_2 = 0$. Such cases are unlikely for the whole dataset, but they can readily occur for (small) covers of PND itemsets[3]. Let us discuss here the legal interpretation of those conditions. First, consider the SD measure. When $m_2 = 0$, all decisions for individuals in the cover are negative. Is that a discriminatory practice? It may be so. If in the context there are $a = 999$ individuals of the protected group and $c = 1$ individuals of the unprotected group, then the protected group suffers from a higher burden of the always-negative decision. The single individual of the unprotected group, called a *token*, may have been assigned a negative decision to create the false appearance of equality and prevent charges of discrimination. This illegal practice is known as *reverse tokenism* (whereas tokenism [10] consists of granting a benefit to a few members of a minority group to create the false appearance of inclusiveness). A measure of the burden suffered by the protected group is the difference $a/m_1 - p_{pro}$ between the proportion of the protected group among the individuals with negative decision (a/m_1) and the proportion p_{pro} of the protected group in the overall dataset (i.e., p_{pro} is $supp(V_N = protected)$). The ratio a/m_1 coincides with n_1/n, since $m_2 = 0$. Thus, SD can be extended as follows:

$$SD = \begin{cases} n_1/n - p_{pro} & \text{if } m_2 = 0 \\ n_1/n - b/m_2 & \text{otherwise} \end{cases}$$

This looks intuitive: when the proportion b/m_2 of individuals from the protected group among the selected ones is undefined, we simply consider the expected proportion p_{pro}. With a similar reasoning, ED is extended to contexts with only individuals of the unprotected group:

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

where p_- is the proportion of negative decisions in the overall dataset (i.e., p_- is $supp(V_{N-1} = -)$). Finally, the extension of RD deals with contexts of only unprotected people (when $n_1 = 0$) and of only protected people (when $n_2 = 0$):

$$RD = \begin{cases} p_- - p_2 & \text{if } n_1 = 0 \\ p_1 - p_- & \text{if } n_2 = 0 \\ p_1 - p_2 & \text{otherwise} \end{cases}$$

In addition to covering relevant legal cases, such extensions will be crucial in linking α-protection to t-closeness.

IV. 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 [2] requires that the support of any non-empty q-block is at least k. t-closeness [3] 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 some distance 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 PD and decision attributes

[3]The original definition of α-protection (w.r.t. ED) is stated for the contingency table of a classification rule [5], assuming that the support of the rule is non-zero. This means that $a > 0$ in the 4-fold contingency table, hence $n_1 > 0$ and then ED is well-defined. The problem of discovering discrimination in contexts with only members of the protected group, is ignored in [5].

(which are binary) to sensitive attributes. Moreover, under this assumption, known distance measures between distributions collapse to variational distance. Let us recall the notion of t-closeness from [3].

Definition 4.1: 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 having sensitive value $\star$, if $|p - p_\star| \leq t$. A relational table is t-close if all q-blocks are t-close.

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 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. IV-A. 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. 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. IV-B to show that t-closeness imply $bd(t)$-protection, for some bounding function $bd()$.

A. 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 ([3], [11]), for which generalizing two or more values of a QI attribute to a common value (or removing completely the attribute from the dataset) leads to a dataset that is t'-close with $t' \leq t$.

Lemma 4.2: Consider a t-close relational table. The cover of any subset of a QI itemset is t-close.

The generalization property does not hold instead for α-protection, due to a form of the Simpson's paradox when comparing (by difference or ratios) two proportions.

Example 4.3: A real life example of the Simpson's paradox occurred in a legal case [12] 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 7 applicants admitted to department A, 2 women out of 4, and 5 men out of 6. The selection difference SD is then $4/10 - 2/7 = 0.11$. When considering applicants to department B, SD is $6/10 - 1/2 = 0.10$. Then, for all PND itemsets with exactly one item, i.e., *dept=A* and *dept=B*, the SD measure is bounded by 0.11. However, for the empty PND itemset (denoting applicants to any department), the selection difference is higher, since it amounts at $10/20 - 3/9 = 0.17$.

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

PND itemset *dept=A*
$SD = 4/10 - 2/7 = 0.11$

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

PND itemset *dept=B*
$SD = 6/10 - 1/2 = 0.10$

PND itemset empty
(both departments)
$SD = 10/20 - 3/9 = 0.17$

Fig. 2. Example of the Simpson's paradox.

B. From t-closeness to α-protection

We have observed that t-closeness proof conditions are stronger than the ones of α-protection. This motivates looking for some implication between the two analytical methods. Assume a given relational table, with fixed PND, PD and decision attributes. The problem we will investigate is: *does there exists 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 $bd(t)$-protective for an appropriate bound function $bd()$. Let us start with the SD measure (as extended in Sect. III-C).

Theorem 4.4: Fix as QIs the set of PND attributes plus the decision attribute, and as sensitive attribute the PD attribute. Let p_{pro} be the fraction of the protected group in a relational table. If the table is t-close then it is $bd(t)$-protective w.r.t. SD, where $bd(t) = min\{2t, p_{pro} + t\}$.

Intuitively, this result states that a dataset does not contain discrimination (more than a threshold $bd(t)$) if it is not possible to be more confident on the membership of an individual to the protected group (more than a threshold t) in a privacy attack assuming as QIs the set of PND attributes plus the decision 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 when compared to the proportion of the group in the overall dataset.

The upper bound $bd(t)$ provided by Thm. 4.4 is sharp.

Example 4.5: Consider a dataset with uniform distribution of the PD attribute, i.e., $p_{pro} = 0.5$. Following Thm. 4.4, we fix such an attribute as sensitive. The dataset is clearly t-close for $t = 0.5$. Consider now a PND itemset having the following contingency table:

	decision		
group	-	+	
protected	a	0	a
unprotected	0	1	1
	c	1	$a+1$

$SD = a/(a+1) - 0/1$ is arbitrarily close to the bound $1 = bd(t) = min\{2 \cdot 0.5, 0.5 + 0.5\}$ for a sufficiently large a.

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

dept=A, admitted=no
$|2/3 - 1/2| = 0.17$
dept=A, admitted=yes
$|2/7 - 1/2| = 0.21$

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

dept=B, admitted=no
$|5/8 - 1/2| = 0.125$
dept=B, admitted=yes
$|1/2 - 1/2| = 0.0$

Fig. 3. Variational distance for q-blocks of the table in Fig. 2, with *dept* and *admitted* as QI attributes and *sex* as sensitive attribute.

Later on, in Sect. VI we will discuss the case of real datasets. The converse of Thm. 4.4 does not hold in general. A counter-example is provided by the Simpson's paradox table from Example 4.3.

Example 4.6: The table in Fig. 3 is a rearranging of the rows and columns in Fig. 2. The distribution of the PD attribute in the overall dataset is uniform: 10 men and 10 women, hence $p_{pro} = 0.5$. From Example 4.3, we know that the dataset is 0.17-protective w.r.t. the SD measure, since the maximal value of SD over PND itemsets is 0.17. Fix now as QIs the attributes *dept* and *decision*, and as sensitive the attribute *sex*. The dataset is not 0.085-close, where $0.085 = bd^{-1}(0.17) = max\{0.17/2, 0.17 - 0.5\}$. The q-block of the QI itemset *dept=A, admitted=yes* includes 2 women and 5 men, with a variational distance of $|2/7 - 10/20| = 0.21$. As shown in Fig. 3, the dataset is t-close only for $t \geq 0.21$.

The conclusion of Thm. 4.4 extends to the other difference measures: ED and RD. The role of the sensitive attribute is now taken by the decision attribute.

Theorem 4.7: Fix as QIs the set of PND attributes plus the PD attribute, and as sensitive attribute the decision attribute. Let p_- be the fraction of the negative decision in a relational table. If the table is t-close then it is $bd(t)$-protective w.r.t. ED and RD, where $bd(t) = min\{2t, p_- + t\}$.

The intuitive interpretation of this result is that a dataset does not contain discrimination (more than a threshold $bd(t)$) if it is not possible to be more confident (than a threshold t) on the decision assigned to an individual (decision is the sensitive attribute here) by exploiting the differences in the fraction of positive and negative decisions between the protected and the unprotected group.

V. Discrimination data protection: dMondrian

As an application of Thms. 4.4,4.7, we can resort to inference control methods for t-closeness to the purpose of controlling the degree of α-protection of a dataset. This provides us with a means to prevent discrimination inference attacks, such as in the examples from the introduction. We consider non-perturbative methods which rely on partial reductions in details of data. 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. A well-known multidimensional recoding model for k-anonymity was proposed in [6], together with a simple and efficient greedy algorithm called *Mondrian*. Alg. 1 reports an adaption of the algorithm, called *dMondrian*, for data protection against discrimination inference. For lack of space, we report and describe only the method *Anonymize()* dealing with the assumptions of Thm. 4.7, namely when the role of QIs is played by PND and PD attributes, and the role of the sensitive attribute is played by the decision attribute.

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

```
1: if no d-allowable cut for P then
2:     return PND_ranges(P)
3: else
4:     V ← choose_PND_dimension()
5:     v ← find_median(P, V)
6:     P1 ← {t ∈ P | t[V] ≤ v }
7:     P2 ← {t ∈ P | t[V] > v }
8:     return Anonymize(P1) ∪ Anonymize(P2)
9: end if
```

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 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. 4.7 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 left unchanged by *dMondrian*. 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.

Let us now discuss how cuts are defined, starting from their definition in the original *Mondrian* algorithm [6]. We assume that the domain $dom(V)$ of any PND attribute V is ordered. This is immediate for continuous attributes, while it requires an additional input from the user for categorial attributes. A (multidimensional) cut $V \leq v$ is *allowable* [6] if it partitions a k-anonymous set of tuples into two sets (respectively, tuples t with $t[V] \leq v$ and tuples with $t[V] > v$) that are both k-anonymous. Notice that a cut $V \leq v$ is allowable iff the cut $V \leq v_m$ is allowable, where v_m is the median value of V 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* [3], 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.

Definition 5.1: Let p_- be the fraction of the negative decision in a relational table. Let $\mathcal{P}$ be a subset of tuples. A cut $V \leq v$ is *d-allowable* for $\mathcal{P}$, where V is a PND attribute and $v \in dom(V)$, if called $\mathcal{P}_1 = \{t \in P \mid t[V] \leq v\}$ and $\mathcal{P}_2 = \{t \in P \mid t[V] > v\}$, the 4-fold contingency tables of $\mathcal{P}_1$ and $\mathcal{P}_2$ (see Fig. 1) satisfy both $|p_1 - p_-| \leq t$ and $|p_2 - p_-| \leq t$.

Differently from k-anonymity, if $V \leq v$ is d-allowable, then $V \leq v_m$ 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, namely the number of subsets of tuples to be checked for d-allowable cuts, is a tractable $O(|\mathcal{R}|log|\mathcal{R}|)$, where $\mathcal{R}$ is the input table. Finally, when more than one PND attribute has a d-allowable cut, we adhere to the *Mondrian* heuristics of choosing the attribute with the widest (normalized) range of values (see function *choose_PND_dimension()* in Alg. 1).

Summarizing, for an input table $\mathcal{R}$ the call *Anonymize($\mathcal{R}$)* returns a t-close dataset, hence $bd(t)$-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 t_0-close where $t_0 = max\{|p_1 - p_-|, |p_2 - p_-|\}$ where p_1, p_2 are computed for the contingency table of the whole $\mathcal{R}$. Therefore, for $t \leq t_0$ we can only conclude that *Anonymize($\mathcal{R}$)* is $bd(t_0)$-protective.

VI. Experiments

We have implemented the *dMondrian* algorithm in Java, adopting some optimizations well-suited for divide & conquer algorithms. In particular, the input relational table is stored 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. Moreover, the calculation of medians adopts a counting sort rather than a quicksort when the domain of values of an attribute is small. The experiments reported in this section consider two classical datasets available from the UCI Machine Learning repository.[4] 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 bad/good credit rating assigned to the bank account holder. 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.

Fig. 4 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 simply the value of a discrimination measure (RD and SD) for that cover. The x-axis is the *maximum variational distance* τ for the protected and the unprotected group, where:

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

for RD. For the SD measure, $\tau = max\{|a/m_1 - p_{pro}|, |b/m_2 - p_{pro}|\}$. The bounds on discrimination measures imposed by variational distance as stated in Thm. 4.4, 4.7 are also shown in Fig. 4 – 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), 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 SD) the bounds are reached for low values of τ, or when looking at contexts with very low minimum support (this case is not shown for lack of space). Intuitively, this is due to the fact that variation of the protected group from the average is always lower or equal than variation from the unprotected group.

Fig. 4 (right) shows the scatter plot for the RD measure after the *German credit* dataset has been sanitized by *dMondrian* for the input parameter $t = 0.20$. As expected, there is no closed PND itemset with maximum variational distance $\tau > 0.20$ nor with RD $> 2 \cdot 0.20 = 0.40$. Actually, the maximum RD turns out to be 0.287. Fig. 5 (left and center) shows the minimum and maximum RD values present in the *German credit* and *Adult* datasets after being sanitized by *dMondrian* for a given parameter t. Whilst the RD measure in the original dataset can be as high as 1 (see Fig. 4 left), in the processed dataset, RD is at most $bd(t)$, where $bd()$ is the bound function in Thm. 4.7.

Let us now measure the utility of sanitized datasets by means of the standard discernibility metric[5]. Fig. 5 (right) shows the utility loss due to data sanitization by *dMondrian* at the variation of the input parameter t. Most of the degradation occurs for very high values of t, in particular for $t \geq 0.65$. Then, for $t \in [0.3, 0.6]$, the discernibility metric degrades more slowly. For low values of t, namely $t \leq 0.3$, the degradation is maximum. In summary, the analyst has to trade-off the benefits of a formal bound on discrimination measures achieved by data sanitization (see Fig. 5 (left and center)) with the loss of data quality that generalization introduces (see Fig. 5 (right)). This is analogous to what occurs in data sanitization for privacy protection. *Our approach provides the necessary tools for applying the trade-off analysis in the scenario of discrimination data sanitization.* Some options can help improving the trade-off. Fig. 5 also shows the utility metrics of the variants of *German credit* and *Adult* where the PND attributes credit_amount and age respectively are not discretized *a-priori*. Rather, it is *dMondrian* that provides a "discrimination-oriented" supervised discretization. This approach yields a visible improvement of dataset utility.

[4] http://archive.ics.uci.edu/ml

[5] Some literature on discrimination prevention (e.g., [7], [13]) measures the degree of discrimination in a dataset of predictions made by a classifier trained from a sanitized dataset. This approach, however, does not provide a measure of the utility of the sanitized dataset, but rather a measure of the reduction of the bias of a classification algorithm in "learning to discriminate".

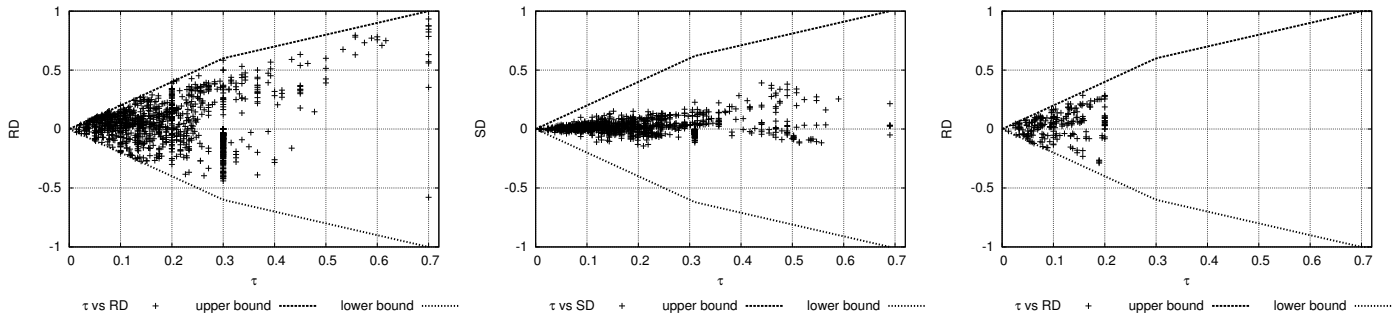

Fig. 4. *German credit* dataset. Scatter plots of maximum variational distance (τ) vs discrimination measures. Each point is a closed PND itemset with minimum support of 20 (i.e., 2%). Left and center: original dataset. Right: dataset processed by *dMondrian* with parameter $t = 0.20$.

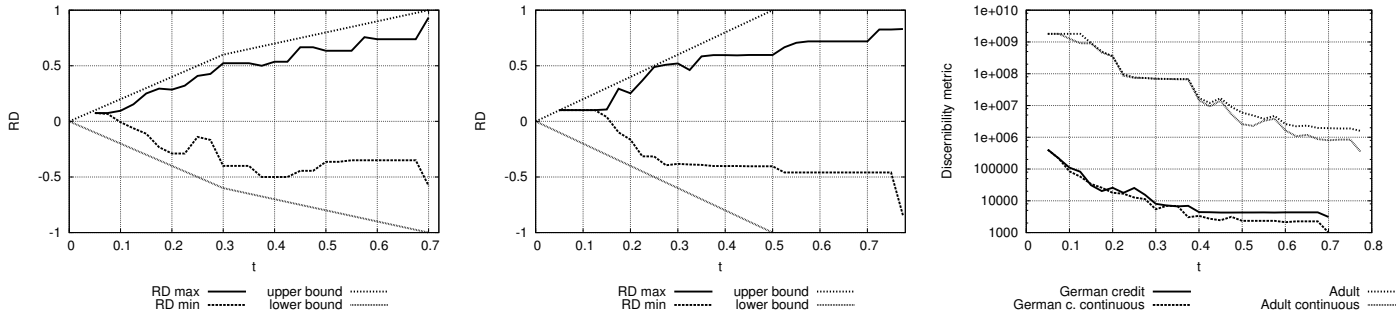

Fig. 5. Left and center: extreme values of scatter plots after processing *German credit* (left) and *Adult* (center) with *dMondrian* for a given parameter t. Right: utility metrics. In datasets with suffix "continuous", attributes are not discretized *a-priori*.

Regarding efficiency, because d-allowable cuts are more restrictive than allowable cuts, the search space of *dMondrian* is in the worst case the same of the *tMondrian* algorithm for privacy data sanitization. Moreover, checking d-allowable cuts requires a single pass over the subset $\mathcal{P}$ to compute the required 4-fold contingency tables. This is the same complexity as for checking whether a cut is allowable. Fig. 6 (left) shows the elapsed execution times of *dMondrian* for the experimental datasets on a commodity PC with Intel Core i5-2410 @ 2.30 GHz running Windows 7. The search space, and hence the elapsed time, is readily monotonic increasing with the parameter t, and with the size of the dataset (*Adult* is 42$\times$ the size of *German credit*). Letting the discretization of continuous attributes to be done at sanitization time may require up to 1.2$\times$ the time of *a-priori* discretized attributes.

The overall affordable elapsed times allow for conducting a thorough trade-off analysis given the plots on protection and utility of sanitized datasets for the full (or, for a large) range of t values. Fix a discrimination measure, say RD. For a value α, the plots in Fig. 5 (left and center) can be used to find out the largest t such that the maximum RD value for the dataset processed by *dMondrian* with parameter t is at most α. Then the plots in Fig. 5 (right) can be used to calculate the quality metrics of such processed dataset. The plots in Fig. 6 (center and right) show the result of this procedure, namely the quality of the anonymized dataset at the variation of its maximum RD and SD values. The better utility of SD compared to RD is a direct consequence of the observed fact that the distance of proportions between the protected group and the whole population is lower or equal than between the protected and the unprotected group – see Fig. 4 (left, center).

VII. Related work

Approaches for building classifiers that do not make discriminatory decisions may be based on data sanitization of the training set [4]. Existing techniques for data sanitization adopt perturbative approaches by changing values of the PD attribute or of the decision attribute. The approaches in [13], [14] massage the dataset by promoting (from - to + decision value) some individuals of the protected group and/or demoting (from + to - decision value) individuals of the unprotected group using some heuristics. [13] adopts the prediction confidence of a classifier for ranking individuals in the protected and in the unprotected groups. [14] ranks individuals on the basis of the number of contexts of possible discrimination they appear in. None of the approaches provides a formal guarantee on the level of α-protection of the sanitized dataset, as in the bounds of Thms. 4.4,4.7. As additional limitations, [13] considers the RD measure only at the grain of the whole dataset, i.e., only a single context of possible discrimination is guaranteed to be sanitized; and [14] deals with nominal attributes only, because it heavily relies on association rule mining. However, because of the intrinsic limitations of non-perturbative methods, we think that an hybrid approach trading off massaging with generalization is a promising future work.

[15] 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 subtree 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

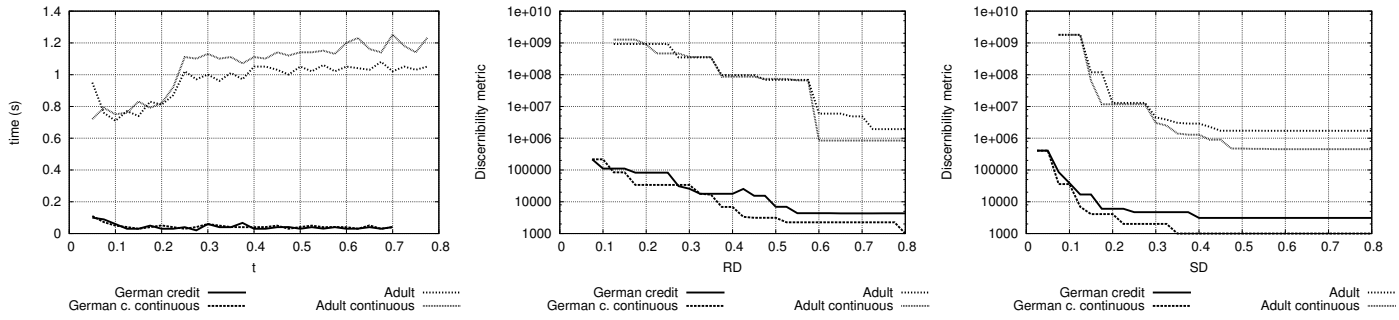

Fig. 6. Left: elapsed execution times of *dMondrian*. Center and right: utility at the variation of maximum RD and SD after processing with *dMondrian*.

this. First, [15] assumes that the (equivalent of the) PND-itemset $\mathbf{B}$ in Def. 3.2 has a minimum support, 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. Second, and more importantly, if the original dataset is α-protective, and the transformed dataset is only α'-protective with $\alpha' > \alpha$, then this is still consistent with the fact that the original dataset is t-close and $\alpha < \alpha' \leq bd(t)$, where $bd(t)$ is the bound from Thms. 4.4, 4.7.

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

VIII. Conclusions

The contribution of this paper was twofold. First, we 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 $bd(t)$-protection for a bound function $bd()$ depending on the reference discrimination measure. Second, the discovered implication allowed us for adapting the *Mondrian* multidimensional generalization algorithm to discrimination data sanitization with a formal bound on the α-protection of the sanitized dataset. No previous approach on discrimination-aware data mining can provide such a guarantee. These results represent a first step towards a more general understanding of the interplay between data anonymization and non-discrimination research in data mining. A promising 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] J. Domingo-Ferrer, "A survey of inference control methods for privacy-preserving data mining," in *Privacy-Preserving Data Mining - Models and Algorithms*. Springer, 2008, vol. 34, pp. 53–80.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.

[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity." in *Proc. of ICDE 2007*. IEEE Computer Society, 2007, pp. 106–115.

[4] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, 2013, to appear, doi:10.1017/S0269888913000039.

[5] S. Ruggieri, D. Pedreschi, and F. Turini, "Data mining for discrimination discovery," *ACM Trans. on Knowledge Discovery from Data*, vol. 4, no. 2, p. Article 9, 2010.

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

[7] T. Calders and I. Žliobaitytė, "Why unbiased computational processes can lead to discriminative decision procedures," in *Discrimination and Privacy in the Information Society*, ser. Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer, 2012, pp. 43–57.

[8] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining & Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010.

[9] S. Ruggieri, D. Pedreschi, and F. Turini, "DCUBE: Discrimination discovery in databases," in *Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010)*. ACM, 2010, pp. 1127–1130.

[10] R. M. Kanter, "Some effects of proportions on group life: Skewed sex ratios and responses to token women," *American Journal of Sociology*, vol. 82, no. 5, pp. 965–990, 1977.

[11] N. Li, T. Li, and S. Venkatasubramanian, "Closeness: A new privacy measure for data publishing," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 7, pp. 943–956, 2010.

[12] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, "Sex bias in graduate admissions: Data from Berkeley," *Science*, vol. 187, no. 4175, pp. 398–404, 1975.

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

[14] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE Trans. on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1445–1459, 2013.

[15] ——, "A study on the impact of data anonymization on anti-discrimination," in *Proc. of the IEEE ICDM 2012 Int. Workshop on Discrimination and Privacy-Aware Data Mining (DPADM)*. IEEE Computer Society, 2012, pp. 352–359.

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