Anti-discrimination Analysis Using Privacy Attack Strategies

Salvatore Ruggieri[1], Sara Hajian[2], Faisal Kamiran[3], and Xiangliang Zhang[4]

[1] Università di Pisa, Italy
[2] Universitat Rovira i Virgili, Spain
[3] Information Technology University of the Punjab, Pakistan
[4] King Abdullah University of Science and Technology, Saudi Arabia

Abstract. Social discrimination discovery from data is an important task to identify illegal and unethical discriminatory patterns towards protected-by-law groups, e.g., ethnic minorities. We deploy privacy attack strategies as tools for discrimination discovery under hard assumptions which have rarely tackled in the literature: indirect discrimination discovery, privacy-aware discrimination discovery, and discrimination data recovery. The intuition comes from the intriguing parallel between the role of the *anti-discrimination authority* in the three scenarios above and the role of an *attacker* in private data publishing. We design strategies and algorithms inspired/based on Frèchet bounds attacks, attribute inference attacks, and minimality attacks to the purpose of unveiling hidden discriminatory practices. Experimental results show that they can be effective tools in the hands of anti-discrimination authorities.

1 Introduction

Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category. Human rights laws prohibit discrimination on several grounds, such as sex, age, marital status, sexual orientation, race, religion or belief, membership of a national minority, disability or illness. Anti-discrimination authorities (equality enforcement bodies, regulation boards, consumer advisory councils) monitor, provide advice, and report on discrimination compliances based on investigations and inquiries. *Data* under investigation are studied by them with the main objective of *discrimination discovery*, which consists of unveiling contexts of discriminatory practices in a dataset of historical decision records. Discrimination discovery is a fundamental task in understanding past and current trends of discrimination, in judicial dispute resolution in legal trials, in the validation of micro-data or of aggregated data before they are publicly released. As an example of the last case, consider an employer noticing 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. Such a behavior, known as *statistical* discrimination [12], should be foreseen before data are publicly released.

Existing approaches for discrimination discovery [12, 13] are designed with two assumptions: (1) the dataset under studying explicitly contains an attribute denoting the protected-by-law social group under investigation, and (2) the dataset has not been pre-processed prior to discrimination discovery. A first major source of complexity is to tackle the case that (1) does not hold – a problem known as *indirect discrimination discovery*, where indirect discrimination refers to apparently neutral practices that take into account personal attributes correlated with indicators of race, gender, and other protected grounds and that result in discriminatory effects on such protected groups. For example, even without race records of credit applicants, racial discrimination may occur in the practice of *redlining*: applicants living in a certain neighborhood are frequently denied, as most of people living in that neighborhood belong to the same ethnic minority. A second source of complexity, ignored in the literature so far, occurs when data contain attributes denoting protected groups but such data have been pre-processed to control the (privacy) risks of revealing confidential information, i.e., assumption (2) does not hold. If the anti-discrimination authority cannot be trusted, the original data cannot be accessed, and then discrimination discovery must be performed on the processed data. We name such a case *privacy-aware discrimination discovery*. A further case in which (2) may not hold occurs when data is pre-processed to hide discriminatory decisions to the anti-discrimination authority. Since the authority has to recover the original decisions as part of its investigation, we name such a case *discrimination data recovery*.

We follow the intriguing parallel between the role of the anti-discrimination authority in discrimination data analysis and the role of an *attacker* in privacy-preserving data publishing [1,4,5] – an unauthorized (possibly malicious) entity. Several attack strategies have been proposed in the literature, which model the reasonings of an attacker and its background knowledge. *Conceptually, the role of an anti-discrimination authority is similar to the one of an attacker.* In the case of indirect discrimination discovery, the authority has to infer personal data of individuals in the dataset under investigation, namely whether she belongs to a protected group or not (this step is necessary in order to measure the degree of discrimination in decisions). We substantiate this view by showing how combinatorial attacks based on Frèchet bounds inference [3] can be deployed to this purpose. In the case of privacy-aware discrimination discovery, the parallel is even more explicit: the anti-discrimination authority has to reason as an attacker to find out as much information as possible on the membership of individuals in the protected group. We will investigate a form of attribute inference attacks for discrimination discovery from a bucketized dataset [11]. Finally, in the case of discrimination data recovery the anti-discrimination authority has the objective of re-constructing original decisions from a perturbed dataset, which, again, is a typical task of privacy attackers. By exploiting an analogy with optimality attacks [14], we will devise an approach to reconstruct a dataset that has been sanitized by means of the approach in [9]. The parallels highlighted open a new research direction consisting of applying the vast amount of methodologies and algorithms of privacy protection for discrimination data analysis.

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

$p_1 = a/n_1$
$p_2 = c/n_2$
$p = m_1/n$
$RD = p_1 - p_2$

Fig. 1. Discrimination table.

This paper is organized as follows. Section 2 formalizes the three scenarios mentioned above. Section 3 recalls basic notions of discrimination analysis. The adaptation of privacy attack approaches and algorithms to each scenario is presented in Sections 4-6. Section 7 reports experimental results. Finally, conclusions report on related work and summarize our contributions.

2 Problem Scenarios

We assume two actors: a *data owner* and an *anti-discrimination authority*. The data owner releases to the anti-discrimination authority some data either in the form of micro-data, e.g., one or more relational or multidimensional tables, or in the form of aggregate data, e.g., one or more contingency tables. The anti-discrimination authority has access to additional information, called the *background knowledge*, that is exploited to unveil contexts of possible discrimination from the released data. The case when attributes to identify protected groups are part of the released data and data are without modification is known as *direct discrimination*. This is well-studied [12,13], and in this paper our main emphasis will be on the alternative case, consisting of one of the following scenarios.

Scenario I: Indirect discrimination discovery. The released data do not include attributes that explicitly identify protected-by-law groups. The task of the anti-discrimination authority is to unveil contexts of discrimination from the released data by exploiting background knowledge (e.g., correlations between attributes) to link the unknown attributes to attributes present in the data.

Scenario II: Privacy-aware discrimination discovery. The released data include attributes that explicitly identify protected-by-law groups, but the data were pre-processed by the data owner by applying a privacy-preserving inference control method to perturb such attributes. The anti-discrimination authority has the task of unveiling contexts of discrimination by exploiting background knowledge (e.g., aggregate counts on members of the protected group) and the awareness of the inference control algorithm used to pre-process the data.

Scenario III: Discriminatory data recovery. The released data were pre-processed by the data owner by applying a discrimination prevention inference control method that perturbed the data to hide discriminatory decisions. The task of the anti-discrimination authority is to reconstruct the original data by exploiting, again, background knowledge (e.g., amount of hidden discrimination) and the awareness of the inference control algorithm. Starting from the reconstructed dataset, standard direct discrimination discovery techniques can then be adopted to unveil contexts of discrimination.

3 Measures of Group Discrimination

A critical problem in the analysis of discrimination is precisely to quantify the degree of discrimination suffered by a given group (say, an ethnic group) in a given context (say, a geographic area and/or an income range) with respect to a decision (say, credit denial). To this purpose, several discrimination measures have been defined over a 4-fold contingency table, as shown in Fig. 1, where: the *protected* group is a social group which is suspected of being discriminated against; the *decision* is a binary attribute recording whether a benefit was granted (value "+") or not (value "-") to an individual; the *total population* denotes a context of possible discrimination, such as individuals from a specific city, job sector, income, or combination thereof.

We call the 4-fold contingency table of Fig. 1 a *discrimination table*. Different outcomes between groups are measured in terms of the proportion of people in each group with a specific outcome. Fig. 1 considers the proportions of benefits denied for the protected group (p_1), the unprotected group (p_2) and the overall population (p). Differences and rates of these proportions can model the legal principle of group under-representation of the protected group in positive outcomes or, equivalently, of over-representation in negative outcomes [12]. For space reasons, we restrict to consider only *risk difference* ($\mathrm{RD} = p_1 - p_2$), which quantifies the marginal chance of the protected group of being given a negative decision. Once provided with a threshold α between "legal" and "illegal" degree of discrimination, we can isolate contexts of possible discrimination [13].

Definition 1 (α-protection). *A discrimination table is α-protective (w.r.t. the RD measure) if $RD \leq \alpha$. Otherwise, it is α-discriminatory.*

Direct discrimination discovery consists of finding α-discriminatory tables from a subset of past decision records. The original approach [13] performs a search in the space of discrimination tables of frequent (closed) itemsets. Fix a relational table whose attributes include GROUP, with values PROTECTED and UNPROTECTED, and DEC, with values + and -. An itemset is a set of items of the form $A = v$, where A is an attribute and $v \in dom(A)$, the domain of A. As usual in the literature, we write $A_1 = v_1, \ldots, A_k = v_k$ instead of $\{A_1 = v_1, \ldots, A_k = v_k\}$. Let $\mathbf{B}$ be an itemset without items over GROUP and DEC. The discrimination table associated to $\mathbf{B}$ regards the tuples in the cover of $\mathbf{B}$ as the total population. Therefore, n in Fig. 1 is the number of tuples satisfying $\mathbf{B}$ (i.e., its absolute support), and the cell values a, b, c and d are the counts of those also satisfying the cell coordinates. For instance, a is the support of the itemset "$\mathbf{B}$, GROUP=PROTECTED, DEC=$-$".

4 Scenario I: Indirect Discrimination Discovery

The release of some aggregate data over a statistical database may lead to inferences on unpublished aggregates. In particular, the inference of bounds on entries in a 4-fold contingency table, given their marginals, trace back to the

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

rel. group	decision -	decision +	
g1	$\hat{a}$	$\hat{b}$	$\hat{n}_1$
g2	$\hat{c}$	$\hat{d}$	$\hat{n}_2$
	m_1	m_2	n

group	rel. group g1	rel. group g2	
protected	e	f	n_1
unprotected	g	h	n_2
	$\hat{n}_1$	$\hat{n}_2$	n

Fig. 2. Indirect discrimination. Left: unknown contingency table. Center: known contingency table. Right: background knowledge contingency table.

1940's – and they are known as Frèchet bounds. They have been generalized to multidimensional contingency tables in the early 2000's [3]. We adopt an itemset based notation for contingency table cell entries. Let us denote by n_X the support of an itemset X in the dataset $\mathcal{R}$ under analysis: $n_X = |\{t \in \mathcal{R} | X \subseteq t\}|$. Consider now an itemset X of the form $A_1 = v_1, A_2 = v_2$, and Y of the form $A_2 = v_2, A_3 = v_3$. The itemset XY is $A_1 = v_1, A_2 = v_2, A_3 = v_3$ and the itemset $X \cap Y$ is $A_2 = v_2$. The Frèchet bounds for the support of XY are the following [3, Theorem 4]:

$$min\{n_X, n_Y\} \geq n_{XY} \geq max\{n_X + n_Y - n_{X \cap Y}, 0\} \tag{1}$$

Let us exploit Frèchet bounds to model indirect discrimination discovery by means of background knowledge on attributes (cor-)related to membership to the protected group. Consider Fig. 2. Our problem is as follows: we want to derive bounds on a discrimination measure for an unknown contingency table (left) given a known/released contingency table (center) and some additional information contained in a background knowledge contingency table (right). The known contingency table shows data on an attribute that is related to the membership to the protected group through the background knowledge contingency table. The higher the correlation the closer the (known) discrimination measures for such an attribute are to the (unknown) discrimination measures for the protected group. The unknown value a can be decomposed into the number a_1 of individuals of the group g1 plus the number a_2 of individuals the group g2. Thus, $a_1 = n_{XY}$, where X is REL. GROUP=G1, DEC=- and Y is GROUP=PROTECTED, REL. GROUP=G1. The Frèchet bounds for a_1 yield:

$$min\{\hat{a}, e\} \geq a_1 \geq max\{\hat{a} + e - \hat{n}_1, 0\} = max\{e - \hat{b}, 0\}$$

and, with similar reasonings, those for a_2 yield: $min\{\hat{c}, f\} \geq a_2 \geq max\{\hat{c} + f - \hat{n}_2, 0\} = max\{f - \hat{d}, 0\}$. Therefore, for $a = a_1 + a_2$, we have the bounds:

$$min\{\hat{a}, e\} + min\{\hat{c}, f\} \geq a \geq max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\} \tag{2}$$

These bounds have an intuitive reading. Of the n_1 individuals in the protected group, e belong to group g1 and f belong to group g2. Consider the lower bounds. At most $min\{\hat{b}, e\}$ of those e (resp., $min\{\hat{d}, f\}$ of f) have a positive decision. Therefore, the number a is at least $max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\}$. Consider now the upper bounds. At most e (resp, f) individuals of the protected

rel. group	decision -	decision +	
g1	2	0	2
g2	6	18	24
	8	18	26

group	rel. group g1	rel. group g2	
pro.	1	0	1
unp.	1	24	25
	2	24	26

Fig. 3. Sample known and background contingency tables.

group are in the g1 group (resp., g2 group), which, in turn, has at most $\hat{a}$ (resp., $\hat{c}$) negative decisions. Summarizing, the background knowledge necessary to derive the bounds for a consists of the distribution of the protected group into individuals of groups g1 and g2, namely values e and f in the background knowledge of Fig. 2. With similar means, one derives bounds for c:

$$min\{\hat{a}, g\} + min\{\hat{c}, h\} \geq c \geq max\{g - \hat{b}, 0\} + max\{h - \hat{d}, 0\}$$

Since n_1 and n_2 are in the background knowledge and m_1 is in the known contingency table, bounds for the proportions $p_1 = a/n_1$, $p_2 = c/n_2$, and $p = m_1/n$ can be readily computed. Finally, we derive a lower bound for RD:

$$RD \geq RDlb = \frac{max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\}}{n_1} - \frac{min\{\hat{a}, g\} + min\{\hat{c}, h\}}{n_2}$$

Example 1. Consider the known and background knowledge tables in Fig. 3. The Frèchet bounds on a (number of protected individuals with negative decisions) and c (number of unprotected individuals with negative decisions) are:

$$1 = min\{2, 1\} + min\{6, 0\} \geq a \geq max\{1 - 0, 0\} + max\{0 - 18, 0\} = 1$$
$$7 = min\{2, 1\} + min\{6, 24\} \geq c \geq max\{1 - 0, 0\} + max\{24 - 18, 0\} = 7$$

We have $p_1 = 1/1$, $p_2 = 7/25 = 0.28$, and then $RD = p_1 - p_2 = 0.72$.

Notice that since Frèchet bounds are sharp [3], the bounds on discrimination measures are sharp as well. Although we described the case of a single attribute related to the protected attribute, the approach can be repeated for two or more related attributes, and the best bounds can be retained at each step. The overall approach is formalized in Algorithm 1, named *FrèchetDD* for Frèchet bounds-based Discrimination Discovery. The algorithm takes as input a relational table $\mathcal{R}$, background knowledge contingency tables $\mathcal{BK}$ and a threshold α for indirect discrimination discovery of α-discriminatory contingency tables. For each closed itemset $\mathbf{B}$, the algorithm infers bounds *ctu* for its unknown contingency table. At the beginning (line 3), such bounds are the widest possible – from 0 to the support $n_{\mathbf{B}}$ of $\mathbf{B}$. For every item $A = v$, where A is not already in $\mathbf{B}$ and such that a contingency table *ctbg* relating the protected group to $A = v$ in the context $\mathbf{B}$ is available in the background knowledge (line 6), the Frèchet bounds are calculated starting from such background contingency table and from a contingency table *ctk* that is computable from $\mathcal{R}$ (line 7), as described earlier in this section. The

Algorithm 1 *FrèchetDD*$(\mathcal{R}, \mathcal{BK}, \alpha)$.

1: $\mathcal{C} \leftarrow$ { frequent closed itemsets of $\mathcal{R}$ w/o GROUP and DEC items }
2: **for** $\mathbf{B} \in \mathcal{C}$ **do**
3: ctu $= ([0, n_{\mathbf{B}}], [0, n_{\mathbf{B}}], [0, n_{\mathbf{B}}], [0, n_{\mathbf{B}}])$
4: $\mathcal{I} = \{A = v \mid$ no A-item is in $\mathbf{B}\}$
5: **for** $A = v \in \mathcal{I}$ **do**
6: **if** ctbg $= ct(\mathbf{B}, (\text{GROUP=PRO.}, \text{GROUP=UNPRO.}), (A = v, A \neq v)) \in \mathcal{BK}$ **then**
7: ctk $= ct(\mathbf{B}, (A = v, A \neq v), (\text{DEC=-}, \text{DEC=+}))$
8: ctu$'$ $\leftarrow$ Frèchet bounds from ctk and ctbg
9: ctu $\leftarrow min$(ctu, ctu$'$)
10: **end if**
11: **end for**
12: RDlb $\leftarrow$ RD lower bound from ctu
13: **if** RDlb $\geq \alpha$ **then**
14: output $\mathbf{B}$
15: **end if**
16: **end for**

bounds are used to update *ctu* (line 9). After all items are considered, the final bounds *ctu* can be adopted for computing a lower bound on the discrimination measure at hand, RD in our case, to be checked against the threshold α (line 13). The computational complexity of Algorithm 1 is $O(|\mathcal{C}| \cdot |\mathcal{BK}|)$, i.e., the product of the size of closed itemsets by the size of the background knowledge.

A remarkable instance of indirect discrimination discovery is *redlining*, a practice banned in the U.S. consisting of denying credit on the basis of residence.

Example 2. Consider a released contingency table in Fig. 4 (right) regarding benefits granted and denied in a neighborhood specified by a ZIP code. In highly segregated cities, it may be very likely that specific neighborhoods, such as ZIP=100, are mostly populated by a specific race, say, a black minority. In such a case, the ZIP code acts as a proxy of the race of the population. Fig. 4 (left) shows the contingency table for the possibly discriminated group of black people living in the specific neighborhood ZIP=100. Entries of such a table may be unknown, due to the fact that the race of individuals is not recorded in the dataset. Fix the itemset X to ZIP=100,DEC=$-$, and Y to BLACK,ZIP=100. From the released contingency in Fig. 4 (right), we know that $n_X = \hat{a}$ and $n_{X \cap Y} = \hat{n}_1$. Assume now to have, as a background knowledge, the number n_Y of black people living in the neighborhood ZIP=100. Notice that, $n_Y = n_1$. Moreover, $n_{XY} = a$. The Frèchet bounds (1) are:

$$min\{\hat{a}, n_1\} \geq n_{XY} = a \geq max\{\hat{a} + n_1 - \hat{n}_1, 0\} = max\{n_1 - \hat{b}, 0\}$$

Dividing by n_1, we get $min\{\hat{a}/n_1, 1\} \geq p_1 \geq max\{1 - \hat{b}/n_1, 0\}$. Since $c = m_1 - a$ and $n_2 = n - n_1$, bounds for $p_2 = c/n_2$ can be derived. The exact value $p = n_1/n$ is also known. Summarizing, ranges can be derived on all proportions in Fig. 1, and, *a fortiori*, on any discrimination measure based on them.

group	decision -	decision +	
black,zip=100	a	b	n_1
others	c	d	n_2
	m_1	m_2	n

group	decision -	decision +	
zip=100	$\hat{a}$	$\hat{b}$	$\hat{n}_1$
others	$\hat{c}$	$\hat{d}$	$\hat{n}_2$
	m_1	m_2	n

Fig. 4. Unknown (left) and known (right) contingency tables.

5 Scenario II: Privacy-aware Discrimination Discovery

In this scenario, the released dataset includes an attribute that explicitly identifies the protected group. However, since such an attribute is considered sensitive[5], data were pre-processed by the data owner using a privacy-preserving inference control method to diminish the correlation between such an attribute and other non-sensitive attributes. There could be different purposes for data sanitization: (1) to protect individuals' sensitive information; (2) to use data privacy as an excuse for hiding discriminatory practices. In both cases, the anti-discrimination authority has to unveil discrimination from the sanitized data.

There is a vast amount of privacy-preserving inference control methods. We investigate the scenario for one of the most popular ones, the *bucketization* method [15]. Bucketization disassociates the sensitive attributes from the non-sensitive attributes. The output of bucketization consists of two tables: a non-sensitive table (e.g., Fig. 5 left) and a sensitive table (e.g., Fig. 5 center). The non-sensitive table contains the entire non-sensitive attributes information, in addition to a group id GID (when tuples are partitioned into groups, a unique GID is assigned to each group). The sensitive table contains the sensitive values that appear in a specific group. Bucketization is a lossy join decomposition using the group id. For instance, tuple r_1 in group GID=1 has probability 25% of referring to a Muslim, Christian, Jewish, or Other individual, but it is impossible to determine which case actually holds. Thus, for the bucketized version $\mathcal{R}'$ of a dataset $\mathcal{R}$, the correlation between sensitive attribute and non-sensitive attributes is diminished. Note that in each group of our example table, every sensitive value is distinct and so the group size is equal to the parameter l in the l-diversity privacy model [11]. We assume that l is the cardinality of the attribute denoting protected and unprotected groups, e.g., the number of religions in our example.

In this context, privacy-aware discrimination discovery can be formalized as the problem of deriving bounds on a discrimination measure for an unknown contingency table (see Fig. 2 left) given the bucketized dataset $\mathcal{R}'$. Consider a subset of n tuples from $\mathcal{R}'$ for which a contingency table has to be derived. The value m_1 is known (and also $m_2 = n - m_1$) because it consists of the number of tuples with negative decision. We assume that, as background knowledge, the number n_1 of tuples regarding protected group individuals is also known (and, *a fortiori*, $n_2 = n - n_1$). Starting from those aggregate values, bounds on cell

[5] Protected group membership and private/sensitive information highly overlap [2], as e.g., for *religion*, *health status*, *genetic information* and *political opinions* attributes.

ID	Education	Job	Dec	GID
r_1	Bachelors	Engineer	-	1
r_2	Bachelors	Engineer	+	1
r_3	Doctorate	Engineer	+	1
r_4	Bachelors	Writer	+	1
r_5	Master	Engineer	+	2
r_6	Doctorate	Writer	+	2
r_7	Bachelors	Dancer	-	2
r_8	Master	Dancer	-	2
r_9	Master	Dancer	-	3
r_{10}	Master	Lawyer	+	3
r_{11}	Bachelors	Engineer	-	3
r_{12}	Bachelors	Dancer	-	3

GID	Religion
1	Muslim
1	Christian
1	Jewish
1	Other
2	Muslim
2	Christian
2	Jewish
2	Other
3	Muslim
3	Christian
3	Jewish
3	Other

	decision		
education=bachelors	-	+	
religion=muslim	a	b	3
religion≠muslim	c	d	3
	4	2	6

Fig. 5. Non-sensitive (left) and sensitive (center) tables. Right: sample unknown c.t.

values of the contingency table can be obtained by Frèchet bounds. Here, we propose to refine such bounds by exploiting the fact that in every bucket there is one and only one individual of the protected group. This yields the following bounds on a:

$$\Sigma_i min\{1, n^i_-\} \geq a \geq n_1 - \Sigma_i min\{1, n^i_+\} \tag{3}$$

where i ranges over group id's, n^i_- (resp., n^i_+) is the number of individuals with negative (resp., positive) decision with GID=i – this is available from the non-sensitive table. The bounds for c are easily derivable from those from a by noting that $c = m_1 - a$, since m_1 (the number of tuples with negative decision) is known. Similarly for $b = n_1 - a$, and for $d = n_2 - c$. Starting from them, bounds for p_1, p_2, p and discrimination measures defined over them can be computed.

Example 3. Consider the set of tuples from Fig. 5 (left) such that EDUCATION=BACHELORS. There are 6 such tuples: 4 with negative decision (r_1, r_7, r_{11}, r_{12}) and 2 with positive decision (r_2, r_4). Moreover, assume to know by background knowledge that $n_1 = 3$ out of the 6 tuples regard Muslims. This gives rise to the unknown contingency table in Fig. 5 (right). It turns out that $n^1_- = 1$, $n^2_- = 1$ and $n^3_- = 2$; and that $n^1_+ = 2$, $n^2_+ = 0$ and $n^3_+ = 0$. Therefore, we have:

$$min\{1,1\} + min\{1,1\} + min\{1,2\} = 3 \geq a \geq$$
$$2 = 3 - (min\{1,2\} + min\{1,0\} + min\{1,0\})$$

Frèchet bounds for Fig. 5 (right) would yield the strictly larger interval $min\{4,3\}$ $= 3 \geq a \geq 1 = max\{4+3-6,0\}$. Since $a + c = 4$, we derive $2 \geq c \geq 1$. Thus, $p_1 = a/n_1 \in [2/3, 3/3]$, $p_2 = c/n_2 \in [1/3, 2/3]$ and then RD $= p_1 - p_2 \in [0, 2/3]$.

Given a bucketized dataset $\mathcal{R}'$ and background knowledge $\mathcal{BK}$, Algorithm 2, whose name is *PADD* for Privacy-Aware Discrimination Discovery, formalizes the search of itemsets $\mathbf{B}$ with a lower bound for RD greater or equal than α. We assume that $\mathcal{BK}$ may also include a further lower bound $lb(a)$ for a, obtained e.g., from answers to a survey or from allegations of discrimination against the data owner. The complexity of *PADD* is linear in the number of closed itemsets.

Algorithm 2 $PADD(\mathcal{R}', \mathcal{BK}, \alpha)$.

1: $\mathcal{C} \leftarrow$ { frequent closed itemsets of $\mathcal{R}'$ w/o GROUP and DEC items }
2: **for** $\mathbf{B} \in \mathcal{C}$ **do**
3: $n \leftarrow n_{\mathbf{B}}$
4: $n_1 \leftarrow n_{\mathbf{B},\text{GROUP=PROTECTED}}$ // found in $\mathcal{BK}$
5: $m_1 \leftarrow n_{\mathbf{B},\text{DEC=-}}$ // compute from $\mathcal{R}'$
6: $a \in [a_l, a_u]$, with $a_u = min\{n_1, m_1, \Sigma_i min\{1, n^i_-\}\}$,
7: $a_l = max\{n_1 + m_1 - n, 0, n_1 - \Sigma_i min\{1, n^i_+\}, lb(a)\}$ // $lb(a)$ found in $\mathcal{BK}$
8: $c \in [c_l, c_u]$ with $c_u = m_1 - a_l$, $c_l = m_1 - a_u$
9: RDlb $\leftarrow a_l/n_1 - c_u/(n - n_1)$
10: **if** RDlb $\geq \alpha$ **then**
11: output $\mathbf{B}$
12: **end if**
13: **end for**

6 Scenario III: Discriminatory Data Recovery

To hide discrimination practices, data owners may apply discrimination prevention methods on datasets before publishing. For example, discrimination may be suppressed in the released data with minimal distortion of the decision attribute, i.e., by relabeling of some tuples to make the released dataset unbiased w.r.t. a protected group. Such discrimination prevention strategies are analogous to mechanisms of anonymization for data publication, where data anonymization is framed as a constrained optimization problem: produce the table with the smallest distortion that also satisfies a given set of privacy requirements. Such an attempt at minimizing information loss provides a loophole for attackers. The *minimality attack* [14] is one of the strategies to recover the private data from optimally anonymized data, given the non-sensitive information of individuals in the released dataset, the privacy policy, and the algorithm used for anonymization. The target of an anti-discrimination authority is precisely to reconstruct the original data from the released data, and then apply direct discrimination discovery techniques on the reconstructed data to unveil discrimination. In this sense, strategies such as minimality attacks can be readily re-proposed as a means in support of discrimination discovery.

We assume that the released dataset $\mathcal{R}'$ is changed minimally w.r.t. the original dataset $\mathcal{R}$ to suppress historical discriminatory practices. For instance, the *massaging* approach [9] changes a minimal number of tuples by promoting (from $-$ to $+$) or demoting (from $+$ to $-$) decision values. By "minimal" here it is meant that a number of changes is performed such that the RD measure for the released dataset is 0. We assume that the anti-discrimination authority knows, as background knowledge, the original value of RD, which we call *discrimination intensity* (*DiscInt*). More realistically, such a value can be estimated on the basis of declarations made by individuals who claim to have been discriminated against. We exploit the observation proposed in [10] that discrimination affects the tuples close to the decision boundary of a classifier. To determine the decision boundary, we rank tuples of the protected and unprotected groups separately

Algorithm 3 *DataRecovery($\mathcal{R}'$, DiscInt)*

1: $M \leftarrow 0.01 \cdot |protected\ group| \cdot |unprotected\ group| \,/\, |\mathcal{R}'|$
2: **for** 1 To $(DiscInt \cdot 100)$ **do**
3: $(pr, dem) \leftarrow Rank(\mathcal{R}')$
4: Change the decision of top M tuples of pr with DEC=- to DEC=+
5: Change the decision of top M tuples of dem with DEC=+ to DEC=-
6: $\mathcal{R}' \leftarrow \mathcal{R}'$ with new decision values of pr, dem
7: **end for**
8: **return** $\mathcal{R}'$

Algorithm 4 *Rank($\mathcal{R}'$)*

1: Learn a ranker L of DEC=+ using $\mathcal{R}'$ as training data
2: $pr \leftarrow$ unprotected group tuples in $\mathcal{R}'$ with DEC=-
ordered descending w.r.t. the scores by L
3: $dem \leftarrow$ protected group tuples in $\mathcal{R}'$ DEC=+
ordered ascending w.r.t. the scores by L
4: **return** (pr, dem)

w.r.t. their positive decision probabilities accordingly to a classifier trained from $\mathcal{R}'$. We change the decision values of the tuples in the decision boundaries of the protected and unprotected groups to recover the original decision labels of $\mathcal{R}$. Algorithms 3 and 4 provide the pseudocode of this discriminatory data recovery process. Procedure *DataRecovery* takes as inputs the released data $\mathcal{R}'$ and the discrimination intensity *DiscInt*. The recovery is iteratively performed to recover one percent of released data in each step[6], rather than performing the entire data recovery in a single step. The reason is that the released data with altered attributes could lead to inaccurate calculation of probability scores. The gradual data recovery process improves the quality of data continuously, and thus provides more and more accurate probability scores.

Example 4. Let us assume that an employment bureau released its historical recruitment data as shown in Fig. 6. We, as an anti-discrimination authority, suspect of hidden discriminatory patterns in the released data due to complains about the biasness of this company w.r.t. sex of applicants. However, the bureau has changed minimally the original data to suppress historical discriminatory practices of the company. We have then to recover the discriminatory data. Assume the background knowledge that $DiscInt = 40\%$. We first calculate the positive decision probabilities for all tuples by adopting a probabilistic classifier (e.g., Naive Bayes), and then order the tuples of males and females w.r.t. these probability scores separately, as shown in Fig. 6. $DiscInt = 40\%$ in these 10 tuples implies that two tuples were relabeled for suppressing the sex discrimination, i.e., one male (resp. female) tuples was relabeled to negative (resp. positive)

[6] The number of modifications M at each step is determined as follow. Let n_1 (resp., n_2) be the size of the protected (resp., unprotected) group in $\mathcal{R}'$. A total of $M \cdot DiscInt \cdot 100$ tuples are demoted (resp., promoted) to move from RD = 0 to RD = $(M \cdot DiscInt \cdot 100)/n_1 + (M \cdot DiscInt \cdot 100)/n_2 = DiscInt$. By solving the equation, we get $M = 0.01 \cdot n_1 \cdot n_2/(n_1 + n_2)$ where $n_1 + n_2 = |\mathcal{R}'|$.

Sex	Ethnicity	Degree	Job Type	Dec	Prob
m	native	h.s.	board	+	98%
m	native	h.s.	board	+	98%
m	native	univ.	board	+	89%
m	*non-nat.*	*h.s.*	*health*	*-*	*47%*
m	non-nat.	univ.	health	-	30%

Sex	Ethnicity	Degree	Job Type	Dec	Prob
f	native	h.s.	board	+	93%
f	native	none	health	+	76%
f	*native*	*h.s.*	*edu.*	*+*	*51%*
f	non-nat.	univ.	edu.	-	2%
f	non-nat.	univ.	edu.	-	2%

Fig. 6. Sample job-application relation with positive decision probability scores.

decision. The procedure *DataRecovery* (steps 4, 5) selects tuples close to the decision boundaries as candidates for correction. In our example, those with *Prob* around 50% (shown in red in Fig. 6) will have decision values changed: the male (resp., female) tuple is promoted from $-$ to $+$ (resp., demoted from $+$ to $-$).

7 Experiments

In this section, we report experiments on three classical datasets available from the UCI Machine Learning repository (http://archive.ics.uci.edu/ml): *German credit*, which consists of 1000 tuples with attributes on bank account holders applying for credit; *Adult*, which contains 48848 tuples with census attributes on individuals; and *Communities and Crimes*, which contains 1994 tuples and describes the criminal behavior of different communities in the U.S.

Scenario I: Indirect discrimination discovery. We experimented the Frèchet bounds approach of Algorithm 1 on the *German credit* and *Adult* datasets. For the former dataset, the personal status attribute, denoting the protected group of non-single females, was removed before applying the algorithm. For the latter dataset, the same approach was taken for the protected group of non-Whites. Closed itemsets are computed by setting a minimum support threshold of 20, i.e., 2%, for *German credit* and of 48, i.e., 0.1%, for *Adult*. We simulate the availability of background knowledge contingency tables (*ctbg* in Algorithm 1) by computing them from the original dataset. In order to evaluate the impact of the size of the available background knowledge, only a random number ni of the items in the set $\mathcal{I}$ (see line 4 of Algorithm 1) are actually looked up. We experiment with $ni = 1$, i.e., the anti-discrimination authority has knowledge of only one related item, with $ni = 5$, and with an optimistic $ni = 30$. Fig. 7 (top) shows the top 10K contingency tables w.r.t. the lower bound on the RD measure computed by Algorithm 1 for the *German credit* and the *Adult* datasets. The plots report the distributions of the contingency tables for which the lower bound is greater or equal than a given threshold α. It is shown the total number of such contingency tables (labels *total*) and the number of them for which the lower bound coincides with the upper bound (labels *exact*), namely, those for which Frèchet bounds are exact. Two facts can be concluded. First, if the inferred lower bound is higher than 0.3, then it is exact with high probability (95% or higher). Second, the higher is ni the higher are the inferred lower bounds. Fig. 7 (middle) shows the recall of the approach, namely the proportion of contingency tables

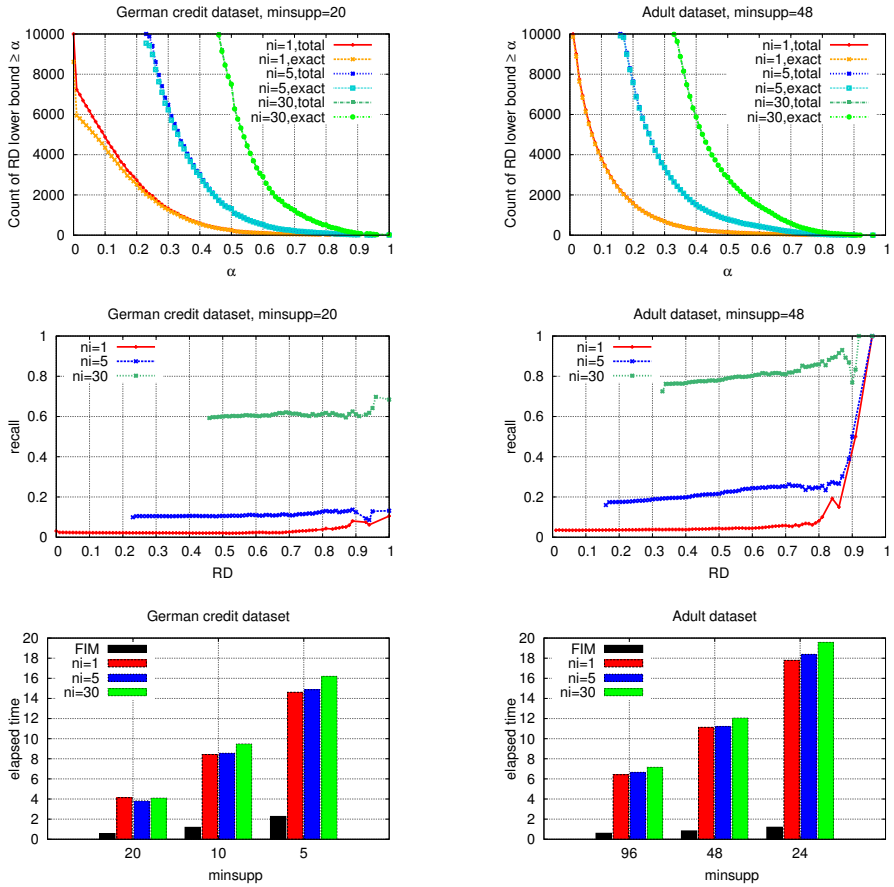

Fig. 7. Scenario I: precision (top), recall (middle), elapsed time (bottom) of *FrèchetDD*.

with a given RD value of v that have been actually inferred a lower bound of v. The plots provide an estimate of the effectiveness of the indirect discrimination discovery approach for a given amount of background knowledge (ni). Finally, Fig. 7 (bottom) shows the elapsed times for the various experiments, including the time (denote by FIM) required for extracting closed itemsets. The time required by Algorithm 1 mainly depends on the number of closed itemsets, while the size of the background knowledge (the ni parameter) has a residual impact.

Scenario II: Privacy-aware discrimination discovery. We experimented with a subset of 200 tuples (resp., 14160) from the *German credit* (resp., *Adult*) dataset, randomly selected with a balanced distribution of the personal status (resp., race) attribute. Such a distribution is required to apply l-diversity data sanitization, where $l = 4$ (resp., $l = 2$) is the number of values of the personal status (resp., race) attribute, including the protected group of non-single females (resp., non-Whites). Tuples have been partitioned into groups of l elements with

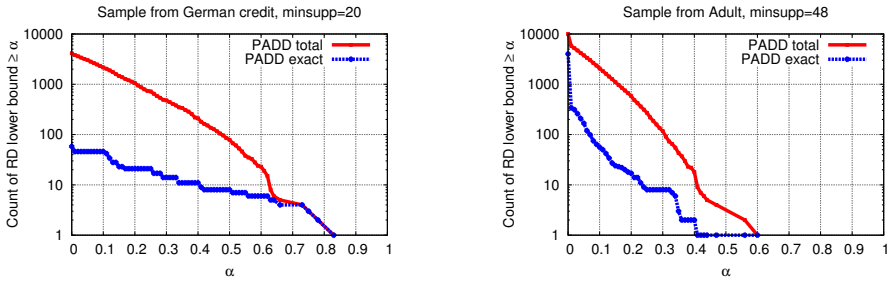

Fig. 8. Scenario II: precision of *PADD*.

distinct personal status (resp., race). Fig. 8 shows the distributions of the RD lower bound for the top 10K contingency tables processed by Algorithm 2. The lower bound $lb(a)$ at line 7 is randomly generated in the interval from 0 to the actual value of a. Contrasting the plots with Fig. 7, we observe that the number and the exactness of the bounds inferred for RD is much lower than in the case of scenario I – notice that the plots in Fig. 8 are logscale in the y-axis. This is expected since the assumptions on the background knowledge exploitable in this scenario are much weaker than in scenario I. The anti-discrimination authority is assumed to know only the number of protected group individuals in the context under analysis as well as a lower bound on those with negative decision. In scenario I, correlation with groups whose decision value is precisely known is instead assumed. Nevertheless, scenario I and II are not mutually exclusive, and a hybrid approach could be applied to improve the inferred bounds.

Scenario III: Discriminatory data recovery. We conducted experiments on the *Adult* dataset, with protected group females, and on the *Crimes and Communities* dataset, with protected group blacks. As background knowledge, we assume to know that discrimination intensity is *DiscInt*=43% in *Crimes and Communities*, and *DiscInt*=19.45% in *Adult*. These numbers can be calculated from the original datasets. We proceeded with suppressing these differences by the method of massaging [9] before releasing the datasets. We then adopted the reverse engineering approach of Algorithm 3 to reconstruct the original data.

The original dataset $\mathcal{R}$ can be used as ground truth for performance comparison. We measure the performances of Algorithm 3 by means of *Recall* and *Precision*. The *Recall* calculates how much massaged tuples were corrected, while the *Precision* measures how much corrected tuples were among those actually massaged. Algorithm 3 recovers data by iterations. In order to evaluate the performance at each iteration step, we compute recall and precision at the t-th step by $Recall = (\sum_{i=1}^{t} C_i)/(DiscInt \cdot |\mathcal{R}|)$ and $Precision = (\sum_{i=1}^{t} C_i)/(2 \cdot t \cdot M)$, respectively, where C_i is the number of tuples whose decision values are successfully corrected at the i-th step, and M is as in Algorithm 3. These sequential performance measures are shown in Fig. 9. The figure shows that our proposed method gives very promising results by reconstructing the *Adult* and the *Crimes and Communities* datasets (massaged to suppress 19.45% and 43% *DisctInt* resp.) with high precision and recall. We can observe that the method recov-

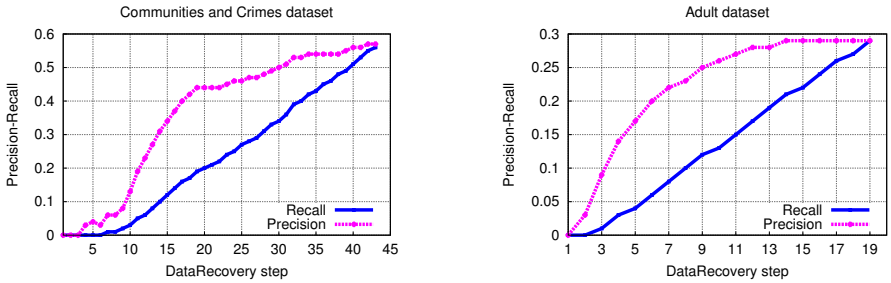

Fig. 9. Scenario III: performances of *DataRecovery*.

ers the *Communities and Crimes* dataset with 59% precision and can assist the authorities to identify the suppressed discriminatory patterns. The recovery process is relatively less accurate over the *Adult* dataset due to a higher imbalance between protected and unprotected groups. Fig. 9 also shows the advantage of stepwise data recovery and refined probability score calculation. Our recovery algorithm continues to be more precise in the identification of perturbed tuples on the later recovery steps. This gradual and significant improvement in the performance can be attributed to the calculation of probability scores over the intermediary recovered and relatively corrected data.

8 Conclusions

Related work. Discrimination analysis is a multi-disciplinary problem, involving sociological causes, legal argumentations, economic models, statistical techniques [12]. More recently, the issue of anti-discrimination has been considered from a data mining perspective. Some proposals are oriented to using data mining to measure and discover discrimination [13]; other proposals [6, 9] deal with preventing data mining from becoming itself a source of discrimination. Summaries of contributions in discrimination-aware data mining are collected in [2,12]. The term privacy-preserving data mining (PPDM) was coined in 2000, although related work on inference control and statistical disclosure control (SDC) started in the 1970s. A detailed description of different PPDM and SDC methods can be found in [1, 5, 8]. Data are sanitized prior to publication and analysis (according to some privacy criterion). In some cases, however, an attacker can still re-identify sensitive information from the sanitized data using varying amounts of skill, background knowledge, and effort. Summaries of contributions and taxonomies of different privacy attacks strategies are collected in [1,4]. Moreover, the problem of achieving simultaneous discrimination prevention and privacy protection in data publishing and mining was recently addressed in [7]. However, to the best of our knowledge, this is the first work that exploits tools from the privacy literature to the purpose of discovering discriminatory practices under hard conditions such as those of three scenarios considered.

Conclusion. The actual discovery of discriminatory situations and practices, hidden in a dataset of historical decision records, is an extremely difficult

task. The reasons are as follows: First, there are a huge number of possible contexts may, or may not, be the theater for discrimination. Second, the features that may be the object of discrimination are not directly recorded in the data (scenario I). Third, the original data has previously been pre-processed due to privacy constraints (scenario II) or for hiding discrimination (scenario III). In this paper, we proposed new discrimination discovery methods inspired by the privacy attack strategies for the three scenarios above. The results of this paper can be considered a promising step towards the systematic application of techniques from the well explored area of privacy-preserving data mining to the emerging and challenging area of discrimination discovery.

References

1. Chen, B.C., Kifer, D., LeFevre, K., Machanavajjhala, A.: Privacy-preserving data publishing. Foundations and Trends in Databases 2(1-2), 1–167 (2009)
2. Custers, B.H.M., Calders, T., Schermer, B.W., Zarsky, T.Z. (eds.): Discrimination and Privacy in the Information Society, Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 3. Springer (2013)
3. Dobra, A., Fienberg, S.E.: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Proc. of the National Academy of Sciences 97(22), 11185–11192 (2000)
4. Domingo-Ferrer, J.: A survey of inference control methods for privacy-preserving data mining. In: Aggarwal, C.C., Yu, P.S. (eds.) Privacy-Preserving Data Mining, Advances in Database Systems, vol. 34, pp. 53–80. Springer (2008)
5. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. 42(4), Article 14 (2010)
6. Hajian, S., Domingo-Ferrer, J.: A methodology for direct and indirect discrimination prevention in data mining. IEEE Trans. on Knowledge and Data Engineering 25(7), 1445–1459 (2013)
7. Hajian, S., Domingo-Ferrer, J., Farràs, O.: Generalization-based privacy preservation and discrimination prevention in data publishing and mining. Data Mining and Knowledge Discovery pp. 1–31 (2014), doi:10.1007/s10618-014-0346-1
8. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., de Wolf, P.P.: Statistical Disclosure Control. Wiley (2012)
9. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1–33 (2012)
10. Kamiran, F., Karim, A., Zhang, X.: Decision theory for discrimination-aware classification. In: Proc. IEEE ICDM 2012. pp. 924–929 (2012)
11. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: *L*-diversity: Privacy beyond *k*-anonymity. ACM Trans. on Knowledge Discovery from Data 1(1), Article 3 (2007)
12. Romei, A., Ruggieri, S.: A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review pp. 1–57 (2014), doi:10.1017/S0269888913000039
13. Ruggieri, S., Pedreschi, D., Turini, F.: Data mining for discrimination discovery. ACM Trans. on Knowledge Discovery from Data 4(2), Article 9 (2010)
14. Wong, R.C.W., Fu, A.W.C., Wang, K., Pei, J.: Minimality attack in privacy preserving data publishing. In: Proc. of VLDB 2007. pp. 543–554 (2007)
15. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: Proc. of VLDB 2006. pp. 139–150 (2006)