A Study on the Impact of Data Anonymization on Anti-discrimination

Sara Hajian and Josep Domingo-Ferrer (*Fellow, IEEE*)
Department of Computer Engineering and Math., UNESCO Chair in Data Privacy
Universitat Rovira i Virgili
Email: {sara.hajian, josep.domingo}@urv.cat

***Abstract*—In last years, data mining has raised some concerns related to privacy invasion of the individuals and potential discrimination based on the extracted patterns and profiles. Efforts at fighting against these risks have led to developing privacy preserving data mining (PPDM) techniques and anti-discrimination techniques in data mining. However, there is an evident gap between the large body of research in data privacy technologies and the recent early results on anti-discrimination technologies. This context presents a study on the relation between data anonymization from privacy technologies literature and anti-discrimination. We discuss how different data anonymization techniques have impact on discriminatory biased datasets.**

***Keywords*-Privacy; Anti-discrimination; Data anonymization; Generalization; Suppression; Classification rules**

I. Introduction

Data mining is an increasingly important technology for extracting useful knowledge hidden in large collections of data. There are, however, negative social perceptions about data mining, among which potential privacy invasion and potential discrimination. Privacy invasion occurs when the values of published sensitive attributes can be linked to specific individuals (or companies). Discrimination is unfair or unequal treatment of people based on membership to a category, group or minority, without regard to individual characteristics. In parallel to the development of privacy legislation [7], anti-discrimination legislation has undergone a remarkable expansion [3], [8], and it prohibits discrimination against *protected groups* on the grounds of race, color, religion, nationality, sex, marital status, age and pregnancy, and in a number of settings, like credit and insurance, personnel selection and wages, and access to public services.

The problem of privacy protection in data mining has been extensively studied in the last decade, under the name of privacy preserving data mining (PPDM, [2], [17]). The main goal of PPDM is to develop algorithms and techniques for modifying the original data in some way, so that the private data/knowledge remain private even after the mining process. PPDM has become increasingly popular because it allows sharing sensitive data for analysis purposes. Necessary steps of PPDM are: i) define the privacy model to prevent a specific kind of attacks (*e.g.* k-anonymity against record linkage attacks); ii) apply proper anonymization techniques (*e.g.* generalization) to satisfy the requirements of the privacy model; iii) measure data quality loss as a side effect of data distortion (the measure can be general or regarding specific data mining tasks).

The issue of anti-discrimination has recently been considered from a data mining perspective [19]. A substantial part of the existing literature on anti-discrimination in data mining is oriented to *discovering* and *measuring* discrimination [19], [20], [21], [25]. Other contributions deal with *preventing* discrimination. Data mining approaches for discrimination discovery have followed the legal principle of *under-representation* to unveil contexts of possible discrimination against protected groups (*e.g.*, minority race). This is done by extracting classification rules from a dataset of historical decision records (inductive part); then, rules are ranked according to some *legally-grounded* measures (*e.g.* $slift$ [20]) of discrimination (deductive part).

Beyond discrimination discovery, preventing knowledge-based decision support systems from making discriminatory decisions is a more challenging issue. In fact, *discrimination prevention in data mining* (DPDM) consists of extracting models that do not lead to discriminatory decisions even if trained from a dataset containing them. Discrimination prevention approaches can be classified according to the phase of the data mining process in which they operate: pre-processing methods [13], [14], [10], [11] transform data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision models can be mined from the transformed data; in-processing methods [15], [5], [27] modify data mining algorithms in such a way that the resulting models do not contain unfair decisions; finally, post-processing [20] pays attention to modifying the resulting data mining models, instead of cleaning the original dataset or changing the data mining algorithms. In this paper, we concentrate on discrimination prevention based on pre-processing.

Considering the literature, there is an evident gap between the large body of research in data privacy technologies and the recent early results on anti-discrimination technologies. In this paper, for the first time, we study the relation between data anonymization and anti-discrimination. In other words, we analyze how different data anonymization techniques (*e.g.*, generalization) have impact on anti-discrimination (*e.g.*, discrimination prevention). When we anonymize the original data to achieve the requirement of the privacy model (*e.g.*, k-anonymity), what will be happen to the

discriminatory bias contained in the original data? Our main motivation to do this study is finding answer for three important questions. First, can providing protection against privacy attacks also achieve anti-discrimination? Second, can we adapt and use some of the data anonymization techniques (*e.g.*, generalization) for discrimination prevention? Third, can we design methods based on data anonymization to make the original data protected against both privacy and discrimination risks.

The rest of this article is organized as follows. Section II introduces basic definitions and concepts used throughout the paper. Data anonymization techniques and discrimination measures are presented in Section III and IV, respectively. In Section V, we study the impact of different data anonymization techniques on anti-discrimination. Finally, Section VI summarizes conclusions and identifies future research topics in this context.

II. Basic notions

Given the data table $\mathcal{DB}(A_1, \cdots, A_n)$, a set of attributes $\mathcal{A} = \{A_i, \cdots, A_n\}$, and a record/tuple $t \in \mathcal{DB}$, $t[A_i, \cdots, A_j]$ denotes the sequence of the values of $A_i, \cdots, A_j$ in t, where $\{A_1, \cdots, A_j\} \subseteq \{A_i, \cdots, A_n\}$. Let $\mathcal{DB}[A_i, \cdots, A_j]$ be the projection, maintaining duplicate records, of attributes $A_i, \cdots, A_j$ in $\mathcal{DB}$. Let $|\mathcal{DB}|$ be the cardinality of $\mathcal{DB}$, that is, the number of records it contains.

The attributes $\mathcal{A}$ in a data table $\mathcal{DB}$ can be classified into several categories. *Identifiers* are attributes that uniquely identify individuals in the database, like *Passport number*. A *quasi-identifier* (QI) is a set of attributes that, in combination, can be linked to external identified information for re-identifying an individual; for example, *Zipcode*, *Age* and *Sex* form a quasi-identifier because together they are likely to be linkable to single individuals in external public identified data sources (like the electoral roll). *Sensitive attributes* (S) are those that contain sensitive information, such as *Disease* or *Salary*. *Non-sensitive attributes* are all attributes that do not fall into the previous three categories. Civil rights laws [3], [8], [26], explicitly identify the attributes to be protected against discrimination. For instance, U.S. federal laws [26] prohibit discrimination on the basis of race, color, religion, nationality, sex, marital status, age and pregnancy. In our context, we consider these attributes as potentially discriminatory (PD) attributes. Let DA be a set of PD attributes in $\mathcal{DB}$ specified by law. Comparing privacy legislation [7] and anti-discrimination legislation [8], [26], depend on the context, PD attributes can overlap with QI attributes (*e.g. Sex*, *Age*, *Marital_status*) and/or sensitive attributes (*e.g. Religion* in some applications) and/or non-sensitive attributes. A *class* attribute $A_c \in \mathcal{A}$ is a fixed attribute of $\mathcal{DB}$, also called *decision* attribute reporting the outcome of a decision made of an individual record. An example is attribute *Credit_approved*, which can be *yes* or *no*.

A domain D_{A_i} is associated with each attribute A_i to indicate the set of values that the attribute can assume. These set of domains of attributes in the original data set is called *ground*. An *item* is an expression $A_i = q$, where $A_i \in \mathcal{A}$ and $q \in D_{A_i}$, *e.g. Sex=female*. A *class item* $A_i = q$ is an item where $A_i = A_c$ and $q \in D_{A_c}$, *e.g. Credit_approved=no*. An *itemset* X is a collection of one or more items, *e.g.* $\{Sex = female, Age = 30\}$. We consider $A_i = q, \forall q \in D_{A_i}$ to be a PD item, where $A_i \in DA$, *e.g.* $Race = q$ is a PD item for any race q, where $DA = \{Race\}$. This definition is compatible with the law. For instance, the U.S. Equal Pay Act [26] states that: "a selection rate for **any** race, sex, or ethnic group which is less than four-fifths of the rate for the group with the highest rate will generally be regarded as evidence of adverse impact". An item $A_i = q, \forall q \in D_{A_i}$ is potentially non-discriminatory (PND) item if $A_i \notin DA$, *e.g.* $Age = 30$ where $DA = \{Race\}$. A PD itemset is an itemset containing only PD items, which we also call it protected-by-law (or protected, for short) groups. A PND itemset is an itemset containing only PND items.

The *support* of an itemset X in a data table $\mathcal{DB}$ is the number of records that contain X, *i.e.* $supp_{\mathcal{DB}}(X) = |\{t_i \in \mathcal{DB} | X \subseteq t_i\}|$. A *classification rule* is an expression $r : X \rightarrow C$, where C is a class item and X is an itemset containing no class item, *e.g.* $\{Sex = Black,\ City = NYC\} \rightarrow Credit\_approved = no$. The itemset X is called the premise of the rule. The *confidence* of a classification rule, $conf_{\mathcal{DB}}(X \rightarrow C)$, measures how often the class item C appears in records that contain X. Hence, if $supp_{\mathcal{DB}}(X) > 0$ then

$$conf_{\mathcal{DB}}(X \rightarrow C) = \frac{supp_{\mathcal{DB}}(X, C)}{supp_{\mathcal{DB}}(X)} \quad (1)$$

Confidence ranges over $[0, 1]$. We omit the subscripts in $supp_{\mathcal{DB}}(\cdot)$ and $conf_{\mathcal{DB}}(\cdot)$ when there is no ambiguity. Also, the notation readily extends to negated itemsets $\neg X$. A *frequent classification rule* is a classification rule with support and confidence greater than respective specified lower bounds. Let $\mathcal{FR}$ be the set of frequent classification rules extracted from $\mathcal{DB}$.

III. Data Anonymization Techniques

Different privacy models have been proposed in literature to prevent different kinds of privacy attacks [9], [12]. Here, we introduce the important models in each group of attacks:

K-anonymity [22], [24] requires that at least k released records match each value combination of the QI. The goal of k-anonymity is to prevent re-identification through record linkage attacks between the released data and external identified data sources.

L-diversity [18] requires at least l distinct values for the sensitive attribute in each group of QI. The goal of l-diversity is to prevent attribute linkage attacks which occurs when

attacker is able to link a record owner to a sensitive attribute in the released data.

T-closeness [16] requires the distribution of a sensitive attribute in any group on QI to be close to the distribution of the attribute in the overall table to prevent attribute linkage attack.

Differential privacy [6] requires that removal/addition of a single record from/to a database do not significantly affect the answer returned for a certain query.

Typically, the original data table does not satisfy the requirement of the respective privacy model (*e.g.,* k-anonymity) and, before being published, it must be modified through data anonymization techniques. There are different categories of data anonymization techniques: generalization, suppression, permutation, and perturbation. Generalization and suppression [23], [22] are the most common data anonymization techniques to achieve the requirement of different privacy models. Then, in this paper we focus on generalization and suppression techniques.

A generalization replaces QI attribute values with a generalized version of them using the generalization taxonomy tree of QI attributes, *e.g.* Figure 1. Five possible generalization schemes [9] are summarized in the following:

In *full-domain generalization*, all values in an attribute are generalized to the same level of the taxonomy tree. For example, consider Figure 1, if *Lawyer* and *Engineer* are generalized to *Professional*, then it also requires generalizing *Dancer* and *Writer* to *Artist*. In *subtree generalization*, at a nonleaf node, either all child values or none are generalized. For example, consider Figure 1, if *Engineer* is generalized to *Professional*, it also requires generalizing *Lawyer* to *Professional*, but *Dancer* and *Writer* can remain ungeneralized.

Sibling generalization is similar to the subtree generalization, except that some siblings may remain ungeneralized. For example, consider Figure 1, if *Engineer* is generalized to *Professional*, *Lawyer* can remain ungeneralized. In all of the above schemes, if a value is generalized, all its instances are generalized. Such schemes are called *global recoding*. In *cell generalization*, also known as *local recoding*, some instances of a value may remain ungeneralized while other instances are generalized. For example, consider Figure 1, *Female* in one record in a data table is generalized to *Any-sex*, while *Female* in another record can remain ungeneralized. *Multi-dimensional generalization* flexibly allows two QI groups, even having the same value, to be independently generalized into different parent groups. For example, consider Figure 1, ⟨*Engineer, Male*⟩ can be generalized to ⟨*Engineer, Any-sex*⟩ while ⟨*Engineer, Female*⟩ can be generalized to ⟨*Professional, Female*⟩.

A suppression consists in suppressing some values of the QI attributes for some (or all) records. Three possible suppression schemes are *record suppression*, *value suppression* and *cell suppression*. Record suppression refers to suppressing an entire record. Value suppression refers to suppressing every instance of a given value in a table. Cell suppression refers to suppressing some instances of a given value in a table.

IV. Discrimination Measures

The legal principle of under-representation has inspired existing approaches for discrimination discovery based on rule/pattern mining. Several legal concepts (*e.g.*, direct and indirect discrimination) and reasoning have been translated into rule filtering and deduction [20]. Direct discrimination occurs when the input data contain PD attributes, *e.g.*, *Race*. Indirect discrimination occurs when the input does not contain PD attributes, but discriminatory decisions against protected groups might be indirectly made because of the availability of some background knowledge; for example, discrimination against black people might occur if the input data contain *Zipcode* as attribute (but not *Race*) and one knows that the specific zipcode is mostly inhabited by black people. Since in this paper we assume that the input data contain protected groups, which is a reasonable assumption for attributes such as *Sex*, *Age* and *Pregnancy/marital status*, we concentrate on *direct discrimination*. In the following we omit the word "direct" for brevity.

Given DA and starting from a database $\mathcal{DB}$ of historical decision records, a frequent classification rule $X \rightarrow C$ in $\mathcal{FR}$ is PD when $X = A, B$ with A is a non-empty PD itemset, B a PND itemset and C a class item denying some benefit. In fact, A is under-represented in context of B with respect to the corresponding positive decision $\neg C$. For example, given $DA = \{Sex\}$, rule $\{Sex = female,\ City = NYC\} \rightarrow\ Credit\_approved = no$ is a PD rule about deny credit (the decision C) to women (in the protected group A) among those living in NYC (the context B). Then, the degree of under-representation should be measured over each PD rule using a legally-grounded measures a by family of the legally-grounded measures, such as those introduced in Pedreschi et al. [20]:

Definition 1. *Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{DB}$ with $conf(B \rightarrow C) > 0$. The* extended lift [1] (elift) *of the rule is*

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} \qquad (2)$$

In fact, *elift* is the ratio of the proportions of benefit denial, *e.g.* credit denial, between the protected groups and all people who were not granted the benefit, *e.g.* women versus all men and women who were denied credit, in the given context, *e.g.* those who live in NYC.

Definition 2. *Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{DB}$ with $conf(\neg A, B \rightarrow C) > 0$. The*

[1] Discrimination occurs when a higher proportion of people not in the group is able to comply.

selection lift (slift) [2] *of the rule is*

$$slift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(\neg A, B \rightarrow C)} \quad (3)$$

In fact, the *slift* is the ratio of the proportions of benefit denial, *e.g.* credit denial, between the protected and unprotected groups, *e.g.* women and men resp., in the given context, *e.g.* those who live in NYC. A special case of *slift* occurs when we deal with nonbinary attributes, for instance when comparing the credit denial ratio of blacks with the ratio for other groups of the population. This yields a third measure called *contrasted lift (clift)* which, given A as a single item $a = v_1$ (*e.g.* black race), compares it with the most favored item $a = v_2$ (*e.g.* white race).

Whether the rule is to be considered discriminatory can be assessed by thresholding one of the above measures as follows:

Definition 3. *Let f be one of the measures from Definitions 1-2, and $\alpha \in R$ be a fixed threshold*[3] *and let A be a PD itemset. A PD classification rule $c = A, B \rightarrow C$ is α-protective w.r.t. f if $f(c) < \alpha$. Otherwise, c is α-discriminatory.*

Building on Definition 3, we introduce the notion of α-protection for a data table.

Definition 4 (α-protective data table)**.** *Let $\mathcal{DB}(A_1, \cdots, A_n)$ be a data table, DA a set of PD attributes associated with it, and f be one of the measures in Definitions 1-2. $\mathcal{DB}$ is said to satisfy α-protection or to be α-protective w.r.t. DA and f if each PD frequent classification rule $c : A, B \rightarrow C$ extracted from $\mathcal{DB}$ is α-protective, where A is a PD itemset and B is a PND itemset.*

Releasing an α-protective version of an original data table is desirable to prevent discrimination with respect to DA. If the original data table is not α-protective w.r.t. DA, it must be modified before being published by applying a proper data distortion method (*i.e.* pre-processing approach). The existing pre-processing discrimination prevention methods are based on data perturbation, either by modifying class attribute values [13], [11] or by modifying PD attribute values [11] of the training data. One of the drawbacks of data perturbation is that it is sometimes not accepted by researchers, because they do not trust the results obtained on perturbed data [4]. Hence we focus here on generalization and suppression.

[2]Discrimination occurs when a group is treated "less favorably" than others.

[3]α states an acceptable level of discrimination according to laws and regulations. For example, the four-fifths rule of U.S. Federal Legislation sets $\alpha = 1.25$.

Table I
PRIVATE DATA TABLE WITH BIASED DECISION RECORDS

ID	Sex	Job	Age	Credit_approved
1	Male	Engineer	35	Yes
2	Male	Engineer	38	Yes
3	Male	Lawyer	38	No
4	Female	Writer	30	No
5	Male	Writer	30	Yes
6	Female	Dancer	31	No
7	Female	Dancer	32	Yes

V. Data Anonymization Techniques and Anti-discrimination

In this section, we study how different generalization and suppression schemes have impact on anti-discrimination. In other words, when we anonymize $\mathcal{DB}$ to achieve the requirement of the privacy model, *e.g.* k-anonymity, w.r.t. QI, what will be happen to α-protection of $\mathcal{DB}$ w.r.t. DA? The problem could be investigated with respect to different possible relations between PD attributes and other attributes (*i.e.* QI, sensitive and non-sensitive attributes) in $\mathcal{DB}$. In this context, we consider the general case where all attributes are QI expect the class/decision attribute. Then, each QI attribute can be PD or not. In summary, the following relations are assumed: (1) $QI \cap C = \emptyset$, (2) $DA \subseteq QI$. As mentioned in Section II, PD attributes can overlap with QI and/or sensitive and/or non-sensitive attributes. Considering all attributes as QI so that $DA \subseteq QI$ can cover all the above cases.

Example 1. Table I presents raw customer credit data, where each record represents a customer's specific information. *Sex*, *Job*, and *Age* can be taken as QI attributes. The class attribute has two values, *Yes* and *No*, to indicate whether or not the customer has received credit. Suppose the privacy model is k-anonymity and $k = 2$. Table I does not satisfy 2-anonymity w.r.t. $QI = \{Sex, Job, Age\}$.

Example 2. Continuing Example 1, suppose $DA = \{Sex\}$, $\alpha = 1.2$ and $f = slift$. Table I does not satisfy 1.2-protection w.r.t. f and DA since for frequent PD rule c equal to $\{Sex = female\} \rightarrow Credit\_approved = no$ we have $slift(c) = \frac{2/3}{1/4} = 2.66$. Then Table I is biased w.r.t. women.

A. Global recoding generalizations and anti-discrimination

In this section, by presenting different scenarios we will show that using global recoding generalizations (*i.e.* full-domain generalization, subtree generalization and sibling generalization) to achieve k-anonymity w.r.t. QI in $\mathcal{DB}$ can lead to different situations regarding the α-protection of $\mathcal{DB}$ w.r.t. DA.

Global recoding generalizations not offering α-protection. It can happen in different scenarios. First, consider a data table $\mathcal{DB}$ with the same attributes as the one in Table I, but many more records, and let $DA = \{Job\}$ and

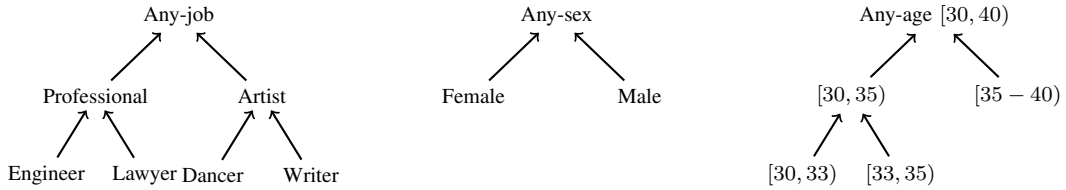

Figure 1. Generalization taxonomy tree for Sex, Job and Age attributes

$QI = \{Sex, Job, Age\}$. Suppose $\mathcal{DB}$ is biased with respect to dancers or a subgroup of dancers, *e.g.* dancer who are women (*i.e.* $\mathcal{DB}$ does not satisfy α-protection w.r.t. $DA = \{Job\}$). Generalizing all instances of 30, 31 and 32 values to the the same generalized value $[30, 35)$ to achieve k-anonymity w.r.t. QI in $\mathcal{DB}$ using full-domain generalization, subtree or sibling generalization cannot achieve α-protection w.r.t. $DA = \{Job\}$, based on Definition 4. Second, consider a data table $\mathcal{DB}$ with the same attributes as the one in Table I, but many more records, and let $DA = \{Job\}$ and $QI = \{Sex, Job, Age\}$. Suppose $\mathcal{DB}$ is biased with respect to dancers. Generalizing all instances of *Dancer* and *Writer* values to the same generalized value *Artist* to achieve k-anonymity in $\mathcal{DB}$ w.r.t. QI using full-domain generalization or subtree generalization, might cause the *Artist* node to inherit the biased nature of *Dancer*. Then, this generalization cannot achieve α-protection w.r.t. $DA = \{Job\}$. Third, consider a data table $\mathcal{DB}$ with the same attributes as the one in Table I, but many more records, and let $DA = \{Age\}$ and $QI = \{Sex, Job, Age\}$. Suppose $\mathcal{DB}$ is not biased (*i.e.* $\mathcal{DB}$ is α-protective) with respect to DA. It means that all PD frequent rules w.r.t. DA extracted from it are not α-discriminatory. However, $\mathcal{DB}$ might contain PD rules which are α-discriminatory and not frequent, *e.g.* $\{Age = 30, Sex = Male\} \rightarrow Credit\_approved = no$, $\{Age = 31, Sex = Male\} \rightarrow Credit\_approved = no$, $\{Age = 32, Sex = Male\} \rightarrow Credit\_approved = no$. Generalizing all instances of 30, 31 and 32 values to the same generalized value $[30, 35)$ to achieve k-anonymity w.r.t. QI in $\mathcal{DB}$ using full-domain generalization, subtree or sibling generalization, can cause new frequent PD rules to appear which might be α-discriminatory and discrimination will show up after generalization, *e.g.* $\{Age = [30 - 35), Sex = Male\} \rightarrow Credit\_approved = no$.

Global recoding generalizations offering α-protection. Consider Table I and let $DA = \{Sex\}$ and $QI = \{Sex, Job, Age\}$. Suppose Table I is biased with respect to women or any subgroup of women, *e.g.* women who are 30 years old and/or who are dancer (*i.e.* Table I does not satisfy α-protection w.r.t. $DA = \{Sex\}$). Generalizing all instances of *Female* values to the same generalized value *Any-sex* to achieve k-anonymity w.r.t QI in Table I can also achieve α-protection w.r.t. $DA = \{Sex\}$, based on Definition 4.

Summarizing, using global recoding generalizations to achieve the requirement of the privacy model (*i.e.* k-anonymity), depend on the generalization, can make original data less or more protected against discrimination.

Table II
DIFFERENT TYPES OF CELL GENERALIZATION

ID	Sex	Job	Age	Credit_approved
1	**(1) Male ⇒ any-sex**	Engineer	35	**Yes**
2	Male	Engineer	38	Yes
3	**(2) Male ⇒ any-sex**	Lawyer	38	**No**
4	Female	Writer	30	No
5	Male	Writer	30	Yes
6	**(3) Female ⇒ any-sex**	Dancer	31	**No**
7	**(4) Female ⇒ any-sex**	Dancer	32	**Yes**

B. Local recoding generalizations and anti-discrimination

In this section, by analyzing different scenarios, we will show how using local recoding generalization, *i.e.* cell generalization, to achieve k-anonymity w.r.t. QI in $\mathcal{DB}$ has impact on α-protection of $\mathcal{DB}$ w.r.t. DA. As mentioned in Section III, in contrast to global recoding generalizations, in cell generalization some instances of a value may remain ungeneralized while other instances are generalized.

Consider Table I and let $DA = \{Sex\}$ and $\alpha = 1.2$. Table I does not satisfy 1.2-protection w.r.t. $f = slift$ and DA since for frequent PD rule c equal to $\{Sex = female\} \rightarrow credit\_approved = no$ by using the definitions of confidence and $slift$ (Expressions (1) and (3), resp.), we have $slift(c) = \frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)} = \frac{2/3}{1/4} = 2.66$. Table I also does not satisfy 1.2-protection w.r.t. $f = elift$ and DA since for PD rule c by using the definitions of confidence and $elift$ [4] (Expressions (1) and (2), resp.) we have $elift(c) = \frac{supp(A,C)/supp(A)}{supp(C)/|\mathcal{DB}|} = \frac{2/3}{3/7} = 1.55$.

Generalizing some instances of *Male* and/or *Female* values to the same generalized value *Any-sex* to achieve k-anonymity w.r.t. $QI = \{Job, Sex, Age\}$ in Table I using cell generalization can lead to different impacts on 1.2-protection of Table I w.r.t. $DA = \{Sex\}$. The impact depends on the value of class attribute (*e.g. Yes* or *No*) of each record in which the value of PD attribute (*e.g. Female* or *Male*) is generalized. Table II shows four types of cell generalization that can happen to achieve k-anonymity in Table I with

[4] when B (PND itemset) in PD rule is empty, $elift$ reduces to the standard lift [19]

numbers (1), (2), (3) and (4). Below, we analyze the impact of each type on 1.2-protection of Table I w.r.t. DA.

- Type (1). Generalizing an instance of *Male* value to the generalized value *Any-sex* while the value of *Credit_approved* attribute in the record is *Yes* cannot make Table I more or less 1.2-protective w.r.t. $f = elift$ since it cannot change the value of $elift(c)$ but it can make Table I more 1.2-protective w.r.t. $f = slift$ since this type of cell generalization can decease the value of $slift(c)$, which is in this example $slift(c) = \frac{2/3}{1/3} = 2$. This type of cell generalization increases the denominator of equation $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$ while keeping the numerator unaltered.
- Type (2). Generalizing an instance of *Male* value to the generalized value *Any-sex* while the value of *Credit_approved* attribute in the record is *No* cannot make Table I more or less 1.2-protective w.r.t. $f = elift$ since it cannot change the value of $elift(c)$ but it can make Table I less 1.2-protective w.r.t. $f = slift$ since it can increase the value of $slict(c)$. This type of cell generalization decreases the denominator of equation $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$ while keeping the numerator unaltered.
- Type (3). Generalizing an instance of *Female* value to the generalized value *Any-sex* while the value of *Credit_approved* attribute for the record is *No* can make Table I more 1.2-protective w.r.t. $f = elift$ since it can decrease the value of $elift(c)$, which is in this example $elift(c) = \frac{1/2}{3/7} = 1.16$. This type of cell generalization decreases the numerator of equation $\frac{supp(A,C)/supp(A)}{supp(C)/|\mathcal{DB}|}$ while keeping the denominator unaltered. In addition, this generalization can also make Table I more 1.2-protective w.r.t. $f = slift$ since it can decrease the value of $slift(c)$, which is in this example $slift(c) = \frac{1/2}{1/4} = 2$. This type of cell generalization decreases the numerator of equation $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$ while keeping the denominator unaltered.
- Type (4). Generalizing an instance of *Female* value to the generalized value *Any-sex* while the value of *Credit_approved* attribute for the record is *Yes* can make Table I less 1.2-protective w.r.t. both $f = elift$ and $f = slift$ since it can increase the values of $elift(c)$ and $slift(c)$, which are in this example $elift(c) = \frac{2/2}{3/7} = 2.33$ and $slift(c) = \frac{2/2}{1/4} = 4$, respectively. This type of cell generalization increases the numerator of equations $\frac{supp(A,C)/supp(A)}{supp(C)/|\mathcal{DB}|}$ and $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$, respectively, while keeping the denominators unaltered.

Summarizing, using cell generalization to achieve the requirement of privacy model (*e.g.* k-anonymity), depend on how many records in each above types modified, can make original data table less or more protected against discrimination. In addition, only the generalization of type (3) can make the original data table α-protective w.r.t. both $f = elift$ and $f = slift$ if enough number of records are modified.

C. Multidimensional generalizations and anti-discrimination

By presenting different scenarios, we also study the impact of using multidimensional generalizations to achieve k-anonymity w.r.t. QI in $\mathcal{DB}$ on α-protection of $\mathcal{DB}$ w.r.t. DA and we observe the similar trend as cell generalization. For the sake of brevity and due to similarity with Section V-B, we do not recall the details here.

D. Suppression and anti-discrimination

In this section, by presenting different scenarios we will show that using suppression techniques (*i.e.* record suppression, value suppression and cell suppression) to achieve k-anonymity w.r.t. QI in $\mathcal{DB}$ can lead to different situations regarding the α-protection of $\mathcal{DB}$ w.r.t. DA. As shown in Section V-B, Table I does not satisfy 1.2-protection w.r.t. $DA = \{Sex\}$ and both $f = slift$ and $f = elift$ since for PD rule c equal to $\{Sex = female\} \rightarrow credit\_approved = no$ we have $slift(c) = 2.66$ and $elift(c) = 1.55$.

Suppressing an entire record to achieve k-anonymity in Table I w.r.t. $QI = \{Job, Sex, Age\}$ using record suppression can lead to different impacts on the 1.2-protection of Table I w.r.t. $DA = \{Sex\}$. The impact depends on the value of PD attribute (*e.g. Female* or *Male*) and the value of class attribute (*e.g. Yes* or *No*) in the suppressed record. Table III shows four types of record suppression which can happen to achieve k-anonymity w.r.t. QI in Table I with numbers (1), (2), (3) and (4). Below, we analyze the impact of each type on α-protection of Table I w.r.t. DA.

- Type (1). Suppressing an entire record with the value of *Male* in *Sex* attribute and the value of *Yes* in *Credit_approved* attribute can make Table I more 1.2-protective w.r.t. $f = elift$ since it can decrease the value of $elift(c)$, which is in this example $elift(c) = \frac{2/3}{3/6} = 1.33$. This type of record suppression increases the denominator of equation $\frac{supp(A,C)/supp(A)}{supp(C)/|\mathcal{DB}|}$ while keeping the numerator unaltered. In addition, this suppression can also make Table I more 1.2-protective w.r.t. $f = slift$ since it can decease the value of $slift(c)$, which is in this example $slift(c) = \frac{2/3}{1/3} = 2$. This type of record suppression increases the denominator of equation $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$ while keeping the numerator unaltered.
- Type (2). Suppressing an entire record with the value of *Male* in *Sex* attribute and the value of *No* in *Credit_approved* attribute can make Table I less 1.2-protective w.r.t. both $f = elift$ and $f = slift$ since it can increase the values of $elift(c)$ and $slift(c)$. This type of record suppression decreases

Table III
DIFFERENT TYPES OF RECORD SUPPRESSION

ID	Sex	Job	Age	Credit_approved
1	**(1) Male**	Engineer	35	**Yes**
2	Male	Engineer	38	Yes
3	**(2) Male**	Lawyer	38	**No**
4	Female	Writer	30	No
5	Male	Writer	30	Yes
6	**(3) Female**	Dancer	31	**No**
7	**(4) Female**	Dancer	32	**Yes**

the denominator of equations $\frac{supp(A,C)/supp(A)}{supp(C)/|\mathcal{DB}|}$ and $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$, respectively, while keeping the numerators unaltered.

- Type (3). Suppressing an entire record with the value of *Female* in *Sex* attribute and the value of *No* in *Credit_approved* attribute cannot make Table I more or less 1.2-protective w.r.t. $f = elift$ since it cannot change the value of $elift(c)$ substantially, which is in this example $elift(c) = \frac{1/2}{2/6} = 1.5$. This is because, this type of record suppression decreases the numerator of equation $\frac{supp(A,C)/supp(A)}{supp(C)/|\mathcal{DB}|}$ while also decreasing the denominator of it. However, this type of record suppression can make Table I more 1.2-protective w.r.t. $f = slift$ since it can decrease the value of $slift(c)$, which is in this example $slift(c) = \frac{1/2}{1/4} = 2$. This suppression decreases the numerator of $\frac{supp(A,C)/supp(A)}{supp(\neg A,C)/supp(\neg A)}$ while keeping the denominator unaltered.
- Type (4). Suppressing an entire record with the value of *Female* in *Sex* attribute and the value of *Yes* in *Credit_approved* attribute can make Table I less 1.2-protective w.r.t. both $f = elift$ and $f = slift$ since it can increase the value of $elift(c)$ and $slift(c)$, which are in this example $elift(c) = \frac{2/2}{3/6} = 2$ and $slift(c) = \frac{2/2}{1/4} = 4$, respectively.

Summarizing, using record suppression, depend on how many records in each above types suppressed, can make original data table less or more protected against discrimination after achieving privacy protection. In addition, only record suppression of type (1) can make original data table α-protective w.r.t. both $f = elift$ and $f = slift$ if enough number of records are suppressed.

As mentioned in Section III, value suppression refers to suppressing every instance of a given value in a data table. Then, depend on which attribute values suppressed after achieving privacy protection, value suppression can offer α-protection or not. Cell suppression refers to suppressing some instances of a given value in a data table. Then, similar to cell generalization, depend on suppressed cells contained which values and the respective records contained which class values, cell suppression can make original data less or more protected against discrimination. Finally, Table IV summarizes the results we obtain in this paper.

VI. CONCLUSION

In this paper, we have investigated the relation between data anonymization techniques and anti-discrimination to answer an important question. How providing protection against privacy attacks by using data anonymization techniques has impact on discriminatory bias contained in the original data? By presenting and analyzing different scenarios, we learn that we cannot protect original data against privacy attacks without taking into account anti-discrimination requirements (*i.e.* α-protection). It is because data anonymization techniques can work against anti-discrimination. In addition, we found that since data anonymization techniques in the special conditions (*e.g.* specific full-domain generalization) can also make data protected against discrimination, we can adapt and use some of them for discrimination prevention and also by considering anti-discrimination requirements during anonymization, we can have solutions for generating privacy- and discrimination- protected datasets.

As future work, we plan to study the relation between data anonymization techniques and anti-discrimination considering indirect discrimination. In addition, we plan to design methods based on data anonymization to make original datasets protected against both privacy and discrimination threats.

ACKNOWLEDGMENT

This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007-65406-C03-01 "E-AEGIS", TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 project "DwB". The second author is partially supported as an ICREA Acadèmia researcher. The authors are with the UNESCO Chair in Data Privacy, but the views expressed in this paper do not necessarily reflect the position of UNESCO nor commit that organization.

REFERENCES

[1] C.C. Aggarwal and P.S. Yu (eds.). *Privacy Preserving Data Mining: Models and Algorithms*. Springer, 2008

[2] R. Agrawal and R. Srikant. Privacy preserving data mining. In *SIGMOD 2000*, pp. 439-450, ACM, 2000.

[3] Australian Legislation. (a) Equal Opportunity Act – Victoria State, (b) Anti-Discrimination Act – Queens– land State, 2008. http://www.austlii.edu.au.

[4] P. Bleninger, J. Drechsler and G. Ronning, "Remote data access and the risk of disclosure from linear regression: an empirical study", in *Privacy in Statistical Databases-PSD 2010*, LNCS 6344, Springer, 2010, pp. 220-233.

Table IV
THE SUMMARIZED RESULTS

Data Anonymization techniques	Achieve α-protection	Against α-protection	No impact
Global recoding generalizations	✓	✓	✓
Cell generalization/Cell suppression Type (1)	✓		✓
Cell generalization/Cell suppression Type (2)		✓	✓
Cell generalization/Cell suppression Type (3)	✓		
Cell generalization/Cell suppression Type (4)		✓	
Multidimensional generalization	✓	✓	✓
Record suppression Type (1)	✓		
Record suppression Type (2)		✓	
Record suppression Type (3)	✓		✓
Record suppression Type (4)		✓	
Value suppression	✓		✓

[5] T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *DMKD Journal*, 21(2):277-292, 2010.

[6] C. Dwork, "Differential privacy", in *33rd International Colloquium on Automata, Languages and Programming-ICALP 2006, Part II*, LNCS 4052, Springer, 2006, pp. 1-12.

[7] European Union Legislation. Directive 95/46/EC, 2012.

[8] European Union Legislation, (a) Race Equality Directive, 2000; (b) Employment Equality Directive, 2000; (c) Equal treatment of persons, 2009.

[9] B. C. M. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu. *Introduction to Privacy-Preserving Data Publishing: Concepts and Techniques*. Chapman & Hall/CRC, 2010.

[10] S. Hajian, J. Domingo-Ferrer and A. Martínez-Ballesté. Rule protection for indirect discrimination prevention in data mining. In *MDAI 2011*, LNCS 6820, pp. 211-222, 2011.

[11] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE TKDE*, 22 March 2012. IEEE computer Society Digital Library.

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

[13] F. Kamiran and T. Calders. Classification without discrimination. In *IEEE IC4 2009*, 2009.

[14] F. Kamiran and T. Calders. Classification with no discrimination by preferential sampling. In *Proc. of the 19th Machine Learning conference of Belgium and The Netherlands*, 2010.

[15] F. Kamiran, T. Calders and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM 2010*, pp. 869-874. IEEE, 2010.

[16] N. Li, T. Li and S. Venkatasubramanian, "T-closeness: privacy beyond k-anonymity and l-diversity", in *Proceedings of the IEEE ICDE 2007*, 2007.

[17] Y. Lindell and B. Pinkas, "Privacy preserving data mining", in *Advances in Cryptology-CRYPTO'00*, LNCS 1880, Springer, 2000, pp. 36-53.

[18] A. Machanavajjhala, J. Gehrke, D. Kiefer and M. Venkitasubramanian, "L-diversity: privacy beyond k-anonymity", in *Proceedings of the IEEE ICDE 2006*, 2006.

[19] D. Pedreschi, S. Ruggieri and F. Turini. Discrimination-aware data mining. In *KDD 2008*, pp. 560-568. ACM, 2008.

[20] D. Pedreschi, S. Ruggieri and F. Turini. Measuring discrimination in socially-sensitive decision records. In *SDM 2009*, pp. 581-592. SIAM, 2009.

[21] S. Ruggieri, D. Pedreschi and F. Turini. Data mining for discrimination discovery. *ACM TKDD*, 4(2) Article 9, ACM, New York, 2010.

[22] P. Samarati. Protecting respondents' identities in microdata release. *IEEE TKDE*, 13(6):1010-1027, 2001.

[23] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information", in *Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS 98)*, Seattle, WA, June 1998, p. 188.

[24] L. Sweeney, "k-Anonymity: a model for protecting privacy" *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5): 557-570, 2002.

[25] B. L. Thanh, S. Ruggieri and F. Turini, "k-NN as an implementation of situation testing for discrimination discovery and prevention", in *KDD 2011*, pp. 502-510. ACM, 2011.

[26] United States Congress, *US Equal Pay Act*, 1963.

[27] I. Zliobaite, F. Kamiran, T.Calders, "Handling Conditional Discrimination", *ICDM 2011*, pp. 992-1001. IEEE, 2011.