Noname manuscript No.
(will be inserted by the editor)

Generalization-based Privacy Preservation and Discrimination Prevention in Data Publishing and Mining

Sara Hajian · Josep Domingo-Ferrer · Oriol Farràs

Received: date / Accepted: date

Abstract Living in the information society facilitates the automatic collection of huge amounts of data on individuals, organizations, etc. Publishing such data for secondary analysis (*e.g.* learning models and finding patterns) may be extremely useful to policy makers, planners, marketing analysts, researchers and others. Yet, data publishing and mining do not come without dangers, namely privacy invasion and also potential discrimination of the individuals whose data are published. Discrimination may ensue from training data mining models (*e.g.* classifiers) on data which are biased against certain protected groups (ethnicity, gender, political preferences, etc.).

The objective of this paper is to describe how to obtain data sets for publication that are: i) privacy-preserving; ii) unbiased regarding discrimination; and iii) as useful as possible for learning models and finding patterns. We present the first generalization-based approach to simultaneously offer privacy preservation and discrimination prevention. We formally define the problem, give an optimal algorithm to tackle it and evaluate the algorithm in terms of both general and specific data analysis metrics (*i.e.* various types of classifiers and rule induction algorithms). It turns out that the impact of our transformation on the quality of data is the same or only slightly higher than the impact of achieving just privacy preservation. In addition, we show how to extend our approach to different privacy models and anti-discrimination legal concepts.

Keywords Data mining · Anti-discrimination · Privacy · Generalization

1 Introduction

In the information society, massive and automated data collection occurs as a consequence of the ubiquitous digital traces we all generate in our daily life. The availability of such wealth of data makes its publication and analysis highly desirable for a variety of purposes,

Universitat Rovira i Virgili, Dept. of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona, Catalonia
Tel.: +34-977558109
Fax: +34-977559710
E-mail: {sara.hajian,josep.domingo,oriol.farras}@urv.cat

including policy making, planning, marketing, research, etc. Yet the real and obvious benefits of data publishing and mining have a dual, darker side. There are at least two potential threats for individuals whose information is published: *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. On the legal side, parallel to the development of privacy legislation [13], anti-discrimination legislation has undergone a remarkable expansion [4,14], and it now 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.

On the technology side, efforts at guaranteeing privacy have led to developing statistical disclosure control (SDC, [46,22]) and privacy preserving data mining (PPDM, [3,30]). SDC and PPDM have become increasingly popular because they allow publishing and sharing sensitive data for secondary analysis. Different privacy models and their variations have been proposed to trade off the utility of the resulting data for protecting individual privacy against different kinds of privacy attacks. k-Anonymity [39,43], l-diversity [32], t-closeness [29] and differential privacy [10] are among the best-known privacy models. Detailed descriptions of different PPDM models and methods can be found in [1,16]. The issue of anti-discrimination has recently been considered from a data mining perspective [34]. Some proposals are oriented to using data mining to discover and measure discrimination [34,35, 38,31,12,6]; other proposals [7,18,24–26,48] deal with preventing data mining from becoming itself a source of discrimination. In other words, *discrimination prevention in data mining* (DPDM) consists of ensuring that data mining models automatically extracted from a data set are such that they do not lead to discriminatory decisions even if the data set is inherently biased against protected groups. For a survey of contributions to discrimination-aware data analysis see [8].

Although PPDM/SDC and DPDM have different goals, they have some technical similarities. Necessary steps of PPDM/SDC are: i) define the privacy model (*e.g.* k-anonymity); ii) apply a proper anonymization technique (*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 tailored to specific data mining tasks). Similarly, necessary steps for DPDM include: i) define the non-discrimination model according to the respective legal concept (*i.e.* α*-protection* according to the legal concept of *direct discrimination*); ii) apply a suitable data distortion method to satisfy the requirements of the non-discrimination model; iii) measure data quality loss as in the case of DPDM. 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 data mining.

1.1 Motivating example

Table 1 presents raw customer credit data, where each record represents a customer's specific information. *Sex*, *Race*, and working hours named *Hours* can be taken as *quasi-identifier* attributes. The class attribute has two values, *Yes* and *No*, to indicate whether the customer has received credit. Assume that *Salary* is a sensitive/private attribute and groups of *Sex* and *Race* attributes are *protected*. The credit giver wants to publish a privacy-preserving and non-discriminating version of Table 1. To do that, she needs to eliminate two types of threats against her customers:

Table 1 Private data set with biased decision records

ID	Sex	Race	Hours	Salary	Credit_ approved
1	Male	White	40	High	Yes
2	Male	Asian-Pac	50	Medium	Yes
3	Male	Black	35	Medium	No
4	Female	Black	35	Medium	No
5	Male	White	37	Medium	Yes
6	Female	Amer-Indian	37	Medium	Yes
7	Female	White	35	Medium	No
8	Male	Black	35	High	Yes
9	Female	White	35	Low	No
10	Male	White	50	High	Yes

- *Privacy threat, e.g., record linkage*: If a record in the table is so specific that only a few customers match it, releasing the data may allow determining the customer's identity (record linkage attack) and hence the salary of that identified customer. Suppose that the adversary knows that the target identified customer is white and his working hours are 40. In Table 1, record $ID = 1$ is the only one matching that customer, so the customer's salary becomes known.
- *Discrimination threat*: If credit has been denied to most female customers, releasing the data may lead to making biased decisions against them when these data are used for extracting decision patterns/rules as part of the automated decision making. Suppose that the *minimum support* (ms) required to extract a classification rule from the data set in Table 1 is that the rule be satisfied by at least 30% of the records. This would allow extracting the classification rule $r : Sex = female \rightarrow Credit\_approved = no$ from these data. Clearly, using such a rule for credit scoring is discriminatory against female customers.

1.2 Paper contributions and overview

We argue that both threats above must be addressed at the same time, since providing protection against only one of them might not guarantee protection against the other. An important question is how we can provide protection against both privacy and discrimination risks without one type of protection working against the other and with minimum impact on data quality. In [20], the authors investigated this problem in the context of knowledge/pattern publishing. They proposed a combined pattern sanitization framework that yields both privacy and discrimination protected patterns, while introducing reasonable (controlled) pattern distortion. In this paper, we investigate for the first time the problem of discrimination- and privacy-aware *data* publishing, *i.e.* transforming the data, instead of patterns, in order to simultaneously fulfill privacy preservation and discrimination prevention in data mining. Our approach falls into the pre-processing category: it sanitizes the data *before* they are used in data mining tasks rather than sanitizing the knowledge patterns extracted by data mining tasks (post-processing). Very often, knowledge publishing (publishing the sanitized patterns) is not enough for the users or researchers, who want to be able to mine the data themselves. This gives researchers greater flexibility in performing the required data analyses.

We introduce an anti-discrimination model that can cover every possible nuance of discrimination w.r.t. multiple attributes, not only for specific protected groups within one attribute. Note that the existing pre-processing discrimination prevention methods are based on data perturbation, either by modifying class attribute values [24, 18, 19] or by modifying

PD attribute values [18,19] of the training data. One of the drawbacks of these techniques is that they cannot be applied (are not preferred) in countries where data perturbation is not legally accepted (preferred), while generalization is allowed; *e.g.* this is the case of Sweden and other Nordic countries (see p. 24 of [41]). Moreover, generalization not only can make the original data privacy-protected but can also simultaneously make the original data both discrimination- and privacy-protected. In our earlier work [21], we explored under which conditions several data anonymization techniques could also help preventing discrimination. The techniques examined included suppression and several forms of generalization (global recoding, local recoding, multidimensional generalizations). In this paper, the approach is quite different: rather than examining a set of techniques, we focus on the (novel) problem of achieving simultaneous discrimination prevention and privacy protection in data publishing and mining. Specifically, we leverage the conclusions of our previous study [21] to choose the best possible technique for solving the problem: the high applicability of generalization in data privacy justifies its choice.

We present an optimal algorithm that can cover different *legally-grounded* measures of discrimination to obtain all full-domain generalizations whereby the data are discrimination and privacy protected. The "minimal" generalization (*i.e.*, the one incurring the least information loss according to some criterion) can then be chosen. In addition, we evaluate the performance of the proposed approach and the data quality loss incurred as a side effect of the data generalization needed to achieve both discrimination and privacy protection. Data quality loss is measured in terms of both general and specific data analysis metrics (*i.e.* various types of classifiers and rule induction algorithms). We compare this quality loss with the one incurred to achieve privacy protection only. Finally, we present how to extend our approach to satisfy different privacy models and anti-discrimination legal concepts.

The article is organized as follows. Section 2 introduces basic definitions and concepts used throughout the paper. Privacy and anti-discrimination models are presented in Section 3 and 4, respectively. In Section 5, we formally define the problem of simultaneous privacy and anti-discrimination data protection. Our proposed approach and an algorithm for discrimination- and privacy-aware data publishing and mining are presented in Sections 5.1 and 5.2. Section 6 reports experimental work. An extension of the approach to alternative privacy-preserving requirements and anti-discrimination legal constraints is presented in Section 7. Finally, Section 8 summarizes conclusions and identifies future research topics.

2 Basic notions

Given the data table $\mathcal{DB}(A_1, \cdots, A_n)$, a set of attributes $\mathcal{A} = \{A_1, \cdots, A_n\}$, 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 $\{A_i, \cdots, A_j\} \subseteq \{A_1, \cdots, A_n\}$. 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 database $\mathcal{DB}$ 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*, *Birthdate* and *Gender* 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*. Let S be a set of sensitive attributes in $\mathcal{DB}$. Civil rights laws [4, 14,44], explicitly identify the attributes to be protected against discrimination. For instance,

Table 2 Main acronyms used throughout the paper

	Definition
$\mathcal{BK}$	Background knowledge
CA	Classification accuracy
CM	Classification metric
DA	A set of PD attributes in $\mathcal{DB}$
$\mathcal{DB}$	Data table
DR	Discernibility ratio
DT	Domain tuple
GH	Generalization height
LA	legally-grounded attributes
PD	Potentially discriminatory
PND	Potentially non-discriminatory
QI	Quasi-identifier

U.S. federal laws [44] 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). Let DA be a set of PD attributes in $\mathcal{DB}$ specified by law. Comparing privacy legislation [13] and anti-discrimination legislation [14,44], PD attributes can overlap with QI attributes (*e.g. Sex*, *Age*, *Marital_status*) and/or sensitive attributes (*e.g. Religion* in some applications). 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. Table 2 lists the main acronyms used throughout the paper.

An *item* is an expression $A_i = q$, where $A_i \in \mathcal{A}$ and $q \in D_{A_i}$, *e.g. Race=black*. 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.* $\{Race = black, Hours = 40\}$. In previous works on anti-discrimination [34,35,38,24,18,19,25,48], the authors propose discrimination discovery and prevention techniques w.r.t. specific protected groups, *e.g.* black and/or female persons. However, this assumption fails to capture the various nuances of discrimination since minority or disadvantaged groups can be different in different contexts. For instance, in a neighborhood with almost all black people, whites are a minority and may be discriminated. Then we consider $A_i = q$ to be a PD item, for every $q \in D_{A_i}$, where $A_i \in DA$, *e.g.* $Race = q$ is a PD item for any race q, where $DA = \{Race\}$. This definition is also compatible with the law. For instance, the U.S. Equal Pay Act [44] 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 $A_i = q$ with $q \in D_{A_i}$ is *potentially*[1] *non-discriminatory* (PND) if $A_i \notin DA$, *e.g.* $Hours = 35$ where $DA = \{Race\}$. A PD itemset is an itemset containing only PD items, which we also call 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.* $\{Race = Black,\ Hours = 35\} \rightarrow\ Credit\_approved = no$. 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. We omit the subscripts in

[1] The use of PD (resp., PND) attributes in decision making does not necessarily lead to (or exclude) discriminatory decisions [38].

$supp_{\mathcal{DB}}(\cdot)$ 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/or confidence greater than respective specified lower bounds. Support is a measure of statistical significance, whereas confidence is a measure of the strength of the rule. In this paper we consider frequent rules w.r.t. the support measure.

3 Privacy model

To prevent record linkage attacks through quasi-identifiers, Samarati and Sweeney [40,42] proposed the notion of k-anonymity.

Definition 1 **(k-anonymity)** Let $\mathcal{DB}(A_1, \cdots, A_n)$ be a data table and $QI = \{Q_1, \cdots, Q_m\} \subseteq \{A_1, \cdots, A_n\}$ be a quasi-identifier. $\mathcal{DB}$ is said to satisfy k-anonymity w.r.t. QI if each combination of values of attributes in QI is shared by at least k tuples (records) in $\mathcal{DB}$.

A data table satisfying this requirement is called k-anonymous. The set of all tuples in $\mathcal{DB}$ for each sequence of values in $\mathcal{DB}[QI]$ is called *frequency set*. Typically, the original data table does not satisfy k-anonymity and, before being published, it must be modified through an anonymization method. Samarati and Sweeney [40,39,43] gave methods for k-anonymization based on *generalization*. Computational procedures alternative to generalization have thereafter been proposed to attain k-anonymity, like microaggregation [9]. Nonetheless, generalization remains not only the main method for k-anonymity, but it can also be used to satisfy other privacy models (*e.g.*, l-diversity in [32], t-closeness in [29] and differential privacy in [33]). Generalization replaces QI attribute values with a generalized version of them. Let D_i and D_j be two domains. If the values of D_j are the generalization of the values in domain D_i, we denote $D_i \leq_D D_j$. A many-to-one *value generalization function* $\gamma : D_i \rightarrow D_j$ is associated with every D_i, D_j with $D_i \leq_D D_j$.

Generalization is based on a *domain generalization hierarchy* and a corresponding *value generalization hierarchy* on the values in the domains. A domain generalization hierarchy is defined to be a set of domains that is totally ordered by the relationship $\leq_D$. We can consider the hierarchy as a chain of nodes, and if there is an edge from D_i to D_j, it means that D_j is the *direct generalization* of D_i. Let Dom_i be a set of domains in a domain generalization hierarchy of a quasi-identifier attribute $Q_i \in QI$. For every $D_i, D_j, D_k \in Dom_i$ if $D_i \leq_D D_j$ and $D_j \leq_D D_k$, then $D_i \leq_D D_k$. In this case, domain D_k is an *implied generalization* of D_i. The maximal element of Dom_i is a singleton, which means that all values in each domain can be eventually generalized to a single value. Figure 1 left shows possible domain generalization hierarchies for the *Race*, *Sex* and *Hours* attributes in Table 1. Value generalization functions associated with the domain generalization hierarchy induce a corresponding value-level tree, in which edges are denoted by γ, *i.e* direct value generalization, and paths are denoted by γ^+, *i.e.* implied value generalization. Figure 1 right shows a value generalization hierarchy with each value in the *Race*, *Sex* and *Hours* domains, *e.g.* Colored = γ(black) and Any-race $\in \gamma^+$(black). For a $QI = \{Q_1, \cdots, Q_n\}$ consisting of multiple attributes, each with its own domain, the domain generalization hierarchies of the individual attributes $Dom_1, \cdots$,Dom_n can be combined to form a multi-attribute *generalization lattice*. Each vertex of a lattice is a domain tuple $DT = \langle N_1, \cdots, N_n \rangle$ such that $N_i \in Dom_i$, for $i = 1, \cdots, n$, representing a multi-attribute domain generalization. An example for *Sex* and *Race* attributes is presented in Figure 2.

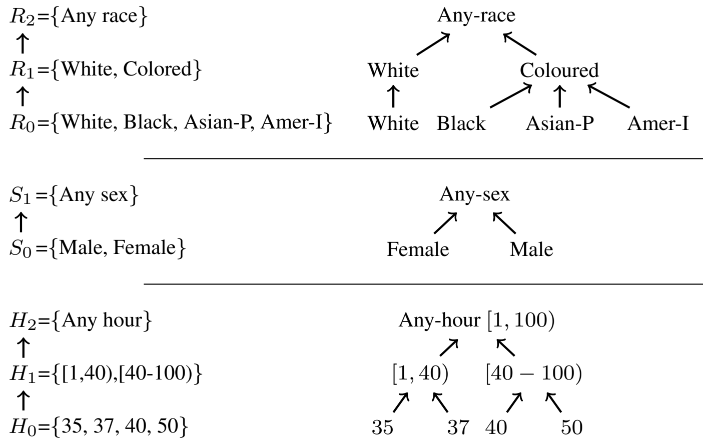

Fig. 1 An example of domain (left) and value (right) generalization hierarchies of Race, Sex and Hours attributes

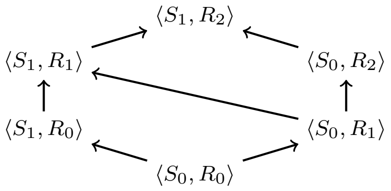
$\langle S_1, R_2\rangle$
$\langle S_1, R_1\rangle$
$\langle S_0, R_2\rangle$
$\langle S_1, R_0\rangle$
$\langle S_0, R_1\rangle$
$\langle S_0, R_0\rangle$

Fig. 2 Generalization lattice for the Race and Sex attributes

Definition 2 (Full-domain generalization) Let $\mathcal{DB}$ be a data table having a quasi-identifier $QI = \{Q_1, \cdots, Q_n\}$ with corresponding domain generalization hierarchies $Dom_1, \cdots,$ Dom_n. A full-domain generalization can be defined by a domain tuple $DT = \langle N_1, \cdots, N_n\rangle$ with $N_i \in Dom_i$, for every $i = 1, \cdots, n$. A full-domain generalization with respect to DT maps each value $q \in D_{Q_i}$ to some $a \in D_{N_i}$ such that $a = q$, $a = \gamma(q)$ or $a \in \gamma^+(q)$.

Full-domain generalization guarantees that all values in an attribute are generalized to the same domain[2]. For example, consider Figure 1 right and assume that values 40 and 50 of *Hours* are generalized to $[40 - 100)$; then 35 and 37 must be generalized to $[1, 40)$. A full-domain generalization w.r.t. domain tuple DT is k-anonymous if it yields k-anonymity for $\mathcal{DB}$ with respect to QI. In the literature, different generalization-based algorithms have been proposed to k-anonymize a data table. They are optimal [39,27,5] or minimal [23,17,45]. Although minimal algorithms are in general more efficient than optimal ones, we choose an optimal algorithm (*i.e.* Incognito [27]) because, in this first work about combining privacy preservation and discrimination prevention in data publishing, it allows us to study the worst-case toll on efficiency of achieving both properties. Incognito is a well-known suite of optimal bottom-up generalization algorithms to generate all possible k-anonymous full-

[2] In full-domain generalization if a value is generalized, all its instances are generalized. There are alternative generalization schemes, such as multi-dimensional generalization or cell generalization, in which some instances of a value may remain ungeneralized while other instances are generalized.

domain generalizations[3]. In comparison with other optimal algorithms, Incognito is more scalable and practical for larger data sets and more suitable for categorical attributes. Incognito is based on two main properties satisfied for k-anonymity:

- **Subset property**. If $\mathcal{DB}$ is k-anonymous with respect to QI, then it is k-anonymous with respect to any subset of attributes in QI (the converse property does not hold in general).
- **Generalization property**. Let P and Q be nodes in the generalization lattice of $\mathcal{DB}$ such that $D_P \leq_D D_Q$. If $\mathcal{DB}$ is k-anonymous with respect to P, then $\mathcal{DB}$ is also k-anonymous with respect to Q (monotonicity of generalization).

Example 2 Continuing the motivating example (Section 1.1), consider Table 1 and suppose $QI = \{Race, Sex\}$ and $k = 3$. Consider the generalization lattice over QI attributes in Figure 2. Incognito finds that Table 1 is 3-anonymous with respect to domain tuples $\langle S_1, R_1 \rangle$, $\langle S_0, R_2 \rangle$ and $\langle S_1, R_2 \rangle$.

4 Non-discrimination model

The legal notion of under-representation has inspired existing approaches for discrimination discovery based on rule/pattern mining [34], taking into account different legal concepts (*e.g.*, direct and indirect discrimination and genuine occupational requirement). Direct discrimination occurs when the input data contain PD attributes, *e.g.*, *Sex*, while for indirect discrimination it is the other way round. In the rest, we assume that the input data contain protected groups, which is a reasonable assumption for attributes such as *Sex*, *Age* and *Pregnancy/marital status* and we omit the word "direct" for brevity. In Section 7, we describe how to deal with other legal concepts, *e.g.*, indirect discrimination.

Given a set DA of potentially discriminatory attributes and starting from a data set $\mathcal{D}$ of historical decision records, the idea is to extract frequent classification rules of the form $A, B \rightarrow C$ (where A is a non-empty PD itemset and B a is PND itemset), called PD rules, to unveil contexts B of possible discrimination, where the non-empty protected group A suffers from over-representation with respect to the *negative* decision C (C is a class item reporting a negative decision, such as credit denial, application rejection, job firing). In other words, A is under-represented with respect to the corresponding positive decision $\neg C$. As an example, rule *Sex=female, Hours=35* $\rightarrow$ *Credit_approved=no* is a PD rule about denying credit (the decision C) to women (the protected group A) among those working 35 hours per week (the context B), with DA=$\{$*Sex*$\}$. In other words, the context B determines the subsets of protected groups (*e.g.,* subsets of women working 35 hours per week).

The degree of under-representation should be measured over each PD rule using a *legally-grounded* measure [4], such as those introduced in Pedreschi *et al.* [35] and shown in Figure 3. *Selection lift (slift)* [5] is the ratio of the proportions of benefit denial between the protected and unprotected groups, *e.g.* women and men resp., in the given context. *Extended*

[3] Although algorithms using multi-dimensional or cell generalizations (*e.g.* the Mondrian algorithm [28]) cause less information loss than algorithms using full-domain generalization, the former suffer from the problem of data exploration [16]. This problem is caused by the co-existence of specific and generalized values in the generalized data set, which make data exploration and interpretation difficult for the data analyst.

[4] On the legal side, different measures are adopted worldwide; see [37] for parallels between different measures and anti-discrimination acts.

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

lift (elift) [6] 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. A special case of *slift* occurs when we deal with non-binary 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. Race=black*), compares it with the most favored item $a = v_2$ (*e.g. Race=white*). The last measure is the *odds lift (olift)*, the ratio between the odds of the proportions of benefit denial between the protected and unprotected groups. The discrimination measures mentioned so far in this paragraph are formally defined in Figure 3. Whether a rule is to be considered discriminatory according to a specific discrimination measure can be assessed by thresholding the measure as follows.

Classification rule: $c = A, B \to C$

B	C	$\neg C$	
A	a_1	$n_1 - a_1$	n_1
$\neg A$	a_2	$n_2 - a_2$	n_2

$$a_1 = supp(A, B, C)$$
$$a_2 = supp(\neg A, B, C)$$
$$n_1 = supp(A, B)$$
$$n_2 = supp(\neg A, B)$$

$$p_1 = a_1/n_1 \quad p_2 = a_2/n_2 \quad p = (a_1 + a_2)/(n_1 + n_2)$$

$$elift(c) = \frac{p_1}{p}, \quad slift(c) = \frac{p_1}{p_2}, \quad olift(c) = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

Fig. 3 Discrimination measures

Definition 3 (**α-protective/discriminatory rule**) Let f be one of the measures in Figure 3, $\alpha \in \mathbb{R}$ a fixed threshold[7], A a PD itemset and B a PND itemset with respect to DA. A PD classification rule $c = A, B \to C$ is α-protective with respect to 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 Figure 3. $\mathcal{DB}$ is said to satisfy α-protection or to be α-protective w.r.t. DA and f if each PD frequent classification rule $c : A, B \to C$ extracted from $\mathcal{DB}$ is α-protective, where A is a PD itemset and B is a PND itemset.

Example 3 Continuing the motivating example, suppose $DA = \{Sex\}$, $\alpha = 1.2$ and $ms = 20\%$. Table 1 does not satisfy 1.2-protection w.r.t. both $f = slift$ and $f = elift$, since for a frequent rule c equal to *Sex=female, Salary = medium → credit_approved=no* we have

$$slift(c) = \frac{p_1}{p_2} = \frac{a_1/n_1}{a_2/n_2} = \frac{2/3}{1/3} = 2$$

[6] Discrimination of a group occurs when a higher proportion of people not in the group is able to comply with a qualifying criterion.

[7] α states an acceptable level of discrimination according to laws and regulations. For example, the U.S. Equal Pay Act [44] 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". This amounts to using *clift* with $\alpha = 1.25$.

and

$$elift(c) = \frac{p_1}{p} = \frac{a_1/n_1}{(a_1 + a_2)/(n_1 + n_2)} = \frac{2/3}{3/6} = 1.33.$$

Note that α-protection in $\mathcal{DB}$ not only prevents discrimination against the main protected groups w.r.t. DA (*e.g.*, women) but also against any subsets of protected groups w.r.t. $\mathcal{A} \backslash DA$ (*e.g.*, women who have medium salary and/or work 36 hours per week) [36]. Releasing an α-protective (unbiased) version of an original data table is desirable to prevent discrimination with respect to DA. If the original data table is biased w.r.t. DA, it must be modified before being published (*i.e.* pre-processed).

5 Simultaneous privacy preservation and discrimination prevention

We want to obtain anonymized data tables that are protected against record linkage and also free from discrimination, more specifically α-protective k-anonymous data tables defined as follows.

Definition 5 (**α-protective k-anonymous data table**) Let $\mathcal{DB}(A_1, \cdots, A_n)$ be a data table, $QI = \{Q_1, \cdots, Q_m\}$ a quasi-identifier, DA a set of PD attributes, k an anonymity threshold, and α a discrimination threshold. $\mathcal{DB}$ is α-protective k-anonymous if it is both k-anonymous and α-protective with respect to QI and DA, respectively.

We focus on the problem of producing a version of $\mathcal{DB}$ that is α-protective k-anonymous with respect to QI and DA. The problem could be investigated with respect to different possible relations between categories of attributes in $\mathcal{DB}$. k-Anonymity uses quasi-identifiers for re-identification, so we take the worst case for privacy in which any attribute can be part of a quasi-identifier, except the class/decision attribute. We exclude the latter attribute from QI because we assume that the decision made on a specific individual is not *publicly* linked to his/her identity by the decision maker, *e.g.* banks do not publicize who was granted or denied credit. On the other hand, each QI attribute can be PD or not. In summary, the following relations are assumed: (1) $QI \cap C = \emptyset$, (2) $DA \subseteq QI$. Taking the largest possible QI makes sense indeed. The more attributes are included in QI, the more protection k-anonymity provides and, in general, the more information loss it causes. Thus, we test our proposal in the worst-case privacy scenario. On the discrimination side, as explained in Section 5.2, the more attributes are included in QI, the more protection is provided by α-protection.

5.1 The generalization-based approach

To design a method, we need to consider the impact of data generalization on discrimination.

Definition 6 Let $\mathcal{DB}$ be a data table having a quasi-identifier $QI = \{Q_1, \cdots, Q_n\}$ with corresponding domain generalization hierarchies $Dom_1, \cdots, Dom_n$. Let DA be a set of PD attributes associated with $\mathcal{DB}$. Each node N in Dom_i is said to be PD if Dom_i corresponds to one of the attributes in DA and N is not the singleton of Dom_i. Otherwise node N is said to be PND.

Definition 6 states that not only ungeneralized nodes of PD attributes are PD but also the generalized nodes of these domains are PD. For example, in South Africa, about 80% of the population is Black, 9% White, 9% Colored and 2% Asian. Generalizing the *Race* attribute in a census of the South African population to $\{$*White, Non-White*$\}$ causes the *Non-White* node to inherit the PD nature of *Black, Colored and Asian*. We consider the singleton nodes as PND because generalizing all instances of all values of a domain to single value is PND, *e.g.* generalizing all instances of male and female values to *any-sex* is PND.

Example 4 Continuing the motivating example, consider $DA = \{Race, Sex\}$ and Figure 1 left. Based on Definition 6, in the domain generalization hierarchy of *Race*, *Sex* and *Hours*, R_0, R_1, S_0 are PD nodes, whereas R_2, S_1, H_0, H_1 and H_2 are PND nodes.

When we generalize data (*i.e.* full-domain generalization) can we achieve α-protection? By presenting two main scenarios we show that the answer can be yes or no depending on the generalization:

- When the original data table $\mathcal{DB}$ is biased versus some protected groups w.r.t. DA and f (*i.e.*, there is at least one frequent rule $c : A \rightarrow C$ such that $f(c) \geq \alpha$, where A is a PD itemset w.r.t. DA), a full-domain generalization can make $\mathcal{DB}$ α-protective if it includes the generalization of the respective protected groups (*i.e.* A).
- When the original data table $\mathcal{DB}$ is biased versus a subset of the protected groups w.r.t. DA (*i.e.*, there is at least one frequent rule $c : A, B \rightarrow C$ such that $f(c) \geq \alpha$, where A is a PD itemset and B is a PND itemset w.r.t. DA), a full-domain generalization can make $\mathcal{DB}$ α-protective if any of the following holds: (1) it includes the generalization of the respective protected groups (*i.e.* A); (2) it includes the generalization of the attributes which define the respective subsets of the protected groups (*i.e.* B); (3) it includes both (1) and (2).

Then, given the generalization lattice of $\mathcal{DB}$ over QI, where $DA \subseteq QI$, there are some candidate nodes for which $\mathcal{DB}$ is α-protective (*i.e.*, α-protective full-domain generalizations). Specifically, we can state the following correctness result.

Theorem 1 *Let $\mathcal{DB}$ be a data table having a quasi-identifier $QI = \{Q_1, \cdots, Q_n\}$ with corresponding domain generalization hierarchies $Dom_1, \cdots, Dom_n$. Let DA be a set of PD attributes in $\mathcal{DB}$. If $\mathcal{FG}$ is the set of k-anonymous full-domain generalizations with respect to $\mathcal{DB}$, QI and $Dom_1, \cdots, Dom_n$, there is at least one k-anonymous full-domain generalization in $\mathcal{FG}$ that is α-protective with respect to DA.*

Proof There exists $g \in \mathcal{FG}$ with respect to $\langle N_1, \cdots, N_n \rangle$ such that for all $i = 1, \cdots, n$, $N_i \in Dom_i$ is the singleton of Dom_i of attribute $Q_i \in DA$. Then g is also α-protective based on Definition 4, since all instances of each value in each DA are generalized to a single most general value. □

Observation 1 *k-Anonymity and α-protection can be achieved simultaneously in $\mathcal{DB}$ by means of full-domain generalization.*

Example 5 Continuing Example 2, suppose $f = elift/slift$ and consider the generalization lattice over QI attributes in Figure 2. Among three 3-anonymous full-domain generalizations, only $\langle S_1, R_1 \rangle$ and $\langle S_1, R_2 \rangle$ are also 1.2-protective with respect to $DA = \{Sex\}$.

Our task is to obtain α-protective k-anonymous full-domain generalizations. The naive approach is the sequential way: first, obtain k-anonymous full-domain generalizations and then restrict to the subset of these that are α-protective. Although this would solve the problem, it is a very expensive solution: discrimination should be measured for each k-anonymous full-domain generalization to determine whether it is α-protective. In the next section we present a more efficient algorithm that takes advantage of the common properties of α-protection and k-anonymity.

5.2 The algorithm

In this section, we present an optimal algorithm for obtaining all possible full-domain generalizations with which $\mathcal{DB}$ is α-protective k-anonymous.

5.2.1 Foundations

Observation 2 (Subset property of α-protection) *From Definition 4, observe that if $\mathcal{DB}$ is α-protective with respect to DA, it is α-protective w.r.t. any subset of attributes in DA. The converse property does not hold in general.*

For example, if Table 1 is 1.2-protective w.r.t $DA = \{Sex, Race\}$, Table 1 must also be 1.2-protective w.r.t. $DA = \{Sex\}$ and $DA = \{Race\}$. Otherwise put, if Table 1 is not 1.2-protective w.r.t. $DA = \{Sex\}$ or it is not 1.2 protective w.r.t. $DA = \{Race\}$, it cannot be 1.2-protective w.r.t. $DA = \{Sex, Race\}$. This is in correspondence with the subset property of k-anonymity. Thus, α-protection w.r.t. all strict subsets of DA is a necessary (but not sufficient) condition for α-protection w.r.t. DA. Then, given generalization hierarchies over QI, the generalizations that are not α-protective w.r.t. a subset DA' of DA can be discarded along with all their descendants in the hierarchy. To prove the generalization property of α-protection, we need a preliminary well-known mathematical result, stated in the following lemma.

Lemma 1 *Let $x_1, \cdots, x_n, y_1, \cdots, y_n$ be positive integers and let $x = x_1 + \cdots + x_n$ and $y = y_1 + \cdots + y_n$. Then*

$$\min_{1\le i\le n}\left\{\frac{x_i}{y_i}\right\} \le \frac{x}{y} \le \max_{1\le i\le n}\left\{\frac{x_i}{y_i}\right\}.$$

Proof Without loss of generality, suppose that $\frac{x_1}{y_1} \le \cdots \le \frac{x_n}{y_n}$. Then

$$\frac{x}{y} = \frac{y_1}{y}\frac{x_1}{y_1} + \cdots + \frac{y_n}{y}\frac{x_n}{y_n} \le \left(\frac{y_1}{y} + \cdots + \frac{y_n}{y}\right)\frac{x_n}{y_n} \le \frac{x_n}{y_n}.$$

The other inequality is proven analogously. □

Proposition 1 (Generalization property of α-protection) *Let $\mathcal{DB}$ be a data table and P and Q be nodes in the generalization lattice of DA with $D_P \le_D D_Q$. If $\mathcal{DB}$ is α-protective w.r.t. to P considering minimum support $ms = 1$ and discrimination measure* elift *or* clift*, then $\mathcal{DB}$ is also α-protective w.r.t. to Q.*

Proof Let $A^1, \ldots, A^n$ and A be itemsets in P and Q, respectively, such that $\{A^1, \ldots, A^n\} = \gamma^{-1}(A)$. That is, A is the generalization of $\{A^1, \ldots, A^n\}$. Let B be an itemset from attributes in $QI \setminus DA$, and C a decision item. For simplicity, assume that $supp(A^i, B) > 0$ for $i = 1, \ldots, n$. According to Section 4, for the PD rule $c : A, B \to C$,

$$elift(c) = \frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}} \quad \text{and} \quad clift(c) = \frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(X,B,C)}{supp(X,B)}},$$

where X is the most favored itemset in Q with respect to B and C. Since $supp(A, B) = \sum_i supp(A^i, B)$, and $supp(A, B, C) = \sum_i supp(A^i, B, C)$, by Lemma 1 we obtain that

$$\frac{supp(A, B, C)}{supp(A, B)} \le \max_i \frac{supp(A^i, B, C)}{supp(A^i, B)}.$$

Hence if none of the rules $A^i, B \to C$ are α-discriminatory with respect to the measure *elift*, then the rule $A, B \to C$ is not α-discriminatory. Now we consider the measure *clift*. Let Y be the most favored itemset in P with respect to the itemsets B and the item C. By following an analogous argument, we obtain that

$$\frac{supp(X,B,C)}{supp(X,B)} \geq \frac{supp(Y,B,C)}{supp(Y,B)}.$$

Therefore if none of the rules $A^i, B \to C$ are α-discriminatory with respect to the measure *clift*, then c is not α-discriminatory. □

For example, considering $DA = \{Race\}$ and $f = elift$ or $f = clift$, based on the generalization property of k-anonymity, if Table 1 is 3-anonymous w.r.t. $\langle R_0, H_0\rangle$, it must be also 3-anonymous w.r.t. $\langle R_1, H_0\rangle$ and $\langle R_0, H_1\rangle$. However, based on the generalization property of α-protection, if Table 1 is 1.2-protective w.r.t. $\langle R_0, H_0\rangle$, it must be also 1.2-protective w.r.t. $\langle R_1, H_0\rangle$, which contains the generalization of the attributes in DA, but not necessarily w.r.t. $\langle R_0, H_1\rangle$ (the latter generalization is for an attribute not in DA). Thus, we notice that the generalization property of α-protection is weaker than the generalization property of k-anonymity, because the former is only guaranteed for generalizations of attributes in $DA \subseteq QI$, whereas the latter holds for generalizations of any attribute in QI. Moreover, the generalization property has a limitation. Based on Definition 4, a data table is α-protective w.r.t. DA if all PD frequent rules extracted from the data table are not α-discriminatory w.r.t. DA. Hence, a data table might contain PD rules which are not α-protective and not frequent, *e.g. Race=White, Hours=35* $\to$ *Credit_approved=no*, *Race=White, Hours=37* $\to$ *Credit_approved=no*, *Race=White, Hours=36* $\to$ *Credit_approved=no*, where $DA = \{Race\}$. However, after generalization, frequent PD rules can appear which might be α-discriminatory and discrimination will show up, *e.g. Race=White, Hours=[1-40)* $\to$ *Credit _approved=no*. This is why the generalization property of α-protection requires that α-protection w.r.t. P hold for all PD rules, frequent and infrequent (this explains the condition $ms = 1$ in Proposition 1). The next property allows improving the efficiency of the algorithm for obtaining α-protective k-anonymous data tables by means of full-domain generalizations. Its proof is straightforward.

Proposition 2 (Roll-up property of α-protection) *Let $\mathcal{DB}$ be a data table with records in a domain tuple DT, let DT' be a domain tuple with $DT \leq_D DT'$, and let $\gamma : DT \to DT'$ be the associated generalization function. The support of an itemset X in DT' is the sum of the supports of the itemsets in $\gamma^{-1}(X)$.*

5.2.2 Overview

We take Incognito as an optimal anonymization algorithm based on the above properties and extend it to generate the set of all possible α-protective k-anonymous full-domain generalizations of $\mathcal{DB}$. Based on the subset property for α-protection and k-anonymity, the algorithm, named α-protective Incognito, begins by checking single-attribute subsets of QI, and then iterates by checking k-anonymity and α-protection with respect to increasingly larger subsets, in a manner reminiscent of [2]. Consider a graph of candidate multi-attribute generalizations (nodes) constructed from a subset of QI of size i. Denote this subset by C_i. The set of direct multi-attribute generalization relationships (edges) connecting these nodes is denoted by E_i.

The i-th iteration of α-protective Incognito performs a search that determines first the k-anonymity status and second the α-protection status of table $\mathcal{DB}$ with respect to each candidate generalization in C_i. This is accomplished using a modified bottom-up breadth-first search, beginning at each node in the graph that is not the direct generalization of some other node. A modified breadth-first search over the graph yields the set of multi-attribute generalizations of size i with respect to which $\mathcal{DB}$ is α-protective k-anonymous (denoted by S_i). After obtaining the entire S_i, the algorithm constructs the set of candidate nodes of size $i+1$ (C_{i+1}), and the edges connecting them (E_{i+1}) using the subset property.

5.2.3 Description

Algorithm 1 describes α-protective Incognito. In the i-th iteration, the algorithm determines the k-anonymity status of $\mathcal{DB}$ with respect to each node in C_i by computing the frequency set in one of the following ways: if the node is root, the frequency set is computed using $\mathcal{DB}$. Otherwise, for non-root nodes, the frequency set is computed using all parents' frequency sets. This is based on the roll-up property for k-anonymity. If $\mathcal{DB}$ is k-anonymous with respect to the attributes of the node, the algorithm performs two actions. First, it marks all direct generalizations of the node as k-anonymous. This is based on the generalization property for k-anonymity: these generalizations need not be checked anymore for k-anonymity in the subsequent search iterations. Second, if the node contains at least one PD attribute and $i \leq \tau$ (where τ is the discrimination granularity level, see definition further below), the algorithm determines the α-protection status of $\mathcal{DB}$ by computing the *Check α-protection*(i, $node$) function (see Algorithm 2). If $\mathcal{DB}$ is α-protective w.r.t. the attributes of the node, the algorithm marks as α-protective k-anonymous all direct generalizations of the node which are α-protective according to the generalization property of α-protection. The algorithm will not check them anymore for α-protection in the subsequent search iterations. Finally, the algorithm constructs C_{i+1} and E_{i+1} by considering only nodes in C_i that are marked as α-protective k-anonymous.

The discrimination granularity level $\tau \leq |QI|$ is one of the inputs of α-protective Incognito. The larger τ, the more protection regarding discrimination will be achieved. The reason is that, if the algorithm can check the status of α-protection in $\mathcal{DB}$ w.r.t. nodes which contain more attributes (*i.e.*, finer-grained subsets of protected groups), then more possible local niches of discrimination in $\mathcal{DB}$ are discovered. However, a greater τ leads to more computation by α-protective Incognito, because α-protection of $\mathcal{DB}$ should be checked in more iterations. In fact, by setting $\tau < |QI|$, we can provide a trade-off between efficiency and discrimination protection.

As mentioned above, Algorithm 2 implements the *Check α-protection*(i, $node$) function to check the α-protection of $\mathcal{DB}$ with respect to the attributes of the node. To do it in an efficient way, first the algorithm generates the set of l-itemsets of attributes of $node$ with their support values, denoted by I_l, and the set of $(l+1)$-itemsets of attributes of $node$ and class attribute, with their support values, denoted by I_{l+1}, where $l = i$ is the number of items in the itemset. In SQL language, I_l and I_{l+1} are obtained from $\mathcal{DB}$ by issuing a suitable query. This computation is only necessary for root nodes in each iteration; for non-root nodes, I_l and I_{l+1} are obtained from I_l and I_{l+1} of parent nodes based on the roll-up property of α-protection. Then, PD classification rules (*i.e.* PD_{groups}) with the required values to compute each f in Figure 3 (*i.e.* n_1, a_1, n_2 and a_2) are obtained by scanning I_{l+1}. During the scan of I_{l+1}, PD classification rules $A, B \rightarrow C$ (*i.e.* PD_{groups}) are obtained with the respective values $a_1 = supp(A, B, C)$, $n_1 = supp(A, B)$ (note that $supp(A, B)$ is in I_l), $a_2 = supp(\neg A, B, C)$ (obtained from I_{l+1}), and $n_2 = supp(\neg A, B)$ (obtained

Algorithm 1 α-PROTECTIVE INCOGNITO

```
Input: Original data table $\mathcal{DB}$, a set $QI = \{Q_1, \cdots, Q_n\}$ of quasi-identifier attributes, a set of domain generalization
    hierarchies $Dom_1, \cdots, Dom_n$, a set of PD attributes $DA$, $\alpha$, $f$, $k$, C={Class item}, $ms$=minimum support, $\tau \leq$
    $|QI|$
Output: The set of $\alpha$-protective $k$-anonymous full-domain generalizations
 1: $C_1$={Nodes in the domain generalization hierarchies of attributes in $QI$}
 2: $C_{PD} = \{\forall C \in C_1$ s.t. $C$ is PD$\}$
 3: $E_1$={Edges in the domain generalization hierarchies of attributes in $QI$}
 4: $queue$=an empty queue
 5: for $i = 1$ to $n$ do
 6:   //$C_i$ and $E_i$ define a graph of generalizations
 7:   $S_i$=copy of $C_i$
 8:   {roots}={all nodes $\in C_i$ with no edge $\in E_i$ directed to them}
 9:   Insert {roots} into $queue$, keeping $queue$ sorted by height
10:   while $queue$ is not empty do
11:     $node$ = Remove first item from $queue$
12:     if $node$ is not marked as $k$-anonymous or $\alpha$-protective $k$-anonymous then
13:       if $node$ is a root then
14:         $frequencySet$= Compute the frequency set of $\mathcal{DB}$ w.r.t. attributes of $node$ using $\mathcal{DB}$.
15:       else
16:         $frequencySet$= Compute the frequency set of $\mathcal{DB}$ w.r.t. attributes of $node$ using the parents' frequency sets.
17:       end if
18:       Use $frequencySet$ to check $k$-anonymity w.r.t. attributes of $node$
19:       if $\mathcal{DB}$ is $k$-anonymous w.r.t. attributes of $node$ then
20:         Mark all direct generalizations of $node$ as $k$-anonymous
21:         if $\exists N \in C_{PD}$ s.t. $N \subseteq node$ and $i \leq \tau$ then
22:           if $node$ is a root then
23:             $MR$= CHECK $\alpha$-PROTECTION($i$, $node$) of $\mathcal{DB}$ w.r.t. attributes of $node$ using $\mathcal{DB}$.
24:           else
25:             $MR$= CHECK $\alpha$-PROTECTION($i$, $node$) of $\mathcal{DB}$ w.r.t. attributes of $node$ using parents' $I_l$ and
                $I_{l+1}$
26:           end if
27:           Use $MR$ to check $\alpha$-protection w.r.t. attributes of $node$
28:           if $MR = case_3$ then
29:             Mark all direct generalizations of $node$ that contain the generalization of $N$ as $k$-anonymous
                $\alpha$-protective
30:           else if $MR = case_1$ then
31:             Delete $node$ from $S_i$
32:             Insert direct generalizations of $node$ into $queue$, keeping $queue$ ordered by height
33:           end if
34:         end if
35:       else
36:         Steps 31-32
37:       end if
38:     else if $node$ is marked as $k$-anonymous then
39:       Steps 21-36
40:     end if
41:   end while
42:   $C_{i+1}, E_{i+1} = GraphGeneration(S_i, E_i)$
43: end for
44: Return projection of attributes of $S_n$ onto $\mathcal{DB}$ and $Dom_1$, ..., $Dom_n$
```

from I_l). By relaxing τ we can limit the maximum number of itemsets in I_l and I_{l+1} that are generated during the execution of α-protective Incognito.

After obtaining PD_{groups} with the values a_1, a_2, n_1 and n_2, Algorithm 2 computes the *Measure_disc* (α, ms, f) function (see Algorithm 3). This function takes f as a parameter and is based on the generalization property of α-protection. If $f = slift$ or $f = olift$ and if there exists at least one frequent group $A, B \rightarrow C$ in PD_{groups} with $slift(A, B \rightarrow C) \geq \alpha$, then $MR = case_1$ (*i.e.* $\mathcal{DB}$ is not α-protective w.r.t. attributes of $node$). Otherwise, $MR = case_2$ (*i.e.* $\mathcal{DB}$ is α-protective w.r.t. attributes of $node$). If $f = elift$ or $f = clift$, the generalization property of α-protection is satisfied, so if there exists at least one frequent group $A, B \rightarrow C$ in PD_{groups} with $elift(A, B \rightarrow C) \geq \alpha$, then $MR =$

Algorithm 2 CHECK α-PROTECTION $(i, node)$

```
$l = i$
$I_l$={$l$-itemsets containing attributes of $node$}
$I_{l+1}$={$(l+1)$-itemsets containing attributes of $node$ and class item $C$}
for each $R \in I_{l+1}$ do
    $X = R \backslash C$
    $a_1 = supp(R)$
    $n_1 = supp(X)$    // $X$ found in $I_l$
    $A$=largest subset of $X$ containing protected groups w.r.t. $DA$
    $T = R \backslash A$
    $Z = \neg A \cup T$
    $a_2 = supp(Z)$    // Obtained from $I_{l+1}$
    $n_2 = supp(Z \backslash C)$    // Obtained from $I_l$
    Add $R: A, B \rightarrow C$ to $PD_{groups}$ with values $a_1$, $n_1$, $a_2$ and $n_2$
end for
Return $MR$=MEASURE_DISC($\alpha$, $ms$, $f$)
```

Algorithm 3 MEASURE_DISC(α, ms, f)

```
if $f = slift$ or $olift$ then
    if $\exists$ a group $(A, B \rightarrow C)$ in $PDgroup$ which is frequent w.r.t. $ms$ and $\alpha$-discriminatory w.r.t. $f$ then
        Return $MR = Case_1$    // $\mathcal{DB}$ is not $\alpha$-protective w.r.t. attributes of $node$
    else
        Return $MR = Case_2$    // $\mathcal{DB}$ is $\alpha$-protective w.r.t. attributes of $node$
    end if
end if
if $f = elift$ or $clift$ then
    if $\exists$ a group $(A, B \rightarrow C)$ in $PDgroup$ which is frequent w.r.t. $ms$ and $\alpha$-discriminatory w.r.t. $f$ then
        Return $MR = Case_1$    // $\mathcal{DB}$ is not $\alpha$-protective w.r.t. attributes of $node$
    else if $\exists$ a group $(A, B \rightarrow C)$ in $PDgroup$ which is infrequent w.r.t. $ms$ and $\alpha$-discriminatory w.r.t. $f$ then
        Return $MR = Case_2$    // $\mathcal{DB}$ is $\alpha$-protective w.r.t. attributes of $node$
    else if $f = clift$ and $\exists$ a group $(A, B \rightarrow C)$ in $PDgroup$ which is infrequent w.r.t. $ms$ whose confidence is
    lower than the confidence of the most favored item considered in the computation of $clift$ then
        Return $MR = Case_2$    // $\mathcal{DB}$ is $\alpha$-protective w.r.t. attributes of $node$
    else
        Return $MR = Case_3$    // $\mathcal{DB}$ is $\alpha$-protective w.r.t. attributes of $node$ and subsets of its generalizations
    end if
end if
```

$case_1$. Otherwise if there exists at least one infrequent group $A, B \rightarrow C$ in PD_{groups} with $elift(A, B \rightarrow C) \geq \alpha$, then $MR = case_2$. Otherwise if all groups in PD_{groups}, frequent and infrequent, have $elift(A, B \rightarrow C) < \alpha$, $MR = case_3$.

It is worth mentioning that in the i-th iteration of α-protective Incognito, for each node in C_i, first k-anonymity will be checked and then α-protection. This is because the algorithm only checks α-protection for the nodes that contain at least one PD attribute, while k-anonymity is checked for all nodes. Moreover, in some iterations, the algorithm does not check α-protection if $\tau < |QI|$.

6 Experimental analysis

Our first objective is to evaluate the performance of α-protective Incognito (Algorithm 1) and compare it with Incognito. Our second objective is to evaluate the quality of unbiased anonymous data output by α-protective Incognito, compared to that of the anonymous data output by plain Incognito, using both general and specific data analysis metrics. We implemented all algorithms using Java and IBM DB2. All experiments were performed on an Intel Core i5 CPU with 4 GB of RAM. The software included Windows 7 Home Edition and DB2 Express Version 9.7. We considered different values of f, DA, k, α and τ in our experiments.

6.1 Data sets

Adult data set: This data set is also known as Census Income and it can be retrieved from the UCI Repository of Machine Learning Databases [15]. Adult has 6 continuous attributes and 8 categorical attributes. The class attribute represents two income levels, $\leq$50K or $>$50K. There are 45,222 records without missing values, pre-split into 30,162 and 15,060 records for training and testing. We ran experiments on the training set. We used the same 8 categorical attributes used in [17], shown in Table 3, and obtained their generalization hierarchies from the authors of [17]. For our experiments, we set $ms = 5\%$ and 8 attributes in Table 3 as QI, and $DA_1 = \{Race, Gender, Marital\_status\}$, $DA_2 = \{Race, Gender\}$ and $DA_3 = \{Race, Marital\_status\}$. The smaller ms, the more computation and the more discrimination discovery. In this way, we considered a very demanding scenario in terms of privacy (all 8 attributes were QI) and anti-discrimination (small ms).

Table 3 Description of the Adult data set

Attribute	#Distinct values	#Levels of hierarchies
Education	16	5
Marital_status	7	4
Native_country	40	5
Occupation	14	3
Race	5	3
Relationship	6	3
Sex	2	2
Work-class	8	5

German Credit data set: We also used this data set from [15]. It has 7 continuous attributes, 13 categorical attributes, and a binary class attribute representing low or high credit risk. There are 666 and 334 records, without missing values, for the pre-split training and testing, respectively. This data set has been frequently used in the anti-discrimination literature [34,24]. We used the 11 categorical attributes, shown in Table 4. For our experiments, we set $ms = 5\%$ and 10 attributes in Table 4 as QI, and $DA_1 = \{Gender, Marital\_status, Foreign\_worker\}$, $DA_2 = \{Gender, Marital\_status\}$ and $DA_3 = \{Gender, Foreign\_worker\}$.

Table 4 Description of the German Credit data set

Attribute	#Distinct values	#Levels of hierarchies
Account-status	4	3
Credit-history	5	3
Load-purpose	11	4
Savings-account	5	4
Employment	5	4
Marital-status	4	3
Sex	2	2
Existing-credits	4	3
Job	4	3
Foreign worker	2	2

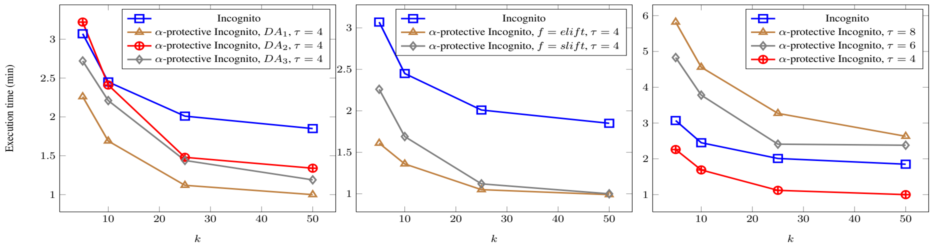

Fig. 4 Adult data set: Performance of Incognito and α-protective Incognito for several values of k, τ, f and DA. Unless otherwise specified, $f = slift$ and $DA = DA_1$ and $\alpha = 1.2$.

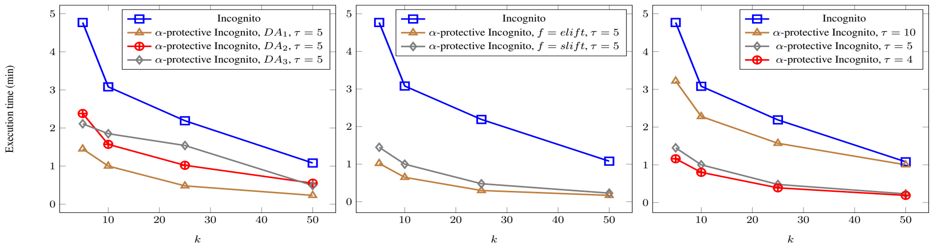

Fig. 5 German credit dataset: Performance of Incognito and α-protective Incognito for several values of k, τ, f and DA. Unless otherwise specified, $f = slift$, $DA = DA_1$ and $\alpha = 1.2$.

6.2 Performance

Figures 4 and 5 report the execution time of α-protective Incognito, for different values of τ, DA, f, k in comparison with Incognito in Adult and German Credit, respectively. We observe that for both data sets, as the size of k increases, the performance of both algorithms improves. This is mainly because, as the size of k increases, more generalizations are pruned as part of smaller subsets, and less execution time is needed. On the German Credit data set, α-protective Incognito is always faster than Incognito. On the Adult data set, α-protective Incognito is slower than Incognito only if the value of τ is very high (*e.g.* $\tau = 6$ or $\tau = 8$). The explanation is that, with α-protective Incognito, more generalizations are pruned as part of smaller subsets by checking both k-anonymity and α-protection, and less execution time is needed. The difference between the performance of the two algorithms gets smaller when k increases. In addition, because of the generalization property of α-protection with respect to *elift*, α-protective Incognito is faster for $f = elift$ than for $f = slift$. However, this difference is not substantial since, as we mentioned in Section 5.2, α-protection should consider all frequent and infrequent PD rules.

In summary, since α-protective Incognito provides extra protection against discrimination compared to Incognito, the cost can be sometimes a longer execution time, especially when the value of τ is very high, near $|QI|$. However, our results show that in most cases α-protective Incognito is even faster than Incognito. This is a remarkable result, because discrimination discovery is an intrinsically expensive task (as discrimination may be linked to a large number of attribute and value combinations).

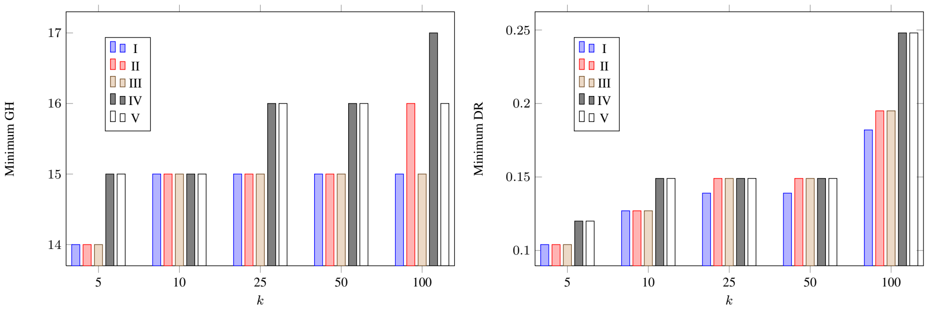

Fig. 6 Adult data set: General data quality metrics. Left, generalization height (GH). Right, discernibility ratio (DR). Results are given for k-anonymity (I); and α-protection k-anonymity with DA_2, $\alpha = 1.2$ (II); DA_2, $\alpha = 1.6$ (III); DA_1, $\alpha = 1.2$ (IV); DA_1, $\alpha = 1.6$ (V). In all cases $f = slift$, $DA_1 = \{Race, Gender, Marital\_status\}$, and $DA_2 = \{Race, Gender\}$.

6.3 Data quality

Privacy preservation and discrimination prevention are one side of the problem we tackle. The other side is retaining information so that the published data remain practically useful. Data quality can be measured in general or with respect to a specific data analysis task (*e.g.* classification).

First, we evaluate the data quality of the protected data obtained by α-protective Incognito and Incognito using standard general metrics: the generalization height [27,39] and discernibility [5]. The generalization height (GH) is the height of an anonymized data table in the generalization lattice. Intuitively, it corresponds to the number of generalization steps that were performed. The discernibility metric charges a penalty to each record for being indistinguishable from other records. For each record in equivalence QI class qid, the penalty is $|\mathcal{DB}[qid]|$. Thus, the discernibility cost is equivalent to the sum of the $|\mathcal{DB}[qid]|^2$. We define the discernibility ratio (DR) as DR= $\frac{\sum_{qid}|\mathcal{DB}[qid]|^2}{|\mathcal{DB}|^2}$. Note that: i) $0 \leq$ DR ≤ 1; ii) lower DR and GH mean higher data quality. From the list of full-domain generalizations obtained from Incognito and α-protective Incognito, respectively, we compute the minimal full-domain generalization w.r.t. both GH and DR for each algorithm and compare them.

Second, we measure the data quality of the anonymous data obtained by α-protective Incognito and Incognito for a classification task using the classification metric CM from [23]. CM charges a penalty for each record generalized to a qid group in which the record's class is not the majority class. Lower CM means higher data quality. From the list of full-domain generalizations obtained from Incognito and α-protective Incognito, respectively, we compute the minimal full-domain generalization w.r.t. CM for each algorithm and we compare them. In addition, to evaluate the impact of our transformations on the accuracy of a classification task, we first obtain the minimal full-domain generalization w.r.t. CM to anonymize the training set. Then, the same generalization is applied to the testing set to produce a generalized testing set. Next, we build a classifier on the anonymized training set and measure the classification accuracy (CA) on the generalized records of the testing set. For classification models we use the well-known decision tree classifier J48 from the Weka software package [47]. We also measure the classification accuracy on the original data without anonymization. The difference represents the cost in terms of classification accuracy for achieving either both privacy preservation and discrimination prevention or privacy preservation only.

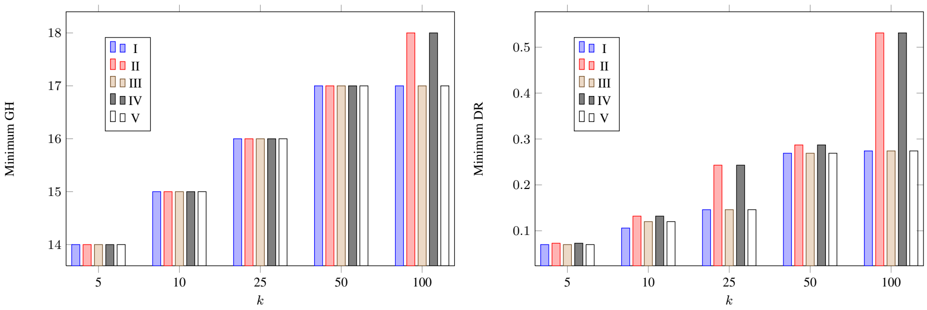

Fig. 7 German credit dataset: General data quality metrics. Left, generalization height (GH). Right, discernibility ratio (DR). Results are given for k-anonymity (I); and α-protection k-anonymity with DA_2, $\alpha = 1.2$ (II); DA_2, $\alpha = 1.6$ (III); DA_1, $\alpha = 1.2$ (IV); DA_1, $\alpha = 1.6$ (V). In all cases $f = slift$, $DA_1 = \{Gender, Marital\_status, Foreign\_worker\}$, and $DA_2 = \{Gender, Marital\_status\}$.

Figures 6 and 7 summarize the data quality results using general metrics for different values of k, DA and α, where $f = slift$ and $\tau = |QI|$ in Adult and German Credit, respectively. We found that the data quality of k-anonymous tables (*i.e.* in terms of GH and DR) without α-protection is equal or slightly better than the quality of k-anonymous tables with α-protection. This is because the α-protection k-anonymity requirement provides extra protection (*i.e.*, against discrimination) at the cost of some data quality loss when DA and k are large and α is small. As k increases, more generalizations are needed to achieve k-anonymity, which increases GH and DR. We performed the same experiment for other discrimination measures f, and we observed a similar trend (details omitted due to lack of space). As discussed in Section 5.2.3, $\tau = |QI|$ indicates the worst-case anti-discrimination scenario.

The left-hand side charts of Figures 8 and 9 summarize the data quality results using the classification metric (CM) for different values of k, DA and α, where $f = slift$ and $\tau = |QI|$. It can be seen that the information loss is higher in the German credit data set than in the Adult data set, due to the former being more biased (that is, having more α-discriminatory rules). However, the relevant comparison is not *between* data sets, but rather *within* each data set. In this respect, we notice that, for each data set, the data quality of k-anonymous tables (*i.e.* in terms of CM) without α-protection is equal or slightly better than the quality of k-anonymous tables with α-protection. This is because the α-protection k-anonymity requirement provides extra protection (*i.e.*, against discrimination) at the cost of some data quality loss when DA and k are large. The right-hand side charts of Figures 8 and 9 summarize the impact of achieving k-anonymity or α-protection k-anonymity on the percentage classification accuracy (CA) of J48 for different values of k, DA and α, where $f = slift$. We observe a similar trend as for CM. The accuracies of J48 using k-anonymous tables without α-protection are equal or only slightly better than the accuracies of J48 using k-anonymous tables with α-protection.

We also extend our results to alternative data mining algorithms. Tables 5 and 6 show for each data set, respectively, the accuracy for various types of classifiers, including decision trees (J48), naïve Bayes, logistic regression, and rule induction (RIPPER and PART) obtained from original data, 50-anonymous and 50-anonymous 1.2-protective. In either data set, we do not observe a significant difference between the accuracy of the classifiers obtained from the k-anonymous and the k-anonymous α-protective version of original data tables. These results support the conclusion that the transformed data obtained by our ap-

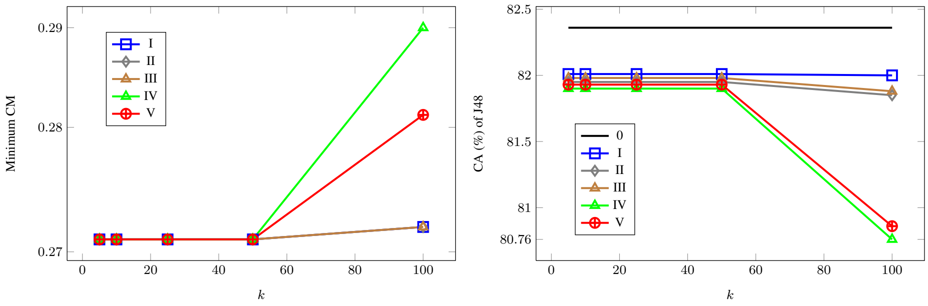

Fig. 8 Adult data set: Data quality for classification analysis. Left, classification metric (CM). Right, classification accuracy, in percentage (CA). Results are given for the original data (0); k-anonymity (I); and α-protection k-anonymity with DA_2, $\alpha = 1.2$ (II); DA_2, $\alpha = 1.6$ (III); DA_1, $\alpha = 1.2$ (IV); DA_1, $\alpha = 1.6$ (V). In all cases $f = slift$, $DA_1 = \{Race, Gender, Marital\_status\}$, and $DA_2 = \{Race, Gender\}$.

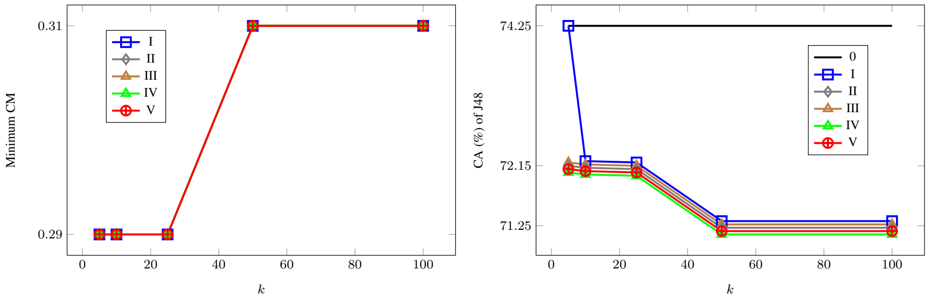

Fig. 9 German credit dataset: Data quality for classification analysis. Left, classification metric (CM). Right, classification accuracy, in percentage (CA). Results are given for the original data (0); k-anonymity (I); and α-protection k-anonymity with DA_2, $\alpha = 1.2$ (II); DA_2, $\alpha = 1.6$ (III); DA_1, $\alpha = 1.2$ (IV); DA_1, $\alpha = 1.6$ (V). In all cases $f = slift$, $DA_1 = \{Gender, Marital\_status, Foreign\_worker\}$, and $DA_2 = \{Gender, Marital\_status\}$.

proach are still usable for learning models and finding patterns, while minimizing both privacy and discrimination threats.

Table 5 Adult dataset: accuracy for various types of classifiers. In all cases $f = slift$, $DA = \{Race, Gender, Marital\_status\}$, and $\tau = |QI|$.

Classifier	original data table	50-anony. data table	50-anony. data table with 1.2-protec.
J48	82.36	82.01	82.01
Naïve Bayes	79.40	82.01	82.01
Logistic regression	82.95	82.08	82.08
RIPPER	82.7	81.34	81.34
PART	82.48	82.01	82.01

7 Extensions

We consider here alternative privacy models and anti-discrimination requirements.

Table 6 German credit dataset: accuracy for various types of classifiers. In all cases $f = slift$, $DA = \{Gender, Marital\_status, Foreign\_worker\}$, and $\tau = |QI|$.

Classifier	**original data table**	**50-anony. data table**	**50-anony. data table with 1.2-protec.**
J48	74.25	71.25	71.25
Naïve Bayes	71.25	71.25	71.25
Logistic regression	72.15	71.25	71.25
RIPPER	76.94	71.25	71.25
PART	66.46	71.25	70.95

7.1 Alternative privacy models

7.1.1 Attribute disclosure

k-Anonymity can protect the original data against record linkage attacks, but it cannot protect the data against attribute linkage (disclosure). In the attack of attribute linkage, the attacker may not precisely identify the record of the specific individual, but could infer his/her sensitive values (*e.g.*, salary, disease) from the published data table $\mathcal{DB}$. In contrast to k-anonymity, the privacy models in attribute linkage assume the existence of sensitive attributes in $\mathcal{DB}$ such that $QI \cap S = \emptyset$. Some models have been proposed to address this type of threat. The most popular ones are l-diversity and t-closeness. The general idea of these models is to diminish the correlation between QI attributes and sensitive attributes (see [32,29] for formal definitions). As shown in [32,29], by using full-domain generalizations over QI, we can obtain data tables protected against attribute disclosure. Considering attribute disclosure risks, we focus on the problem of producing an anonymized version of $\mathcal{DB}$ which is protected against attribute disclosure and free from discrimination (*e.g.*, α-protective l-diverse data table). We study this problem considering the following possible relations between QI, DA and S:

- $DA \subseteq QI$: It is possible that the original data are biased in the subsets of the protected groups which are defined by sensitive attributes (*e.g.* women who have medium salary). In this case, only full-domain generalizations which include the generalization of protected groups values can make $\mathcal{DB}$ α-protective. This is because the generalization is only performed over QI attributes.
- $DA \subseteq S$: A full-domain generalization over QI can make the original data α-protective only if $\mathcal{DB}$ is biased in the subsets of protected groups which are defined by QI attributes. In other scenarios, *i.e.*, when data are biased versus some protected groups or subsets of protected groups which are defined by sensitive attributes, full-domain generalizations over QI cannot make $\mathcal{DB}$ α-protective. One possible solution is to generalize attributes which are both sensitive and PD (*e.g.*, *Religion* in some applications), even if they are not in QI.

Observation 3 *If $DA \subseteq QI$, l-diversity/t-closeness and α-protection can be achieved simultaneously in $\mathcal{DB}$ by means of full-domain generalization.*

Since the subset and generalization properties are also satisfied for l-diversity and t-closeness, to obtain all full-domain generalizations with which data are α-protective and protected against attribute disclosure, we take α-protective Incognito and make the following changes: 1) every time a data table is tested for k-anonymity, it is also tested for l-diversity or t-closeness; 2) every time a data table is tested for α-protection, it is tested w.r.t. attributes of $node$ and sensitive attributes. This can be done by simply updating the *Check α-protection* function. Just as the data quality of k-anonymous data tables without l-diversity or t-closeness is slightly better than the quality of k-anonymous data tables with l-diversity or t-closeness, we expect a similar slight quality loss when adding l-diversity or t-closeness to k-anonymity α-protection.

7.1.2 Differential privacy

Differential privacy is a privacy model that provides a worst-case privacy guarantee in the presence of arbitrary external information. It protects against any privacy breaches resulting from joining different databases. Informally, differential privacy requires that the output of a data analysis mechanism be approximately the same, even if any single record in the input database is arbitrarily added or removed (see [10] for a formal definition). There are several approaches for designing algorithms that satisfy differential privacy. One of the best-known approaches is *Laplace noise addition*. After the query function is computed on the original data set $\mathcal{DB}$, Laplace-distributed random noise is added to the query result, where the magnitude of the noise depends on the sensitivity of the query function and a privacy budget. The sensitivity of a function is the maximum difference of its outputs from two data tables that differ only in one record. We define a differentially private data table as an anonymized data table generated by a function (algorithm) which is differentially private. There are some works available in literature studying the problem of differentially private data release [11]. The general structure of these approaches is to first build a contingency table of the original raw data over the database domain. After that, noise is added to each frequency count in the contingency table to satisfy differential privacy. However, as mentioned in [33], these approaches are not suitable for high-dimensional data with a large domain because when the added noise is relatively large compared to the count, the utility of the data is significantly decreased. In [33], a generalization-based algorithm for differentially private data release is presented. It first probabilistically generates a generalized contingency table and then adds noise to the counts. Thanks to generalization, the count of each partition is typically much larger than the added noise. In this way, generalization helps achieving a differentially private version of $\mathcal{DB}$ with higher data utility. Considering the differential privacy model, we focus on the problem of producing a private version of $\mathcal{DB}$ which is differentially private and free from discrimination with respect to DA. Since the differentially private version of $\mathcal{DB}$ is an approximation of $\mathcal{DB}$ generated at random, we have the following observation.

Observation 4 *Making original data table $\mathcal{DB}$ differentially private using Laplace noise addition can make $\mathcal{DB}$ more or less α-protective w.r.t. DA and f.*

Given the above observation and the fact that generalization can help to achieve differential privacy with higher data quality, we propose to obtain a noisy generalized contingency table of $\mathcal{DB}$ which is also α-protective. To do this, one solution is to add uncertainty to an algorithm that generates all possible full-domain generalizations with which $\mathcal{DB}$ is α-protective. As shown in [33], for higher values of the privacy budget, the quality of differentially private data tables is higher than the quality of k-anonymous data tables, while for smaller value of the privacy budget it is the other way round. Therefore, we expect that differential privacy plus discrimination prevention will compare similarly to the k-anonymity plus discrimination prevention presented in the previous sections of this paper.

7.2 Alternative anti-discrimination legal concepts

Unlike privacy legislation, anti-discrimination legislation is very sparse and includes different legal concepts, *e.g.* direct and indirect discrimination and the so-called genuine occupational requirement.

7.2.1 Indirect discrimination

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 *R*ace) and one knows that the specific zipcode is mostly inhabited by black people [8] (*i.e.*, there is high correlation between *Zipcode* and *Race* attributes). Then, if the protected groups do not exist in the original data table or have been removed from it due to privacy or anti-discrimination constraints, indirect discrimination still remains possible. Given DA, we define background knowledge as the correlation between DA and PND attributes which are in $\mathcal{DB}$:

$$\mathcal{BK} = \{A_i \rightarrow A_x | A_i \in \mathcal{A}, A_i \text{ is PND and } A_x \in DA\}$$

Given $\mathcal{BK}$, we define IA as a set of PND attributes in $\mathcal{DB}$ which are highly correlated to DA, determined according to $\mathcal{BK}$. Building on Definition 3, we introduce the notion of non-redlining α-protection for a data table.

Definition 7 (Non-redlining α-protected data table) Given $\mathcal{DB}(A_1, \cdots, A_n)$, DA, f and $\mathcal{BK}$, $\mathcal{DB}$ is said to satisfy non-redlining α-protection or to be non-redlining α-protective w.r.t. DA and f if each PND frequent classification rule $c : D, B \rightarrow C$ extracted from $\mathcal{DB}$ is α-protective, where D is a PND itemset of IA attributes and B is a PND itemset of $\mathcal{A}\backslash IA$ attributes.

Given DA and $\mathcal{BK}$, releasing a non-redlining α-protective version of an original table is desirable to prevent indirect discrimination against protected groups w.r.t. DA. Since indirect discrimination against protected groups originates from the correlation between DA and IA attributes, a natural countermeasure is to diminish this correlation. Then, an anonymized version of an original data table protected against indirect discrimination (*i.e.* non-redlining α-protective) can be generated by generalizing IA attributes. As an example, generalizing all instances of 47677, 47602 and 47678 zipcode values to the same generalized value 476** can prevent indirect discrimination against black people living in the 47602 neighborhood.

Observation 5 *If $IA \subseteq QI$, non-redlining α-protection can be achieved in $\mathcal{DB}$ by means of full-domain generalization.*

Consequently, non-redlining α-protection can be achieved with each of the above-mentioned privacy models based on full-domain generalization of $\mathcal{DB}$ (*e.g.* k-anonymity), as long as $IA \subseteq QI$. Fortunately, the subset and generalization properties satisfied by α-protection are also satisfied by non-redlining α-protection. Hence, in order to obtain all possible full-domain generalizations with which $\mathcal{DB}$ is indirect discrimination- and privacy-protected, we take α-protective Incognito and make the following changes: 1) add $\mathcal{BK}$ as the input of the algorithm and determine IA w.r.t. $\mathcal{BK}$, where PD attributes are removed from $\mathcal{DB}$;

[8] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:62011CJ0385:EN:Not

2) every time a data table is tested for α-protection, test it for non-redlining α-protection instead. Considering the above changes, when combining indirect discrimination prevention and privacy protection, we expect similar data quality and algorithm performance as we had when combining direct discrimination prevention and privacy protection.

7.2.2 Genuine occupational requirement

The legal concept of genuine occupational requirement refers to detecting that part of the discrimination which may be explained by other attributes [48], named legally-grounded attributes; *e.g.*, denying credit to women may be explainable if most of them have low salary or delay in returning previous credits. Whether low salary or delay in returning previous credits is an acceptable legitimate argument to deny credit is for the law to determine. Given a set LA of legally-grounded attributes in $\mathcal{DB}$, there are some works which attempt to cater technically to them in the anti-discrimination protection [31,48,12]. The general idea is to prevent only unexplainable (bad) discrimination. Loung *et al.* [31] propose a variant of the k-nearest neighbor (k-NN) classification which labels each record in a data table as discriminated or not. A record t is discriminated if: i) it has a negative decision value in its class attribute; and ii) the difference between the proportions of k-nearest neighbors of t w.r.t. LA whose decision value is the same of t and belong to the same protected-by-law groups as t and the ones that do not belong to the same protected groups as t is greater than the discrimination threshold. This implies that the negative decision for t is not explainable on the basis of the legally-grounded attributes, but it is biased by group membership. We say that a data table is protected only against unexplainable discrimination w.r.t. DA and LA if the number of records labeled as discriminated is zero (or near zero). An anonymized version of an original data table which is protected against unexplainable discrimination can be generated by generalizing LA and/or DA attributes. Given a discriminated record, generalizing LA and/or DA attributes can decrease the difference between the two above-mentioned proportions. Hence, an anonymized version of an original data table which is privacy-protected and protected against unexplainable discrimination can be obtained using full-domain generalization over QI attributes as long as $DA \subseteq QI$ and $LA \subseteq QI$.

8 Conclusions

We have investigated the problem of discrimination- and privacy-aware data publishing and mining, *i.e.*, distorting an original data set in such a way that neither privacy-violating nor discriminatory inferences can be made on the released data sets, while maximizing the usefulness of data for learning models and finding patterns. To study the impact of data generalization (*i.e.* full-domain generalization) on discrimination prevention, we applied generalization not only for making the original data privacy-protected but also for making them protected against discrimination. We found that a subset of k-anonymous full-domain generalizations with the same or slightly higher data distortion than the rest (in terms of general and specific data analysis metrics) are also α-protective. Hence, k-anonymity and α-protection can be combined to attain privacy protection and discrimination prevention in the published data set. We have adapted to α-protection two well-known properties of k-anonymity, namely the subset and the generalization properties. This has allowed us to propose an α-protective version of Incognito, which can take as parameters several legally-grounded measures of discrimination and generate privacy- and discrimination-protected full-domain generalizations. We have evaluated the quality of data (*i.e.* in terms of various types of classifiers and rule induction algorithms) output by this algorithm, as well as its execution time. Both turn out to be nearly as good as with plain Incognito, so the toll paid to obtain α-protection is very reasonable. Finally, we have sketched how our approach can be extended to satisfy alternative privacy guarantees or anti-discrimination legal constraints. Detailed implementations of these extensions are left for future work.

Acknowledgments and disclaimer

The authors wish to thank Kristen LeFevre for providing the implementation of the Incognito algorithm and Guillem Rufian-Torrell for helping in the implementation of the algorithm proposed in this paper. This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TIN2011-27076-C03-01 “CO-PRIVACY”, TIN2012-32757 “ICWT” and CONSOLIDER INGENIO 2010 CSD2007-00004 “ARES”, and by the European Comission under FP7 projects “DwB” and “INTER-TRUST”. The second author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia. The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which 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. Fast algorithms for mining association rules in large databases. In *Proc. of the 20th Intl. Conf. on Very Large Data Bases*, pp. 487-499. VLDB, 1994.
3. R. Agrawal and R. Srikant. Privacy preserving data mining. In *SIGMOD 2000*, pp. 439-450. ACM, 2000.
4. Australian Legislation. (a) Equal Opportunity Act – Victoria State, (b) Anti-Discrimination Act – Queensland State, 2008.
5. R. J. Bayardo and R. Agrawal. Data privacy through optimal k-anonymization. In *ICDE 2005*, pp. 217-228. IEEE, 2005.
6. B. Berendt and S. Preibusch. Exploring discrimination: a user-centric evaluation of discrimination-aware data mining. In *IEEE 12th International Conference on Data Mining Workshops-ICDMW 2012*, pp. 344-351. IEEE Computer Society, 2012.
7. T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277-292, 2010.
8. B. Custers, T. Calders, B. Schermer and T. Z. Zarsky (eds.). *Discrimination and Privacy in the Information Society - Data Mining and Profiling in Large Databases.* Studies in Applied Philosophy, Epistemology and Rational Ethics 3. Springer, 2013.
9. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195-212, 2005.
10. C. Dwork. Differential privacy. In *ICALP 2006*, LNCS 4052, pp. 112. Springer, 2006.
11. C. Dwork. A firm foundation for private data analysis. *Comm. of the ACM*, 54(1):8695, 2011.
12. C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. Fairness through awareness. In *ITCS 2012*, pp. 214-226. ACM, 2012.
13. European Union Legislation. Directive 95/46/EC, 1995.
14. European Union Legislation, (a) Race Equality Directive,2000/43/EC, 2000; (b) Employment Equality Directive, 2000/78/EC, 2000; (c) Equal Treatment of Persons, European Parliament legislative resolution, P6_TA(2009)0211, 2009.
15. A. Frank and A. Asuncion. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2010. `http://archive.ics.uci.edu/ml/datasets`
16. 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.
17. B. C. M. Fung, K. Wang, and P. S. Yu. Top-Down Specialization for Information and Privacy Preservation. In *ICDE 2005*, pp. 205-216. IEEE, 2005.

18. 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. Springer, 2011.
19. S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445-1459, 2013.
20. S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In *IEEE 12th International Conference on Data Mining Workshops-ICDMW 2012*, pp. 360-369. IEEE Computer Society, 2012.
21. S. Hajian and J. Domingo-Ferrer. A study on the impact of data anonymization on anti-discrimination. In *2012 IEEE 12th International Conference on Data Mining Workshops-ICDMW 2012*, pp. 352-359. IEEE Computer Society, 2012.
22. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.
23. V. S. Iyengar. Transforming data to satisfy privacy constraints. In *SIGKDD 2002*, pp.279288. ACM, 2002.
24. F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge Information Systems*, 33(1): 1-33, 2011.
25. F. Kamiran, T. Calders and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM 2010*, pp. 869-874. IEEE, 2010.
26. T. Kamishima, S. Akaho, H. Asoh and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD*, LNCS 7524, pp. 35-50. Springer, 2012.
27. K. Lefevre, D. J. Dewitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *SIGMOD 2005*, pp. 49-60. ACM, 2005.
28. K. Lefevre, D. J. Dewitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *ICDE 2006*, p. 25. IEEE, 2006.
29. N. Li, T. Li and S. Venkatasubramanian. t-Closeness: privacy beyond k-anonymity and l-diversity. In *IEEE ICDE 2007*, pp. 106-115. IEEE, 2007.
30. Y. Lindell and B. Pinkas. Privacy preserving data mining. In *Advances in Cryptology-CRYPTO'00*, LNCS 1880, Springer, 2000, pp. 36-53.
31. B. L. Loung, 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.
32. A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-Diversity: privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), Article 3, 2007.
33. N. Mohammed, R. Chen, B. C. M. Fung, and P. S. Yu. Differentially private data release for data mining. In *KDD 2011*, pp. 493-501. ACM, 2011.
34. D. Pedreschi, S. Ruggieri and F. Turini. Discrimination-aware data mining. In *KDD 2008*, pp. 560-568. ACM, 2008.
35. D. Pedreschi, S. Ruggieri and F. Turini. Measuring discrimination in socially-sensitive decision records. In *SDM 2009*, pp. 581-592. SIAM, 2009.
36. D. Pedreschi, S. Ruggieri and F. Turini. Integrating induction and deduction for finding evidence of discrimination. In *ICAIL 2009*, pp. 157-166. ACM, 2009.
37. D. Pedreschi, S. Ruggieri and F. Turini. The discovery of discrimination. In *Discrimination and Privacy in the Information Society* (eds. B. H. M. Custers, T. Calders, B. W. Schermer, and T. Z. Zarsky), volume 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics, pp. 4357. Springer, 2013.
38. S. Ruggieri, D. Pedreschi and F. Turini. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(2), Article 9, 2010.
39. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010-1027, 2001.
40. 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.
41. Statistics Sweden. Statistisk rjandekontroll av tabeller, databaser och kartor (Statistical disclosure control of tables, databases and maps, in Swedish). Örebro: Statistics Sweden, 2001 (downloaded Feb. 5, 2013). `http://www.scb.se/statistik/_publikationer/OV9999_2000I02_BR_X97P0102.pdf`
42. L. Sweeney. Datafly: A system for providing anonymity in medical data. In *Proc. of the IFIP TC11 WG11.3 11th International Conference on Database Security XI: Status and Prospects*, pp. 356-381, 1998.
43. L. Sweeney. k-Anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557-570, 2002.
44. United States Congress, *US Equal Pay Act*, 1963.
45. K. Wang, P. S. Yu and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In *ICDM 2004*, pp. 249-256. IEEE, 2004.
46. L. Willenborg and T. de Waal. *Elements of Statistical Disclosure Control*. Springer, 1996.
47. I. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2nd edition, 2005.
48. I. Zliobaite, F. Kamiran and T. Calders. Handling conditional discrimination. In *ICDM 2011*, pp. 992-1001. IEEE, 2011.