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

Discrimination- and Privacy-aware Patterns

Sara Hajian · Josep Domingo-Ferrer · Anna Monreale · Dino Pedreschi · Fosca Giannotti

Received: date / Accepted: date

Abstract Data mining is gaining societal momentum due to the ever increasing availability of large amounts of human data, easily collected by a variety of sensing technologies. We are therefore faced with unprecedented opportunities and risks: a deeper understanding of human behavior and how our society works is darkened by a greater chance of privacy intrusion and unfair discrimination based on the extracted patterns and profiles. Consider the case when a set of patterns extracted from the personal data of a population of individual persons is released for a subsequent use into a decision making process, such as, e.g., granting or denying credit. First, the set of patterns may reveal sensitive information about individual persons in the training population and, second, decision rules based on such patterns may lead to unfair discrimination, depending on what is represented in the training cases.

Although methods independently addressing privacy or discrimination in data mining have been proposed in the literature, in this context we argue that privacy and discrimination risks should be tackled *together*, and we present a methodology for doing so while publishing frequent pattern mining results. We describe a set of pattern sanitization methods, one for each discrimination measure used in the legal literature, to achieve a fair publishing of frequent patterns in combination with two possible privacy transformations: one

S. Hajian and J. Domingo-Ferrer
Universitat Rovira i Virgili, Department of Computer Engineering and Maths, UNESCO Chair in Data Privacy, Av. Països Catalans 26 - 43007 Tarragona, Catalonia. Tel.: +34 977558270
E-mail: {sara.hajian,josep.domingo}@urv.cat

A. Monreale and D. Pedreschi
Università di Pisa, Dipartimento di Informatica, Largo Pontecorvo, 3 - 56127 Pisa, Italy. Tel.: +39 050 2213162
E-mail: {annam,pedre}@di.unipi.it

F. Giannotti
ISTI-CNR, Pisa, Italy. Tel.: +39 050 315 299
E-mail: fosca.giannotti@isti.cnr.it

based on k-anonymity and one based on differential privacy. Our proposed pattern sanitization methods based on k-anonymity yield both privacy- and discrimination-protected patterns, while introducing reasonable (controlled) pattern distortion. Moreover, they obtain a better trade-off between protection and data quality than the sanitization methods based on differential privacy. Finally, the effectiveness of our proposals is assessed by extensive experiments.

Keywords Frequent patterns · Anti-discrimination · Privacy · Data mining

1 Introduction

Data mining is an increasingly important set of analytical processes that allow extracting useful knowledge hidden in large collections of data, especially human and social data sensed by the ubiquitous technologies that support most human activities in our age. As a matter of fact, the new opportunities to extract knowledge and understand human and social complex phenomena increase hand in hand with the risks of violation of fundamental human rights, such as privacy and non-discrimination.

Privacy refers to the individual right to keep personal sensitive information confidential while *discrimination* refers to unfair or unequal treatment of people based on membership to a category, group or minority, without regard to individual merit. Human rights laws not only have concern about data protection [16] but also prohibit discrimination [5,17] against protected groups on the grounds of ethnicity, 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. Clearly, preserving the great benefits of data mining within a privacy-aware and discrimination-aware technical system would lead to a wider social acceptance of a multitude of new services and applications based on the knowledge discovery process.

The problem of protecting privacy within data mining has been extensively studied since the 1970s, when Dalenius was the first to formulate the statistical disclosure control problem [12]. Research on data anonymization has carried on ever since in the official statistics community, and several computational procedures were proposed during the 1980s and 1990s, based on random noise addition, generalization, suppression, microaggregation, etc. (see [30] for a compendium). In the statistics community the approach was first to anonymize and then measure how much anonymity had been achieved, by either computing the probability of re-identification or performing record linkage experiments. In the late 1990s, researchers in the database community stated the k-anonymity model [50,52]: a data set is k-anonymous if its records are indistinguishable by an intruder within groups of k. The novelty of this approach was that the anonymity target was established *ex ante* and then computational procedures were used to reach that target. The computational procedures initially proposed for *k-anonymity* were generalization and suppression; microaggregation was proposed later as a natural alternative [13]. In 2000, the database com-

munity [3] coined the new term privacy-preserving data mining (PPDM) and proposed anonymization via random noise addition.

Differential privacy [14] is a more recent anonymity model that holds much promise: it seeks to render the influence of the presence/absence of any individual on the released outcome negligible. The computational approach initially proposed to achieve differential privacy was Laplace noise addition, although other approaches have recently been proposed [51].

On the other hand, the issue of anti-discrimination has been recently considered from a data mining perspective [45]. Some proposals are oriented to the discovery and measurement of discrimination using data mining [45,46, 49,41], while others deal with preventing data mining from becoming itself a source of discrimination, due to automated decision making based on discriminatory models extracted from inherently biased datasets [24,32,10,56,34]. In fact, *discrimination prevention in data mining* (DPDM) consists of extracting models that do not lead to discriminatory decisions even if trained from a dataset containing them.

Up to now, PPDM and DPDM have been studied in isolation. We argue in this paper that, in significant data mining processes, privacy and anti-discrimination protection should be addressed *together*. Consider the case in which a set of patterns extracted (mined) from the personal data of a population of individual persons is released for subsequent use in a decision making process, such as, e.g., granting or denying credit. First, the set of patterns may reveal sensitive information about individual persons in the training population. Second, decision rules based on such patterns may lead to unfair discrimination, depending on what is represented in the training cases. The following example illustrates this point. Assume a credit institution, e.g., a bank, wants to release among its employees the rules to grant/deny credit, for the purpose of supporting future decision making. Assume that such rules have been mined from decision records accumulated during the past year in a certain city, such as those illustrated in Table 1. Consider two options:

- *Protection against the privacy threat only.* Only rules used in at least k different credit applications are published, in order to protect the applicants' privacy according to k-anonymity. This would allow releasing a rule such that $Sex = female \rightarrow Credit\_approved = no$ if k or more female applicants have been denied credit. Clearly, using such a rule for credit scoring is discriminatory against women.
- *Protection against the discrimination threat only.* Discriminatory rules are sanitized, in order to prevent discrimination against female applicants. However, one could publish high-support high-confidence rules such as $Job = veterinarian,\ salary > 15000 \rightarrow Credit\_approved = yes$ and $Job = veterinarian \rightarrow Credit\_approved = yes$. Assuming that the first rule holds for 40 people and the second one for 41, their release would reveal that there is only one veterinarian in the city that has been granted credit even if s/he makes no more than €15000 a year. This is a potential privacy violation,

that is, a probable disclosure of the applicant's identity, and therefore of his/her income level.

Table 1 A data table of personal decision records

Sex	**Job**	**Credit_history**	**Salary**	**Credit_approved**
Male	Writer	None taken	... €	Yes
Female	Lawyer	Paid-duly	... €	No
Male	Veterinarian	Paid-delay	...€	Yes
...	...	...	...	...

This simple example shows that protecting both privacy and non-discrimination is needed when disclosing a set of patterns. The next question is: why not simply apply known techniques for PPDM and DPDM one after the other? We show in this paper that this straightforward sequential approach does not work in general: we have no guarantee that applying a DPDM method after a PPDM one preserves the desired privacy guarantee, because the DPDM sanitization of a set of patterns may destroy the effect of the earlier PPDM sanitization (and the other way round). We therefore need a holistic method capable of addressing the two goals together, so that we can safely publish the patterns extracted from a dataset of personal information, while keeping the distortion of the extracted patterns as low as possible. A truly trustworthy technology for knowledge discovery should face both privacy and discrimination threats as two sides of the same coin. This line of reasoning also inspires in the General Data Protection Regulation proposed in 2012 by the European Commission, currently in process of approval by the European Parliament. This introduces measures to control profiling and discrimination within a broader concept of privacy and personal data[1].

The contributions of this paper, towards the above stated aim, are summarized as follows. First, we define a natural scenario of pattern mining from personal data records containing sensitive attributes, potentially discriminatory attributes and decision attributes. So, we characterize the problem statement of publishing a collection of patterns which is at the same time both privacy-protected and discrimination-free. Second, we propose new pattern sanitization methods for discrimination prevention when publishing frequent patterns. Third, we propose the notion of α-protective k-anonymous (i.e., discrimination- and privacy-protected) patterns to thwart both privacy and discrimination threats in a collection of published frequent patterns. Fourth, we present a combined pattern sanitization algorithm to obtain a α-protective k-anonymous version of the original pattern sets. Fifth, we study how to produce frequent patterns that guarantee both differential privacy and α-protection. Finally, we theoretically and empirically show that the proposed

[1] Article 20, General Data Protection Regulation, unofficial consolidated version provided by the Rapporteur, 22 October 2013. `http://www.janalbrecht.eu/fileadmin/material/Dokumente/DPR-Regulation-inofficial-consolidated-LIBE.pdf`

algorithms are effective at protecting against both privacy and discrimination threats while introducing reasonable (controlled) pattern distortion. Figure 1 depicts a taxonomy that describes what producing social-aware pattern sets means and briefly highlights the main topics addressed in this paper.

In our earlier conference paper [26], we introduced the idea of achieving simultaneous privacy and anti-discrimination protection in frequent pattern discovery. The work in that previous paper has been extended in many ways in this paper, as listed below:

- We extend the approach in [26] to the various measures of discrimination used in the legal literature. We propose new algorithms which are parametric to different measures of discrimination.
- The proposal in [26] does not take into consideration the so-called genuine requirement, i.e., the fact that some apparent discriminatory rule may be actually explained by other admissible factors that are not specified explicitly in the rule (*explainable discrimination*). We propose a new algorithm to make the frequent pattern protected against unexplainable discrimination only.
- We extend our ideas in [26] to rule-based classifiers. We measure the impact of our fairness transformation in terms of accuracy loss of the original and sanitized classifiers.
- We show how simultaneous anti-discrimination and privacy can be achieved in frequent pattern discovery while satisfying differential privacy, as an alternative.

The rest of this article is organized as follows. Section 2 reviews related work at great length. Section 3 presents the background material used in the paper: basic concepts used throughout the paper (Section 3.1); the methods that we use to obtain privacy-preserving pattern sets (Section 3.2); and the notion of discrimination-protected frequent patterns (Section 3.3). Section 4 states the main problem addressed in this paper. Section 5 describes our methods and algorithms to make the collection of patterns discrimination-free. In Section 6 we introduce our solution to solve the problem of simultaneous privacy and anti-discrimination pattern protection. Section 7 reports the evaluation of our sanitization methods. In Section 8 we study and discuss the realistic use of our sanitizations by analyzing the trade-off between protection and data utility. Finally, Section 9 concludes the paper and identifies future research topics in this context.

2 Related work

Privacy. In the last fifteen years plenty of privacy models and their variations have been proposed for protecting individual privacy while maintaining under control the trade-off between privacy and quality of resulting data or models. k-Anonymity [50,52], l-diversity [42], t-closeness [39], differential privacy [14] and crowd-blending [22] are examples of such models. A detailed description of

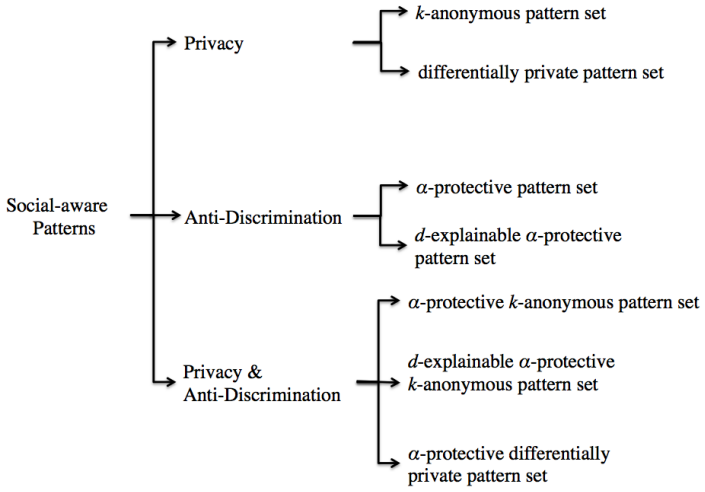

Fig. 1 Taxonomy of Social-aware Pattern Sets

different PPDM models and methods can be found in [1,21]. One key challenge in PPDM originates from the following privacy question [36]: do the data mining results themselves violate privacy? In other words, may the disclosure of extracted patterns reveal sensitive information? Some works on this problem are available, under the name of privacy-aware data analysis, that focus on the privacy of the individuals whose data is collected in the database [19,4,7,20, 38,37,55,8]. They study the problem of how to run a particular data mining algorithm on databases while satisfying the requirement of a particular privacy model.

Among the above-mentioned privacy models, k-anonymity and differential privacy have been studied in privacy-aware data analysis. In [19] the problem of sanitizing decision trees is studied and a method is given for directly building a k-anonymous decision tree from a private data set. The proposed algorithm is basically an improvement of the classical decision tree building algorithm, combining mining and anonymization in a single process (in-processing approach). In [4] the anonymity problem is addressed in the setting of frequent patterns. The authors define the notion of k-anonymous patterns and propose a methodology to guarantee the k-anonymity property in a collection of published frequent patterns (post-processing approach).

Differential privacy model has been used in [20,7,38,55,8]. In [20], an algorithm is proposed for building a classifier while guaranteeing differential privacy. In [38] the authors tackle the problem of extracting the top-k differentially private frequent patterns. This work is inspired by Bhaskar *et al.* [7], who proposed an approach for releasing the top k itemsets of a predefined length m. Recently, in [55] Zeng *et al.* proposed an approach that is based on the idea of truncating the transactions in order to introduce less noise during the generation of differential privacy patterns. Lastly, in [8] Bonomi addresses

the problem of mining differentially private *sequential* patterns that are in the form of a sequence rather than a simple subset of items.

In this paper, to provide patterns with both privacy and anti-discrimination protection, we investigate the combination of anti-discrimination sanitization for frequent patterns with the k-anonymity privacy transformation presented in [4] and the differential privacy transformation proposed in [38].

Anti-discrimination. According to current legislation, discrimination occurs when a group is treated "less favorably" [5] than others, or when "a higher proportion of people not in the group is able to comply" [17] with a qualifying criterium. As mentioned above, the issue of anti-discrimination has recently been considered from a data mining perspective [45], under the name of discrimination-aware data analysis. A substantial part of the existing literature on anti-discrimination in data mining is oriented to *discovering* and *measuring* discrimination. Other contributions deal with *preventing* discrimination. Summaries of contributions in discrimination-aware data analysis are collected in [11]. Pedreschi *et al.* [45,46,49,47] have introduced data mining approaches for discrimination discovery. The approaches have followed the legal principle of *under-representation* to unveil contexts of possible discrimination against *protected-by-law* groups (e.g., women). This is done by extracting classification rules from a dataset of historical decision records (inductive part); then, rules are ranked according to some *legally-grounded* measures of discrimination (deductive part).

In our framework we use these measures introduced in Pedreschi *et al.* [46] for measuring the degree of discrimination. In [56,34] the problem of handling conditional discrimination is addressed. It takes into account the concept of genuine requirement to detect that part of discrimination which may be explained by other (i.e., legally grounded) attributes. Loung *et al.* [41] implemented a concept of situation testing for measuring discrimination by checking the factors that may influence the decision outcome. The concept is almost the same as that of the above explainable discrimination. In [15], the concept of fairness in classification is addressed. This approach is also more related to the above notion of situation testing and explainable discrimination. The possible connection between fair classification and differential privacy is also discussed. Recently, in [54] Dwork *et al.* address the problem of fair classification that achieves both *group fairness*, i.e., the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole, and *individual fairness*, i.e., similar individuals should be treated similarly.

Beyond discrimination discovery, preventing knowledge-based decision support systems from making discriminatory decisions is a more challenging issue. In the motivating example, the credit granting organization might systematically have denied credit to women during the past year. If this biased historical dataset is used as training data by an automated credit granting system to learn classification rules, the learned rules will also show biased behavior toward women.

Discrimination prevention approaches can be classified according to the phase of the data mining process in which they operate: *pre-processing*, *in-processing* and *post-processing* methods.

Pre-processing methods. The strategy used in the methods belonging to this category consists in controlling distortion of the training set. In particular, they transform data in such a way that the discriminatory biases contained in the original data are removed so that no unfair decision models can be mined from the transformed data [31, 24, 25, 41, 15, 54, 6]. The property of the pre-processing approach is that it does not require changing the standard data mining algorithms and allows data publishing rather than knowledge publishing.

In-processing methods. The strategy used in the approaches in this category [32, 10, 56, 34, 35] modifies data mining algorithms in such a way that the resulting models do not contain unfair decisions. Calders and Verwer [10] consider three approaches to deal with naive Bayes models, two of which consist in modifying the learning algorithm: training a separate model for each protected group; and adding a latent variable to model the class value in the absence of discrimination. Kamiran *et al.* [32] modify the entropy-based splitting criterion in decision tree induction to account for attributes denoting protected groups. Kamishima *et al.* [35] measure the indirect causal effect of variables modeling grounds of discrimination on the independent variable in a classification model by their mutual information. Then, they apply a regularization (i.e., a change in the objective minimization function) to probabilistic discriminative models, such as logistic regression.

Post-processing methods. The third strategy is to post-process the extracted data mining models, instead of cleaning the original dataset or changing the data mining algorithms. Pedreschi *et al.* [46] alter the confidence of classification rules inferred by the CPAR algorithm. Calders and Verwer [10] act on the probabilities of a naive Bayes model. Kamiran *et al.* [32] re-label the class predicted at the leaves of a decision tree induced by C4.5. Lastly, Kamiran *et al.* [33] propose correcting predictions of probabilistic classifiers that are close to the decision boundary, given that (statistical) discrimination may occur when there is no clear feature supporting a positive or a negative decision.

Our anti-discrimination transformation belongs to the last category and considers the sanitization of frequent patterns. These are very simple structures that can be used as basic bricks to build more complex mining models such as association rules, classification or clustering models. In particular, our approach changes the support of the patterns to guarantee anti-discrimination protection. Moreover, our final goal is more complex because we want to produce frequent patterns that guarantee both privacy and anti-discrimination. This makes our work very different from the current state of the art.

Simultaneous privacy & anti-discrimination. To the best of our knowledge the first work addressing the problem of simultaneously achieving anti-discrimination and privacy in data mining is our conference paper [26]. In that paper we in-

troduced preliminary ideas on publishing frequent patterns while preventing discrimination and protecting privacy. Here, we build further on the results we obtained in [26] and we propose new algorithms to solve the same problem by taking into consideration other discrimination, privacy and data utility measures (see Section 1 for a list of the extensions in this paper w.r.t. [26]).

3 Background

In this section we introduce some notions that are useful to understand our contributions.

3.1 Basic definitions

Let $\mathcal{I} = \{i_1, \ldots, i_n\}$ be a set of items, where each *item* i_j has the form *attribute=value* (e.g., *Sex=female*). An *itemset* $X \subseteq \mathcal{I}$ is a collection of one or more items, e.g., {*Sex=female, Credit_history=none-taken*}. A *database* is a collection of data objects (records) and their attributes; more formally, a (transaction) database $\mathcal{D} = \{r_1, \ldots, r_m\}$ is a set of data records or transactions where each $r_i \subseteq \mathcal{I}$. Civil rights laws [5,17] explicitly identify the groups to be protected against discrimination, such as minorities and disadvantaged people, e.g., women. In our context, these groups can be represented as items, e.g., *Sex=female*, which we call potentially discriminatory (PD) items. A collection of PD items can be represented as an itemset, e.g., {*Sex=female, Foreign_worker=yes*}, which we call PD itemset or protected-by-law (or protected for short) groups, denoted by DI_b. An itemset X is PND if $X \cap DI_b = \emptyset$, e.g., {*credit_history=none-taken*} is a PND itemset where DI_b:{*Sex=female*}.

The attributes in a database $\mathcal{D}$ can be classified in several non-disjoint categories. *Identifiers* are attributes that uniquely identify individuals in the database, like *Passport number*. A *quasi-identifier* (QI) is a minimal set of attributes that can be joined with external information to re-identify individual records; e.g., *Sex*, *Zip* and *birthdate* can be linked to external public information to reveal the name of the corresponding individual. *Sensitive attributes* are those that contain sensitive information, such as *Disease* or *Salary*. *PD attributes* are those that can take PD items as values; for instance, *Ethnicity* and *Gender* where DI_b:{*Sex=female, Ethnicity=black*}. PD attributes can overlap with QI attributes (e.g., *Sex*, *Age*, *Marital_status*) and/or sensitive attributes (e.g., *Religion* in some applications).

A *decision (class) attribute* is one taking as values *yes* or *no* to report the outcome of a decision made on an individual; an example is the attribute *credit_approved*, which can be *yes* or *no*. The *support* of an itemset X in a database $\mathcal{D}$ is the number of records that contain X, i.e. $supp_{\mathcal{D}}(X) = |\{r_i \in \mathcal{D} | X \subseteq r_i\}|$, where $|\ .\ |$ is the cardinality operator. Given a support threshold σ, an itemset X is called *σ-frequent* in a database $\mathcal{D}$ if $supp_{\mathcal{D}}(X) \geq \sigma$. A σ-frequent itemset is also called *σ-frequent pattern*. The collection of all σ-frequent patterns in $\mathcal{D}$ is denoted by $\mathcal{F}(\mathcal{D}, \sigma)$. The frequent pattern mining

problem is formulated as follows: given a database $\mathcal{D}$ and a support threshold σ, find all σ-frequent patterns, i.e., the collection $\mathcal{F}(\mathcal{D}, \sigma)$. Several algorithms have been proposed for finding $\mathcal{F}(\mathcal{D}, \sigma)$. In this paper we use the Apriori algorithm [2], which is a very common choice.

From patterns it is possible to derive classification rules. A *classification rule* is an expression $X \rightarrow C$, where C is a class item and X is an itemset containing no class item, e.g., *Sex=female, Credit_history=none-taken* $\rightarrow$ *Credit_approved=no*. The itemset X is called the premise of the rule. The *confidence* of a classification rule, $conf_{\mathcal{D}}(X \rightarrow C)$, measures how often the class item C appears in records that contain X. Hence, if $supp_{\mathcal{D}}(X) > 0$ then

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

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

3.2 Privacy-aware frequent patterns

In this section we describe the two approaches that we will use in the sequel to obtain patterns that guarantee privacy protection.

3.2.1 k-Anonymous frequent patterns

Atzori *et al.* in [4] introduce the notion of k-anonymous frequent patterns and a method to obtain a k-anonymous version of an original pattern set. The notion of k-anonymous patterns is defined as follows: a collection of patterns is k-anonymous if each pattern p in it is k-anonymous (i.e., $supp(p) = 0$ or $supp(p) \geq k$) as well as any further pattern whose support can be inferred from the collection.

Atzori *et al.* introduce a possible attack that exploits non-k-anonymous patterns whose support can be inferred from the collection and propose a framework for sanitizing patterns and block this kind of attacks.

Example 1 Consider again the motivating example and take $k = 8$. The two patterns p_1:{*Job=veterinarian, Credit_approved=yes*} and p_2: {*Job=veterinarian, Salary > 15000, Credit_approved=yes*} are 8-anonymous because $supp(p_2) = 40 > 8$ and $supp(p_1) = 41 > 8$. However, an attacker can exploit a non-8-anonymous pattern {*Job = veterinarian,* $\neg$ *(Salary > 15000), Credit_approved=yes*}, whose support he infers from $supp(p_1) - supp(p_2) = 41 - 40 = 1$. □

In order to check whether a collection of patterns is k-anonymous, in [4] the *inference channel* concept is introduced. Informally, an inference channel is any collection of patterns (with their respective supports) from which it is possible to infer non-k-anonymous patterns.

Definition 1 Given a database $\mathcal{D}$ and two patterns I and J, with $I = \{i_1, \ldots, i_m\}$ and $J = I \cup \{a_1, \ldots, a_n\}$, the set $C_I^J = \{\langle X, supp_{\mathcal{D}}(X)\rangle | I \subseteq X \subseteq J\}$ constitutes an inference channel for the non-k-anonymous pattern $p = I \cup \{\neg a_1, \ldots, \neg a_n\}$ if $0 < supp_{\mathcal{D}}(C_I^J) < k$ where

$$supp_{\mathcal{D}}(C_I^J) = \sum_{I \subseteq X \subseteq J} (-1)^{|X \setminus I|} supp_{\mathcal{D}}(X). \quad (2)$$

See [4] for details. An example of inference channels is given by any pattern such as $p : \{b\}$ which has a superset $p_s : \{b, d, e\}$ such that $0 < C_p^{p_s} < k$. In this case the pair $\langle p, supp(p)\rangle, \langle p_s, supp(p_s)\rangle$ constitutes an inference channel for the non-k-anonymous pattern $\{a, \neg b, \neg c\}$, whose support is given by $supp(b) - supp(b, d) - supp(b, e) + supp(b, d, e)$. Then, we can formally define the collection of k-anonymous pattern set as follows.

Definition 2 (k-Anonymous pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$ and an anonymity threshold k, $\mathcal{F}(\mathcal{D}, \sigma)$ is k-anonymous if (1) $\nexists p \in \mathcal{F}(\mathcal{D}, \sigma)$ s.t. $0 < supp(p) < k$, and (2) $\nexists p_1$ and $p_2 \in \mathcal{F}(\mathcal{D}, \sigma)$ s.t. $0 < supp_{\mathcal{D}}(C_{p_1}^{p_2}) < k$, where $p_1 \subset p_2$.

To generate a k-anonymous version of $\mathcal{F}(\mathcal{D}, \sigma)$, Atzori *et al.* [4] proposed to first detect inference channels violating k-anonymity in $\mathcal{F}(\mathcal{D}, \sigma)$ (Definition 1) and then block them in a second step. The pattern sanitization method blocks an inference channel C_I^J due to a pair of patterns $I = \{i_1, \ldots, i_m\}$ and $J = \{i_1, \ldots, i_m, a_1, \ldots, a_n\}$ in $\mathcal{F}(\mathcal{D}, \sigma)$ by increasing the support of I by k to achieve $supp(C_J^I) \geq k$. In addition, to avoid contradictions among the released patterns, the support of all subsets of I is also increased by k.

Example 2 Let us resume Example 1 and take $k = 8$. An inference channel due to patterns p_1 and p_2 can be blocked by increasing the support of pattern p_1:{*Job=veterinarian, Credit_approved=yes*} and all its subsets by 8. In this way, the non-8-anonymous pattern {*Job=veterinarian, ¬(Salary > 15000), Credit_approved=yes*} is 8-anonymous. □

The privacy pattern sanitization method can avoid generating new inference channels as a result of its transformation. In this way, we can obtain a k-anonymous version of $\mathcal{F}(\mathcal{D}, \sigma)$.

3.2.2 Differentially private frequent patterns

Li *et al.* in [38] recently proposed the *PrivBasis* algorithm for producing a differentially private version of the top K frequent patterns. Differential privacy provides a worst-case privacy guarantee. It guarantees that an adversary learns nothing about an individual, regardless of whether the individual's record is present or absent in the data. 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.

Definition 3 (Differential privacy, [14]) A randomized algorithm $\mathcal{A}$ is ϵ-differentially private if for all datasets $\mathcal{D}$, $\mathcal{D}'$ that differ in one individual (i.e., data of one person), and for all $S \subseteq Range(\mathcal{A})$, it holds that $Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^{\epsilon} Pr[\mathcal{A}(\mathcal{D}') \in S]$.

Differential privacy is composable according to the following lemma.

Lemma 1 (Sequential composition, [14]) *If there are M randomized algorithms $\mathcal{A}_1, \cdots, \mathcal{A}_M$, whose privacy guarantees are $\epsilon_1, \cdots, \epsilon_M$-differential privacy, respectively, then any function g of them, $g(\mathcal{A}_1, \cdots, \mathcal{A}_M)$, is $\sum_{i=1}^{M} \epsilon_i$-differentially private.*

We refer to ϵ as the privacy budget of a privacy-aware data analysis algorithm. When an algorithm involves multiple steps, each step uses a portion of ϵ so that the sum of these portions is no more than ϵ. There are several approaches for designing algorithms that satisfy ϵ-differential privacy. Here, we mention two of them. The first approach is the *Laplacian mechanism*. It computes a function F on the dataset $\mathcal{D}$ in a differentially privately way, by adding to $F(\mathcal{D})$ Laplace-distributed random noise. The magnitude of the noise depends on the *sensitivity* S_F of F:

$$S_F = \max_{(\mathcal{D},\mathcal{D}')} |F(\mathcal{D}) - F(\mathcal{D}')|$$

where $(\mathcal{D}, \mathcal{D}')$ is any pair of datasets that differ in one individual and belong to the domain of F. Formally, the Laplacian mechanism $\mathcal{A}_F$ can be written as

$$\mathcal{A}_F(\mathcal{D}) = F(\mathcal{D}) + Lap(\frac{S_F}{\epsilon})$$

where $Lap(\beta)$ denotes a random variable sampled from the Laplace distribution with scale parameter β. The second approach is the *exponential mechanism* proposed by McSherry and Talwar in [43], which can work on any kind of data. It computes a function F on a dataset $\mathcal{D}$ by sampling from the set of all possible outputs in the range of F according to an exponential distribution, with outputs that are "more accurate" being sampled with higher probability. This approach requires specifying a utility function $u : \mathcal{D} \times \mathcal{R} \rightarrow \mathbb{R}$, where the real valued score $u(\mathcal{D}, t)$ indicates how accurate it is to return t when the input dataset is $\mathcal{D}$. Higher scores means better utility outputs which should be returned with higher probabilities. For any function u, an algorithm $\mathcal{A}_F$ that chooses an output t with probability proportional to $exp(\frac{\epsilon u(\mathcal{D},t)}{2S_u})$ satisfies ϵ-differential privacy, where S_u is the sensitivity of utility function.

The notion of ϵ-differentially private frequent pattern set can be defined as follows.

Definition 4 (ϵ-differentially private frequent pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$ and a differential privacy budget ϵ, $\mathcal{F}(\mathcal{D}, \sigma)$ is ϵ-differentially private if it can be obtained using a randomized algorithm $\mathcal{A}$ satisfying ϵ-differential privacy.

Since *PrivBasis* returns the top K frequent patterns, we observe that if one desires to publish all patterns with support above a given threshold σ, i.e., $\mathcal{F}(\mathcal{D}, \sigma)$, one can first compute the value K such that the K-th most frequent pattern has a support value $\geq \sigma$ and the $K+1$-th pattern has support $< \sigma$, and then use *PrivBasis* with parameter K. The approach falls in the in-processing category and it uses a novel notion of basis sets. A σ-basis set $B = \{B_1, B_2, ..., B_w\}$, where each B_i is a set of items, has the property that any pattern with support value higher than σ is a subset of some basis B_i. The algorithm constructs a basis set while satisfying differential privacy and uses this set to find the most frequent patterns. The *PrivBasis* algorithm consists of the following steps:

1. Obtain λ, the number of unique items that are involved in the K most frequent patterns.
2. Obtain F, the λ most frequent items among the set I of all items in $\mathcal{D}$.
3. Obtain P, the set of the most frequent pairs of items among F.
4. Construct σ-basis set $B = \{B_1, B_2, ..., B_w\}$, using F and P.
5. Obtain noisy support values of patterns in candidate set $C(B) = \bigcup_{i=1}^{w} \{p | p \subseteq B_i\}$; one can then select the top K patterns from $C(B)$.

The privacy budget ϵ is divided among Steps 1, 2, 3, 5 above. Step 4 does not access the dataset $\mathcal{D}$, and only uses the outputs of earlier steps. Step 1 uses the exponential mechanism to sample j from $\{1, 2, \cdots, K\}$ with the utility function $u(\mathcal{D}, j) = (1 - |f_K - fitem_j|)|\mathcal{D}|$ where $f_K = \sigma$ is the support value of K-th most frequent pattern and $fitem_j$ is the support value of the j-th most frequent item. That is, Step 1 chooses j such that the j-th most frequent item has frequency closest to that of the K-th most frequent pattern. The sensitivity of the above utility function is 1, because adding or removing a record can affect f_K by at most $1/|\mathcal{D}|$ and $fitem_j$ by at most $1/|\mathcal{D}|$.

Step 2 differentially privately selects the λ most frequent items among the $|I|$ items and Step 3 differentially privately selects a set of the most frequent pairs of items among λ^2 patterns. Both steps use repeated sampling without replacement, where each sampling step uses the exponential mechanism with the support value of each pattern as its utility. Step 4 constructs B that covers all maximal cliques in (F, P) (see [38] for details). Step 5 computes the noisy support values of all patterns in $C(B)$ as follows. Each basis B_i algorithm divides all possible records into $2^{|B_i|}$ mutually disjoint bins, one corresponding to each subset of B_i. For each pattern $p \subseteq B_i$, the bin corresponding to p consists of all records that contain all items in p, but no item in $B_i \backslash p$. Given a basis set B, adding noise $Lap(w/\epsilon)$ to each bin count and outputting these noisy counts satisfies ϵ-differential privacy. The array element $b[i][p]$ stores the noisy count of the bin corresponding to pattern p and basis B_i. For each basis B_i, adding or removing a single record can affect the count of exactly one bin by exactly 1. Hence the sensitivity of publishing all bin counts for one basis is 1; and the sensitivity of publishing counts for all bases is w. From these noisy bin counts, one can recover the noisy support values of all patterns in $C(B)$ by summing up the respective noisy bin counts in $b[i][p]$.

We note that due to the possibility of drawing a negative noise from the Laplace distribution, *PrivBasis* can obtain noisy bin counts in $b[i][p]$ which are negative. This can lead to two problems: 1) some patterns could have negative support values; and 2) we can obtain a collection of frequent patterns with contradictions among them. More specifically, a pattern could have a noisy support value which is smaller than the noisy support values of its superset patterns. In order to avoid the contradictions among the released patterns published by *PrivBasis*, it is enough to post-process the noisy bin counts in $b[i][p]$ by rounding each count to the nearest non-negative integer. This should be done after Line 11 of Algorithm 1 in [38]. Note that, as proven by Hay *et al.* [29], a post-processing of differentially private results does non change the privacy guarantee. Hence, the algorithm *PrivBasis* will remain differentially private by the above changes. In the following we use *PrivBasis* with the above update. The problem of achieving consistency constraints among the noisy count values has been also addressed in [29].

3.3 Discrimination protected frequent patterns

Given DI_b and starting from a dataset $\mathcal{D}$ of historical decision records, the authors of [45] propose to extract frequent classification rules of the form $A, B \rightarrow C$, called PD rules, to unveil contexts B of possible discrimination, where the non-empty protected group $A \subseteq DI_b$ suffers from over-representation regarding the *negative* decision C (C is a class item reporting a negative decision, such as credit denial, application rejection, job firing, and so on). In other words, A is under-represented regarding the corresponding positive decision $\neg C$. As an example, rule *Sex=female, Job=veterinarian* $\rightarrow$ *Credit_approved=no* is a PD rule about denying credit (the decision C) to women (the protected group A) among those who are veterinarians (the context B), where DI_b:{*Sex=female*}. A classification rule of the form $X \rightarrow C$ is called PND rule if X is a PND itemset. As an example, rule *Credit_history=paid-delay, Job=veterinarian* $\rightarrow$ *Credit_approved=no* is a PND rule, where DI_b:{*Sex=female*}. Starting from the above definition of PD and PND classification rule, we define when a frequent pattern is PD.

Definition 5 Given protected groups DI_b, a frequent pattern $p \in \mathcal{F}(\mathcal{D}, \sigma)$ is said to be a PD if: (1) p contains a class item C denying some benefit, i.e., $C \subset p$, and (2) $\exists p' \subset p$ s.t. $p' \subseteq DI_b$.

In other words, a frequent pattern $p : \{A, B, C\}$ is a PD pattern if a PD classification rule $A, B \rightarrow C$ can be derived from it. As an example, pattern {*Sex=female, Job=veterinarian, Credit_approved=no*} is a PD pattern, where DI_b :{*Sex=female*}. Then, the degree of under-representation should be measured over each PD rule by one of the *legally-grounded* measures introduced in Pedreschi *et al.* [46]. We recall below the mathematical definition of these measures (because it is the one we will be using); see [46] and references therein for more information on the legal definition of these measures and on how the mathematical definition is obtained from the legal one.

Definition 6 Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$ with $conf(\neg A, B \rightarrow C) > 0$. The selection lift[2] (slift) of the rule is

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

In fact, $slift$ is the ratio of the proportions of benefit denial, e.g., credit denial, between the protected and unprotected groups, e.g., women and men resp., in the given context, e.g., applicants with no credit taken so far. 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., black ethnicity), compares it with the most favored item $a = v_2$ (e.g., white ethnicity).

Definition 7 Let $a = v_1, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$, and $v_2 \in dom(a)$ with $conf(a = v_2, B \rightarrow C)$ minimal and non-zero. The *contrasted lift (clift)* of the rule is

$$clift(a = v_1, B \rightarrow C) = \frac{conf(a = v_1, B \rightarrow C)}{conf(a = v_2, B \rightarrow C)} \tag{4}$$

Definition 8 Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$ with $conf(B \rightarrow C) > 0$. The *extended lift* [3] *(elift)* of the rule is

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} \tag{5}$$

In fact, *elift* is the ratio of the proportions of benefit denial, e.g., credit denial, between the protected groups and all people who were not granted the benefit in the given context, e.g. women versus all men and women who were denied credit, in the given context, e.g., those who live in NYC. Although the measures introduced so far are defined in terms of ratios, measures based on the difference of confidences have been considered on the legal side as well.

Definition 9 Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$. The *difference measures* are defined as

$$slift_d(A, B \rightarrow C) = conf(A, B \rightarrow C) - conf(\neg A, B \rightarrow C) \tag{6}$$

$$elift_d(A, B \rightarrow C) = conf(A, B \rightarrow C) - conf(B \rightarrow C) \tag{7}$$

Difference-based measures range over $[-1, 1]$. Lastly, the following measures are also defined in terms of ratios and known as chance measures.

[2] Discrimination on the basis of an attribute value happens if a person with that attribute value is treated less favorably than a person with another value.

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

Definition 10 Let $A, B \rightarrow C$ be a PD classification rule extracted from $\mathcal{D}$. The *chance measures* are defined as

$$slift_c(A, B \rightarrow C) = \frac{1 - conf(A, B \rightarrow C)}{1 - conf(\neg A, B \rightarrow C)} \tag{8}$$

$$elift_c(A, B \rightarrow C) = \frac{1 - conf(A, B \rightarrow C)}{1 - conf(B \rightarrow C)} \tag{9}$$

For $slift$ and $elift$, the values of interest (potentially indicating discrimination) are those greater than 1; for $slift_d$ and $elift_d$, they are those greater than 0; and for $slift_c$ and $elift_c$, they are those less than 1. On the legal side, different measures are adopted worldwide. More details about parallels between different measures and anti-discrimination acts are presented in [48]. For example, UK law mentions mostly $slift_d$. The EU court of justice refers to $slift$ while US laws courts mainly refer to $slift_c$. The discrimination measures mentioned so far in this paragraph are formally defined in Fig. 2.

Classification rule: $r = A, B \rightarrow C$

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

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

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

$$elift(r) = \frac{p_1}{p}, \quad elift_d(r) = p_1 - p, \quad elift_c(r) = \frac{1 - p_1}{1 - p}$$

$$slift(r) = \frac{p_1}{p_2}, \quad slift_d(r) = p_1 - p_2, \quad slift_c(r) = \frac{1 - p_1}{1 - p_2}$$

Fig. 2 Discrimination measures

Whether the rule has to be considered discriminatory or not can be assessed by thresholding one of the above measures as follows.

Definition 11 Let f be one of the measures in Definitions 8 or 9. Given protected groups DI_b and $\alpha \in R$, a fixed threshold[4], a PD classification rule $r : A, B \rightarrow C$, where C denies some benefit and $A \subseteq DI_b$, is α-protective w.r.t. f if $f(r) < \alpha$; otherwise, c is α-discriminatory. If f is one of the measures in Definition 10, a rule r is α-protective w.r.t. f if $f(r) > \alpha$; otherwise, c is α-discriminatory.

Example 3 Let $f = slift$, $\alpha = 1.25$ and DI_b :{*Sex=female*}. Assume that, in the data set of Table 1, the total number of veterinarian women applicants

[4] α states an acceptable level of discrimination according to laws and regulations. For example, the U.S. Equal Pay Act [53] 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 the number of veterinarian women applicants who are denied credit are 34 and 20, respectively, and the total number of veterinarian men applicants and the number of veterinarian men applicants who are denied credit are 47 and 19, respectively. The PD classification rule r :*Sex=female, Job=veterinarian* $\rightarrow$ *Credit_approved=no* extracted from Table 1 is 1.25-discriminatory, because $slift(r) = \frac{20/34}{19/47} = 1.45$. □

Based on Definitions 5 and 11, we introduce the notions of α-protective and α-discriminatory patterns.

Definition 12 Let f be one of the measures in Definitions 8-9. Given protected groups DI_b and $\alpha \in R$ a fixed threshold, a PD pattern $p : \{A, B, C\}$, where C denies some benefit and $A \subseteq DI_b$, is α-protective w.r.t. f if the classification rule $r : A, B \rightarrow C$ is α-protective. Otherwise, p is α-discriminatory.

Example 4 Continuing Example 3, a PD pattern p :{*Sex = female, Job = veterinarian, Credit_approved = no*} extracted from Table 1, is 1.25-discriminatory because rule r is 1.25-discriminatory, where r is *Sex = female, Job = veterinarian* $\rightarrow$ *Credit_approved = no*. □

Based on Definition 12, we introduce the notion of discrimination protected pattern set.

Definition 13 (α-protective pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$, discrimination measure f, a discrimination threshold α, and protected groups DI_b, $\mathcal{F}(\mathcal{D}, \sigma)$ is α-protective w.r.t. DI_b and f if $\nexists p \in \mathcal{F}(\mathcal{D}, \sigma)$ s.t. p is an α-discriminatory pattern.

4 Problem statement

The problem we want to study in this paper is how to produce a set of frequent patterns that simultaneously guarantees privacy and anti-discrimination protection while keeping the pattern distortion under control.

To address this problem, we propose a two-step approach. In the first step, the original set of patterns is transformed for privacy protection. In the second step, the private set is sanitized to avoid discriminatory patterns, and this is achieved without violating the privacy guarantee reached in the first step.

Concerning the privacy-preserving transformation, we study two options: in the first option, we rely on the k-anonymity model as discussed in Section 3.2.1; in the second option, we consider the approach proposed in [38] based on the differential privacy model. However, our experiments show that in this context differential privacy tends to cause much more pattern distortion than k-anonymity.

Concerning the anti-discrimination transformation, in the sequel we propose an approach to produce α-protective patterns (Section 5). Moreover, we analyze the interplay between the privacy and the anti-discrimination transformations to identify the impact that these sanitizations have on each other.

Our findings will highlight that the composition of the two steps is not trivial: applying first anti-discrimination sanitization and then privacy sanitization does not work, because privacy sanitization can yield patterns that are not α-protective (Section 6).

5 Discrimination-aware frequent patterns

In this section, we first present our proposed methods and algorithms to obtain discrimination protected patterns. Then, we introduce the notion of unexplainable discrimination protection and finally, we provide an approach to achieve this protection.

5.1 Achieving a α-protective pattern set

In order to generate a discrimination-protected (i.e., an α-protective) version of $\mathcal{F}(\mathcal{D}, \sigma)$, we propose an approach including two steps. First, detecting α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ regarding the discriminatory measure f, DI_b and α as discussed in Section 3.3. We propose Algorithm 1 for detecting α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$. The algorithm starts by obtaining the subset $\mathcal{D}_{PD}$ which contains the PD patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ found according to C and DI_b (Line 4). For each pattern $p : \{A, B, C\}$ in $\mathcal{D}_{PD}$, where $A \subseteq DI_b$, the value of f (one of the measures in Definitions 8-9) regarding its PD rule $r : X \rightarrow C$, where $X = A, B$, is computed to determine the subset $\mathcal{D}_D$ which contains the α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ (Lines 5-13). After obtaining $\mathcal{D}_D$, the second step of our approach is sanitization for each pattern in $\mathcal{D}_D$, in order to make it α-protective. In the sequel, we study and propose a pattern

Algorithm 1 DETECTING α-DISCRIMINATORY PATTERNS

1: Inputs: Database $\mathcal{D}$, $\mathcal{FP} := \mathcal{F}(D, \sigma)$, DI_b, discrimination measure f, α, C =class item with a negative decision value
2: Output: $\mathcal{D}_D$: α-discriminatory patterns in $\mathcal{FP}$
3: **Function** DETDISCPATT($\mathcal{FP}$, $\mathcal{D}$, DI_b, f, α, C)
4: $\mathcal{D}_{PD} \leftarrow$ All patterns $\langle p : A, B, C, supp(p)\rangle \in \mathcal{FP}$ with $p \cap C \neq \emptyset$ and $p \cap DI_b \neq \emptyset$
5: **for all** $p \in \mathcal{D}_{PD}$ **do**
6: $X = p \backslash C$
7: $r = X \rightarrow C$
8: Compute $f(r)$ using $\mathcal{FP}$ and $\mathcal{D}$ where f is one of the measures from Definitions 8-9
9: **if** $f(r) \geq \alpha$ **then**
10: Add p in $\mathcal{D}_D$
11: **end if**
12: **end for**
13: **return** $\mathcal{D}_D$
14: **End Function**

sanitization solution for each possible measure of discrimination f.

5.1.1 Anti-discrimination pattern sanitization for slift and its variants

According to Definition 12, to make an α-discriminatory pattern $p : \{A, B, C\}$ α-protective where $f = slift$, we should enforce the following inequality

$$slift(A, B \rightarrow C) < \alpha \tag{10}$$

where $A \subseteq DI_b$ and C denies some benefit. By using the definitions of confidence and $slift$ (Expressions (1) and (3), resp.), Inequality (10) can be rewritten as

$$\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(\neg A,B,C)}{supp(\neg A,B)}} < \alpha. \tag{11}$$

Then, it is clear that Inequality (10) can be satisfied by decreasing the left-hand side of Inequality (11) to a value less than the discriminatory threshold α, which can be done in the following way:

– *Anti-discrimination pattern sanitization where $f = slift$.* Increase the support of the pattern $\{A, B\}$ and all its subsets by a specific value Δ_{slift} to satisfy Inequality (11). This increment decreases the numerator of equation $\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(\neg A,B,C)}{supp(\neg A,B)}}$ while keeping the denominator unaltered.

Modifying the support of the subsets of respective patterns accordingly is needed to avoid contradictions (maintain compatibility) among the released patterns. In fact, *anti-discrimination pattern sanitization* makes pattern p α-protective by decreasing the proportion of people in the protected group and given context who were not granted the benefit (e.g., decreasing the proportion of veterinarian women applicants who were denied credit). Let us compute the value Δ_{slift} to be used in anti-discrimination pattern sanitization where $f = slift$. The support of the pattern $\{A, B\}$ should be increased to satisfy Inequality (11):

$$slift(A, B \rightarrow C) = \frac{\frac{supp(A,B,C)}{supp(A,B)+\Delta_{slift}}}{\frac{supp(\neg A,B,C)}{supp(\neg A,B)}} < \alpha.$$

The above equality can be rewritten as

$$\Delta_{slift} > \frac{supp(A, B, C) \times supp(\neg A, B)}{supp(\neg A, B, C) \times \alpha} - supp(A, B). \tag{12}$$

Hence, taking Δ_{slift} equal to the ceiling of the right-hand side of Equation (12) suffices to make $p : \{A, B, C\}$ α-protective w.r.t. $f = slift$. Considering the definitions of $slift_d$ and $slift_c$ (Expressions (6) and (8), resp.), a similar method can make pattern p α-protective w.r.t. $f = slift_d$ and $f = slift_c$. The value of Δ_{slift_d} and Δ_{slift_c} can be computed in the same way as we compute Δ_{slift}.

Example 5 Continuing Examples 3 and 4, pattern $p : \{Sex = female,\ Job = veterinarian,\ Credit\_approved = no\}$ can be made 1.25-protective by increasing the support of pattern $\{Sex=female,\ Job=veterinarian\}$ and all its subsets by $\Delta_{slift} = 6$, which is the value resulting from Inequality (12). □

As we define in Section 3.3, $clift$ is a special case of $slift$ and it has the same formula (see Definitions 6 and 7). Then, a similar anti-discrimination pattern sanitization can make an α-discriminatory $p : \{a = v_1, B, C\}$ α-protective where $f = clift$. The value of Δ_{clift} is computed in the following way

$$\Delta_{clift} = \left\lceil \frac{supp(a = v_1, B, C) \times supp(a = v_2, B)}{supp(a = v_2, B, C) \times \alpha} - supp(a = v_1, B) \right\rceil. \quad (13)$$

5.1.2 Anti-discrimination pattern sanitization for elift and its variants

According to Definition 12, to make an α-discriminatory pattern $p : \{A, B, C\}$ α-protective where $f = elift$, we should enforce the following inequality

$$elift(A, B \rightarrow C) < \alpha \quad (14)$$

where $A \subseteq DI_b$ and C denies some benefit. By using the definitions of confidence and $elift$ (Expressions (1) and (5), resp.), Inequality (14) can be rewritten as

$$\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}} < \alpha. \quad (15)$$

Then, it is clear that Inequality (14) can be satisfied by decreasing the left-hand side of Inequality (15) to a value less than the discriminatory threshold α. A similar anti-discrimination pattern sanitization proposed for $f = slift$ cannot make pattern p α-protective w.r.t. $f = elift$ because increasing the support of pattern $\{A, B\}$ and all its subsets by a specific value can decrease the numerator of equation $\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}}$ and decrease the denominator of it as well. Then, making pattern $p : \{A, B, C\}$ α-protective w.r.t. $f = elift$ is possible using an alternative pattern sanitization method, namely anti-discrimination pattern sanitization where $f = elift$. To apply this method, increase the support of the pattern $\{B, C\}$ and all its subsets by a specific value Δ_{elift} to satisfy Inequality (15). This increment increases the denominator of equation $\frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)}{supp(B)}}$ while keeping the numerator unaltered.

In fact, the above method makes pattern p α-protective w.r.t. $elift$ by increasing the proportion of people in the given context who were not granted the benefit (e.g., increasing the proportion of veterinarian applicants who were denied credit while the proportion of veterinarian women applicants who were denied credit is unaltered). Let us compute the value Δ_{elift} to be used in

anti-discrimination pattern sanitization where $f = elift$. The support of the pattern $\{B, C\}$ should be increased to satisfy Inequality (15):

$$elift(A, B \rightarrow C) = \frac{\frac{supp(A,B,C)}{supp(A,B)}}{\frac{supp(B,C)+\Delta_{elift}}{supp(B)+\Delta_{elift}}} < \alpha.$$

Since the value of α is higher than 1 and $\frac{supp(A,B,C)}{supp(A,B)} \leq \alpha$, from the above equality we obtain

$$\Delta_{elift} > \frac{\alpha \times supp(A,B) \times supp(B,C) - supp(A,B,C) \times supp(B)}{supp(A,B,C) - \alpha \times supp(A,B)}. \quad (16)$$

Hence, taking Δ_{elift} equal to the ceiling of the right-hand side of Equation (16) suffices to make $p : \{A, B, C\}$ α-protective w.r.t. $f = elift$. Considering the definitions of $elift_d$ and $elift_c$ (Expressions (6) and (8), resp.), a similar method can make pattern p α-protective w.r.t. $f = elift_d$ and $f = elift_c$. The values of Δ_{elift_d} and Δ_{elift_c} can be computed in the same way as Δ_{elift}.

5.1.3 Discrimination analysis

An essential property of a valid anti-discrimination pattern sanitization method is not to produce new discrimination as a result of the transformations it performs. The following theorem shows that all the methods described above satisfy this property.

Theorem 1 *Anti-discrimination pattern sanitization methods for making $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t. f do not generate new discrimination as a result of their transformations, where f is one of the measures from Definition 6-9.*

Proof It is enough to show that anti-discrimination pattern sanitization methods to make each α-discriminatory pattern in $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t. f cannot make α-protective patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ α-discriminatory. Consider two PD patterns $p_1 : \{A, B, C\}$ and $p_2 : \{A', B', C\}$, where $A, A' \subseteq DI_b$ and $p_1 \neq p_2$. The following possible relations between p_1 and p_2 are conceivable:

- $A = A'$ and $B \neq B'$, special case: $B' \subset B$
- $A \neq A'$ and $B = B'$, special case: $A' \subset A$
- $A \neq A'$ and $B \neq B'$, special case: $A' \subset A$ and $B' \subset B$

In all the above special cases (i.e., $p_2 \subset p_1$), making p_1 α-protective w.r.t. f involves increasing $supp(A', B')$ by Δ_{slift}, Δ_{clift} or Δ_{slift_d} where $f = slift$, $f = clift$ or $f = slift_d$, resp., and involves increasing $supp(B', C)$ and $supp(B')$ by Δ_{elift}, Δ_{elift_d} where $f = elift$ or $f = elift_d$, respectively. This cannot make p_2 less α-protective w.r.t. f; actually, it can even make p_2 more protective because increasing $supp(A', B')$ can decrease $slift(A', B' \rightarrow C)$ and $slift_d(A', B' \rightarrow C)$ and increasing $supp(B', C)$ and $supp(B')$ can decrease $elift(A', B' \rightarrow C)$ and $elift_d(A', B' \rightarrow C)$. On the other hand, making p_2 α-protective w.r.t. f cannot make p_1 less or more protective since there is no

overlap between the modified patterns to make p_2 α-protective and the patterns whose changing support can change $f(A, B \rightarrow C)$. Otherwise (no special cases), making p_1 (resp. p_2) α-protective w.r.t. f cannot make p_2 (resp. p_1) less or more protective since there is no overlap between the modified patterns to make p_1 (resp. p_2) α-protective w.r.t. f and the patterns whose changing support can change $f(A', B' \rightarrow C)$ (resp. $f(A, B \rightarrow C)$). Hence, the theorem holds. □

Therefore, using the proposed anti-discrimination pattern sanitization methods, we can obtain an α-protective version of $\mathcal{F}(\mathcal{D}, \sigma)$ w.r.t. f. We propose Algorithm 2 for doing so. The algorithm performs anti-discrimination pattern sanitization to make each α-discriminatory pattern p in $\mathcal{F}(D, \sigma)$ α-protective w.r.t. f. The value of Δ_f is computed for each α-discriminatory pattern regarding the value of f.

Algorithm 2 Anti-discrimination pattern sanitization

Inputs: Database $\mathcal{D}$, $\mathcal{FP} := \mathcal{F}(D, \sigma)$, $\mathcal{D}_D$, DI_b, discrimination measure f, α, C = class item with negative decision value
Output: $\mathcal{FP}^*$: α-protective version of $\mathcal{FP}$
Function AntiDiscPattSanit($\mathcal{FP}$, $\mathcal{D}$, $\mathcal{D}_D$, DI_b, f, α, C)
for all $p : \{A, B, C\} \in \mathcal{D}_D$ **do**
  $X = p \backslash C$
  Compute Δ_f for pattern p regarding the value of f using $\mathcal{D}$, $\mathcal{FP}$ and α
  if $\Delta_f \geq 1$ **then**
    if $f = slift$ or $f = clift$ or $f = slift_d$ **then**
      $p_t = X$
    else if $f = elift$ or $elift_d$ **then**
      $Y = p \cap DI_b$
      $p_t = p \backslash Y$
    end if
  end if
  $\mathcal{D}_s = \{p_s \in \mathcal{FP} | p_s \subseteq p_t\}$
  for all $\langle p_s, supp(p_s) \rangle \in \mathcal{D}_s$ **do**
    $supp(p_s) = supp(p_s) + \Delta_f$
  end for
end for
return $\mathcal{FP}^* = \mathcal{FP}$
End Function

5.2 Extension of the discrimination protection to closed frequent patterns

Since the authors of [4] provide a solution to detect inference channels starting from closed patterns rather than frequent patterns, we also discuss here how the above anti-discrimination sanitization methodology can be adapted for closed frequent patterns. The latter patterns were introduced in [44] as a concise and lossless representation of all frequent patterns; in fact, they contain the same information without redundancy. Intuitively, a closed pattern groups together all its subsets that have its same support.

Definition 14 A closed pattern is defined as a pattern whose supersets have a strictly smaller support. Given a database $\mathcal{D}$ and a minimum support threshold σ, the set of frequent closed patterns is denoted: $\mathcal{C}l(\mathcal{D},\sigma) = \{\langle X, supp_{\mathcal{D}}(X)\rangle \in \mathcal{FP} | \nexists Y \supset X \ s.t. \ \langle Y, supp_{\mathcal{D}}(X)\rangle \in \mathcal{FP}\}$.

It is easy to understand that if we have closed patterns as input we need to apply a slightly different procedure to detect discrimination (Algorithm 1). The change concerns Line 8, where one of the possible discrimination measures f is used to decide whether a pattern r is α-discriminatory or not. Given a pattern $\{A, B, C\}$ and $f = slift$, we need to satisfy Inequality (11), which requires knowledge of the support of $\{A, B, C\}$, $\{A, B\}$, $\{\neg A, B, C\}$ and $\{\neg AB\}$.

We can have two cases: (a) all above itemsets are closed; (b) some of them are not closed. In case (a) everything remains as in the case of frequent patterns because the information to compute $slift$ can be directly found in the list of itemsets $\mathcal{C}l$. In case (b) one of the itemsets is not a closed pattern in $\mathcal{C}l$. It turns out, though, that we can derive its support from the closed itemsets. Suppose for example that $\{A, B\}$ is not closed, then we simply recover all its possible supersets $\{X_1, \ldots, X_n\}$ and their support values and compute $supp(A, B)$ as the highest support of its supersets, i.e., $supp(A, B) = \max_i(supp(X_i))$.

In this way we are able to compute $slift$ of a pattern p and apply the sanitization described in Algorithm 2. However, we highlight that the sanitization of closed patterns can generate ghost patterns, i.e., it can transform non-closed patterns in closed ones, while maintains all the original closed patterns. Indeed, suppose $\{A, B\}$ is not closed, then we have that $\exists Y \supset \{A, B\}$ such that $supp(A, B) = supp(Y)$. Increasing the support value of $\{A, B\}$ by Δ_{slift} we eliminate the equality and the pattern $\{A, B\}$ becomes closed.

The case of $f = elift$ is very similar. Given a pattern $\{A, B, C\}$ and $f = elift$, we need to satisfy Inequality (15), which requires knowledge of the support of $\{A, B, C\}$, $\{A, B\}$, $\{B, C\}$ and $\{B\}$. Again, if those itemsets are closed the methodology of discrimination detection is the same. If some itemset is not closed, then we can use the methodology described above to derive its support. In particular, if $\{B, C\}$ or $\{B\}$ are not closed, ghost closed itemsets may appear when increasing the value of the supports of $\{B, C\}$ and $\{B\}$ by Δ_{elift}, as required to satisfy Inequality (15).

While in the case of frequent patterns, our pattern sanitization methods for anti-discrimination never generate ghost frequent patterns, we cannot guarantee that in the case of closed patterns no ghost closed patterns are generated.

5.3 Achieving a d-explainable α-protective pattern set

Another legal concept we take into account in this paper is the *genuine requirement*. This requirement refers to discrimination that may be partly explained by attributes *not* in DI_b ([56,34]), e.g., denying credit to women may be explainable by the fact that most of them have low salary or delay in returning previous credits. Whether low salary or delay in returning previous credits is a legally admissible argument to deny credit is for the law to determine.

In our context, a legally-grounded group DI_e is a PND itemset which is legally admissible in a discrimination litigation, e.g., {*Credit_history=paid-delay*}. Given DI_e and DI_b, discrimination against protected groups is explained if there is a high correlation between DI_b and DI_e and also between DI_e and class item C.

As an example, discrimination against women in a given context is explainable by their delay in returning previous credits if two conditions hold: i) the majority of women in the given context have delay in returning previous credits; ii) the delay in returning previous credits is an objective reason to deny credit. To determine which α-discriminatory patters are explainable and which ones are not, we use the notion of d-instance [47].

We say that a PD classification rule $r' : A, B \rightarrow C$ is an instance of PND rule $r : D, B \rightarrow C$, if rule r holds with the same or higher confidence, namely $conf(r) \geq conf(r')$, and a case (record) satisfying PD itemset A in context B satisfies PND itemset D as well, namely $conf(A, B \rightarrow D) = 1$. The two conditions can be relaxed as follows.

Definition 15 Let $d \in [0, 1]$. A classification rule $r' : A, B \rightarrow C$ is a d-instance of PND rule $r : D, B \rightarrow C$ if both conditions below are true:

- Condition 1: $conf(r) \geq d \cdot conf(r')$.
- Condition 2: $conf(r'' : A, B \rightarrow D) \geq d$.

For high values of d (that is, 1 or near 1) Condition 1 shows high correlation between the class item (e.g., credit denial) and the legally-grounded group (e.g., delay in returning previous credits) and Condition 2 shows high correlation between the legally-grounded group (e.g., delay in returning previous credits) and the protected group (e.g., women) in a given context (In the example of most women being delayed in returning previous credits, the delay would justify denial of new credit and this could not be construed as a discrimination against women).

Example 6 Continuing Examples 3 and 4, let DI_e : { *Credit_history = paid-delay*} and $d = 0.9$. Assume that in the dataset of Table 1, the total number of veterinarian applicants who are delayed in returning previous credits and the total number of veterinarian applicants who are delayed in returning previous credits and are denied credit are 64 and 52, respectively, and the total number of women applicants who are veterinarian and are delayed in returning previous credits is 31. A PD classification rule r' :*Sex=female, Job=veterinarian* $\rightarrow$ *Credit_approved=no* extracted from Table 1 is a 0.9-instance of a PND classification rule r :*Credit_history=paid-delay, Job=veterinarian* $\rightarrow$ *Credit_approved=no* because (1) $\frac{52}{64} \geq 0.9 \cdot \frac{20}{34}$ and (2) $\frac{31}{34} \geq 0.9$. Thus, both conditions of Definition 15 are satisfied. □

Based on Definitions 15 and 5, we introduce the notions of d-explainable and d-unexplainable frequent patterns.

Definition 16 (d-explainable/unexplainable pattern) Let $d \in [0, 1]$. An α-discriminatory pattern $p' : \{A, B, C\}$, where C denies some benefit and

$A \subseteq DI_b$, is a d-explainable pattern if a PD classification rule $r' : A, B \rightarrow C$ is a d-instance of a PND classification rule $r : D, B \rightarrow C$, where $D \subseteq DI_e$. Otherwise p' is a d-unexplainable pattern.

Example 7 Continuing Examples 3-6, the 1.25-discriminatory pattern p : {*Sex = female, Job = veterinarian, Credit_approved = no*} extracted from Table 1 is a 0.9-explainable pattern because rule r' is a 0.9-instance of rule r where r is *Credit_history = paid-delay, Job = veterinarian* $\rightarrow$ *Credit_approved = no* and r' is *Sex = female, Job = veterinarian* $\rightarrow$ *Credit_approved = no.* □

From Definition 16, we introduce the notion of d-explainable α-protective pattern set.

Definition 17 (d-explainable α-protective pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$, a discrimination measure f, a discrimination threshold α, an explainable discrimination threshold d, protected groups DI_b and legally-grounded groups DI_e, $\mathcal{F}(\mathcal{D}, \sigma)$ is d-explainable α-protective w.r.t. DI_b, DI_e and f if there is no α-discriminatory pattern $p \in \mathcal{F}(\mathcal{D}, \sigma)$ s.t. p is a d-unexplainable pattern.

In order to make $\mathcal{F}(D, \sigma)$ protected against only unexplainable discrimination (i.e., generating a d-explainable α-protective version of $\mathcal{F}(\mathcal{D}, \sigma)$), we propose an approach including three steps. First, detecting α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ regarding the discriminatory measure f, DI_b and α. Second, detecting d-unexplainable patterns among α-discriminatory patterns obtained in the first step. Third, sanitizing each d-unexplainable pattern to make it α-protective. We propose Algorithm 3 for detecting d-unexplainable patterns in $\mathcal{F}(\mathcal{D}, \sigma)$. The algorithm starts by obtaining the subset $\mathcal{D}_{PND}$ containing the PND patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ found according to C and DI_b (Line 4). Then, the algorithm computes the subset $\mathcal{D}_{instance}$ containing the PND patterns which are legally-grounded according to DI_e (Line 5). Then, the algorithm uses the DetDiscPatt function in Algorithm 1 to determine the subset $\mathcal{D}_D$ which contains α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ (Line 6). Finally, the algorithm computes the subset $\mathcal{D}_{bad}$ containing patterns in $\mathcal{D}_D$ which are d-unexplainable according to $\mathcal{D}_{instance}$ and d (Lines 7-19). After obtaining $\mathcal{D}_{bad}$, the third step is sanitizing each d-unexplainable pattern to make it α-protective. In order to do this, we need to examine the impact of this transformation on d-explainable patterns in $\mathcal{F}(\mathcal{D}, \sigma)$.

Theorem 2 *Anti-discrimination pattern sanitization methods for making $\mathcal{F}(\mathcal{D}, \sigma)$ d-explainable α-protective w.r.t. f do not generate any new d-unexplainable pattern as a result of their transformations, where $f = elift$, $f = elift_d$, and $f = elift_c$.*

Proof It is enough to show that anti-discrimination pattern sanitization methods to make each d-unexplainable pattern in $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t. f cannot make d-explainable patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ d-unexplainable, where $f = elift$, $f = elift_d$ and $f = elift_c$. Consider two PD patterns $p_1 : \{A, B, C\}$ and

Algorithm 3 DETECTING d-UNEXPLAINABLE PATTERNS

Inputs: Database $\mathcal{D}$,$\mathcal{FP} := \mathcal{F}(D, \sigma)$, DI_b, DI_e, explainable discrimination threshold d, discrimination measure f, α, C =class item with negative decision value
Output: $\mathcal{D}_{bad}$: d-unexplainable patterns in $\mathcal{FP}$
Function DETUNEXPLAINPATT($\mathcal{FP}$, $\mathcal{D}$, DI_e, DI_b, f, α, d, C)
$\mathcal{D}_{PND} \leftarrow$ All patterns $\langle p : A, B, C, supp(p)\rangle \in \mathcal{FP}$ with $p \cap C \neq \emptyset$ and $p \cap DI_b = \emptyset$
$\mathcal{D}_{instance} \leftarrow$ All patterns $\langle p : D, B, C, supp(p)\rangle \in \mathcal{D}_{PND}$ with $p \cap DI_e \neq \emptyset$
$\mathcal{D}_D \leftarrow$ **Function** DETDISCPATT($\mathcal{FP}$, $\mathcal{D}$, DI_b, f, α, C)
for all $p : \{A, B, C\} \in \mathcal{D}_D$ **do**
  $X = p \backslash C$
  $r = X \rightarrow C$
  for each $p_d \in \mathcal{D}_{instance}$ **do**
    $X_d = p_d \backslash C$
    $r_d = X_d \rightarrow C$
    if r is a d-instance of r_d **then**
      Add p in $\mathcal{D}_{legal}$
    end if
  end for
end for
$\mathcal{D}_{bad} = \mathcal{D}_D \backslash \mathcal{D}_{legal}$
return $\mathcal{D}_{bad}$
End Function

$p_2 : \{A', B', C\}$, where $A, A' \subseteq DI_b$ and $p_1 \neq p_2$. The following possible relations between p_1 and p_2 are conceivable:

- $A = A'$ and $B \neq B'$, special case: $B' \subset B$
- $A \neq A'$ and $B = B'$, special case: $A' \subset A$
- $A \neq A'$ and $B \neq B'$, special case: $A' \subset A$ and $B' \subset B$

In all the above cases (i.e. special and non-special cases), making d-unexplainable pattern p_1 α-protective w.r.t. f involves increasing $supp(B', C)$ and $supp(B')$ by Δ_{elift}, Δ_{elift_d}, or Δ_{elift_c} where $f = elift$, $f = elift_d$ or $f = elift_c$, respectively. This cannot make d-explainable pattern p_2 d-unexplainable w.r.t. f because there is no overlap between the modified patterns to make p_1 α-protective and the patterns whose changing support can make p_2 d-unexplainable. On the other hand, making d-unexplainable pattern p_2 α-protective cannot make d-explainable pattern p_1 d-unexplainable for the same reason. Hence, the theorem holds. □

Theorem 3 *Anti-discrimination pattern sanitization methods for making $\mathcal{F}(\mathcal{D}, \sigma)$ d-explainable α-protective regarding f might generate new d-unexplainable patterns as a result of their transformations, where $f = slift$, $f = slift_d$, $f = clift$ and $f = slift_c$.*

Proof It is enough to show that anti-discrimination pattern sanitization methods to make each d-unexplainable pattern in $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t. f can make a d-explainable in $\mathcal{F}(\mathcal{D}, \sigma)$ d-unexplainable, where $f = slift$, $f = slift_d$, $f = slift_c$ and $f = elift$. Consider two PD patterns $p_1 : \{A, B, C\}$ and $p_2 : \{A', B', C\}$, where $A, A' \subseteq DI_b$ and $p_1 \neq p_2$. The following possible relations between p_1 and p_2 are conceivable:

- $A = A'$ and $B \neq B'$, special case: $B' \subset B$
- $A \neq A'$ and $B = B'$, special case: $A' \subset A$
- $A \neq A'$ and $B \neq B'$, special case: $A' \subset A$ and $B' \subset B$

In all the above special cases (i.e., $p_2 \subset p_1$), making d-unexplainable pattern p_1 α-protective w.r.t. f involves increasing the support of pattern $\{A', B'\}$ and all its subsets by Δ_{slift}, Δ_{clift}, Δ_{slift_d}, or Δ_{slift_c} where $f = slift$, $f = clift$ $f = slift_d$ or $f = clift_c$, respectively. This can make d-explainable pattern p_2 d-unexplainable because this transformation can cause Condition 2 of Definition 15 to be non-satisfied. This is because there is overlap between the modified patterns to make p_1 α-protective and the patterns whose changing support can change the satisfaction of Condition 2 of Definition 15 regarding pattern p_2. On the other hand, making d-unexplainable pattern p_2 α-protective cannot make d-explainable pattern p_1 d-unexplainable since there is no overlap between the modified patterns to make p_2 α-protective and the patterns whose changing support can change the satisfaction of Conditions 1 and 2 of Definition 15 regarding pattern p_1. Otherwise (no special cases), making p_1 (resp. p_2) α-protective w.r.t. f cannot make p_2 (resp. p_1) d-unexplainable since there is no overlap between the modified patterns to make p_1 (resp. p_2) α-protective w.r.t. f and the patterns whose changing support can make p_2 (resp. p_1) d-unexplainable by changing the satisfaction of Conditions (1) and (2) in Definition 15. Hence, the theorem holds. □

Thus, according to the above theorems, Algorithm 2 cannot make $\mathcal{F}(\mathcal{D}, \sigma)$ d-explainable α-protective regarding *all* possible values of f. To attain this desirable goal, we propose Algorithm 4. This algorithm performs anti-discrimination pattern sanitization to make each d-unexplainable pattern p in $\mathcal{D}_{bad}$ α-protective by calling function AntiDiscPattSanit in Algorithm 2, where $f = elift$, $f = elift_d$ or $f = elift_c$ (Line 6). As shown in Theorem 3, given d-unexplainable pattern $p : \{A, B, C\}$ and d-explainable pattern $p_x : \{A', B', C\}$, where $p_x \subset p$, making p α-protective w.r.t. f might make p_x d-unexplainable, where f is $slift$, $slift_d$, $f = clift$ or $f = slift_c$. For this reason and because pattern p becomes α-protective first (see Algorithm 5 in Section 6.1), the algorithm checks whether making pattern p α-protective makes p_x d-unexplainable. If yes, the algorithm adds p_x to $\mathcal{D}_{bad}$ (Lines 7-14).

6 Simultaneous discrimination-privacy awareness in frequent pattern discovery

In this section, we present how simultaneous anti-discrimination and privacy can be achieved in frequent pattern discovery. We first present our approach to obtain a k-anonymous and α-protective version of an original pattern set. Then, we present our approach to obtain an unexplainable discrimination and privacy protected version of an original pattern set. Lastly, we describe our proposal to transform a set of patterns into its differentially private and α-protective version.

Algorithm 4 UNEXPLAINABLE ANTI-DISCRIMINATION PATTERN SANITIZATION

1: Inputs: Database $\mathcal{D}$, $\mathcal{FP} := \mathcal{F}(D, \sigma)$, $\mathcal{D}_{bad}$, DI_b, d, α, discrimination measure f, C =class item with negative decision value
2: Output: $\mathcal{FP}^*$: d-explainable α-protective version of $\mathcal{FP}$
3: **Function** UNEXPLAINANTIDISCPATTSANIT($\mathcal{FP}$, $\mathcal{D}$, $\mathcal{D}_{bad}$, DI_b, f, α, C)
4: **if** $\mathcal{D}_{bad} \neq \emptyset$ **then**
5: **if** $f = elift$ or $f = elift_d$ or $f = elift_c$ **then**
6: $\mathcal{FP}^* \leftarrow$ **Function** ANTIDISCPATTSANIT($\mathcal{FP}$, $\mathcal{D}$, $\mathcal{D}_{bad}$, DI_b, f, α, C)
7: **else if** $f = slift$ or $f = slift_d$ or $f = clift$ or $f = slift_c$ **then**
8: **for all** $p : \{A, B, C\} \in \mathcal{D}_{bad}$ **do**
9: Lines 5-18 of Algorithm 2
10: **if** $\exists p_x \subset p$ in $\mathcal{FP}$ s.t. $p_x \notin \mathcal{D}_{bad}$ and p_x is d-unexplainable **then**
11: Add p_x in $\mathcal{D}_{bad}$
12: **end if**
13: **end for**
14: **end if**
15: **end if**
16: **return** $\mathcal{FP}^* = \mathcal{FP}$
17: **End Function**

6.1 Achieving an α-protective and k-anonymous pattern set

In order to simultaneously achieve anti-discrimination and privacy in $\mathcal{F}(\mathcal{D}, \sigma)$, we need to generate a discrimination and privacy protected version of $\mathcal{F}(\mathcal{D}, \sigma)$.

Definition 18 (α-protective k-anonymous pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$, anonymity threshold k, discrimination threshold α, protected groups DI_b, and discrimination measure f, $\mathcal{F}(\mathcal{D}, \sigma)$ is α-protective k-anonymous if it is both k-anonymous and α-protective w.r.t. DI_b and f.

We focus on the problem of producing a version of $\mathcal{F}(\mathcal{D}, \sigma)$ that is α-protective k-anonymous w.r.t. DI_b and f. Like most works in k-anonymity [21], we consider a single QI containing all attributes that can be potentially used in the quasi-identifier. The more attributes included in QI, the more protection k-anonymity provides (and usually the more information loss). Moreover, each QI attribute (unless it is the class/decision attribute) can be a PD attribute or not depend on DI_b. To obtain an α-protective k-anonymous version of $\mathcal{F}(\mathcal{D}, \sigma)$, we should first examine the following issues: (1) how making $\mathcal{F}(\mathcal{D}, \sigma)$ k-anonymous impacts on the α-protectiveness of $\mathcal{F}(\mathcal{D}, \sigma)$ w.r.t. f; (2) how making $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t f impacts on the k-anonymity of $\mathcal{F}(\mathcal{D}, \sigma)$. Regarding the first issue, by presenting two scenarios we will show that *privacy pattern sanitization* to achieve k-anonymity in $\mathcal{F}(\mathcal{D}, \sigma)$ can lead to different situations regarding the α-protectiveness of $\mathcal{F}(\mathcal{D}, \sigma)$.

Scenario 1 Table 2 illustrates an example of frequent patterns that come from the data set in Table 1 with $\sigma = 15$. Let DI_b :{*Sex=female*}, $\alpha = 1.25$, $k = 8$ and $f = slift$. The PD pattern p_1 in Table 2 is 1.25-discriminatory, since the value of $slift$ regarding its PD rule is $\frac{32/45}{16/58} = 2.58$ (i.e., this is

Table 2 Scenario 1: Examples of frequent patterns extracted from Table 1

Patterns	*Support*
p_s:{female, veterinarian}	45
p_2 :{female, veterinarian, salary > 15000}	42
p_1 :{female, veterinarian, No}	32
p_n :{male, veterinarian, No}	16
p_{ns} :{male, veterinarian}	58

inferred discrimination against veterinarian women applicants). On the other hand, although the support of each pattern in the collection is higher than k, there is an inference channel between patterns p_s and p_2; note that $supp(p_s) - supp(p_2) = 3$ is smaller than 8 (i.e., one can infer the existence of only three veterinarian women in the city with salary no more than 15000 €). To block the inference channel between p_s and p_2, the following privacy pattern sanitization is performed:

$$supp(p_s) + k, \forall p \subseteq p_s \tag{17}$$

After this transformation, the new support of pattern p_s is 53. However, the supports of p_{ns} and p_n remain unaltered since there are no inference channels between p_{ns} and p_n ($supp(p_{ns}) - supp(p_n) = 42$). Hence, the new value of $slift$ for the PD rule of pattern p_1 is $\frac{\frac{supp(p_1)}{supp(p_s)+k}}{\frac{supp(p_n)}{supp(p_{ns})}}$ which in this example is $\frac{32/(45+8)}{16/58} = 2.19$. That is, the overall value of $slift$ is decreased. Thus, in this scenario, making the collection of patterns in Table 2 k-anonymous can decrease discrimination; if the value of $slift$ became less than α, pattern p_1 would even become α-protective. □

Table 3 Scenario 2: Examples of frequent patterns extracted from Table 1

Patterns	*Support*
p_s :{male, veterinarian}	58
p_2 :{male, veterinarian, salary > 15000}	56
p_1 :{female, veterinarian, No}	23
p_n :{male, veterinarian, No}	26
p_{ns} :{female, veterinarian}	45

Scenario 2 Table 3 illustrates an example of frequent patterns that could come from the data set in Table 1 with $\sigma = 20$. Let DI_s:{*Sex=female*}, $\alpha = 1.25$, $k = 8$, $f = slift$. A PD pattern p_1 in Table 3 is not 1.25-discriminatory since the value of $slift$ regarding its PD rule is $\frac{23/45}{26/58} = 1.14$. On the other hand, although the support of each pattern in the collection is higher than k, there is an inference channel between p_s and p_2; note that $supp(p_s) - supp(p_2) = 2$ is less than 8 (i.e., one can infer the existence of only two veterinarian men in the city with salary no more than 15000 €). To block the inference channel between p_s and p_2, pattern sanitization is performed

according to Expression (17). After this transformation, the new support of pattern p_s is 66 and the supports of p_1 and p_{ns} stay unaltered since there is no inference channel between p_1 and p_{ns} ($supp(p_{ns}) - supp(p_1) = 22$). Hence, the new value of $slift$ for the PD rule of pattern p_1 is $\frac{\frac{supp(p_1)}{supp(p_{ns})}}{\frac{supp(p_n)}{supp(p_s)+k}}$ which in this example is $\frac{23/45}{26/(58+8)} = 1.3$. That is, the overall value of $slift$ is increased. Thus, in this scenario, making the collection of patterns in Table 3 k-anonymous can increase discrimination; in fact, with the numerical values we have used, p_1 stops being 1.25-protective and becomes 1.25-discriminatory. □

To summarize, using privacy pattern sanitization for making $\mathcal{F}(\mathcal{D}, \sigma)$ k-anonymous can make $\mathcal{F}(\mathcal{D}, \sigma)$ more or less α-protective w.r.t. $slift$. We also observe a similar behavior for alternative discrimination measures. Then, achieving k-anonymity in frequent pattern discovery can achieve anti-discrimination or work against anti-discrimination. Hence, detecting α-discriminatory patterns in a k-anonymous version of $\mathcal{F}(\mathcal{D}, \sigma)$ makes sense. Regarding the second issue mentioned at the beginning of this section, we will prove that if $\mathcal{F}(\mathcal{D}, \sigma)$ is k-anonymous, anti-discrimination pattern sanitization methods proposed in Section 5.1 to make it α-protective w.r.t. f cannot make $\mathcal{F}(\mathcal{D}, \sigma)$ non-k-anonymous, i.e., they cannot violate k-anonymity. Since the proposed anti-discrimination pattern sanitization methods are based on adding support, they cannot make a k-anonymous pattern non-k-anonymous; hence, we only need to prove that they cannot generate new inference channels.

Theorem 4 *Anti-discrimination pattern sanitization for making $\mathcal{F}(\mathcal{D}, \sigma)$ α-protective w.r.t. f does not generate inference channels, where f is one of the measures from Definition 8-9.*

Proof For any α-discriminatory pattern $p_a : \{A, B, C\}$, where $A \subseteq DI_b$ and C denies some benefit, anti-discrimination pattern sanitization is performed in one of the following ways regarding the value of f:

$$supp(p_s : \{A, B\}) + \Delta_f, \forall p \subset p_s \tag{18}$$

where $f = slift$, $f = clift$, $f = slift_d$ or $f = slift_c$;

$$supp(p_s : \{B, C\}) + \Delta_f, \forall p \subset p_s \tag{19}$$

where $f = elift$, $f = elift_d$ or $f = elift_c$. Inference channels could appear in two different cases: (a) between the pattern p_s or one of its supersets (p_x s.t. $p_s \subset p_x$), and (b) between the pattern p_s and one of its subsets (p_x s.t. $p_x \subset p_s$). *Case (a).* Since $\mathcal{F}(\mathcal{D}, \sigma)$ is k-anonymous, we have $supp(C^{p_x}_{p_s}) \geq k$. Increasing the support of p_s and its subsets by Δ_f as in Expressions (18-19) causes the value of $supp(C^{p_x}_{p_s})$ to increase by Δ_f, because only the first term of the sum in Expression (2) used to compute $supp(C^{p_x}_{p_s})$ is increased (the support of p_s). Hence, the support of $C^{p_x}_{p_s}$ stays above k. *Case (b).* Since $\mathcal{F}(\mathcal{D}, \sigma)$ is k-anonymous, we have $supp(C^{p_s}_{p_x}) \geq k$. Increasing the support of p_s and its subsets by Δ_f as in Expressions (18-19) means adding the same value Δ_f to each term of the sum in Expression (2) used to compute $supp(C^{p_s}_{p_x})$. Hence, this support of $C^{p_s}_{p_x}$ does not change. Thus, the theorem holds. □

Since using our anti-discrimination pattern sanitization methods for making $\mathcal{F}(\mathcal{D},\sigma)$ α-protective w.r.t. f cannot make $\mathcal{F}(\mathcal{D},\sigma)$ non-k-anonymous, a safe way to obtain an α-protective k-anonymous $\mathcal{F}(\mathcal{D},\sigma)$ w.r.t. f is to apply anti-discrimination pattern sanitization methods to a k-anonymous version of $\mathcal{F}(\mathcal{D},\sigma)$, in order to turn α-discriminatory patterns detected in that k-anonymous version into α-protective patterns w.r.t f. We propose Algorithm 5 to obtain an α-protective k-anonymous version of an original frequent pattern set w.r.t. a discrimination measure f. There are two assumptions in this algorithm: first, the class attribute is binary; second, protected groups DI_b correspond to nominal attributes. Given an original pattern set $\mathcal{F}(\mathcal{D},\sigma)$, denoted by $\mathcal{FP}$ for short, a discriminatory threshold α, an anonymity threshold k, a discrimination measure f, protected groups DI_b and a class item C which denies some benefit, Algorithm 5 starts by obtaining $\mathcal{FP}'$, which is a k-anonymous version of FP (Line 3). It uses Algorithm 3 in [4] to do this. Then, the algorithm uses the DetDiscPatt function in Algorithm 1 to determine the subset $\mathcal{D}_D$ which contains α-discriminatory patterns in $\mathcal{FP}'$ (Line 4). As we

Algorithm 5 Anti-discrimination k-anonymous pattern sanitization

1: Inputs: Database $\mathcal{D}$, $\mathcal{FP} := \mathcal{F}(D,\sigma)$, k, DI_b, discrimination measure f, α, C =class item with negative decision value
2: Output: $\mathcal{FP}''$: α-protective k-anonymous frequent pattern set
3: $\mathcal{FP}' \leftarrow PrivacyAdditiveSanitization(\mathcal{FP},k)$ //Algorithm 3 in [4]
4: $\mathcal{D}_D \leftarrow$ **Function** DetDiscPatt($\mathcal{FP}'$, $\mathcal{D}$, DI_b, f, α, C)
5: **for all** $p \in \mathcal{D}_D$ **do**
6: Compute $impact(p) = |\{p' : A', B', C\} \in \mathcal{D}_D$ s.t. $p' \subset p|$
7: **end for**
8: Sort $\mathcal{D}_D$ by descending $impact$
9: $\mathcal{FP}'' \leftarrow$ **Function** AntiDiscPattSanit($\mathcal{FP}'$, $\mathcal{D}$, $\mathcal{D}_D$, DI_b, f, α, C)
10: Output: $\mathcal{FP}''$

showed in Theorem 1, given two α-discriminatory patterns $p_1 : \{A, B, C\}$ and $p_2 : \{A', B', C\}$, where $p_2 \subset p_1$, making p_1 α-protective w.r.t. f can make also p_2 less discriminatory or even α-protective, depending on the value of α and the support of patterns. This justifies why, among the patterns in $\mathcal{D}_D$, Algorithm 5 transforms first those with maximum impact on making other patterns α-protective w.r.t. f. For each pattern $p \in \mathcal{D}_D$, the number of patterns in $\mathcal{D}_D$ which are subsets of p is taken as the impact of p (Lines 4-7), that is $impact(p)$. Then, the patterns in $\mathcal{D}_D$ will be made α-protective w.r.t. f by descending order of $impact$ (Line 8). Thus, the patterns with maximum $impact(p)$ will be made α-protective first, with the aim of minimizing the pattern distortion. Finally, the algorithm uses the function AntiDiscPattSanit in Algorithm 2 to make each pattern p in $\mathcal{D}_D$ α-protective using anti-discrimination pattern sanitization methods w.r.t. f (Line 9).

6.2 Achieving d-explainable α-protective and k-anonymous pattern set

In order to simultaneously achieve unexplainable discrimination and privacy-awareness in $\mathcal{F}(\mathcal{D}, \sigma)$, we need to generate an unexplainable discrimination and privacy protected version of $\mathcal{F}(\mathcal{D}, \sigma)$:

Definition 19 (d-explainable α-protective k-anonymous pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D}, \sigma)$, an anonymity threshold k, an explainable discrimination threshold d, a discrimination threshold α, protected groups DI_b, legally-grounded groups DI_e, and a discrimination measure f, $\mathcal{F}(\mathcal{D}, \sigma)$ is d-explainable α-protective k-anonymous if it is both k-anonymous and d-explainable α-protective w.r.t. DI_b, DI_e and f.

In order to generate a d-explainable α-protective k-anonymous version of $\mathcal{F}(D, \sigma)$, we need to examine the following issues: (1) how making $\mathcal{F}(D, \sigma)$ k-anonymous impacts on d-explainable and d-unexplainable patterns in $\mathcal{F}(D, \sigma)$; (2) how making $\mathcal{F}(D, \sigma)$ d-explainable α-protective w.r.t. f impacts on the k-anonymity of $\mathcal{F}(D, \sigma)$. We study the first issue by presenting two scenarios.

Table 4 Scenario 3: Examples of frequent patterns extracted from Table 1

Patterns	*Support*
p_s:{`female, veterinarian`}	34
p_2 :{`paid-delay, veterinarian, salary > 15000`}	59
p_1 :{`female, veterinarian, No`}	20
p_n :{`paid-delay, veterinarian, No`}	37
p_{ns} :{`paid-delay, veterinarian`}	64
p_d :{`female, veterinarian, paid-delay`}	31

Scenario 3 Table 4 illustrates an example of frequent patterns that come from the data set in Table 1 with $\sigma = 15$. Let DI_b :{*Sex = female*}, DI_e :{*Credit_history = paid-delay*}, $f = slift$, $\alpha = 1.25$, $k = 8$ and $d = 0.9$. Suppose the PD pattern p_1 in Table 4 is 1.25-discriminatory (i.e., there is inferred discrimination against veterinarian women applicants). However, pattern p_1 is a 0.9-explainable pattern because both conditions of Definition 15 are satisfied ($\frac{supp(p_n)}{supp(p_{ns})} = \frac{37}{64} = 0.57$ is higher than $0.9 \cdot \frac{supp(p_1)}{p_s} = 0.9 \cdot \frac{20}{34} = 0.53$ and $\frac{supp(p_d)}{supp(p_s)} = \frac{31}{34} = 0.91$ is higher than 0.9). Then, p_1 is a d-explainable pattern regarding pattern p_n (i.e., the inferred discrimination against veterinarian women applicants is explainable by their delay in returning previous credits). On the other hand, although the support of each pattern in the collection is higher than k, there is an inference channel between patterns p_{ns} and p_2; note that $supp(p_{ns}) - supp(p_2) = 64 - 59 = 5$ is smaller than 8 (i.e., one can infer the existence of only 5 veterinarians who are delayed in returning previous credits and with salary no more than 15000 €). To block the inference channel between p_{ns} and p_2, the following privacy additive sanitization is performed:

$$supp(p_{ns}) + k, \forall p \subseteq p_{ns} \tag{20}$$

After this transformation, the new support of pattern p_{ns} is 72. The new support value of pattern p_{ns} changes the satisfaction of Condition 1 in Definition 15 in the following way: $\frac{supp(p_n)}{supp(p_{ns})+k} = \frac{37}{64+8} = 0.513$ is less than $0.9 \cdot \frac{supp(p_1)}{p_s} = 0.9 \cdot \frac{20}{34} = 0.53$. Then, in this scenario, making the collection of patterns in Table 4 k-anonymous makes d-explainable pattern p_1 d-unexplainable. □

Table 5 Scenario 4: Examples of frequent patterns extracted from Table 1

Patterns	*Support*
p_s:{female, veterinarian}	30
p_2 :{female, veterinarian, salary > 15000}	29
p_1 :{female, veterinarian, No}	29
p_n :{paid-delay, veterinarian, No}	23
p_{ns} :{paid-delay, veterinarian}	27
p_d :{female, veterinarian, paid-delay}	30

Scenario 4 Table 5 illustrates an example of frequent patterns that come from the data set in Table 1 with $\sigma = 15$. Let DI_b : {*Sex = female*}, DI_e :{*Credit_history = paid-delay*}, $f = slift$, $\alpha = 1.25$, $k = 2$ and $p = 0.9$. Suppose the PD pattern p_1 in Table 5 is 1.25-discriminatory (i.e., there is inferred discrimination against veterinarian women applicants). Pattern p_1 is a d-unexplainable pattern regarding pattern p_n because Condition 1 of Definition 15 does not satisfy ($\frac{supp(p_n)}{supp(p_{ns})} = \frac{23}{27} = 0.85$ is less than $0.9 \cdot \frac{supp(p_1)}{supp(p_s)} = 0.87$) while Condition 2 of Definition 15 is satisfied (i.e., $\frac{supp(p_d)}{supp(p_s)} = \frac{30}{30} = 1$ is higher than 0.9). Then, p_1 is a d-unexplainable pattern (i.e., the inferred discrimination against veterinarian women applicants is not explainable by their delay in returning previous credits). On the other hand, although the support of each pattern in the collection is higher than k, there is an inference channel between patterns p_s and p_2; note that $supp(p_s) - supp(p_2) = 30 - 29 = 1$ is smaller than 2. To block the inference channel between p_s and p_2, privacy pattern sanitization is performed according to Expression (17). After this transformation, the new support value of pattern p_s is 32. The new support value of pattern p_{ns} satisfies Condition 1 of Definition 15 ($\frac{supp(p_n)}{supp(p_{ns})} = \frac{23}{27} = 0.85$ is higher than $0.9 \cdot \frac{supp(p_1)}{supp(p_s)+k} = 0.815$) while it does not change the satisfaction of Condition 2 of Definition 15 ($\frac{supp(p_d)}{supp(p_s)+k} = \frac{30}{32} = 0.93$ is higher than 0.9). Thus, in this scenario, making the collection of patterns in Table 4 k-anonymous turns d-unexplainable pattern p_1 into a d-explainable pattern. □

To summarize, using a privacy pattern sanitization method for making $\mathcal{F}(D, \sigma)$ k-anonymous can make $\mathcal{F}(D, \sigma)$ more or less d-explainable α-protective w.r.t. $slift$. We also observe a similar behavior for alternative discrimination measures. Hence, what makes sense is first to obtain a k-anonymous version of $\mathcal{F}(D, \sigma)$ and then look for d-unexplainable patterns in it. After that, we use

anti-discrimination pattern sanitization methods proposed in Section 5.1 for making $\mathcal{F}(D,\sigma)$ d-explainable α-protective. According to Theorem 4, these methods cannot make $\mathcal{F}(D,\sigma)$ non-k-anonymous as a result of their transformation. The above procedure (first dealing with k-anonymity and then with d-explainability and α-protectiveness) is encoded in Algorithm 6. Given an original pattern set $\mathcal{F}(\mathcal{D},\sigma)$, denoted by $\mathcal{FP}$ for short, a discriminatory threshold α, an anonymity threshold k, a discrimination measure f, an explainable discrimination threshold d, protected groups DI_b, legally-grounded groups DI_e, and a class item C which denies some benefit, Algorithm 6 starts by obtaining $\mathcal{FP}'$, which is a k-anonymous version of $\mathcal{FP}$ (Step 3). It calls Algorithm 3 in [4] to do this. Then, the algorithm uses the function DetUnExplainPatt in Algorithm 3 to determine the subset $\mathcal{D}_{bad}$ which contains d-unexplainable patterns in $\mathcal{FP}'$ (Line 4). Then, the algorithm sorts $\mathcal{D}_{bad}$ by descending order of *impact* to transform first those patterns in $\mathcal{D}_{bad}$ with maximum impact on making other patterns α-protective (Lines 4-8). Finally, the algorithm uses the function UnExplainAntiDiscPattSanit in Algorithm 4 to make each d-unexplainable pattern in $\mathcal{FP}'$ α-protective using anti-discrimination pattern sanitization w.r.t. f (Line 9).

Algorithm 6 Unexplainable discrimination protected and anonymous pattern sanitization

1: Inputs: $\mathcal{FP} := \mathcal{F}(D,\sigma)$, k, DI_b, DI_e, explainable discrimination threshold d, discrimination measure f, α, C =class item with negative decision value
2: Output: $\mathcal{FP}''$: d-explainable α-protective k-anonymous frequent pattern set
3: $\mathcal{FP}' = PrivacyAdditiveSanitization(FP,k)$ //Algorithm 3 in [4]
4: $\mathcal{D}_{bad} \rightarrow$ **Function** DetUnExplainPatt($\mathcal{FP}'$, DI_e, DI_b, f, α, C)
5: **for all** $p \in \mathcal{D}_{bad}$ **do**
6: Compute $impact(p) = |\{p' : A', B', C\} \in \mathcal{D}_{bad}$ s.t. $p' \subset p|$
7: **end for**
8: Sort $\mathcal{D}_{bad}$ by descending *impact*
9: $\mathcal{FP}'' \leftarrow$ **Function** UnExplainAntiDiscPattSanit($\mathcal{FP}'$, $\mathcal{D}_{bad}$, DI_b, f, α, C)
10: Output: $\mathcal{FP}''$

6.3 Discrimination protected and differentially private frequent patterns

In this section, we present how simultaneous anti-discrimination and privacy can be achieved in frequent pattern discovery while satisfying differential privacy (instead of k-anonymity) and α-protection.

To simultaneously achieve anti-discrimination and differential privacy in frequent pattern discovery, we need to generate an α-protective ϵ-differentially private version of $\mathcal{F}(\mathcal{D},\sigma)$:

Definition 20 (α-protective ϵ-differentially private frequent pattern set) Given a collection of frequent patterns $\mathcal{F}(\mathcal{D},\sigma)$, a differential privacy

budget ϵ, a discriminatory threshold α, protected groups DI_b, and a discrimination measure f, $\mathcal{F}(\mathcal{D}, \sigma)$ is α-protective ϵ-differentially private if it is both ϵ-differentially private and α-protective w.r.t. f and DI_b.

Clearly, first we need to examine how to make $\mathcal{F}(\mathcal{D}, \sigma)$ ϵ-differentially private impacts on the α-protectiveness of $\mathcal{F}(\mathcal{D}, \sigma)$ w.r.t. f, where f is one of the measures from Definitions 8-9.

Theorem 5 *The* PrivBasis *approach for making $\mathcal{F}(\mathcal{D}, \sigma)$ ϵ-differential private can make $\mathcal{F}(\mathcal{D}, \sigma)$ more or less α-protective w.r.t. f and DI_b.*

Proof The ϵ-differentially private version of $\mathcal{F}(\mathcal{D}, \sigma)$, denoted by $\mathcal{FP}_d$ for short, generated by *PrivBasis* is an approximation of $\mathcal{F}(\mathcal{D}, \sigma)$. As a side effect of this transformation due to Laplacian or exponential mechanisms, $\mathcal{FP}_d$ might contain patterns that are not in $\mathcal{F}(\mathcal{D}, \sigma)$ (i.e., ghost patterns) and might not contain patterns that are in $\mathcal{F}(\mathcal{D}, \sigma)$ (i.e., missing patterns). Moreover, for patterns that are in both $\mathcal{F}(\mathcal{D}, \sigma)$ and $\mathcal{FP}_d$, $\mathcal{FP}_d$ contains the noisy new support values of original patterns in $\mathcal{F}(\mathcal{D}, \sigma)$. Hence, making $\mathcal{F}(\mathcal{D}, \sigma)$ ϵ-differentially private can lead to different situations regarding the α-protectiveness of $\mathcal{FP}_d$ w.r.t. f and DI_b. We list such situations below:

- There are α-discriminatory patterns in $\mathcal{F}(\mathcal{D}, \sigma)$ w.r.t. f and DI_b which are not in $\mathcal{FP}_d$ (i.e., missing patterns). In this situation, making $\mathcal{F}(\mathcal{D}, \sigma)$ ϵ-differentially private makes $\mathcal{F}(\mathcal{D}, \sigma)$ more α-protective.
- There are α-discriminatory patterns in $\mathcal{FP}_d$ w.r.t. f and DI_b which are not in $\mathcal{F}(\mathcal{D}, \sigma)$ (i.e., ghost patterns). In this situation, making $\mathcal{F}(\mathcal{D}, \sigma)$ ϵ-differentially private makes $\mathcal{F}(\mathcal{D}, \sigma)$ less α-protective.
- There are PD patterns (e.g., $p : \{A, B, C\}$) in both $\mathcal{F}(\mathcal{D}, \sigma)$ and $\mathcal{FP}_d$ that are α-protective (resp. α-discriminatory) w.r.t. f and DI_b in $\mathcal{F}(\mathcal{D}, \sigma)$ and α-discriminatory (resp. α-protective) in $\mathcal{FP}_d$. This is because the new noisy support values of patterns in $\mathcal{FP}_d$ can increase (resp. decrease) the values of $f(A, B \rightarrow C)$. In this situation, making $\mathcal{F}(\mathcal{D}, \sigma)$ ϵ-differentially private makes $\mathcal{F}(\mathcal{D}, \sigma)$ less (resp. more) α-protective.

Hence, the theorem holds. □

Thus, similar to k-anonymity, achieving differential privacy in frequent pattern discovery can achieve anti-discrimination or work against anti-discrimination. Hence, what makes sense is first to obtain an ε-differentially private version of $\mathcal{F}(D, \sigma)$ and then deal with discrimination. An interesting point is that, regardless of whether k-anonymity or ε-differential privacy is used, achieving privacy impacts on anti-discrimination similarly. Based on this observation, we present Algorithm 7 to generate an α-protective ϵ-differentially private version of an original pattern set w.r.t. f and DI_b.

Theorem 6 *Algorithm 7 is ϵ-differentially private.*

Proof The only part in Algorithm 7 that depends on the dataset is Step 3, which is ϵ-differentially private because it uses *PrivBasis*. Starting from Step

Algorithm 7 ANTI-DISCRIMINATION DIFFERENTIALLY PRIVATE PATTERN SANITIZATION

1: Inputs: Database $\mathcal{D}$, K, items I, differential privacy budget ϵ, DI_b, discrimination measure f, α, C =class item with negative decision value
2: Output: $\mathcal{FP}''$: α-protective ϵ-differential private frequent pattern set
3: $(\mathcal{FP}_d, b[i][p]) \leftarrow$ *PrivBasis*($\mathcal{D}$,I,K,ϵ) //Algorithm 3 in [38]
4: $\mathcal{D}_D \leftarrow$ **Function** DETDISCPATT($\mathcal{FP}_d$, $b[i][p]$, DI_b, f, α, C)
5: Lines 5-8 of Algorithm 5 to sort $\mathcal{D}_D$
6: $\mathcal{FP}'' \leftarrow$ **Function** ANTIDISCPATTSANIT($\mathcal{FP}_d$, $b[i][p]$, $\mathcal{D}_D$, DI_b, f, α, C)
7: Output: $\mathcal{FP}''$

4, the algorithm only performs post-processing, and does not access $\mathcal{D}$ again. Indeed, in Step 4 the value of f w.r.t. each PD pattern in $\mathcal{FP}_d$ and in Step 6 the value of Δ_f can be computed using the noisy support values of patterns in $\mathcal{FP}_d$ and the noisy bin counts in array $b[i][p]$ (array $b[i][p]$ replaces parameter $\mathcal{D}$ in functions DETDISCPATT and ANTIDISCPATTSANIT). Adding Δ_f to the noisy support values of respective patterns in Step 6 is post-processing of differentially private results which remain private as proven in [29]. □

Thus, α-protection in $\mathcal{F}(\mathcal{D}, \sigma)$ can be achieved by anti-discrimination pattern sanitization methods proposed in Section 5.1 without violating differential privacy, just as we could do it without violating k-anonymity.

7 Experimental analysis

This section presents the experimental evaluation of the approaches we proposed in this paper. First, we describe the utility measures and then the empirical results and the execution time. In the sequel, $\mathcal{FP}$ denotes the set of frequent patterns extracted from database $\mathcal{D}$ by the Apriori algorithm [2]; $\mathcal{FP}'$ denotes the k-anonymous version of $\mathcal{FP}$ obtained by the privacy pattern sanitization method; $\mathcal{FP}''$ denotes the α-protective k-anonymous version of $\mathcal{FP}$ obtained by Algorithm 5, and $\mathcal{FP}^*$ denotes the α-protective version of $\mathcal{FP}$ obtained by Algorithm 2. We also denote by $\mathcal{TP}$ any transformed pattern set, i.e., either $\mathcal{FP}'$, $\mathcal{FP}''$ or $\mathcal{FP}^*$.

We used the Adult and German credit datasets from the UCI Repository of Machine Learning Databases [18]. These are well-known real-life datasets, containing both numerical and categorical attributes, frequently used in anti-discrimination and PPDM research.

The Adult dataset, also known as Census Income, consists of 48,842 records, split into a "train" part with 32,561 records and a "test" part with 16,281 records. The dataset has 14 attributes (without the class attribute). We used the "train" part to obtain $\mathcal{FP}$ and any transformed pattern set $\mathcal{TP}$ defined above. The prediction task associated to the Adult dataset is to determine whether a person makes more than 50K$ a year based on census and demographic information about people. For our experiments with the Adult dataset, we set DI_b : {*Sex = female, Age = young*} (cut-off for Age = young: 30 years

old). The percentages of records in protected groups are 33% for *Sex = female* and 29% for *Age = young*.

The German credit dataset consists of 1,000 records and 20 attributes (without class attribute) of bank account holders. The class attribute in the German credit dataset takes values representing good or bad classification of the bank account holders. For our experiments with this dataset, we set DI_b : {*Age = old, Foreign worker = yes*} (cut-off for Age = old: 50 years old). The percentages of records in protected groups are 11% for *Age = old* and 96% for *Foreign worker = yes*.

7.1 Utility measures

To assess the information loss incurred to achieve privacy and anti-discrimination in frequent pattern discovery, we use the following measures.

- **Support-altered**. Fraction of original frequent patterns in $\mathcal{FP}$ which have their support changed in any transformed pattern set $\mathcal{TP}$:

$$\frac{|\{\langle I, supp(I)\rangle \in \mathcal{FP} : supp_{\mathcal{FP}}(I) \neq supp_{\mathcal{TP}}(I)\}|}{|\mathcal{FP}|}$$

- **Pattern distortion error**. Average distortion w.r.t. the original support of frequent patterns:

$$\frac{1}{|\mathcal{FP}|} \cdot \sum_{I \in \mathcal{FP}} \left(\frac{supp_{\mathcal{TP}}(I) - supp_{\mathcal{FP}}(I)}{supp_{\mathcal{FP}}(I)}\right)$$

- **Misses Cost (MC)**. The fraction of original frequent patterns which do not appear in the published transformed patterns.
- **Ghost Cost (GC)**. The fraction of frequent patterns appearing in the published transformed patterns that are not in the original pattern set.

The purpose of these measures is assessing the distortion introduced when making $\mathcal{FP}$ α-protective k-anonymous, in comparison with the distortion introduced by either making $\mathcal{FP}$ α-protective or making $\mathcal{FP}$ k-anonymous, separately. The above measures are evaluated considering $\mathcal{TP} = \mathcal{FP}''$, $\mathcal{TP} = \mathcal{FP}'$ and $\mathcal{TP} = \mathcal{FP}^*$. In addition, we measure the impact of our pattern sanitization methods for making $\mathcal{FP}$ k-anonymous, α-protective and α-protective k-anonymous on the accuracy of a classifier using the CMAR (i.e., classification based on multiple association rules) approach [40]. Below, we describe the process of our evaluation:

1. The original data are first divided into training and testing sets, $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$.
2. The original frequent patterns $\mathcal{FP}$ are extracted from the training set $\mathcal{D}_{train}$ by the Apriori algorithm [2].
3. Patterns in $\mathcal{TP}$ which contain the class item are selected as candidate patterns for classification. They are denoted by $\mathcal{TP}_s$. Note that, $\mathcal{TP}$ can be $\mathcal{FP}$, $\mathcal{FP}'$, $\mathcal{FP}^*$ or $\mathcal{FP}''$.

4. To classify each new object (record) in $\mathcal{D}_{test}$, the subset of patterns matching the new record in $\mathcal{TP}_s$ is found.
5. If all the patterns matching the new object have the same class item, then that class value is assigned to the new record. If the patterns are not consistent in terms of class items, the patterns are divided into groups according to class item values (e.g., denying credit and accepting credit). Then, the effects of the groups should be compared to get the strongest group. A strongest group is composed of a set of patterns highly positively correlated and with good support. To determine this, for each pattern $p : \{X, C\}$, the value of $\max \chi^2$ is computed as follows:

$$\max \chi^2 = (\min\{supp(X), supp(C)\} - \frac{supp(X) \times supp(C)}{|\mathcal{D}_{train}|})^2 \times |\mathcal{D}_{train}| \times e \tag{21}$$

where

$$e = \frac{1}{supp(X) \times supp(C)} + \frac{1}{supp(X) \times (|\mathcal{D}_{train}| - supp(C))} + \frac{1}{(|\mathcal{D}_{train}| - supp(X)) \times supp(C)} + \frac{1}{(|\mathcal{D}_{train}| - supp(X))(|\mathcal{D}_{train}| - supp(C))} \tag{22}$$

That is, $\max \chi^2$ computes the upper bound of χ^2 values of the pattern. Then, for each group of patterns, the *weighted* χ^2 measure of the group is computed as $\sum \frac{\chi^2 \times \chi^2}{max\chi^2}$ (see [23]). The class item of the group with maximum *weighted* χ^2 is assigned to the new record.
6. After obtaining the predicted class item of each record in $\mathcal{D}_{test}$, the accuracy of the classifier can be simply computed w.r.t. observed and predicted class items (i.e., contingency table).

To measure the accuracy of a classifier based on a collection of frequent patterns, we perform the above process considering $\mathcal{TP} = \mathcal{FP}$, $\mathcal{TP} = \mathcal{FP}''$, $\mathcal{TP} = \mathcal{FP}'$ and $\mathcal{TP} = \mathcal{FP}^*$. Finally, to measure the impact of making $\mathcal{FP}$ k-anonymous on discrimination, we use the following measures, which are evaluated considering $\mathcal{TP} = \mathcal{FP}'$:

- **Discrimination prevention degree (DPD)**. Percentage of α-discriminatory patterns obtained from $\mathcal{FP}$ that are no longer α-discriminatory in the transformed patterns ($\mathcal{TP}$).
- **New (ghost) discrimination degree (NDD)**. Percentage of α-discriminatory patterns obtained from transformed patterns ($\mathcal{TP}$) that were α-protective in $\mathcal{FP}$.

7.2 Pattern distortion

In this section, we report the pattern distortion scores to make the collection of frequent patterns k-anonymous (Section 7.2.1), α-protective (Section

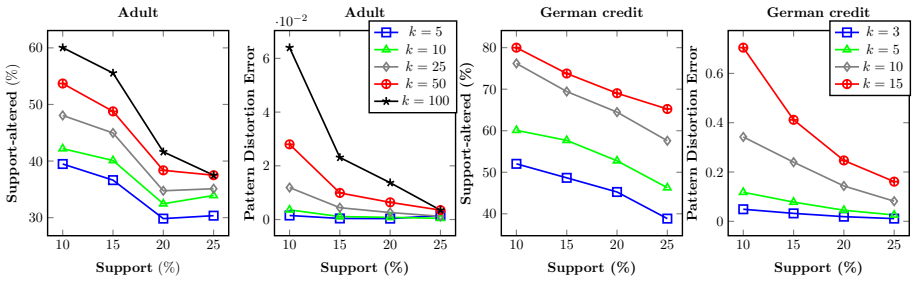

Fig. 3 Pattern distortion scores to make the Adult and German credit dataset k-anonymous

7.2.2) and α-protective k-anonymous (Section 7.2.3). Moreover, we present the pattern distortion scores to make the collection of frequent patterns α-protective and α-protective k-anonymous by using an alternative discrimination measure in Section 7.2.4. The pattern distortion scores to make the patterns d-explainable α-protective and d-explainable α-protective k-anonymous are presented in Section 7.2.5. Finally, Section 7.2.6 reports the pattern distortion scores to make the pattern set ϵ-differentially private and α-protective ϵ-differentially private.

7.2.1 k-anonymous frequent patterns

Fig. 3 shows pattern distortion scores observed after making $\mathcal{FP}$ k-anonymous in Adult and German credit. We show the results for varying values of k and the support σ. It can be seen that the percentage of patterns whose support has changed and the average distortion introduced increase with larger k and with smaller support σ, due to the increasing number of inference channels. Comparing the two datasets, pattern distortion scores in German credit are higher than those in Adult, even taking the same values of k and σ. This is mainly due to the substantial difference in the number of inference channels detected in the two datasets: the maximum number of inference channels detected in Adult is 500, while in German credit it is 2,164.

Moreover, the MC and GC when making the original pattern set k-anonymous are always equal to zero.

7.2.2 α-protective frequent patterns

Fig. 4 shows pattern distortion scores observed after making $\mathcal{FP}$ α-protective in Adult and German credit. We take $f = slift$ and we show the results for varying values of α and support σ. It can be seen that distortion scores increase with smaller α and smaller σ, because the number of α-discriminatory patterns increases. Also in this case the values of the distortion scores for Adult are less than for German credit.

When comparing Fig. 3 and Fig. 4, we observe that the percentage of patterns with changed support and the average distortion introduced are higher

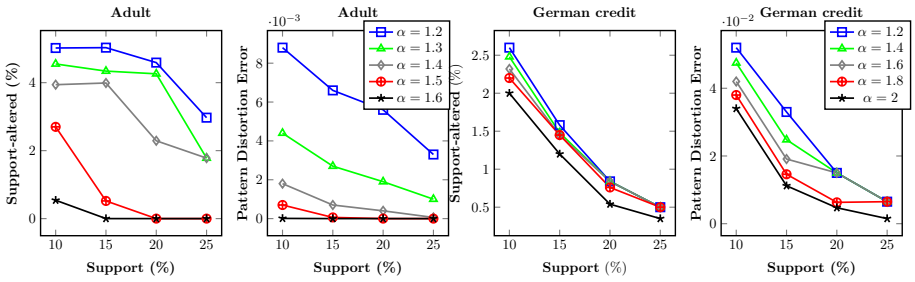

Fig. 4 Pattern distortion scores to make the Adult and German credit dataset α-protective w.r.t. $f = slift$

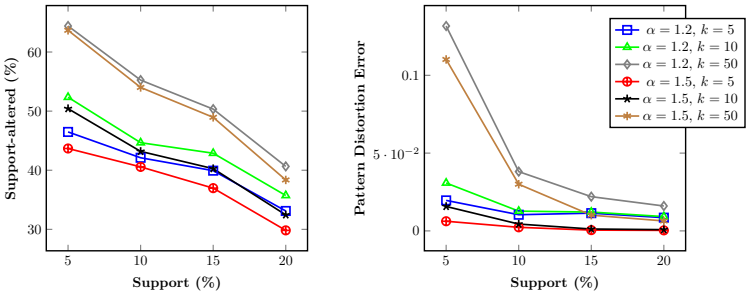

Fig. 5 Pattern distortion scores to make the Adult dataset α-protective k-anonymous w.r.t. $f = slift$

after the application of privacy pattern sanitization in both datasets. In other words, we observe that guaranteeing privacy produces more distortion than guaranteeing anti-discrimination. This is due to the number of inference channels detected in our experiment (for different values of k), which is higher than the number of α-discriminatory patterns detected (for different values of α). Moreover, the MC and GC when making the original pattern set α-protective are always equal to zero.

7.2.3 α-protective k-anonymous frequent patterns

Figs. 5 and 6 show the pattern distortion scores observed after making $\mathcal{FP}$ α-protective k-anonymous in the Adult and German credit datasets, respectively. We take $f = slift$ and we show the results for varying values of k, α and σ. Since the number of inference channels increases with k, the number of α-discriminatory patterns increases as α becomes smaller, and both numbers increase as σ becomes smaller, the percentage of patterns with modified support and the average distortion introduced have the same dependencies w.r.t. k, α and σ. Comparing the two datasets, here we also observe that the values of the distortion scores for the Adult dataset are less than for the German credit dataset.

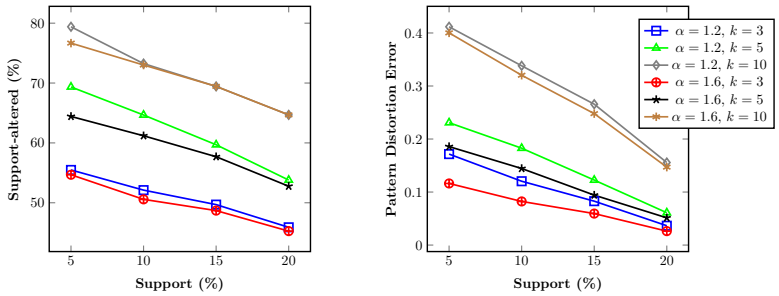

Fig. 6 Pattern distortion scores to make the German credit dataset α-protective k-anonymous w.r.t. $f = slift$

If we compare Figs. 5 and 6 with Fig. 3, we observe only a marginal difference between the distortion introduced when making $\mathcal{FP}$ α-protective k-anonymous, and the distortion introduced when making it only k-anonymous. For instance, in the experiment where we make the collection of patterns (extracted with minimum support 10% from Adult) α-protective k-anonymous with $k = 5$, the percentage of patterns with changed support is 42.1% (in the worst case $\alpha = 1.2$), while when making the pattern set k-anonymous we get a value of 39.48%. In addition, the average distortion introduced is 0.010 (in the worst case $\alpha = 1.2$) to obtain an α-protective k-anonymous version of original patterns, while it is 0.001 to obtain a k-anonymous version of it. As a consequence, we can (empirically) conclude that we provide protection against both the privacy and discrimination threats with a marginally higher distortion w.r.t. providing protection against the privacy threat only. Moreover, the MC and GC when making the original pattern set α-protective k-anonymous are always equal to zero.

7.2.4 Alternative discrimination measures

As introduced in Section 3.3, different legally-grounded discrimination measures are adopted worldwide. The main difference between our proposed pattern sanitization methods comes from the difference between the $slift$ and $elift$ measures. The other measures are variations of these two measures, e.g., $slift$ versus $slift_d$, and the same pattern sanitization can be applied for each variation of $slift$ or $elift$. For this reason, in this section we report the pattern distortion scores to make the frequent patterns extracted from Adult α-protective and k-anonymous α-protective where $f = elift$.

Fig. 7 shows pattern distortion scores observed after making $\mathcal{FP}$ α-protective and α-protective k-anonymous in Adult. We take $f = elift$ and we show the results for varying values of α, k and support σ. As expected, we can observe that these results are in line with the results we obtain where $f = slift$: the distortion scores increase with smaller α, larger k and smaller σ. If we compare the pattern distortion scores of two measures, we observe that the pattern distortion scores for $f = elift$ are slightly different from the ones for $f = slift$.

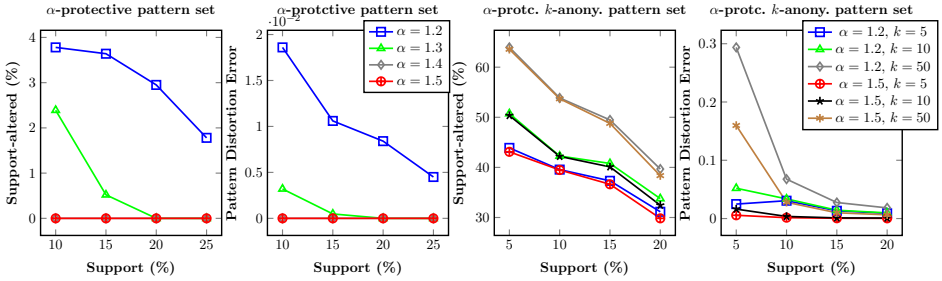

Fig. 7 Pattern distortion scores to make the Adult dataset α-protective and α-protective k-anonymous w.r.t. $f = elift$

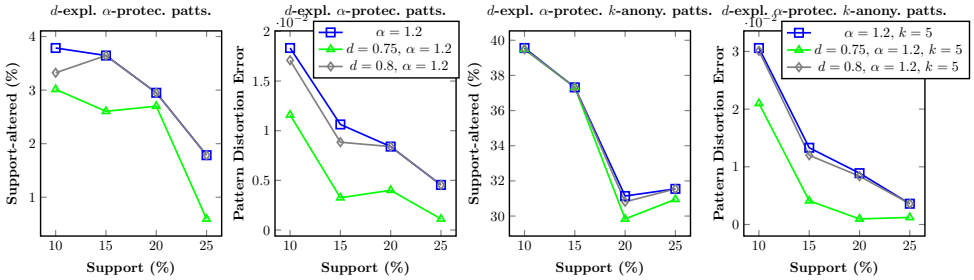

Fig. 8 Pattern distortion scores to make the Adult dataset d-explainable α-protective and d-explainable α-protective k-anonymous w.r.t. $f = elift$

This is due to the fact that, in each measure and for the same value of α, different numbers of α-discriminatory patterns can be discovered. Another reason is the difference in the pattern distortion method and the value of Δ_f that is added to the support of original patterns.

7.2.5 Unexplainable discrimination protection

All the empirical results presented in the previous sections are related to the worst-case scenario in which the original patterns are protected against both explainable and unexplainable discrimination. In other words, we generate α-protective or α-protective k-anonymous versions of the original pattern set. In this section, we report the pattern distortion scores for protecting the original pattern set against only unexplainable discrimination (that is, generating d-explainable α-protective or d-explainable α-protective k-anonymous versions). Fig. 8 shows these results for the Adult dataset, $f = elift$, $k = 5$, $\alpha = 1.2$ and different values of d and σ, In this experiment, we assume that all the itemsets excluding the ones belonging to PD attributes are legally grounded.

As expected, Fig. 8 highlights the fact that protecting the original pattern set against only unexplainable discrimination can lead to less or equal information loss and pattern distortion. This is because pattern sanitization methods transform a smaller number of patterns if the number of d-explainable patterns is greater than zero. It can be seen that distortion scores decrease with

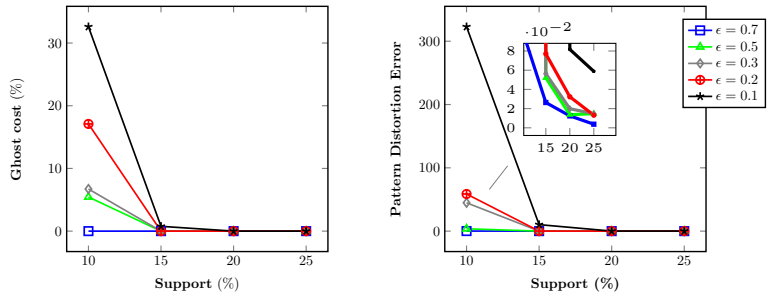

Fig. 9 Pattern distortion scores to make the Adult dataset ϵ-differential private

smaller d and smaller σ, because the number of d-explainable α-discriminatory patterns increases.

7.2.6 Differentially private frequent patterns

We also evaluated the distortion introduced by applying the privacy transformation *PrivBasis* presented in [38] to obtain differentially private patterns in the Adult dataset. We tested different privacy budget values $\epsilon = 0.1, 0.2, 0.3, 0.5, 0.7$. Note that *PrivBasis* algorithm returns the top-K frequent itemsets, so we adapted the algorithm to the frequent pattern mining problem by setting K to be the number of frequent itemsets given a threshold. Fig. 9 shows the pattern distortion scores. We observe that the approach based on k-anonymity introduces less distortion, which is not surprising, because it changes only the support values of subsets of original patterns. In contrast, with the differential privacy approach, the support values of all the patterns are changed. Moreover, *PrivBasis* can generate *fake* frequent patterns, i.e., GC can be greater than zero. In Fig. 9 we only report GC and the pattern distortion error, because for all ϵ values we have that 100% of patterns have the support value perturbed and MC is zero. This is different for the approach based on k-anonymity that changes the support of a low number of patterns and does not generate *fake* patterns for construction.

It can be seen that the average distortion introduced and GC increase as ϵ and the smaller minimum support become smaller, due to the increase in the values of random noise and the increase in the number of frequent patterns, respectively. In summary, we observe a huge difference between the distortion introduced when making $\mathcal{FP}$ k-anonymous, and the distortion introduced when making it ϵ-differentially private.

Fig. 10 shows the pattern distortion scores observed after making $\mathcal{FP}$ α-protective ϵ-differentially private in the Adult. We take $f = elift$ and we show the results for varying values of ϵ, α and minimum support. We observe a marginal difference between the distortion introduced when making $\mathcal{FP}$ α-protective ϵ-differentially private, and the distortion introduced when making it only ϵ-differentially private. These results are in line with the results we obtain for α-protection k-anonymity showing that we provide protection against

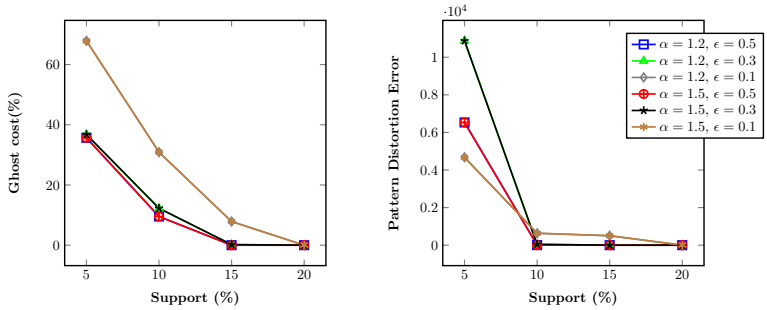

Fig. 10 Pattern distortion scores to make the Adult dataset α-protective ϵ-differentially private

both the privacy and discrimination threats with a marginally higher distortion w.r.t. providing protection against the privacy threat only. However, pattern distortion scores for making $\mathcal{FP}$ α-protective ϵ-differentially private are substantially higher than the ones for making $\mathcal{FP}$ α-protective k-anonymous. Clearly, this is because of the huge difference in the pattern distortion scores to achieve ϵ-differential privacy and k-anonymity.

7.3 Preservation of the classification task

Tables 6 and 7 show the accuracy of classifiers obtained from $\mathcal{FP}$, $\mathcal{FP}'$, $\mathcal{FP}''$ and $\mathcal{FP}^*$ in the Adult and German credit datasets for $f = slift$, $\sigma = 10\%$ and different values of α and k. We do not observe a significant difference between the accuracy of the classifier obtained from an α-protective k-anonymous version of the original pattern set and the accuracy of the classifier obtained from either a k-anonymous or an α-protective version. In addition, the accuracy of the classifier decreases with larger k and with smaller α. When comparing the two datasets, we observe less accuracy for the German credit dataset; this is consistent with the higher distortion observed above for this dataset. Note that the low values of accuracy in Tables 6 and 7 are related to the worst-case scenario (that is, maximum value of k and minimum value of α).

Table 6 Adult dataset: accuracy of classifiers

k	α	$\mathcal{FP}$	$\mathcal{FP}'$	$\mathcal{FP}''$	$\mathcal{FP}^*$
5	1.2	0.744	0.763	0.724	0.691
5	1.5	0.744	0.763	0.752	0.739
50	1.2	0.744	0.751	0.682	0.691
50	1.5	0.744	0.751	0.746	0.739

Table 7 German dataset: accuracy of classifiers

k	α	$\mathcal{FP}$	$\mathcal{FP}'$	$\mathcal{FP}''$	$\mathcal{FP}^*$
3	1.2	0.7	0.645	0.582	0.572
3	1.8	0.7	0.645	0.624	0.615
10	1.2	0.7	0.583	0.561	0.572
10	1.8	0.7	0.583	0.605	0.615

Table 8 Discrimination utility measures after privacy pattern sanitization: Adult (top); German credit (bottom)

	Support = 15%										Support = 20%									
	K = 5		K = 10		K = 25		K = 50		K = 100		K = 5		K = 10		K = 25		K = 50		K = 100	
α	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD
1.2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.3	0	0	0	0	0	25	0	29.41	0	33.33	0	0	0	0	0	0	0	0	0	0
1.4	0	0	0	0	0	12.5	0	12.5	0	30	0	0	0	33.33	0	40	0	40	0	0
1.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	25	0	40
1.6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

	Support = 15%								Support = 20%							
	K = 3		K = 5		K = 10		K = 15		K = 3		K = 5		K = 10		K = 15	
α	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD	DPD	NDD
1.2	0	0	9.8	0	54.9	0	82.35	0	0	0	0	0	16.67	0	58.33	0
1.4	4.17	0	18.75	0	60.42	0	85.40	0	0	0	0	0	16.67	0	58.33	0
1.6	8.51	0	31.91	0	80.85	0	87.23	0	0	0	0	0	16.67	0	58.33	0
1.8	23.91	0	47.83	0	82.61	0	91.3	0	18.18	0	45.45	0	63.64	0	81.82	0
2	25	0	50	0	86.11	0	94.44	0	25	0	25	0	62.5	0	87.5	0

7.4 Degree of discrimination

Finally, Tables 8 show the discrimination degree measures (DPD and NDD) obtained after applying privacy pattern sanitization to frequent patterns extracted from Adult and German credit, respectively, for $f = slift$ and different values of α, k and σ. As stated in Section 6.1, applying privacy pattern sanitization can eliminate or create discrimination. Tables 8 clearly highlight this fact: the values of DPD and NDP increase with k. This is because more inference channels are detected for larger values of k and our method perturbs the support of more patterns. As a consequence, the impact on anti-discrimination may increase. Comparing the results obtained in the two datasets, we observe that in Adult usually the value of NDD is higher than DPD, while in German credit it is the other way round. This shows that in the Adult dataset the privacy pattern sanitization tends to make $\mathcal{FP}$ less α-protective; in German credit the situation is reversed: the privacy pattern sanitization tends to make $\mathcal{FP}$ more α-protective (NDD = 0%).

Moreover, Table 9 shows the percentage of d-explainable patterns among α-discriminatory ones obtained from the original pattern set and a k-anonymous version of it, for the Adult and German credit datasets, $\alpha = 1.2$, $\sigma = 10\%$, $f = slift$ and several values of d and k. In this experiment, we assume that all the itemsets excluding the ones belonging to PD attributes are legally grounded. We can observe that the percentage of d-explainable patterns decreases with larger d. In addition, as stated in Section 6.2, applying privacy pattern transformation to make $\mathcal{FP}$ k-anonymous can make $\mathcal{FP}$ more or less d-explainable α-protective.

Table 9 Percentage of d-explainable patterns detected in $\mathcal{FP}$ and $\mathcal{FP}'$

	Adult										German							
	K = 5		K = 10		K = 25		K = 50		K = 100		K = 3		K = 5		K = 10		K = 15	
d	FP	FP′	FP	FP′	FP	FP′	FP	FP′	FP	FP′	FP	FP′	FP	FP′	FP	FP′	FP	FP′
0.85	70	85	70	70	70	70	70	65	70	30	62.47	2	62.47	2	62.47	2	62.47	2
0.9	15	15	15	15	15	15	15	0	15	0	35.2	0	35.2	0	35.2	0	35.2	0
0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

7.5 Execution time

The execution time of the algorithms presented in this paper increases linearly with the number of original data records $|\mathcal{D}|$ and the number of original frequent patterns $|\mathcal{FP}|$. Note that the number of frequent patterns can be exponential w.r.t. the number of records, which may affect also the execution time of our algorithms. The overall execution time of Algorithms 5 and 7 to achieve both privacy protection and discrimination prevention depends on the execution time of the algorithms to achieve privacy protection (either k-anonymity or differential privacy). Depending on the minimum support, the runtime execution of each of these algorithms takes between 1 to 10 minutes.

Excluding the execution time of the privacy protection step and depending on the number of original frequent patterns and α, the runtime of the final steps (discovering α-discriminatory patterns and anti-discrimination pattern sanitization) takes less than 2 minutes. To improve the performance of these two steps, we pre-compute the contingency table of each PD discriminatory pattern (see Fig. 2) to compute the value of the discrimination measure for each PD pattern efficiently. Thus, overall, in the worst-case scenarios (minimum support 5%) the execution of Algorithms 5 and 7 takes less than 12 minutes, which is a reasonable performance for an off-line process.

8 On the trade-off between protection and data utility

In the previous sections, we showed that we can produce frequent pattern sets that simultaneously guarantee discrimination and privacy protection. Clearly, each step of our pattern transformation introduces some information loss. However, as discussed in the experiments described in Section 7, the pattern distortion introduced to attain simultaneous α-protection and k-anonymous is quite affordable. Furthermore, such distortion is little more than the one incurred to achieve just k-anonymity, the additional distortion caused by anti-discrimination being very low.

In addition, we observe that this sanitization approach does not generate *fake* frequent patterns and does not suppress any real frequent patterns; indeed, we have that both MC and GC are always equal to zero in our k-anonymous and α-protective pattern sets. This property, together with the aforementioned results on pattern distortion, makes the set of transformed

Table 10 The summarized results

Social-aware patterns	**Relationships & Findings**
k-anonymous patterns	1) Making patterns more or less discrimination protected. 2) Acceptable pattern distortion.
differential private patterns	1) Higher distortion than k-anonymous patterns. 2) Making Patterns more or less discrimination protected.
α-protective patterns	1) Less distortion than k-anonymous patterns. 2) No privacy violation w.r.t. both privacy models.
d-explainable α-protective patterns	1) Equal or less distortion than α-protective patterns. 2) Patterns protected against unexplainable discrimination.
α-protective k-anonymous patterns	1) Slightly higher distortion than k-anonymous patterns.
d-explainable α-protective k-anonymous patterns	1) Equal or less distortion than α-protective k-anonymous patterns.
α-protective differential private patterns	1) Slightly higher distortion than differential privacy. 2) Higher distortion than α-protective k-anonymous patterns.

patterns suitable to be used as input for decision making. This fact is confirmed by the results presented in Section 7.3, where we evaluated how the accuracy of a classifier changes with the set of patterns produced by our sanitization. The relatively small accuracy loss incurred seems a price worth paying to satisfy two important human rights, privacy and non-discrimination.

Furthermore, the experiment in Section 7.3 provides a tool to help the decision maker in selecting the set of patterns to be used to obtain the classifier. Indeed, we could assume that, in a realistic scenario, the decision maker can choose among different sets of sanitized patterns with different associated levels of protection and accuracy.

The situation is a bit different when we apply to the original patterns the sanitization that guarantees both differential privacy and α-protection. Our experiments show that this approach introduces too much noise. Therefore, in terms of privacy, using differential privacy seems preferable because it provides a worst-case privacy guarantee. In terms of data utility, the situation is different because, as discussed in Section 3.2.2, if the approach proposed in [38] does not control the generation of negative bin counts, it may result in negative support values and collections of frequent patterns with contradictions among them. However, even if we proposed a post-processing step that solves this issue our findings are that the quality of the pattern set decreases too much with this transformation.

9 Conclusion and future research

In this paper, we have investigated the problem of discrimination and privacy-aware frequent pattern discovery, i.e., the sanitization of the collection of patterns mined from a transaction database in such a way that neither privacy-violating nor discriminatory inferences can be inferred on the released patterns. In particular, for each measure of discrimination used in the legal literature

we proposed a solution for obtaining a discrimination-free collection of patterns. We also proposed an algorithm to take into account the legal concept of genuine requirement to make an original pattern set protected only against unexplainable discrimination. We also found that our discrimination preventing transformations do not interfere with a privacy preserving sanitization based on k-anonymity, thus accomplishing the task of combining the two and achieving a robust (and formal) notion of fairness in the resulting pattern collection. We have also described a version of our proposed framework where we replace k-anonymity with differential privacy. Although our discrimination prevention method can be combined with the differentially private transformations, we have shown in our experiments that doing so can lead to substantially more information loss than with k-anonymity. Table 10 summarizes the different scenarios we considered in this paper, their relationship and our key findings.

Further, we have presented extensive empirical results on the utility of the protected data. Specifically, we evaluate the distortion introduced by our methods and its effects on classification. It turns out that the utility loss caused by simultaneous anti-discrimination and privacy protection is only marginally higher than the loss caused by each of those protections separately. This result supports the practical deployment of our methods. This work is the first one addressing in depth the problem of providing protection against both privacy and discrimination threats in data mining, and it opens several future research avenues. First of all, it would be interesting to study the integration of other differential privacy approaches to verify if they solve the problem of data utility. It would be interesting to study the effect of using non-derivable patterns instead of standard frequent patterns [9]. Another important aspect to be studied is the design of a framework fully unifying anti-discrimination and privacy protection.

Dealing with indirect discrimination is also another topic of interest, which is an issue when the original data contain no attributes that explicitly identify the protected groups, but they contain attributes that are very correlated to protected groups. For example, if ethnicity has been removed but zipcode remains in the data set, and one knows that in a certain zipcode nearly all inhabitants are black, then zipcode could be used to indirectly discriminate black people. This situation would call for joint treatment of privacy protection and indirect discrimination.

We also plan to investigate discrimination- and privacy-aware data publishing, i.e., the transformation of data, instead of patterns, aimed to fulfill the privacy-preserving and discrimination-preventing constraints; we have published preliminary exploratory work on this subject in [27,28].

Finally, we plan as future work a thorough real-world validation of our approaches by developing targeted campaigns to collect real data from different sources, including crowdsourcing/crowdsensing. Doing that is a challenge in itself, due to the difficulty of setting up suitable social experiments to gather high-quality real data and real feedback. However, we think this is an important research line that is motivated by our work presented here.

Acknowledgments and disclaimer

The following funding sources are gratefully acknowledged: Government of Catalonia (ICREA Acadèmia Prize to the second author and grant 2014 SGR 537), Spanish Government (project TIN2011-27076-C03-01 "CO-PRIVACY"), European Commission (projects FP7 "DwB", FP7-SMARTCITIES n. 609042 "PETRA", FP7 "Inter-Trust" and H2020 "CLARUS") and Templeton World Charity Foundation (grant TWCF0095/AB60 "CO-UTILITY"). The authors are with the UNESCO Chair in Data Privacy. The views in this paper are the authors' own and do not necessarily reflect the views of UNESCO or the Templeton World Charity Foundation.

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. M. Atzori, F. Bonchi, F. Giannotti and D. Pedreschi. Anonymity preserving pattern discovery. *VLDB Journal*, 17(4):703-727, 2008.
5. Australian Legislation. (a) Equal Opportunity Act – Victoria State, (b) Anti-Discrimination Act – Queensland State, 2008. `http://www.austlii.edu.au`
6. B. Berendt, S. Preibusch. Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. *Artif. Intell. Law* 22(2): 175-209, 2014.
7. R. Bhaskar, S. Laxman, A. Smith and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD 2010*, pp. 503-512. ACM, 2010.
8. L. Bonomi. Mining Frequent Patterns with Differential Privacy. *PVLDB* 6(12): 1422-1427, 2013
9. T. Calders, B. Goethals: Non-derivable itemset mining. *DMKD* 14(1): 171-206, 2007.
10. T. Calders and S. Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277-292, 2010.
11. 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.
12. T. Dalenius. The invasion of privacy problem and statistics production — an overview. *Statistik Tidskrift*, 12:213-225, 1974.
13. J. Domingo-Ferrer and V. Torra. Ordinal, continuous and heterogeneous k-anonymity through microaggregation. *Data Mining and Knowledge Discovery*, 11(2):195-212, 2005.
14. C. Dwork. Differential privacy. In *ICALP 2006*, LNCS 4052, pp. 112. Springer, 2006.
15. C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. Fairness through awareness. In *ITCS 2012*, pp. 214-226. ACM, 2012.
16. European Union Legislation. Directive 95/46/EC, 1995.
17. 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.
18. 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`
19. A. Friedman, R. Wolff and A. Schuster. Providing k-anonymity in data mining. *VLDB Journal*, 17(4):789-804, 2008.

20. A. Friedman and A. Schuster. Data mining with differential privacy. In *KDD 2010*, pp. 493-502. ACM, 2010.
21. 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.
22. J. Gehrke, M. Hay, E. Lui and R. Pass. Crowd-blending privacy. In *CRYPTO 2012*, pp. 479-496.
23. P.E. Greenwood, M.S. Nikulin. *A guide to chi-squared testing*. Wiley, 1996.
24. 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.
25. 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.
26. S. Hajian, A. Monreale, D. Pedreschi, J. Domingo-Ferrer and F. Giannotti. Injecting discrimination and privacy awareness into pattern discovery. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 360-369. IEEE Computer Society, 2012.
27. 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*, pp. 352-359. IEEE Computer Society, 2012.
28. S. Hajian, J. Domingo-Ferrer, and O. Farràs. Generalization-based privacy preservation and discrimination prevention in data publishing and mining. *Data Mining and Knowledge Discovery*, 28(5-6): 1158-1188, 2014.
29. M. Hay, V. Rastogi, G. Miklau and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of VLDB*, 3(1):1021-1032, 2010.
30. A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte-Nordholt, K. Spicer and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.
31. F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge Information Systems*, 33(1): 1-33, 2011.
32. F. Kamiran, T. Calders and M. Pechenizkiy. Discrimination aware decision tree learning. In *ICDM 2010*, pp. 869-874. IEEE, 2010.
33. F.Kamiran, A. Karim and X. Zhang. Decision theory for discrimination-aware classification. In *ICDM 2012*, pp. 924-929. IEEE, 2010.
34. F. Kamiran, I. Zliobaite, T. Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge Information Systems*, 35(3):613-644, 2013.
35. T. Kamishima, S. Akaho, H. Asoh, J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD*, LNCS 7524, pp. 35-50. Springer, 2012.
36. M. Kantarcioglu, J. Jin and C. Clifton. When do data mining results violate privacy? In *KDD 2004*, pp. 599-604. ACM, 2004.
37. J. Lee and C. Clifton. Differential identifiability. In *KDD 2012*, pp. 1041-1049. ACM, 2012.
38. N. Li, W. H. Qardaji, D. Su and J. Cao. PrivBasis: frequent itemset mining with differential privacy. *Proceedings of VLDB*, 5(11):1340-1351 (2012).
39. 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.
40. W. Li, J. Han and J. Pei. CMAR: accurate and efficient classification based on multiple class-association rules. In *ICDM 2001*, pp. 369-376. IEEE, 2001.
41. 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.
42. 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.
43. F. McSherry and K. Talwar. Mechanism design via differential privacy. In *FOCS 2007*, pp. 94-103. IEEE, 2007.
44. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proc. ICDT 99*, 1999.
45. D. Pedreschi, S. Ruggieri and F. Turini. Discrimination-aware data mining. In *KDD 2008*, pp. 560-568. ACM, 2008.

46. D. Pedreschi, S. Ruggieri and F. Turini. Measuring discrimination in socially-sensitive decision records. In *SDM 2009*, pp. 581-592. SIAM, 2009.
47. D. Pedreschi, S. Ruggieri and F. Turini. Integrating induction and deduction for finding evidence of discrimination. In *ICAIL 2009*, pp. 157-166. ACM, 2009.
48. 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.
49. 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.
50. P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010-1027, 2001.
51. J. Soria-Comas and J. Domingo-Ferrer. Sensitivity-independent differential privacy via prior knowledge refinement. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(6):855-876, 2012.
52. L. Sweeney. k-Anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557-570, 2002.
53. United States Congress, *US Equal Pay Act*, 1963. `http://archive.eeoc.gov/epa/anniversary/epa-40.html`
54. R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, C.Dwork: Learning Fair Representations. *ICML* (3) 2013: 325-333.
55. C. Zeng, J. F. Naughton, J.-Y. Cai. On differentially private frequent itemset mining. *PVLDB* 6(1): 25-36, 2012.
56. I. Zliobaite, F. Kamiran and T. Calders. Handling conditional discrimination. In *ICDM 2011*, pp. 992-1001. IEEE, 2011.