Department of Computer Science Technical Reports

Department of Computer Science

2010

Generalizations with Probability Distributions for Data Anonymization

Mehmet Ercan Nergiz
Purdue University

Suleyman Cetintas
Purdue University

Ahmet Erhan Nergiz
Purdue University

Ferit Akova
Purdue University

Report Number:
10-013

VLDB JOURNAL manuscript No.
(will be inserted by the editor)

Generalizations with Probability Distributions for Data Anonymization

Mehmet Ercan Nergiz · **Suleyman Cetintas** · **Ahmet Erhan Nergiz** · **Ferit Akova**

Received: date / Accepted: date

Abstract Anonymization-based privacy protection ensures that data cannot be traced to an individual. To this end, an anonymizer faces two challenges. First, the output anonymization must satisfy the underlying privacy definition and second, the anonymization needs to contain as much information as possible. One way to address the latter challenge has been to introduce flexibility in value generalizations by enlarging the output domain of the algorithms. This paper presents the most flexible way of releasing generalizations by introducing the family of PDF generalizations. In a PDF generalization, each generalized data value is empowered by probability distribution functions. Such distribution functions capture more statistics on data, thus enable the publisher to have better control over the trade-off between privacy and utilization. We evaluate the PDF approach for ℓ-diversity and δ-presence privacy models and show how to convert a ℓ-diverse or δ-present anonymization to a PDF generalization of higher utility without violating the privacy constraints. Data mining experiments on real world data show that information gained from PDFs increases the utility of the anonymizations.

Keywords Privacy · security · utility · PDF · generalization

M. E. Nergiz
Sabanci University
Tel.: +90 216 483 9000 - 2114
E-mail: ercann@sabanciuniv.edu

S. Cetintas
Purdue University
E-mail: scetinta@purdue.edu

A. E. Nergiz
Purdue University
E-mail: anergiz@purdue.edu

F. Akova
Purdue University
E-mail: akova@purdue.edu

1 Introduction

The tension between the value of publishing personal data and concern over individual privacy, is ever-increasing. Simply removing uniquely identifying information (SSN, name) from data is not sufficient to prevent identification because partially identifying information (quasi-identifiers or QI attributes such as age, gender ...) can still be mapped to individuals by using external knowledge [24].

Table anonymization is one method used to prevent identification. Many different privacy notions that make use of anonymization have been introduced for different adversary models. Among these models, k-Anonymity [20,21], ℓ-diversity [13], t-closeness [11], (α,k)-anonymity [26], anatomization [27] protects sensitive information while δ-presence [15] protects the existence of individuals in shared datasets. Privacy preserving algorithms working on these models applied different generalization techniques (replacing data values with more general values) over data cells to satisfy privacy constraints. *DGH based generalization* technique used in [23,6,9,4,18,16,3] requires user specified domain generalization hierarchies (DGHs or taxonomy trees) to carry out generalizations. DGHs are tree structures defined over each attribute domain and are used to specify to what value a given data value can generalize (in Figure 1, Peru can be generalized to America or *). Moreover, works in [2,10] assumed a total order between the values of each attribute domain and used interval based generalizations which are more flexible (using the total ordering in Figure 1, Peru can be generalized to a range such as [Canada,USA] since Canada $\leq$ Peru $\leq$ USA). Later in [16], *NDGH based generalizations* (generalization through Natural Domain Generalization Hierarchies) were introduced where data values can be replaced with any set of values from the associated domain to provide even more flexibility in generalizations (Peru can be generalized to a set such as $\{$Peru,USA$\}$). in Tables 2 and

1 we show example anonymizations of dataset T; T_d^*, T_i^*, and T_n^* that make use of DGH, interval, and NDGH generalizations respectively. In Section 2, we briefly explain the previously proposed methods and some of the privacy models that we will be referring to in future sections.

The motivation behind the current trend towards flexibility is achieving utility. As the released dataset is required to conform to the underlying privacy requirements, it is also expected to contain as much information as possible. Introducing further flexibility in generalizations has the potential to increase the utility of the anonymizations while still satisfying the underlying privacy metric. As an example, consider the ℓ-diversity privacy metric. ℓ-Diversity requires that the probability that an individual will be mapped to a sensitive value, given the set of his/her quasi-identifiers, is at most $\frac{1}{\ell}$. Given such a definition, in Tables 1 and 2 datasets T_d^*, T_i^*, or T_n^* are all valid 1.33-diverse anonymizations of dataset T. However, T_n^* that benefits from the flexible NDGH based generalizations contains more specific data values thus is more utilized than the others at the same privacy level ℓ.

However, even NDGH based generalization, being the most flexible solution offered so far, has still limitations in expressing generalized information. From the point of view of a third party, a data cell with value $\{$Peru, USA$\}$ is equally likely to be Peru or USA. However, in many cases, supplying the data cells with probability distribution information regarding how likely the data cell takes each specific value gives the publisher more control over the trade-off between privacy and utility. In this paper, we present a new generalization type, generalizations with probability distributions (*PDF generalizations*) in which NDGH generalizations are empowered with probability distribution functions. In a PDF anonymization, a data value, say 'Peru', can be generalized to a value {Peru:0.8,USA:0.2} which implies that data cell is Peru with 0.8 probability and USA with 0.2 probability. T_p^*, and T_{p2}^* in Table 6 are two examples for PDF anonymizations. Such generalizations can be used to better reflect the distribution of the original dataset. More importantly, PDF functions can be set according to different privacy constraints and thus produce anonymizations of variable utilization. In Section 3, we formally define PDF generalizations. The definition subsumes the previous generalization types. We next evaluate the effects of PDF generalizations on both utility and privacy. However, such evaluation of PDF generalizations makes sense only when we assume a privacy model such as k-anonymity, ℓ-diversity, t-closeness, or δ-presence. So our evaluation of utility and privacy will be separate for each privacy definition.

For privacy techniques that apply generalizations on quasi-identifiers (k-anonymity, ℓ-diversity, t-closeness, or δ-presence), the impact of more flexible generalization types (such as PDFs) on utilization can be observed explicitly through statistical measures. In this work, as a utility metric, we use the *KL cost metric* [8] which is based on the KL divergence distance between the tuple distributions in the released data and the original data. In Section 4, we formally state why KL cost metric is a good measure of utility in our domain. We also show how to set PDF distributions in a given anonymization in order to minimize the KL cost (thus maximize utility).

For privacy models in which the existence of individuals in the released datasets is already known by the adversaries (e.g., k-anonymity, ℓ-diversity, t-closeness, $\cdots$), the use of more flexible generalization types does not introduce any privacy violation [16,12]. More specifically, if an NDGH anonymization is ℓ-diverse, then any other PDF anonymization with the same grouping of tuples is also ℓ-diverse (e.g., in Tables 2 and 6, datasets T_d^*, T_i^*, T_n^*, T_p^*, and T_{p2}^* are all 1.33-diverse.) This implies that in terms of privacy, there is no shortcoming of using a more flexible generalization type such as PDF generalizations. Thus in these privacy models, utilization gained by PDFs can always be maximized and the output of any anonymization algorithm can trivially be post processed to return a more utilized PDF anonymization. However, there already exists privacy techniques that better utilize the quasi-identifier attributes while conforming to the same privacy requirements [27,29]. Thus, we discuss PDFs with respect to such models only to evaluate the relation between PDFs, KL cost, and utilization.

For those privacy metrics in which the existence of individuals in the released data is inherently sensitive, switching to a more flexible generalization type might result in privacy loss. As an example, consider the privacy metric δ-presence which ensures that the probability that an individual exists in the released anonymization (existence probability) is bounded by the δ parameter. Again consider the DGH anonymization T_d^* and the PDF anonymization T_p^* of the same dataset T in Tables 3, 4 and 6. Even though the groupings of the tuples are the same, the existence probability of a student is higher in T_p^* since tuples are more likely to represent students in T_p^*. Thus, for a better analysis of PDF generalization type in terms of utilization and privacy loss, in Section 5, we use the δ-presence privacy constraints. We show how to check for the δ-presence property when non-uniform distributions are used for data cells and show how to post process output of the optimal single dimensional δ-presence algorithm, SPALM [15], to make use of PDF generalizations. We present the PDF algorithm, PPALM, which is not optimal with respect to its domain but shows how PDFs can be used, even in a probabilistic adversary model, to increase utilization without violating the underlying privacy constraints.

In Section 6, we evaluate the effect of the new approach on the utilization of the output dataset. We present rule mining and classification results on real world data and show that extra information gained from PDFs can significantly

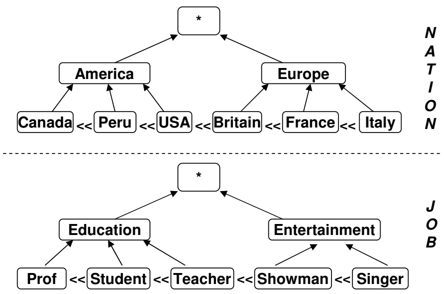

*
America
Europe
Canada << Peru << USA << Britain << France << Italy
N A T I O N
*
Education
Entertainment
Prof << Student << Teacher << Showman << Singer
J O B

Fig. 1 DGH structures for T_d^* and total ordering for T_i^* in Table 2

reduce rule mining and classification error on anonymized datasets without violating the privacy constraints of ℓ-diversity and δ-presence.

2 Preliminary

2.1 Background and Notation

Given a dataset (table) T with size R and number of dimensions C, $T[c][r]$ refers to the value of column c, row r of T with $c \in [1,C]$ and $r \in [1,R]$. $T[c]$ refers to the projection of column c on T.

Generalization of a single value v involves replacing v with a more general value v^* such that v^* implies v and possibly other values from the same domain. Many different ways to generalize a given value have been proposed. We now define DGH, interval and NDGH generalizations.

Definition 1 (Generalization Function) Given a data value v, a generalization function ψ returns the set of all generalizations of v.

We will name the DGH generalization function as ψ_d, interval generalization function as ψ_i, and the NDGH generalization function as ψ_n

Definition 2 (Table Generalization) Given two tables T and T^*, we say T^* is a generalization of T with respect to a given set of attributes QI if and only if $|T^*| = |T|$ and records in T, T^* can be ordered in such a way that for every possible index j, $T^*[i][j] \in \psi(T[i][j])$ if $i \in QI$ and $T^*[i][j] = T[i][j]$ otherwise. We say tuple $t = T[.][j]$ is linked to tuple $t^* = T^*[.][j]$ and write $(t \in T) \rightleftharpoons (t^* \in T^*)$.

In Table 2, all datasets are generalizations of table T given QI={Sex, Job, Nation}. For each table, a generalization function is defined according to the generalization type being used. According to DGH structures given in Figure 1; $\psi_d(\text{USA}) = \{\text{USA, America, *}\}$. T_d^* in Table 2 shows an example DGH based anonymization of T. According to the total ordering given in Figure 1; $\psi_i(\text{USA}) = \{[v_{min}, v_{max}] \mid v_{min} \in \{\text{Canada,Peru,USA}\} \wedge v_{max} \in \{\text{USA,Britain,France,Italy}\}\}$. T_i^* in Table 2 shows one interval based anonymization of T according to the same total ordering. $\psi_n(\text{USA}) = \{S_v \mid \{\text{USA}\} \subseteq S_v \subseteq \{\text{Canada, Peru, USA, Britain, France, Italy}\}\}$. NDGH based anonymizations are the most flexible anonymizations proposed so far. T_n^* in Table 2 shows one NDGH based anonymization of T. Tables T_d^*, T_i^*, and T_n^* have the same grouping of tuples however the generalization type being used enables T_n^* to contain more specific values compared to other tables.

Work in [12] presents three more generalization types, however NDGH still stands as the most flexible. Due to limited space, we do not include the discussion on these and assume NDGH as the baseline for the evaluations in coming sections unless noted otherwise.

While publishing person specific sensitive data, removing only uniquely identifying information (SSN, name) from data has been shown to be insufficient to prevent identification. An adversary can still map partially identifying information, *quasi-identifiers*, (age, gender, $\cdots$) to individuals by using external knowledge. E.g., in Table 1, the Salary attribute of private table T can be considered as *sensitive* attribute. The Sex, Job and Nation attributes are quasi-identifiers (QI_T) since they can be used to identify an individual in the public table PT. Releasing T as it is does not prevent linkage even though it does not contain any uniquely identifying information [24].

In most of the privacy models, the adversary is assumed to know the QI attributes about an individual from some public dataset or background knowledge. While releasing private datasets, we also face two different scenarios according to the adversary's knowledge on the existence of the individual:

- **Existential Certainty:** The adversary knows that the individual is in the private dataset and tries to learn the sensitive information about the individual in the private dataset.
- **Existential Uncertainty:** The adversary does not know the individual is or is not in the private dataset ($T \in PT$). Disclosure of existence or absence of an individual in the private dataset is a privacy violation. (In this case, there need not even be sensitive attributes in the private dataset; e.g., releasing data about diabetic patients.)

ℓ-Diversity [13] and its variants [11,19,26] provide privacy protection for the existential certainty model by limiting the linking of a sensitive value to a specific individual. In this paper, we will be covering the naive version of ℓ-diversity:

Definition 3 (Equivalence Class) The equivalence class of tuple t in dataset T^* is the set of all tuples in T^* with identical quasi-identifiers to t.

The equivalence class of row1 in the anonymized datasets is the set {row1, row2, row3, row4}.

Table 1 ℓ-Diversity Framework: Public and Private Datasets. Private dataset has the same size as the Public dataset.

PT:Public Dataset

Name	Sex	Job	Nation
Chris	M	Student	Canada
Luke	M	Student	USA
Darth	M	Student	USA
George	M	Prof.	USA
Padme	F	Showman	Italy
Laila	F	Singer	Italy
Kim	F	Singer	Italy
Ann	F	Teacher	Britain

T:Private Dataset

Sex	Job	Nation	Salary
M	Student	Canada	$\leq$ 50K
M	Student	USA	$\leq$ 50K
M	Student	USA	$\leq$ 50K
M	Prof.	USA	$>$ 50K
F	Showman	Italy	$>$ 50K
F	Singer	Italy	$>$ 50K
F	Singer	Italy	$>$ 50K
F	Teacher	Britain	$\leq$ 50K

Table 2 1.33 diverse generalizations of T in Table 1

T_d^*:DGH-anonymized Dataset

Sex	Job	Nation	Salary
M	*	America	$\leq$ 50K
M	*	America	$\leq$ 50K
M	*	America	$\leq$ 50K
M	*	America	$>$ 50K
F	*	Europe	$>$ 50K
F	*	Europe	$>$ 50K
F	*	Europe	$>$ 50K
F	*	Europe	$\leq$ 50K

T_i^*:Interval-anonymized Dataset

Sex	Job	Nation	Salary
M	[Pr,St]	[Ca,US]	$\leq$ 50K
M	[Pr,St]	[Ca,US]	$\leq$ 50K
M	[Pr,St]	[Ca,US]	$\leq$ 50K
M	[Pr,St]	[Ca,US]	$>$ 50K
F	[Te,Si]	[Br,It]	$>$ 50K
F	[Te,Si]	[Br,It]	$>$ 50K
F	[Te,Si]	[Br,It]	$>$ 50K
F	[Te,Si]	[Br,It]	$\leq$ 50K

T_n^*:NDGH-anonymized Dataset

Sex	Job	Nation	Salary
M	{Pr,St}	{Ca,US}	$\leq$ 50K
M	{Pr,St}	{Ca,US}	$\leq$ 50K
M	{Pr,St}	{Ca,US}	$\leq$ 50K
M	{Pr,St}	{Ca,US}	$>$ 50K
F	{Te,Sh,Si}	{Br,It}	$>$ 50K
F	{Te,Sh,Si}	{Br,It}	$>$ 50K
F	{Te,Sh,Si}	{Br,It}	$>$ 50K
F	{Te,Sh,Si}	{Br,It}	$\leq$ 50K

Table 3 δ-Presence Framework: Public and Private Datasets. Individuals in Private dataset is a subset of that of the Public dataset. Attribute "Ext" is not part of the public dataset but specifies which tuples are in the private dataset.

PT:Public Dataset

Name	Sex	Job	Nation	Ext
Chris	M	Student	Canada	1
Luke	M	Student	USA	1
Darth	M	Student	USA	1
George	M	Prof.	USA	1
Obi	M	Prof	Canada	0
Padme	F	Showman	Italy	1
Laila	F	Singer	Italy	1
Kim	F	Singer	Italy	1
Ann	F	Teacher	Britain	1
Marie	F	Teacher	Britain	0

T:Private Dataset

Sex	Job	Nation
M	Student	Canada
M	Student	USA
M	Student	USA
M	Prof.	USA
F	Showman	Italy
F	Singer	Italy
F	Singer	Italy
F	Teacher	Britain

Definition 4 **(ℓ-Diversity)** Let r_i be the frequency of the most frequent sensitive attribute in an equivalence class EC_i. An anonymization T^* is ℓ-diverse iff for all equivalence class $EC_i \in T^*$, we have $\frac{r_i}{|EC_i|} \leq \frac{1}{\ell}$.

Table 1 shows an example for the privacy risk in ℓ-diversity framework where the adversary knows PT and wants to link salary information to individuals. Clearly releasing T will result in sensitive info disclosure. (e.g., Showman Padme has salary >50K) All datasets given in Table 2, respect 1.33-diversity. The equivalence class of row1 in the anonymized datasets is the set {row1, row2, row3, row4}. Three out of four individuals in this class have a salary >50K. Thus by seeing one of the 1.33-diverse tables, an adversary can at best link Padme to the first four tuples and will only have 75% confidence that Padme has a salary >50K.

Table 4 PT_d^* is a generalization of PT and T_d^* is a (0,0.80)-present generalizations of T with respect to PT in Table 3. Both generalizations have the same generalization mapping.

PT_d^*:DGH-anonymized Dataset

Sex	Job	Nation	Ext
M	*	America	1
M	*	America	1
M	*	America	1
M	*	America	1
M	*	America	0
F	*	Europe	1
F	*	Europe	1
F	*	Europe	1
F	*	Europe	1
F	*	Europe	0

⇒

T_d^*:DGH-anonymized Dataset

Sex	Job	Nation
M	*	America
M	*	America
M	*	America
M	*	America
F	*	Europe
F	*	Europe
F	*	Europe
F	*	Europe

Another privacy metric, k-anonymity [20,21] provides (partial) privacy protection for the existential certainty model by limiting the linking of a record from a set of released records to a specific individual:

Definition 5 **(k-Anonymity)** A given table T^* is said to satisfy k-anonymity if and only if each combination of values in $T^*[QI_{T^*}]$ appears at least k times in T^*.

k-Anonymity fails to protect privacy when there is not enough diversity over the sensitive values within a given equivalence class. Thus, in our evaluations, we stick to ℓ-diversity as a representative of the class of existential certainty models. However, the technique of k-anonymization is still useful especially when the adversary also knows the underlying anonymization algorithm [25]. We will present a more detailed discussion on existential certainty models in Section 3.2.

It should be noted that the use of different generalization types does not violate ℓ-diversity definition. This makes it difficult to evaluate the privacy/utility relations for more flexible generalization types. Thus we need a probabilistic privacy notion: δ*-Presence* is defined in [15] for existential uncertainty model and introduces a δ metric to evaluate the probabilistic risk of identifying an individual in a private table based on publicly known data:

Definition 6 **(δ-Presence)** Given an external public table PT, and a private table T, we say that $\delta = \{\delta_{min}, \delta_{max}\}$-*presence* holds for a generalization T^* of T, if

$$\delta_{min} \leq \mathscr{P}(t \in T \mid T^*, PT) \leq \delta_{max} \qquad \forall\, t \in PT$$

In such a dataset, we say that a tuple $t \in PT$ is δ*-present* in T^*. Therefore, $\mathscr{P}(t \in T \mid T^*, PT)$ should be between δ_{min} and δ_{max} (the probability that the tuple exists in the private dataset should be between δ_{min} and δ_{max}).

Table 3 shows an example for the privacy risk in the δ-presence framework where the adversary knows PT and wants to identify the tuples in the private dataset T. (Attribute 'Ext' in Tables 3 and 4 is not part of the dataset but shown for ease in discussion. It basically states if the corresponding tuple exists in the private dataset. In other words, information in the private table is shown in the attribute 'Ext' of the public table.) Dataset T_d^* of Table 4 satisfies $(\delta_{min}, 0.8)$-presence for any $\delta_{min} \leq 0.8$. Out of 5 people {Chris, Luke, Darth, George, Obi}, 4 people are in T_d^*. So the probability that Chris (or any others) is in T_d^* is 0.8. This is also true for the females.

2.2 Related Work

Besides those mentioned in the previous sections, there has been other work on releasing more flexible generalizations. Work in [12] presents three generalization types; in SPS, the domain of a given attribute is partitioned into distinct groups and each group can be a generalization of a value inside the group. SPS is more flexible than interval based and DGH based generalizations. GSPS takes into account semantic relations among different values and improves SPS by enforcing constraints on the groups. Similarly, GOPS improves interval based generalizations by capturing semantic relationship among values in an attribute domain. As the generalization types mentioned in Section 2.1, all these generalizations imply a uniform distribution over the values in a given group, thus differ from PDF generalizations that provides arbitrary distributions.

Works in [27,29] propose an anatomization (also known as bucketization [29]) approach for existential certainty model. In anatomy, no QI attribute generalization is done and a distribution for sensitive values satisfying a given privacy standard is returned for groups of tuples. It can be shown that assuming a strict ℓ-diversity framework, anatomization achieves a higher level of utility at the same level of privacy. (As an example, in Table 5, we show a 1.33-diverse anatomization T_1^a which uses the same set of equivalence classes and which is clearly more utilized than anonymizations T_d^* and T_n^*.) We postpone a more detailed comparison of anatomization and PDF anonymization until Section 3.2.

Table 5 1.33-Diverse anatomizations of T in Table 1.

T_1^a

Sex	Job	Nation	Salary
M	Student	Canada	three ≤50K, one >50K
M	Student	USA	
M	Student	USA	
M	Prof.	USA	
F	Showman	Italy	one ≤50K, three >50K
F	Singer	Italy	
F	Singer	Italy	
F	Teacher	Britain	

T_2^a

Sex	Job	Nation	Salary
F	Teacher	Britain	two ≤50K, two >50K
F	Singer	Italy	
M	Student	USA	
M	Prof.	USA	
F	Showman	Italy	two ≤50K, two >50K
F	Singer	Italy	
M	Student	USA	
M	Student	Canada	

Work in [8] proposes publishing marginals (count tables on quasi identifiers) along with anonymized datasets in order to increase the utility of the anonymization while still preserving k-anonymity and ℓ-diversity. Their approach is different from ours in three ways. First, the work is based on existential certainty model and its extension to existential uncertainty without sacrificing its utility guarantees is not trivial (e.g., in Table 3, if an adversary learns that there are exactly two singers in T, he/she will trivially conclude Laila and Kim is in T. Certainly some degree of generalization on the marginals is required. But generalizations on DGHs do not help here. The adversary can derive the same conclusion if he/she knows that there are three entertainers in T. Furthermore, one needs to calculate the existence probabilities given the generalized T^* and the generalized marginals. Such an analysis poses challenges similar to the problem being addressed in this paper.) Second, even if marginals can be generalized, they still reflect the exact count information on the values of a given equivalence class, thus do not allow one to adjust them to fit into a privacy metric. Third, it is not easy to extract information from a released anonymization and the corresponding marginal tables, though as we shall see later in Section 4.1, PDF generalizations can be treated as fuzzy datasets and previously proposed reconstruction techniques can be used to run applications on PDFs.

3 PDF Generalizations

3.1 Formulation

A PDF generalization is basically a distribution defined over the associated domain:

Definition 7 (PDF Generalization Function) A PDF generalization function ψ_p is a function that, when given a value v from a categorical attribute domain $D = \{v_1, \cdots, v_n\}$, returns the set of all distributions f defined over D of the form, $\{f \mid f(v_i) \geq 0 \wedge f(v) > 0 \wedge \sum_{v_i \in D} f(v_i) = 1\}$.

We write a distribution function f in open form as $\{v_1 : f(v_1), \cdots, v_n : f(v_n)\}$ and do not write value entries with zero probability. T_p^* and T_{p2}^* in Table 6 shows different PDF anonymizations of T in Table 1 and 3. We assume for a generalized value v^* in a PDF generalization, $v^*.f$ returns the corresponding distribution function of v^* (e.g., $T_p^*[2][1]$={Pr: 0.25, St:0.75}, $T_p^*[2][1].f(Pr) = 0.25$). Semantically, a PDF generalization v^* represents an atomic value v_i with $v^*.f(v_i)$ probability.

PDF generalizations can also be defined over joint attributes:

Definition 8 (Joint PDF Generalization Function) Let D_i be the domain of categorical attribute a_i. A joint PDF generalization function ψ_p is a function that, when given a value v from $D_1 \times \cdots \times D_m = \{v_1, \cdots, v_n\}$ with $m > 1$, returns the set of all distributions f defined over $D_1 \times \cdots \times D_m$ of the form, $\{f \mid f(v_i) \geq 0 \wedge f(v) > 0 \wedge \sum_{v_i \in D} f(v_i) = 1\}$.

In Table 7, we show an example joint PDF defined over attributes Job and Nation. Even if joint PDF functions are defined over multiple attributes, from a technical point of view, they are no different than a non-joint PDF function defined over an attribute with a large domain. Thus, without loss of generality, we assume all PDFs are non-joint and are defined over single attributes unless otherwise noted. We discuss joint PDFs in more detail in Section 5.5.

NDGH (and other generalization types) implies uniform distribution on possible data values the generalized data stands for. PDF generalizations extend NDGH generalizations with probability distribution information. This makes the previous generalizations to be special cases of PDF generalizations (for a DGH value 'Europe', corresponding PDF value is {Br:0.33,Fr:0.33,It:0.33}). The PDF generalization T_p^* (or T_{p2}^*) obviously contains more information compared to the DGH generalization T_n^*. In coming sections, we investigate how the extra distribution information can be exploited for the sake of data utilization.

3.2 Use of PDF Generalization in Privacy Metrics

We emphasize that by proposing PDF generalizations, we do not define a privacy policy but merely provide an alternative

Table 6 PDF generalizations of T in Tables 1 and 3. Tables serve as examples for both ℓ-diversity and δ-presence. Attribute Salary is part of the dataset in the ℓ-diversity framework but not in the δ-presence framework.

T_p^*:PDF-anonymized Dataset

Sex	Job	Nation	Salary
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	$\leq$ 50K
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	$\leq$ 50K
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	$\leq$ 50K
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	$>$ 50K
F	{Te:0.25,Sh:0.25,Si:0.5}	{Br:0.25,It:0.75}	$>$ 50K
F	{Te:0.25,Sh:0.25,Si:0.5}	{Br:0.25,It:0.75}	$>$ 50K
F	{Te:0.25,Sh:0.25,Si:0.5}	{Br:0.25,It:0.75}	$>$ 50K
F	{Te:0.25,Sh:0.25,Si:0.5}	{Br:0.25,It:0.75}	$\leq$ 50K

T_{p2}^*:PDF-anonymized Dataset

Sex	Job	Nation	Salary
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	$\leq$ 50K
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	$\leq$ 50K
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	$\leq$ 50K
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	$>$ 50K
F	{Te:0.3,Sh:0.3,Si:0.4}	{Br:0.40,It:0.60}	$>$ 50K
F	{Te:0.3,Sh:0.3,Si:0.4}	{Br:0.40,It:0.60}	$>$ 50K
F	{Te:0.3,Sh:0.3,Si:0.4}	{Br:0.40,It:0.60}	$>$ 50K
F	{Te:0.3,Sh:0.3,Si:0.4}	{Br:0.40,It:0.60}	$\leq$ 50K

way for the data publisher to release anonymized data. So the proposed technique can only be evaluated with respect to a privacy model. In this section, we give an overview on how the use of PDF generalizations affects privacy with respect to previous proposed privacy definitions. Following sections give a more detailed discussion.

3.2.1 k-Anonymity, ℓ-Diversity, t-Closeness

These privacy models can *mostly* be considered as existential certainty models. As mentioned before, for such privacy models, different use of generalization types do not affect the amount of privacy provided (against the adversaries of existential certainty model mentioned in Section 2.1) given that the same grouping of tuples is used and the tuples in each equivalence class are indistinguishable [16, 12]. In T_d^*, T_i^*, T_n^* of Table 2, the first and the last 4 tuples are indistinguishable and the set of sensitive values is the same. This is also the case in T_p^* and T_{p2}^* in Table 6. Thus, all tables satisfy the ℓ-diversity constraint at the same privacy level ℓ.

However, as mentioned before, if the sole purpose is maximizing utility, there is a better alternative to PDF generalizations. In fact, assuming total existential certainty, releasing anatomization of datasets is a better approach than releasing PDF generalizations since anatomization better utilizes the QI attributes without disclosing sensitive attributes. (in Table 5, 1.33-diverse anatomization T_1^a which uses the same set of equivalence classes is more utilized than anonymizations T_p^* and T_{p2}^*.) However, recent developments on existential certainty models showed that anatomizations do not make previous approaches obsolete for two main reasons.

- Compared to anonymizations with the same grouping criteria, anatomizations are more prone to other forms of attacks on privacy. Works in [25, 28, 7] present various attacks that are empirically proved to be effective against anatomizations. It is possible but much harder to employ such attacks on generalizations.
- The grouping of tuples, even when releasing anatomizations, still affects utility [5]. A grouping where similar tuples are clustered with each other would produce a better utilized anatomization when there is correlation between sensitive and QI attributes. The ℓ-diversity anonymization algorithm optimizing against a utility metric is a candidate for such grouping. For example, T_1^a (with groups of similar tuples) is surely better utilized than T_2^a (e.g., correlation between sex and salary is preserved in T_1^a). The anatomization process can still benefit from the use of generalization in capturing the statistical closeness of tuples. An optimal (with respect to a cost metric) PDF anonymization could be used to better form groups of statistically close tuples.

PDFs can be viewed as an intermediate step between previous generalization approaches and anatomization. One can benefit from the flexibility of PDF generalizations to limit the privacy breaches caused by anatomizations while achieving a certain level of utility. However, in this paper we do not propose a grouping algorithm, thus we do not elaborate on this issue. In Section 4, we assume an existential certainty model ℓ-diversity to show the relation between the

KL cost, utility and PDFs. Doing so greatly eases discussion on utility since the level of privacy (e.g., ℓ) remains the same through anonymizations with the same set of equivalence classes.

3.2.2 δ-Presence

We can fully benefit from PDFs when we assume an existential uncertainty model such as δ-presence [15]. In this case, different PDF anonymizations (even if they use the same grouping) will provide different privacy levels [1]. Consider the PDF anonymizations T_p^* and T_{p2}^* in Table 6. Even though, both anonymizations have the same grouping, the existence probability for, say a Prof., is higher in T_{p2}^*. Also note that anatomization is no longer a good way of releasing deidentified datasets in a δ-presence framework. E.g., in Tables 3 and 5, if we release T^a, the adversary sees that a male student from Canada is in T^a. As there is no other person than Chris with these QI attributes, the adversary also concludes that Chris should be in T; $\mathscr{P}(Chris \in T \mid PT, T^a) = 1$.

In a δ-presence setting, selection of PDF probabilities directly affects not only the utility but also the privacy level. This property makes δ-presence a perfect candidate to evaluate the utility/privacy trade-off when using PDF generalizations. In Section 5, we show how to calculate existence probabilities given a PDF anonymizations and a public table. We also present an algorithm to create PDF anonymizations that respect δ-presence constraints.

4 PDF and Utilization

As mentioned before, the real advantage of using PDF generalizations is utility. However, quantifying the utility of a given anonymization is not an easy problem merely because utility depends on the target applications that will run on the anonymizations. Many utility metrics have been proposed to address various types of applications [6,2,16,8,4]. Work in [8] presents the first statistical utility metric that is based on the KL-Divergence between the original dataset and its anonymization. In this section, we formally define and use a slightly modified version of this metric, the KL cost metric, to quantify utility. We formally describe why KL cost is a good choice in our domain and also show how to set PDF distributions in a given anonymization to minimize the KL cost. This section discusses only utility. The reader can assume an existential certainty model such as ℓ-diversity throughout this section since no matter what PDFs are used, the privacy level do not change if the grouping of tuples remain the same. We postpone the discussion on privacy in existential certainty models until Section 5 when we evaluate PDFs in a δ-presence framework. Discussion on Section 5 benefits from the theorems on utility presented in this section.

We begin by describing the methodology we use to prepare the anonymous dataset for any application.

4.1 Data Reconstruction

Many of the anonymizations initially are not suitable for most data mining applications. The reason is that such applications assume non overlapping, distinct data cell values. However for many anonymizations, data value generalizations may imply or intersect with each other. (E.g., for DGH anonymizations, USA, America, *; all may occur at the same time as distinct values in a given attribute column.) So we need a process that will convert the heterogeneous (multi-level) anonymizations into homogeneous (leaf-level, atomic) datasets. For this purpose, we adapt the methodology proposed in [16] for PDF generalizations. Anonymized tables are first *reconstructed* before any data mining application is run:

Definition 9 (Reconstruction Function) Reconstruction function REC is a function that when given some multi-level PDF anonymized dataset T^* respecting a generalization function ψ, returns an atomic data set of the same size T^R (e.g., $REC(T^*) = T^R$), such that

$$\mathscr{P}(T^R[c][r] = v) = T^*[c][r].f(v)$$

From now on, we use the notation T_i^R for the reconstruction of T_i^*.

Informally the reconstruction function converts the generalized data entries to one of their atomic values probabilistically. Probabilistic conversion is done uniformly for DGH, interval and NDGH generalizations and according to PDF distributions for PDF generalizations. (For Table 6, $T_p^R[3][1]$ will be US with 0.75 probability. For Table 2, $T_d^R[3][1]$ will be US with 0.33 probability.) The reconstructed data will be suitable for all data mining applications.

4.2 The KL Cost Utility Metric

Since data mining applications run on reconstructed data, effectiveness of the application heavily depends on the similarity of the reconstructed data to the original data. Since an anonymization process does not add any noise, there is always a non-zero probability that the reconstructed data will be the same as the original data. We take this *matching probability* as the measure of utility in our domain and now formally derive it.

[1] For δ-presence, we can define the privacy level as the existence probabilities of public tuples

Let T^* be an anonymization with a set of equivalence classes $\{EC_1, \cdots, EC_\ell\}$. Also let each EC_i has the set of PDF distribution $F_i : \bigcup_{attribute\ a} f_a$. We denote the global set of PDF distributions as $GF : \{F_1, \cdots, F_\ell\}$. (E.g., in Table 6, for the first equivalence class EC_1 in T^*_p $F_1 : \{f_{Sex}, f_{Job}, f_{Nation}\}$ where $f_{Job}(Pr) = 0.25$.) Since each equivalence class is independent of each other, the matching probability of the anonymization T^* of T is the product of the matching probabilities for each equivalence class in T^*:

$$\mathscr{P}_{GF}(T^*) = \prod_{EC_i \in T^*} \mathscr{P}_{F_i}(EC_i)$$

So it is enough to derive the matching probability for each equivalence class EC independently.

Let c^i_a be the number of times an atomic data value v_i from D_a (domain of attribute a) appears in attribute a of EC. Note that for attribute a, the same distribution f_a is used in all tuples of EC. (E.g., if we assume we have the PDF anonymization T^*_p of T in Table 3 and the atomic value v_i is USA, then for the first equivalence class, $c^i_{Nation} = 3$ and $f_{Nation}(v_i) = 0.75$.) Then we have the following theorems:

Theorem 1 *The matching probability for EC is negatively correlated with the following equation defined over EC:*

$$KL(EC) = -\sum_{a=1}^{A} \sum_{v_i \in D_a} c^i_a \cdot \ln f_a(v_i) \tag{1}$$

to which we will refer as the KL cost *of EC.*

Proof See Appendix A

Equation 1 is nothing but $|EC|$ multiplied with the *negative cross-entropy* between the initial value distribution and value distribution of the given anonymization. This is not surprising. As discussed in [8], anonymizations maximizing the negative cross-entropy minimizes KL-divergence with the original value distribution. Statistically, such an anonymization better explains the original data.

4.3 Utility Optimal PDF

We now derive the optimal distribution for a fixed set of equivalence classes in a given anonymization that will minimize KL cost. As mentioned above, we can analyze each equivalence class EC independently:

Theorem 2 *The distribution function* $F : \bigcup_a f_a$ *defined as*

$$f_a(v_i) = \frac{c^i_a}{|EC|} \tag{2}$$

for each value $v_i \in D_a$, minimizes the KL cost, thus maximizes the matching probability for EC.

Proof See Appendix A

Definition 10 (Utility Optimal) A PDF generalization T^* is utility optimal with respect to T and a given set of equivalent classes if probability distribution function for every equivalence class in T^* is defined as in Equation 2.

This means that the utility optimal PDF probability for a data value $v \in D_a$ in an equivalence class EC is the number of times v appears in attribute a of EC divided by the size of EC. (e.g., weight of v in EC) By definition, utility optimal anonymizations maximize the matching probability (e.g., In Tables 3 and 6, T^*_p is utility optimal with respect to T and the corresponding tuple grouping. The first four tuples contain 1 professor and 3 students, so $f_{job} = \{Pr : 0.25, St : 0.75\}$.)

The next theorem states that matching probability monotonically decreases as each f_a gets far away from the utility optimal distribution;

Theorem 3 *For an equivalence class EC, let $F^{(o)} : \bigcup_a f^{(o)}{}_a$ be the utility optimal distribution and let $F^{(1)}$ and $F^{(2)}$ be two other distribution functions with $|f^{(1)}{}_a(v_i) - f^{(o)}{}_a(v_i)| \leq |f^{(2)}{}_a(v_i) - f^{(o)}{}_a(v_i)|$ for all attribute a and for all $v_i \in D_a$ then $\mathscr{P}_{F^{(1)}} \geq \mathscr{P}_{F^{(2)}}$.*

Proof See Appendix A

Theorem 3 gives a way to compare PDF generalizations in terms of utilization. In Tables 2 and 6 matching probability for T^*_{p2} is bigger than that of T^*_n. This is due to the fact that distributions in T^*_{p2} are closer to those of the utility optimal T^*_p (for the first equivalence class, $f_{Job}(Pr)$ is 0.25 for T^*_p, 0.4 for T^*_{p2} and 0.5 for T^*_n.) In Section 5, we use this observation in Theorem 3 to increase utilization in a given anonymization.

Since all other generalization types assume uniform distribution on atomic values of a generalized value, (no matter what the underlying original frequencies of the atomic values are) it is clear that utility optimal PDF generalizations simulate original datasets at least as good as the other generalization types do.

As the reconstructed data becomes similar to the original data, all applications run on reconstructed data increase in accuracy. In Appendix B, we observe the effects of utility optimal PDF generalizations on data mining applications, rule mining and classification, by looking at the example datasets in Table 2. Since the NDGH approach is the most flexible one among previous generalization types, the comparison is carried out between datasets T^*_n and T^*_p from Tables 2 and 6.

5 PDF and Privacy:δ-Presence

In this section, we switch to a probabilistic existential uncertainty model, δ-presence [15]. We focus on how privacy is affected in a δ-presence environment when PDF generalizations are used. We introduce a new δ-presence algorithm WPALM that injects utilization into the datasets without violating the privacy constraints. We then improve WPALM in terms of efficiency with a second algorithm, PPALM.

5.1 PDF δ-Presence Algorithms:WPALM & PPALM

In this section, we empower the previously proposed δ-presence algorithm, SPALM [15], to make use of PDF generalizations. SPALM, when given a public table PT and private table T, returns a DGH anonymization T^* of T which is δ-present with respect to PT and T. The PDF algorithms presented in this section, *WPALM* and *PPALM*, both attempt to increase the utilization of the output anonymization of SPALM further without violating δ-presence privacy constraints (so no privacy loss is encountered). The difference between two PDF algorithms is covered in the next subsections, the discussion in this subsection applies for both of the algorithms. We show experimentally in Section 6 that outputs of WPALM and PPALM are better utilized with respect to KL cost and data mining applications.

Both WPALM and PPALM operate on the SPALM output, which is already δ-present with respect to input datasets. Additionally, both WPALM and PPALM shift PDFs within the output towards utility optimal distribution as long as the δ-presence property is preserved. The resulting anonymization is obviously not optimal with respect to space of all possible PDF outputs, but is statistically at least as good as the SPALM output.

For each equivalence class EC of the SPALM output, WPALM and PPALM shift the value distributions (fs), from uniformity towards utility-optimal distribution step by step. The maximum number of steps is set by the input variable *mxs*. In other words, the distribution of EC becomes utility optimal in *mxs* steps, if neither of the intermediate distributions violates δ-presence. For value v_i of attribute a in EC, let $f^{(u)}$ be the initial (uniform) distribution function. E.g., given that v^* is the generalized value used in EC initially, $f^{(u)}(v_i) = \frac{1}{|\{v \mid v^* \in \psi_d(v)\}|}$ if $v^* \in \psi_d(v_i)$ and zero otherwise. Let $f^{(o)}$ be the utility optimal distribution function (e.g., $f^{(o)}(v_i) = \frac{c_a^i}{|EC|}$). Then the distribution function $f^{(k)}$ being tried in step k is defined as

$$f^{(k)}(v_i) = f^{(u)}(v_i) + k \cdot \frac{f^{(o)}(v_i) - f^{(u)}(v_i)}{mxs} \tag{3}$$

In Figure 2, $f^{(u)}$('Europe')={Italy:.33,Britain:.33,France: .33}, $f^{(o)}$('Europe')={Italy:.75,Britain:.25,France:0}. For $mxs = 3$, $f^{(1)}$('Europe')= {Italy:.47,Britain:.3,France:.22}, and $f^{(2)}$ ('Europe')= {Italy:.61,Britain:.27,France:.11}. By Theorem 3, outputs with $f^{(i)}$ distribution is better utilized than those of with $f^{(j)}$ if $i > j$. So each shift injects utilization into the anonymization.

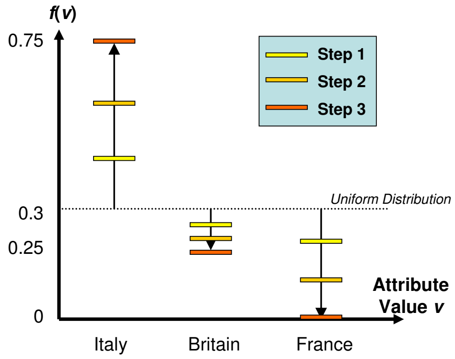


Fig. 2 Shifting the uniform distribution (inherited in data value 'Europe') in T_d^* of Table 2 to the utility optimal distribution in three steps.

In Algorithm 1, we show the pseudocode for the algorithms WPALM and PPALM. The algorithm, in line 2 calls SPALM to get the optimal DGH δ-present anonymization of PT, PT^* (note that $T^* \subset PT^*$). In lines 4-10, the distribution of each equivalence class of the anonymization are shifted towards the utility optimal distribution as long as the presence property is not violated.

The boolean function *isPresent* is called in line 8 to check for the presence property. However, checking for the presence property for non-uniform PDFs is not as simple as in uniform PDFs (e.g, DGH, interval, NDGH generalizations). The next two sections cover how the checking process is carried out for PDF generalizations. WPALM and PPALM differ in their implementation of *isPresent*.

Algorithm 1 WPALM and PPALM

```
Require: public table $PT$; private table $T$, parameter $\delta$, maximum number of shift steps $mxs$.
Ensure: return a PDF generalization of $T$ respecting $(\delta_{min}, \delta_{max})$-presence with KL cost at most that of the optimal full domain generalization.
1:  insert "Ext" attribute into $PT$ according to $T$ as in Table 3.
2:  run SPALM on $PT$, $T$, and $\delta$, let $PT^*$ be the output anonymization of $PT$
3:  for all equivalence class $EC$ in $PT^*$ do
4:      $k = 1$.
5:      while $k \leq mxs$ do
6:          update the distribution function of values as $f^{(k)}$ given in Equation 3.
7:          if !isPresent($PT^*$,$PT$,$\delta_{min}$,$\delta_{max}$) then
8:              undo last updates on $EC$.
9:              go to line 5 to process the next equivalence class.
10:         $k++$.
```

5.2 Checking for the δ-Presence Property

We show in this section how to check if a given PDF anonymization T^* of T is δ-present with respect to a public dataset PT. We first recall how it is done for uniform distributions.

5.2.1 Checking for Uniform Distributions

For a public dataset PT, private dataset T, and its non-overlapping anonymization T^* with some generalization mapping μ, let PT^* be the anonymization of PT with the same mapping μ. (see Table 4). For uniform and non-overlapping generalizations, the existence probabilities can simply be calculated by working on the anonymization PT^*:

Definition 11 (Projected Set) A set of tuples $J \subset PT$ is a projected set of PT if their generalizations form an equivalence class in PT^*. We denote tuple j^* to be their generalization in PT^* (or in T^*).

In Tables 3 and 4, given PT_d^* and PT, {Chris,Luke,Darth, George,Obi} is a projected set with $j^* =$ <M,*,America>. In non-overlapping generalizations, the projected sets do not intersect.

Let J be a projected set in PT and let $n_\sigma = |\{tuple\ j_i \in J \mid j_i[Ext] = \sigma\}|$ then the existence probability for any $j_i \in J$ is given by

$$\mathscr{P}(j_i \in T \mid T^*, PT) = \frac{n_1}{n_0 + n_1}$$

In other words, the existence probability for a tuple is the number of tuples with Ext=1 over the total number of tuples in the equivalence class. This is because, given T^* and PT, among $n_0 + n_1 = |J|$ many tuples, n_1 of them exists in T. (Note that n_1 is the cardinality of j^* in T^*.) Since every tuple is equally likely to appear in the private dataset, the existence probabilities are the same for any tuple of the same projected set.

5.2.2 Checking for Arbitrary Distributions

When we introduce non-uniform probability distributions, the existence probabilities will be different for each tuple in a given projected set. An adversary still knows n_1 tuples are selected among $|J|$ tuples but the likelihood of each tuple is different due to the distribution of the outcome:

Definition 12 (Likelihood Probability) The likelihood probability for a tuple $j \in J$, written as $p_j^{j^*}$, is the probability that $j \in J$ and $j^* \in T^*$ are the same entities. $p_j^{j^*} = \mathscr{P}((j \in PT) \rightleftharpoons (j^* \in T^*)) = \prod_i j^*[i].f(j[i])$.

Given PT of Table 3 and T_p^* of Table 6 $J =$ {Chris,Luke, Darth,George,Obi} is a projected set with $j^* =$ <M, {Pr:0.25, St:0.75}, {Ca:0.25,US:0.75}>. The likelihood probability for Chris (<M,St,US>) is $p_{Chris}^{j^*} = 1 \cdot 0.75 \cdot 0.25 = \frac{3}{16}$.

Definition 13 (Likelihood Set and Existence Set) Let the set of tuples $J = \{j_1, \cdots j_n\}$ be a projected set in PT with respect to some anonymization T^*. The likelihood set for J is defined as $P = \{p_1, \cdots, p_n\}$ where $p_i = p_{j_i}^{j^*}$. Given a set S of likelihoods, we use the notation P_S for the product of likelihoods in S ($P_S = \prod_{p \in S} p$).

The existence set for J contains the existence probabilities of all tuples in J and thus is defined as $EX = \{ex_1, \cdots, ex_n\}$ where $ex_i = \mathscr{P}(j_i \in T \mid T^*, PT)$.

The likelihood set for J in the example above is $P = \{\frac{3}{16}, \frac{9}{16}, \frac{9}{16}, \frac{3}{16}, \frac{1}{16}\}$.

It is very easy and efficient to create the likelihood set for a given projected set. Note, however, that we are interested in the existence set. Given the likelihood set and the number of existent tuples n_1, each element in the existence set can be calculated one by one. The existence probability for any tuple $j_k \in J$ takes the following conditional form:

$$\begin{aligned} ex_k = \mathscr{P}(j_k \in T \mid T^*, PT) &= \frac{\mathscr{P}(j_k \in T \bigwedge PT \mid T^*)}{\mathscr{P}(PT \mid T^*)} \\ = \frac{\sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p_k \in S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S} &= \frac{p_k \cdot \sum\limits_{\substack{S \subset P \wedge |S| = n_1 - 1 \wedge \\ p_k \notin S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S} \end{aligned} \tag{4}$$

In the denominator, set S is a variable traversing all possible subsets of P having size n_1. More precisely, the denominator sums up the probabilities of every possible worlds (e.g., every possible sets of existing tuples) while numerator sums up only those in which j_k is existing (e.g., $j_k \in T$).

Following the above example, the existence probability for Chris is calculated as

$$\begin{aligned} ex_{Chris} &= \mathscr{P}(Chris \in T \mid T_p^*, PT) \\ &= \frac{\frac{3}{16}(\frac{9}{16}\frac{9}{16}\frac{3}{16} + \frac{81}{16^3} + \frac{27}{16^3} + \frac{27}{16^3})}{\frac{729}{16^4} + \frac{243}{16^4} + \frac{243}{16^4} + \frac{81}{16^4} + \frac{81}{16^4}} \\ &= \frac{14}{17} = 0.82 \end{aligned}$$

Similarly, the existence probability for Luke and Darth is 0.94, for George 0.82 and for Obi 0.47 ($EX = \{0.82, 0.94, 0.94, 0.82, 0.47\}$ implying this equivalence class respects (0.47, 0.94)-presence). Note that the existence probabilities for the tuples of the same projected set are not necessarily the same when releasing PDFs.

Algorithm 2 isPresent for WPALM

```
Require: public table $PT$ with attribute Ext; one anonymization of $PT$, $PT^*$; parameter $\delta$.
Ensure: return true iff $PT^*$ satisfies $(\delta_{min}, \delta_{max})$-presence.
for all projected set $J \in PT$ with respect to $PT^*$ do
    for all tuples $j \in J$ do
        calculate existence probability $ex$ for $j$ as given in Equation 4.
        if $ex \leq \delta_{min}$ then
            return false
        if $ex \geq \delta_{max}$ then
            return false
return true;
```

Algorithm 2 shows the implementation of the boolean function *isPresent* for WPALM that makes use of Equation 4 to check for the presence property. Basically, the algorithm calculates the existence probabilities for all tuples and returns true iff all existence probabilities lie within the boundaries of presence constraints.

The minimum and the maximum existence probability in all of the existence sets of PT is sufficient to check for the presence property. However, calculating the *exact* existence probabilities by using Equation 4 is very costly. Many possible groupings of likelihood probabilities need to be multiplied. For a projected set of size $m = n_0 + n_1$ with n_1 present tuples, calculating the existence probability of one tuple will require $\binom{m}{n_1}$ summations on the denominator. For even moderate values of m (and with $n_1 \approx \frac{m}{2}$), calculation of Equation 4 is infeasible even if the likelihood probabilities for the tuples fit into main memory. Next subsection shows how to weaken this problem by presenting an alternative algorithm.

5.3 Speeding Up the Checking Process

In this section, we improve the δ-presence checking process in terms of efficiency and introduce the algorithm PPALM that makes use of the speed up process.

Checking for the δ-presence property can be speed up by the following observations:

1. The existence probability of only two tuples needs to be calculated for checking.
2. The calculation of *exact* existence properties is not needed. Finding upper and lower bounds on the maximum and minimum existence probabilities also works given the bounds are tight enough.

We first show the correctness of item 1. To check for the δ-presence property, it is sufficient to calculate just the maximum and minimum existence probabilities in a given projected set. Theorem 4 states that tuples with maximum and minimum likelihoods have maximum and minimum existence probabilities and it is sufficient to check only these two boundary tuples for δ-presence property.

Theorem 4 *Given a likelihood set $P = \{p^{min}, p^{max}, p_1, \cdots, p_m\}$ and the number of present tuples n_1, let $p^{min} \leq p_i \leq p^{max}$ for $i \in [1-m]$. If $ex^{min} \geq \delta_{min}$ and $ex^{max} \leq \delta_{max}$ then $\delta_{min} \leq ex \leq \delta_{max}$ for any $ex \in EX$.*

Proof See Appendix C

Following the example above, Luke and Obi have the maximum and minimum likelihoods $(\frac{9}{16}, \frac{1}{16})$ respectively. They also have the maximum and minimum existence probability (0.94,0.47). So it is sufficient to calculate the probabilities for Luke and Obi.[2]

We next show the correctness of item 2. The checking process can be fastened by calculating boundaries on the existence probabilities rather than calculating the exact probabilities. The *lower* and *upper bound likelihood sets*, defined below, are used to bound the minimum and maximum existence probabilities:

Definition 14 Given the number of present tuples n_1, let $P = \{p^{min}, p^{max}, p_1, \cdots, p_m\}$ be a likelihood set with $p^{min} < p_i < p^{max}$ for all $i \in [1-m]$. We say $P^{\downarrow} = \{(p^{\downarrow})^{min}, (p^{\downarrow})^{max}, p_1^{\downarrow}, \cdots, p_m^{\downarrow}\}$ is a *lower bound likelihood set* of P if $(p^{\downarrow})^{min} = p^{min}$, $(p^{\downarrow})^{max} = p^{max}$, and $p_i^{\downarrow} = p^{max}$ for all $i \in [1-m]$.

Similarly $P^{\uparrow} = \{(p^{\uparrow})^{min}, (p^{\uparrow})^{max}, p_1^{\uparrow}, \cdots, p_m^{\uparrow}\}$ is an *upper bound likelihood set* of P if $(p^{\uparrow})^{min} = p^{min}$, $(p^{\uparrow})^{max} = p^{max}$, and $p_i^{\uparrow} = p_{min}$ for all $i \in [1-m]$.

Following the example above, the lower bound set of $P = \{\frac{3}{16}, \frac{9}{16}, \frac{9}{16}, \frac{3}{16}, \frac{1}{16}\}$ is $P^{\downarrow} = \{\frac{9}{16}, \frac{9}{16}, \frac{9}{16}, \frac{9}{16}, \frac{1}{16}\}$ and the upper bound set is $P^{\uparrow} = \{\frac{1}{16}, \frac{9}{16}, \frac{1}{16}, \frac{1}{16}, \frac{1}{16}\}$.

The following theorem states that the lower and upper bound likelihood sets can be used to check if the original likelihood set satisfies δ-presence. If the lower and upper boundary sets satisfy the presence property over one of the δ constraint, so does the original likelihood set. However, the reverse is not true.

Theorem 5 *Given the number of present tuples n_1, the likelihood sets $P, P^{\downarrow}, P^{\uparrow}$, and their corresponding existence sets $EX, EX^{\downarrow}, EX^{\uparrow}$;*
$\delta_{min} \leq ex \leq \delta_{max}$ for any $ex \in EX$ if $\delta_{min} \leq (ex^{\downarrow})^{min}$ and $(ex^{\uparrow})^{max} \leq \delta_{max}$.

Proof See Appendix D.

Following the example above, corresponding existence sets $EX^{\downarrow} = \{0.92, 0.92, 0.92, 0.92, \mathbf{0.31}\}$, $EX^{\uparrow} = \{0.75, \mathbf{0.97}, 0.75, 0.75, 0.75\}$. This implies that the original likelihood set P (and the original projected set) satisfies $(0.31, 0.97)$-presence. We found earlier that in fact P satisfies (0.47, 0.94)-presence.

[2] If $\delta_{min} = 0$ or $\delta_{max} = 1$, only one tuple needs to be checked as opposed to two.

The advantage of working on the boundary sets is that checking for the presence property is much more efficient for the boundary sets due to element repetition. Equation 4 takes the following form for the existence probability $(ex^{\downarrow})^{min}$:

$$(ex^{\downarrow})^{min} = \frac{\binom{m+1}{n_1-1}\cdot(p^{\downarrow})^{min}\cdot((p^{\downarrow})^{max})^{n_1-1}}{\binom{m+1}{n_1-1}\cdot(p^{\downarrow})^{min}\cdot((p^{\downarrow})^{max})^{n_1-1}+\binom{m+1}{n_1}\cdot((p^{\downarrow})^{max})^{n_1}} \qquad (5)$$

Equation 5 does not require addition of many likelihood products so it is much faster to compute compared to Equation 4. However, the boundary sets are useful if the lower and upper bounds on the existence probabilities are tight enough. The more each likelihood probability is shifted in the boundary sets, the more existence probabilities deviate from the original probability. Observing this, one can use a trade-off between efficiency and precision on the calculation of existence probabilities. For the lower boundary set, instead of shifting likelihoods up until the max likelihood, we calculate a median as $p^{med} = \frac{p^{min}+p^{max}}{2}$ and shift likelihoods smaller than median to median and the rest to the maximum likelihood. This would give tighter bounds but would be slower to compute as we deal with three different likelihoods instead of two.

Algorithm 3 isPresent for PPALM

```
Require: public table $PT$ with attribute Ext; one anonymization of $PT$, $PT^*$; parameter $\delta$.
Ensure: return true iff $N^*$ satisfies $(\delta_{min}, \delta_{max})$-presence.
for all projected set $J \in PT$ do
    let $n_1$ be the number of tuples in $J$ with $Ext = 1$
    create the likelihood set $P$ for $J$
    create lower and upper bound likelihood sets $P^{\downarrow}, P^{\uparrow}$ of $P$.
    calculate existence probability $(ex^{\downarrow})^{min}$ [$(ex^{\uparrow})^{max}$] for the minimum [maximum] likelihood in $P^{\downarrow}$ [$P^{\uparrow}$] with respect to $n_1$
    if $(ex^{\downarrow})^{min} \leq \delta_{min}$ then
        return false
    if $(ex^{\uparrow})^{max} \geq \delta_{max}$ then
        return false
return true;
```

Algorithm 3 shows the implementation of the boolean function *isPresent* for PPALM that makes use of the speed up process. Basically, the algorithm creates upper and lower bound likelihood sets for the likelihood sets of each projected set in PT with respect to the anonymization and returns true if and only if bound sets satisfy the *partial* presence property.

In Section 6, we show experimentally that both PPALM and WPALM better utilize the anonymizations compared to SPALM without violating the presence constraints. We also compare WPALM and PPALM in terms of efficiency and utilization. Experiments on real world data show that speed up techniques given in this section work with great precision and efficiency in practice.

5.4 Efficiency Analysis

If no pruning can be performed[3], the worst-case complexity of SPALM is $O(\prod_{i=1}^{n} H_i \cdot |PT|)$ where n is the dimensionality of PT and H_i is the height of the ith DGH structure. In addition to this, WPALM and PPALM does a post processing. Post processing checks δ-presence $mxs \cdot G$ many times in the worst case where G is the number of equivalence classes PT^*. The time complexity of the checking process differs for WPALM and PPALM.

To check for δ-presence, for each equivalence class, WPALM iterates through all possible combinations of set of existing tuples. If we assume the sizes of the equivalence classes are uniform, the size of an equivalence class becomes $\frac{|PT|}{G}$. In the worst case, half of the tuples will be existing. Thus, assuming the unit of operation is one multiplication of all likelihoods (e.g., one summand of the denominator in Equation 4), worst case complexity is $\omega(2^{\frac{|PT|}{2G}})$. The total complexity of post processing is then $\omega(mxs \cdot G \cdot 2^{\frac{|PT|}{2G}})$. Even if this is an in-memory calculations, $|PT|$ is much bigger than G, making the post processing infeasible.

For PPALM, checking for δ-presence through Equation 5 takes constant time as we only have two summands. Thus complexity is $O(mxs \cdot G)$. However, we also need to calculate binomials of big numbers in this case and the time needed to calculate binomials dominates multiplication of the likelihoods.

5.5 Joint PDF Anonymizations

Both WPALM and PPALM return a PDF anonymization where each data cell is defined in the associated attribute domain. While doing so is completely privacy preserving, the joint distribution of attributes cannot be captured by the parties that can access the PDF anonymization (e.g., T_p^* of Table 6) but not the public table (e.g., PT of Table 3). Data publisher can increase utility even more by working on the joint distributions of attributes and injecting the information in the public table into the anonymization.

Consider the fifth tuple in T_p^* of Table 6. T_p^* itself implies that the probability that the fifth tuple is a <Te,It> (an Italian teacher) is $\frac{3}{16}$. However, there is no such tuple in the public table PT of Table 3. In other words, T_p^*, PT together implies that the fifth tuple cannot be <Te,It>. Parties that do not have access to PT cannot gain such information.

It is possible for the data releaser to capture joint distributions by combining attributes and return another PDF $T_p^{(j)}$ on the joint distributions. This can be done by post processing the output PDF anonymization $T_p^{(o)}$. Let $\ell_t^{(o)}$ and $\ell_t^{(j)}$

[3] In practice, pruning eliminates majority of the search space.

Table 7

T_{p3}^{*}:*Joint PDF Anonymization*

Sex	Job Nation
M	{(St Ca):$\frac{3}{16}$,(St US):$\frac{9}{16}$,(Pr US):$\frac{1}{16}$,(Pr Ca):$\frac{3}{16}$}
M	{(St Ca):$\frac{3}{16}$,(St US):$\frac{9}{16}$,(Pr US):$\frac{1}{16}$,(Pr Ca):$\frac{3}{16}$}
M	{(St Ca):$\frac{3}{16}$,(St US):$\frac{9}{16}$,(Pr US):$\frac{1}{16}$,(Pr Ca):$\frac{3}{16}$}
M	{(St Ca):$\frac{3}{16}$,(St US):$\frac{9}{16}$,(Pr US):$\frac{1}{16}$,(Pr Ca):$\frac{3}{16}$}
F	{(Sh It):$\frac{3}{10}$,(Si It):$\frac{6}{10}$,(Te Br):$\frac{1}{10}$}
F	{(Sh It):$\frac{3}{10}$,(Si It):$\frac{6}{10}$,(Te Br):$\frac{1}{10}$}
F	{(Sh It):$\frac{3}{10}$,(Si It):$\frac{6}{10}$,(Te Br):$\frac{1}{10}$}
F	{(Sh It):$\frac{3}{10}$,(Si It):$\frac{6}{10}$,(Te Br):$\frac{1}{10}$}

be the likelihood of tuple $t \in PT$ in $T_p^{(o)}$ and $T_p^{(j)}$ respectively. Let J be the projected set t belongs to. Then, we have $\ell_t^{(j)} = \frac{\ell_t^{(o)}}{\sum_{t_i \in J} \ell_{t_i}^{(o)}}$.

As an example, we show, in Table 7, another PDF anonymization T_{p3}^{*} that returns conditional (conditioned on the public table) joint distributions. For the second equivalence class T_p^{*}, we have three distinct tuples with likelihoods $\frac{3}{16}$, $\frac{6}{16}$, and $\frac{1}{16}$. The likelihood of <Te,Br> (thus the PDF probability) is then $\frac{1}{16}/\frac{10}{16} = \frac{1}{10}$. It can easily be shown that the existence probabilities remain the same for tuples of the public dataset. We name this algorithm JPALM (Joint PDF Algorithm) and evaluate its effectiveness in Section 6

Instead of post processing a non-joint PDF, one can also try to find the joint PDF probabilities directly. Since the desired output is still a PDF anonymization, the methodology for such a process would not be any different than the methodology given in the previous sections. The resulting anonymization is still a PDF anonymization in fewer dimensions. More specifically, if the intention is to create a fully joint PDF from a DGH anonymization T_d^{*} with attributes $a_1, \cdots, a_m$ from domains $D_1, \cdots, D_m$, we start with another anonymization T_j^{*} with a single attribute from domain $D_1 \times .... \times D_m$. We can apply the same techniques given in previous sections to create a joint PDF from T_j^{*}. However, it should be noted that working in one dimension does not make the problem any easier. Equations 4 and 5 are independent of the dimensionality meaning finding the existence probabilities in one dimension is as hard as finding them in many dimensions.

Joint PDFs are superior to attribute based PDFs when we have all of the following conditions satisfied:

- The public table PT is not known to all parties that want to make use of the released data and such disclosure of PT is not a privacy concern. We will further elaborate on this issue in Section 5.6.2. [4]
- The public table is known by the data releaser. There are applications of δ-presence where only statistical info on PT (rather than the PT itself) is available to the data publisher [17].
- The joint PDF anonymization can be stored and managed effectively. This is an issue since the joint domain of attributes may be quite large to effectively store a joint PDF anonymization.

Algorithm 4 MPALM

```
Require: public table $PT$; private table $T$, parameter $\delta$, mSPALM is
    minimality attack resistant DGH algorithm.
Ensure: return a PDF generalization of $T$ respecting $(\delta_{min}, \delta_{max})$-
    presence that resists minimality attacks.
1: insert "Ext" attribute into $PT$ according to $T$ as in Table 3.
2: run mSPALM on $PT$, $T$, and $\delta$, let $PT_s^*$ be the output anonymiza-
   tion of $PT$
3: for all equivalence class $EC$ in $PT_s^*$ do
4:     Find the utility optimal distribution with respect to $EC$ and
       update the distribution function of values in $EC$ as the utility
       optimal distribution.
```

5.6 Minimality Attack Resistant PDF Generalizations

5.6.1 Algorithm MPALM

Up until now, we assumed that the adversary has access only to the public table PT. However, it has been shown in [25, 28] that an adversary that also knows the underlying anonymization algorithm can perform *minimality attacks* and learn more by exploiting the utility optimality (see Appendix E). We present, in Algorithm 4, a new PDF algorithm, MPALM and now prove that MPALM is minimality attack resistant.

Theorem 6 *MPALM, given in Algorithm 1, is minimality attack resistant.*

Proof Let PT_p^* be the output generated by MPALM from PT_s^*. Let $A_{alg}(PT_s^*, PT)$ represents whatever can be learned by the adversary from the inputs PT_s^*, PT and the knowledge of the algorithm. Note that since mSPALM is minimality resistant, disclosure of $A_{alg}(PT_s^*, PT)$ does not violate δ-presence. We now prove that $A_{alg}(PT_s^*, PT) \geq A_{alg}(PT_p^*, PT)$ by looking at the post processing. For each equivalence class $EC \in PT_s^*$, in line 4, we shift PDFs towards the utility optimal distribution of EC. Now, since we only use the knowledge in PT_s^* and PT to calculate the utility optimal distribution PT_p^*, PT_p^* can also be calculated from $A_{alg}(PT_s^*, PT)$, thus $A_{alg}(PT_s^*, PT) \geq A_{alg}(PT_p^*, PT)$.

Theorem 6 and its proof state that an adversary that knows the public dataset PT and the underlying algorithm can simulate the post processing in the MPALM algorithm. The main reason for this is that MPALM shifts towards the utility optimal distribution of $EC \in PT$ other than equivalence classes

[4] One can limit the information disclosure on PT by releasing *partial* joint PDFs instead. Some attributes can be treated as joint while others are left as single column PDFs. This would be a trade off between utility and privacy.

Table 8 δ-Present anatomization of PT and T

T_1^a

Sex	Job	Nation	Salary
M	Student	Canada	4
M	Student	USA	
M	Student	USA	
M	Prof.	USA	
M	Prof.	Canada	
F	Showman	Italy	4
F	Singer	Italy	
F	Singer	Italy	
F	Teacher	Britain	
F	Teacher	Britain	

in the private dataset. This also means such an adversary does not learn anything other than a minimality resistant output. Those parties that do not know PT learn from the output of MPALM but cannot conduct minimality attacks without the knowledge of the whole PT.

As a last note we state that the methodology followed by [25] can be used to create a presence algorithm that is minimality resistant. Specifically, a k-anonymization PT^k of the public dataset is found where k is a user input. Most likely, there will be equivalence classes in PT^k that violates δ-presence. For each such class, the Ext. attribute of the tuples is distorted in such a way that the distribution for the rate of the existing tuples to non existing tuples is indistinguishable from that of the ECs satisfying δ-presence.

5.6.2 Drawbacks of MPALM

While resistant to minimality attacks, algorithm MPALM has two drawbacks.

First, releasing MPALM anonymizations discloses information about table PT. This may be a problem in most applications in which another third party owns PT. While the use of PT is permitted in such cases, the distribution of it is generally restricted due to copyrights or privacy regulations. (Both practical examples given in [15, 17] fall in this category. In [15], PT is a voters' dataset which is sold by the government. In [17], a hospital releases data about diabetics and uses data about all patients as PT.) Fortunately, information released by MPALM is bounded by first order statistics. One can also limit the amount of information (with respect to a utility or privacy metric) by adjusting the PDF parameters. As long as, we only use information in PT, the algorithm will remain minimality resistant.

Second, MPALM does not guarantee that the utility of the released dataset will be increased. Note that MPALM shifts towards the distribution of ECs in table PT. This does not necessarily agree with the actual distribution of ECs in table T. As a simple example, suppose we have, in PT, the group of three tuples with a single attribute sex $\{$<F>, <M>, <M>$\}$ and suppose only <F> is present in T. SPALM would output <*> (or <M:0.5,F:0.5>). MPALM would output <M:0.66, F:0.33> which is statistically further away from the actual distribution of T (e.g., <M:0,F:1>). On the other hand, if the existing tuple were <M>, MPALM output would be closer to the original distribution. We show in Section 6, that the latter is generally the case and MPALM significantly increases the output of SPALM.

We also want to note that if releasing PT is permitted, there is another alternative to MPALM. One can use an anatomization approach to integrate the whole PT into the released dataset. The idea is as follows: We run mSPALM as in Section 5.6.1. We release PT after marking the equivalence classes in PT. We also state for each equivalence class EC how many tuples in EC are present in T. (see Table 8) We will name this approach as APALM (Anatomy-based Presence Algorithm). One can construct a similar proof as in Theorem 6 to show that the resulting algorithm is also minimality attack resistant. APALM algorithm has the same weaknesses as MPALM. In terms of utility, there exists scenarios in which using APALM [MPALM] is more effective. As an example suppose we have, in PT, the set of tuples $\{<a_1,b_1>,<a_1,b_2>,<a_2,b_1>\}$ and we have only $<a_1,b_2>$ present in T. The matching probability for APALM, in this case, is 0.33 (we pick one of the tuples in PT randomly) whereas the matching probability for MPALM is 0.22, thus APALM output is more utilized in this case. But if only $<a_1,b_1>$ was present MPALM output would be more utilized with 0.44 probability. We show in Section 6, that there are cases in which APALM achieves a higher utility level compared to MPALM.

6 Experiments

In this section, we experimentally evaluate PDF generalizations. We first experiment with the maximum utilization we can get from PDFs by assuming an ℓ-diversity framework and next explore the trade-off between data utilization and privacy when using PDF algorithms in a δ-presence framework.

6.1 PDF, KL cost, and Utility

This section presents ℓ-diversity experiments to evaluate the maximum utilization one can get from PDFs. The reason we perform ℓ-diversity experiments is not to show the superiority of PDFs in terms of utility over previous approaches. As mentioned before, anatomization approach better preserves utility in existential certainty. Instead, our objectives are

1. to show the correctness of Theorems 2 and 3 empirically by presenting the relation between KL cost and PDFs anonymizations with varying distributions.

2. to show the correctness of Theorem 1 by experimenting the relation between KL cost and data mining accuracy.

We tried "real data" experiments by adapting the Adult dataset [5] and the Census dataset from the UCI Machine Learning Repository [6]. The continuous *age* columns in both datasets were discretized into ten nominal values to facilitate probability distribution calculations. We used the DGH algorithm Incognito [9] to create ℓ-diverse generalizations. Each output is then post processed to form the utility optimal PDF anonymization. while keeping the equivalence classes intact (same process as shown in Tables 1,2, and 6). We also used two additional PDF generalizations, INTER1 and INTER2, that assigns value distributions between uniform (as in DGH) and optimal distribution. Both distributions equally partition the Euclidean distance from uniform to the optimal into three parts. INTER1 is closer to optimal distribution. More precisely, INTER1 and INTER2 are the two intermediate distributions $f^{(2)}$ and $f^{(1)}$ defined in Eqn 3 with $mxs = 3$. Note that we proved in Theorem 3 that INTER1 is statistically more utilized than INTER2 and our purpose now is to show this experimentally. Each anonymization is reconstructed 5 times with different random seeds before mining applications are applied on each of them. We present in the graphs average results of these 5 executions.

We first plot, in Figures 3(a) and 3(b), the KL costs of the four anonymizations. As stated by Theorem 3, KL cost decreases as the distribution of PDFs gets close to the utility optimal distribution in both of the datasets.

We then run association rule mining as a data mining application on the reconstructions. From each reconstruction, we extracted set of rules with support higher than 0.1 and confidence higher than 0.6. (These thresholds were chosen so that every output returns at least one frequent rule.)

Figures 3(c), 3(d), 3(e), 3(f) show precision and recall values of rules mined from anonymizations with respect to the rules in the original data. As stated in Section 4, the utility optimal PDF reconstruction with the minimum KL cost is much closer to the original dataset in terms of rules supported. As PDF distributions get closer to uniform distribution, the precision and recall decreases for nearly all ℓ values. (We also conducted similar experiments for mining class rules. The results were similar, and due to space constraints, we do not present them here.) This set of experiments explicitly shows that the accuracy in data mining on anonymizations negatively correlates with the KL cost.

6.2 PDF for δ-Presence

This section presents experiments regarding privacy - utility relations when using PDF generalizations in a δ-presence framework. Six different δ-presence algorithms are compared with respect to utilization of the output anonymizations and execution time: SPALM, previously proposed δ-presence algorithm [15]; PPALM, PDF δ-presence algorithm (Section 5.3); MPALM, minimality attack resistant PDF algorithm (Section 5.6.1), JPALM, joint PDF algorithm (Section 5.5); APALM, anatomy-based δ-presence algorithm (Section 5.6.2) and WPALM, the weak version of PPALM without the speed up approach (Section 5.2.2). In order to fix the set of equivalence classes, we used the original SPALM algorithm as the sub-procedure in MPALM rather than a minimality attack resistant presence algorithm. Note that all these algorithms have pros and cons compared to each other discussed in the corresponding sections.

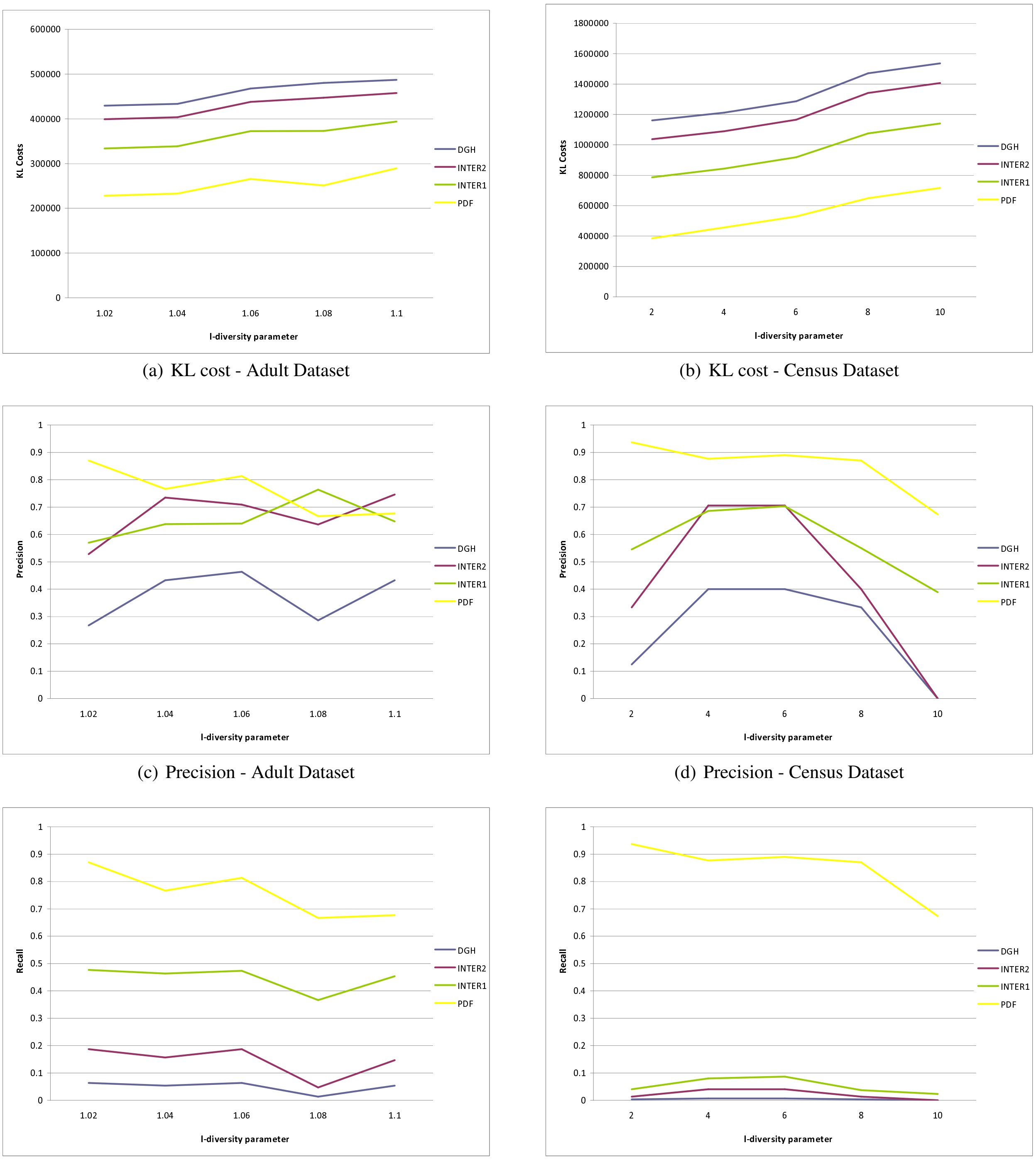
As mentioned in the previous sections; WPALM, PPALM, and JPALM try to shift uniform distributions of data values given in the output of SPALM towards the utility optimal distributions without violating δ-presence. For these set of algorithms, we set the maximum number of steps (*mxs*) to 10 for the experiments. Each shift triggers a check if the presence property still holds. As described in Section 5.2, the checking process is very costly for WPALM (time required by the checking is exponential in the size of the equivalence classes, see Section 5.2). Thus WPALM has to ignore those equivalence classes that cannot be handled in a reasonable time. In our experiments, we ignore the ECs that require the computation of existence probabilities with more than 5 million combinations. We show, in the coming sections, that WPALM is still slower than PPALM even with this assumption.

For the experiments in this section, we need a public dataset PT and a private dataset sampled from T. We use the diabetes datasets prepared and used in [15] which contain a public dataset of size 45,222 tuples and a private table of size 1957. The public dataset is the same as the adult dataset mentioned above. As a second dataset, we used the Census dataset as the public dataset. The private tables are created from the public table based on the statistics taken from [14]. So the private datasets are skewed towards people with similar characteristics. Note that due to the rate of the data sizes $\delta_{min} < 0.043 < \delta_{max}$ needs to hold on the constraints. The δ parameters were chosen so that the effect of δ_{min} and δ_{max} on the evaluation is observed. The experiments were designed to answer the following questions:

1. How effective are the proposed algorithms in terms of data utilization?
2. How efficient is the proposed PDF algorithms with speed up compared to the WPALM & SPALM algorithms?

[5] available at `http://www.ics.uci.edu/~mlearn/MLRepository.html`

[6] available at `http://www.cse.cuhk.edu.hk/taoyf/paper/vldb06.html`



(a) KL cost - Adult Dataset

(b) KL cost - Census Dataset

(c) Precision - Adult Dataset

(d) Precision - Census Dataset

(e) Recall - Adult Dataset

(f) Recall - Census Dataset

Fig. 3 Association Rule Mining Results and KL Costs for ℓ-Diverse PDFs

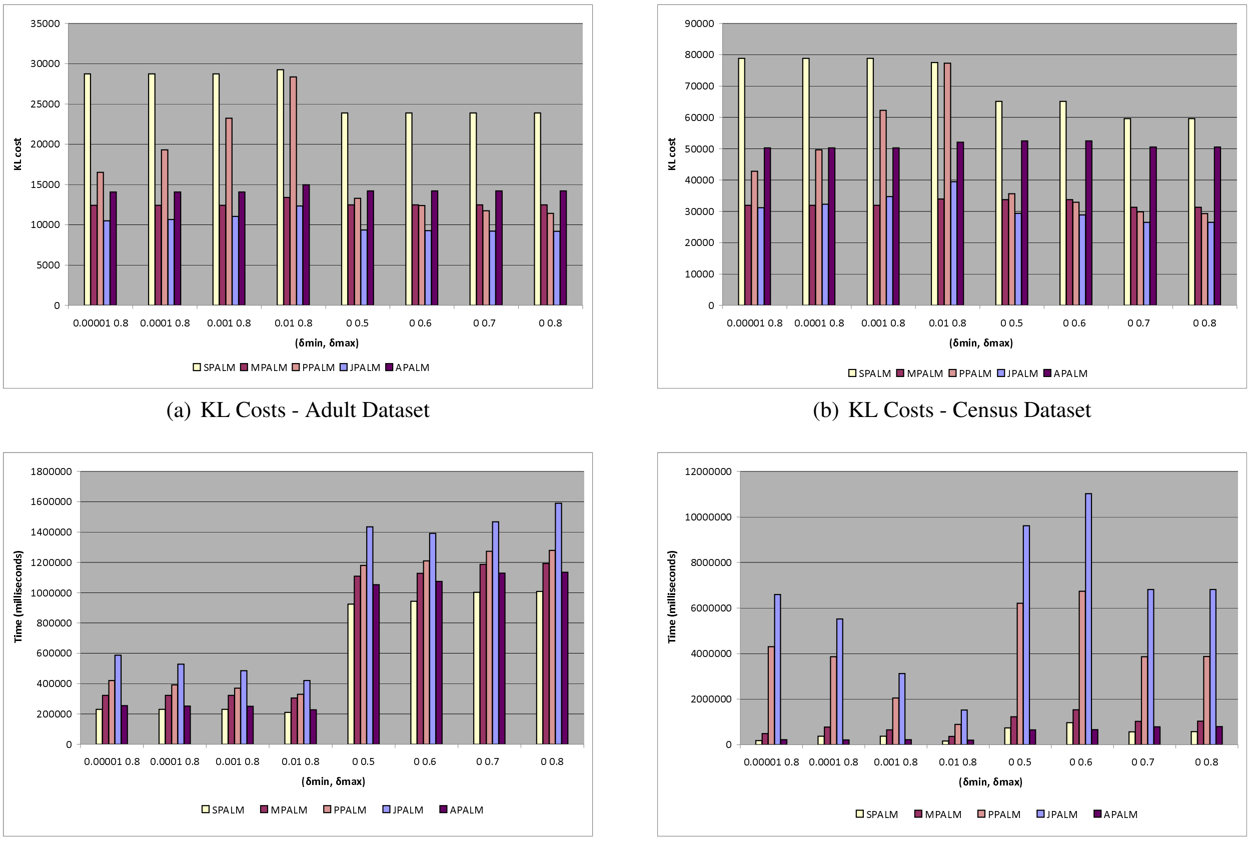
Table 9 Percentage of dataset processed by WPALM for varying δ values

.00001 .8	.0001 .8	.001 .8	.01 .8	0 .5	0 .6	0 .7	0 .8
0.4%	0.4%	0.4%	0.3%	9.4%	9.4%	9.4%	9.4%

6.2.1 *The Effectiveness of PDF algorithms in terms of data utilization*

We conducted experiments to compare the output utilizations of SPALM and PPALM with respect to KL cost metric. As it was stated in Theorem 1 and in Section 6.1, KL cost metric is negatively correlated with the matching probability and the accuracy in rule mining.

Figure 4(a) and 4(b) shows the KL costs of the output anonymizations for SPALM, APALM, and PDF algorithms for various δ_{min} & δ_{max} intervals. In both datasets, compared to the previous SPALM algorithm, PDF algorithms



(a) KL Costs - Adult Dataset

(b) KL Costs - Census Dataset

(c) Execution Times - Adult Dataset

(d) Execution Times - Census Dataset

Fig. 4 KL costs and execution times for SPALM, PPALM, MPALM, JPALM, and APALM

PPALM, JPALM, and MPALM introduce a great increase in utilization by a factor of three at times. MPALM, JPALM, and APALM that inject the information of *PT* into released anonymizations achieve better utility than PPALM. JPALM achieves the highest utility level. MPALM seems to perform better than PPALM for strict privacy requirements. Improvement by PPALM, on the other hand, is more observable for larger δ intervals. The reason for this is that single dimensional assumption for algorithm SPALM is quite inflexible and often fails to add enough information content into the output anonymization even when we lower the δ constraints. This leaves room for PPALM (and consequently JPALM) to shift PDFs toward the utility optimal distribution, thus inject utilization into the anonymization. APALM outputs are a bit less utilized than MPALM with respect to KL-cost, but perform better than PPALM outputs for strict privacy requirements.

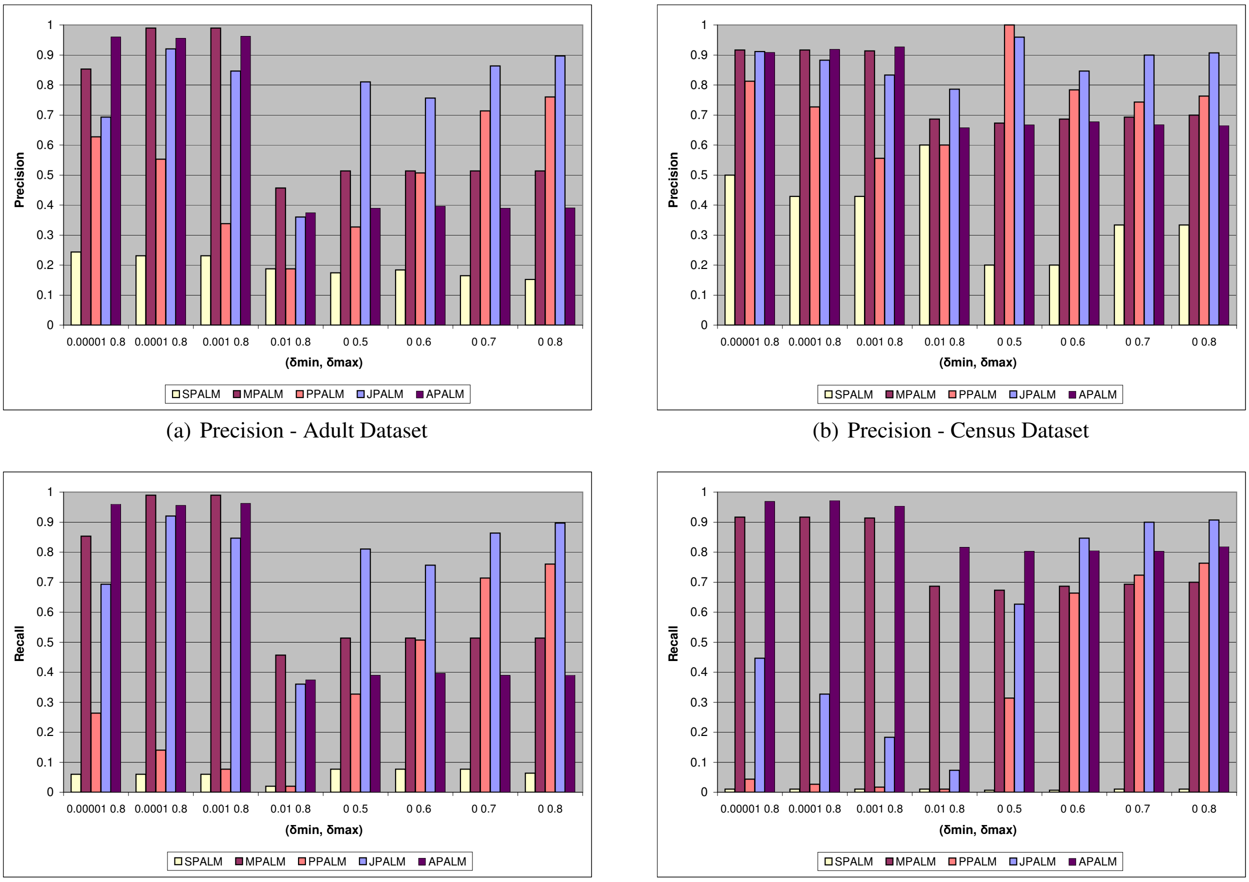
The data mining results given in Figure 5 justify the cost metric results. The error rates in finding association rules from output anonymizations generally correlate with the KL costs of the anonymizations. The only observable exception is the performance of APALM algorithm in Census Dataset. APALM output contains more accurate rules with a higher level of KL-cost compared to MPALM and JPALM algorithms. We want to note that these kind of exceptions are not unusual. As noted in [16], data mining accuracy depends also on other factors (e.g., the attributes distorted) and mining accuracy may not always correlate with what a general purpose utility metric reports.

Table 10 Support counts for frequent itemsets. A:marital-status=Married-civ-spouse, B:sex=Male, C:family-status=Husband, D:native-country=United-States, E:race=White

	Org	SP.	MP.	PP.	JP.	AP.
A	1121	575	1123	844	1118	1121
AB	1030	528	1031	781	1027	1030
ABC	1024	N/A	1028	469	1020	1025
ABCD	944	N/A	934	242	980	933
ABCDE	854	N/A	852	N/A	838	871

The support and confidence levels for frequent rules we found were consistent with Figure 5. As an example, in Table 10 we show the support counts of five frequent itemsets taken from the adult dataset and its anonymizations when $\delta_{min} = 0.00001, \delta_{max} = 0.8$.

Due to large number of tuples not processed, WPALM did not result in a significant improvement in utility over SPALM. Thus, we do not show results for WPALM in the figures.



(a) Precision - Adult Dataset

(b) Precision - Census Dataset

(c) Recall - Adult Dataset

(d) Recall - Census Dataset

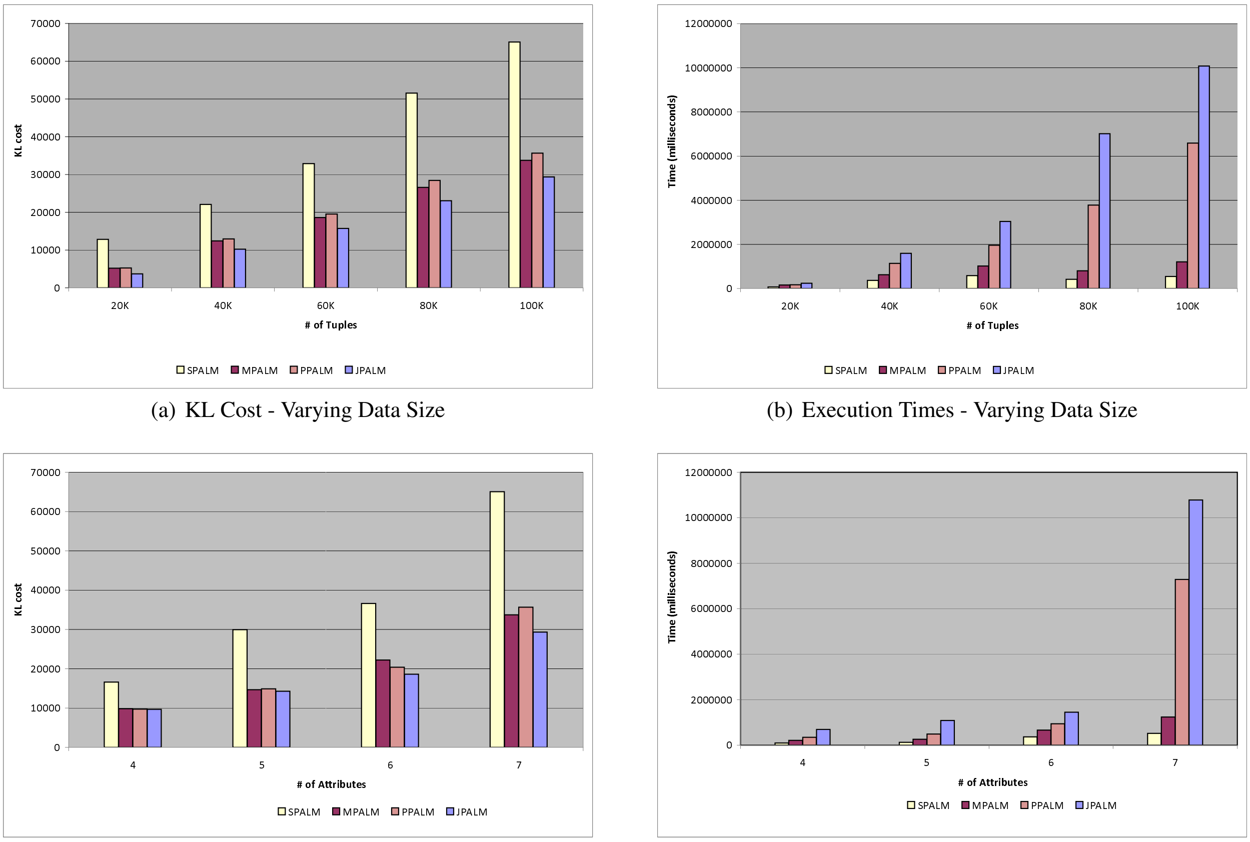
Fig. 5 Data mining results for SPALM, PPALM, MPALM, and JPALM

6.2.2 The Efficiency of PDF Algorithms

We conducted a set of experiments to compare the running times of SPALM, APALM, and the PDF algorithms on a Core2duo 3GHz Linux computer with 3GB of RAM. The running times of these algorithms for various δ_{min} & δ_{max} configurations can be seen in Figures 4(a) and 4(b). As expected, SPALM is the algorithm with the shortest running time requirement, since it acts as a subroutine for the other algorithms. MPALM and APALM requires only one scan of *PT* to post-process SPALM output, thus the execution times for MPALM, APALM, and SPALM are nearly the same. PPALM requires more time than SPALM due to the post processing of shifting distribution towards utility optimal. JPALM post-processes PPALM thus it is the slowest PDF algorithm proposed in this paper. However, additional time cost needed for PPALM and JPALM is within acceptable limits and scales well with the length of the δ intervals.

Even if we do not show in the figures, in most experiments, WPALM requires more execution time compared to PPALM and this is so despite of the fact that it does not process most of the ECs. Table 9 shows the percentage of the database ignored by WPALM. Majority of the tuples (90+%) have not been processed. Besides as we force WPALM to process more equivalence classes, the execution time becomes intractable. As an example, for the experiment where $\delta = (0.01, 0.8)$, (in which WPALM seems to be slightly faster than PPALM) WPALM processes 9 equivalence classes (147 tuples) all of which require around 16000 likelihood multiplications in total. The smallest equivalence class which is not processed by WPALM is of size 38 tuples with 10 existent tuples. Processing an equivalence class of this size would cost WPALM to make around 472 million multiplications. Roughly speaking WPALM would run 1345 times slower to process an additional 0.084% of the whole data. This is due to the exponential complexity of WPALM mentioned in Section 5.4.

Even though ideal WPALM acts as an upper bound for PPALM in terms of utilization, experiments in this section along with the previous section show that WPALM is too inefficient to be practical compared to PPALM. For WPALM to be as utilized as PPALM, an extremely huge amount of execution time is required as the number of combinations that is taken into account during the calculation of existence probabilities grows exponentially in the size of EC groups. In reasonable settings PPALM is faster than WPALM with



(a) KL Cost - Varying Data Size

(b) Execution Times - Varying Data Size

(c) KL Cost - Varying no of Attributes

(d) Execution Times - Varying no of Attributes

Fig. 6 Scalability of SPALM and PPALM

better utilization. So all of these explicitly demonstrates the power of the speed-up technique in reducing the execution time as well as increasing the utilization of the data.

In Figure 6, we show the behavior of PPALM with varying data size and number of dimensions for the Census dataset. KL cost increases with data size because KL cost is not normalized. Also increasing number of dimensions decreases utility due to curse of dimensionality [1]. Regardless of the data size and dimensionality, KL costs of PDF algorithms are less than half of that of SPALM. As for efficiency, PPALM scales well with increasing data size. However, there is a major difference in efficiency between executions with 6 and 7 attributes. The reason for this is that the sudden increase in the number of equivalence classes triggers many additional calls to Equation 5.

7 Future Work

There remain issues that are not addressed in this paper.

While PPALM increases utility by shifting PDFs of a SPALM output, they make use of the grouping of SPALM. A methodology that finds the best grouping within the space of all PDF anonymizations clearly has potential of providing even more utilization. One possible algorithm addressing this is based on hierarchical clustering. The algorithm starts with the tuples in PT, each belonging to a separate cluster. The algorithm then merges close clusters with each other until in each cluster c, the utility optimal PDF generalization of the tuples in c satisfies δ-presence. The closeness here is measured as the negative KL cost of the generalization. The evaluation of such an algorithm is left as future work.

Storage and utilization of PDF anonymizations require unconventional database management systems. Fortunately, uncertainty management in databases is not a new concept, there has already been systems implemented that provide support for uncertain data [22]. PDF reconstructions can also be stored in a conventional database management system. However, that would harm the utility due to the random nature of the reconstruction. Instead, applications running on such databases can make better use of PDFs on the software level while storing the PDF anonymization in multiple tables (e.g., one table holds ids and sensitive attributes and the rest hold the PDF functions so that the join is the PDF anonymization.) Experiments in this paper have been conducted in a similar fashion.

8 Conclusions

We presented PDF generalizations in which the value generalizations are empowered with probability distributions. We showed how PDFs can be used to increase utility without violating the ℓ-diversity and δ-presence privacy constraints. We also presented several optimizations to make the technique practical. We proposed two PDF algorithms WPALM and PPALM to achieve δ-presence. The experiments on real world data showed that the use of PDFs increases utilization without violating the privacy constraints.

References

1. C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB'05: Proceedings of the 31st International Conference on Very Large Data Bases*. VLDB Endowment, 2005, pp. 901–909.
2. R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *ICDE'05: Proceedings of the 21st International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 217–228.
3. J.-W. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in *DASFAA07: The 12th International Conference on Database Systems for Advanced Applications*, Apr. 2007.
4. B. C. M. Fung, K. Wang, and P. S. Yu, "Top-down specialization for information and privacy preservation," in *ICDE'05: Proceedings of the 21st International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 205–216.
5. G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast data anonymization with low information loss," in *VLDB'07: Proceedings of the 33rd International Conference on Very Large Data Bases*. VLDB Endowment, 2007, pp. 758–769.
6. V. S. Iyengar, "Transforming data to satisfy privacy constraints," in *KDD'02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2002, pp. 279–288.
7. D. Kifer, "Attacks on privacy and definetti's theorem," in *SIGMOD '09: Proceedings of the 35th SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 2009, pp. 127–138.
8. D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *SIGMOD'06: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM Press, 2006, pp. 217–228.
9. K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *SIGMOD'05: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2005, pp. 49–60.
10. ——, "Mondrian multidimensional k-anonymity," in *ICDE'06: Proceedings of the 22nd International Conference on Data Engineering*, Atlanta, GA, Apr. 3-7 2006, pp. 25–35. http://dx.doi.org/10.1109/ICDE.2006.101
11. N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE'07: Proceedings of the 23nd International Conference on Data Engineering*, Istanbul, Turkey, Apr. 16-20 2007.
12. T. Li and N. Li, "Optimal k-anonymity with flexible generalization schemes through bottom-up searching," in *PADM'06: IEEE International Workshop on Privacy Aspects of Data Mining*, Hong Kong, Dec. 18 2006.
13. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," in *ICDE'06: Proceedings of the 22nd IEEE International Conference on Data Engineering*, Atlanta Georgia, Apr. 2006.
14. National Institute of Diabetes and Digestive and Kidney Diseases, "National diabetes statistics fact sheet: General information and national estimates on diabetes in the United States," U.S. Department of Health and Human Services, National Institute of Health, Bethesda, MD, Tech. Rep. NIH Publication No. 06–3892, Nov. 2005. http://diabetes.niddk.nih.gov/dm/pubs/statistics/
15. M. E. Nergiz, M. Atzori, and C. Clifton, "Hiding the presence of individuals in shared databases," in *SIGMOD'07: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, Beijing, China, June 11-14 2007.
16. M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," *Data and Knowledge Engineering*, vol. 63, no. 3, pp. 622–645, Dec. 2007. \url{http://dx.doi.org/10.1016/j.datak.2007.03.009}
17. ——, "δ-Presence without complete world knowledge," *IEEE Transactions on Knowledge and Data Engineering*, 2009.
18. M. E. Nergiz, C. Clifton, and A. E. Nergiz, "Multirelational k-anonymity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. 1, 2009.
19. A. Øhrn and L. Ohno-Machado, "Using boolean reasoning to anonymize databases," *Artificial Intelligence in Medicine*, vol. 15, no. 3, pp. 235–254, Mar. 1999. http://dx.doi.org/10.1016/S0933-3657(98)00056-6
20. P. Samarati, "Protecting respondent's identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001.
21. P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*. New York, NY, USA: ACM, 1998, p. 188.
22. S. Singh, C. Mayfield, S. Mittal, S. Prabhakar, S. Hambrusch, and R. Shah, "Orion 2.0: Native support for uncertain data," in *SIGMOD'08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. New York, NY, USA: ACM, 2008, pp. 1239–1242.
23. L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system," in *Journal of the American Medical Informatics Association*. Hanley & Belfus, Inc., 1997.
24. ——, "k-Anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness Knowledge-Based Systems*, vol. 10, no. 5, pp. 557–570, 2002.
25. R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality attack in privacy preserving data publishing," in *VLDB'07: Proceedings of the 33rd International Conference on Very Large Data Bases*. VLDB Endowment, 2007, pp. 543–554.
26. R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing," in *KDD'06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2006, pp. 754–759.
27. X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *VLDB'06: Proceedings of 32nd International Conference on Very Large Data Bases*, Seoul, Korea, Sept. 12-15 2006.
28. L. Zhang, S. Jajodia, and A. Brodsky, "Information disclosure under realistic assumptions: privacy versus optimality," in *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*. New York, NY, USA: ACM, 2007, pp. 573–583.
29. Q. Zhang, N. Koudas, D. Srivastava, and T. Yu, "Aggregate query answering on anonymized tables," *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pp. 116–125, April 2007.

A Utility Optimal Distribution

In this section, we prove Theorems 1, 2, and 3.

We focus on the equivalence class EC and derive the optimal distribution function $F : \{f_1, \cdots, f_A, f_{A+1}\}$ for QI attributes $1 \cdots A$ and (if any) sensitive attribute $A+1$ in EC that will maximize the matching probability for a pdf anonymization T^* of T. Let again c_a^i be the number of times an atomic data value v_i from D_a (domain of attribute a) appears in attribute a of T. Note that for attribute a, the same distribution f_a is used in all tuples of EC. To compact the equations below, we use notation f_a^i in place of $f_a(v_i)$. ; *Theorem 1*: The matching probability for EC is negatively correlated with the following equation defined over EC:

$$KL(EC) = -\sum_{a=1}^{A} \sum_{v_i \in D_a} c_a^i \cdot \ln f_a^i \tag{6}$$

to which we will refer as the *KL cost* of EC.

Proof Given distribution function $F : \bigcup_a f_a$ for the equivalence class EC, matching probability $\mathscr{P}_F$ is given by

$$\mathscr{P}_F = (\prod_{v_i \in D_{A+1}} c_{A+1}^i!) \cdot (\prod_{a=1}^{A} \prod_{v_i \in D_a} (f_a^i)^{c_a^i}) = C_1 \cdot \prod_{a=1}^{A} \prod_{v_i \in D_a} (f_a^i)^{c_a^i}$$

Maximizing $\mathscr{P}_F$ is the same as maximizing $\ln \mathscr{P}_F$;

$$\ln \mathscr{P}_F = C_2 + \sum_{a=1}^{A} \sum_{v_i \in D_a} c_a^i \cdot \ln f_a^i$$

This is nothing but the negative KL cost given in Equation 1.

Theorem 2: The distribution function $F : \bigcup_a f_a$ defined as

$$f_a^i = \frac{c_a^i}{|EC|}$$

for each value $v_i \in D_a$, maximizes the matching probability for EC.

Proof For a fixed equivalence class, the F function that maximizes Equation 1:

$$\max_F(\ln \mathscr{P}_F) = C_2 + \sum_{a=1}^{A} \max_{f_a}(\sum_{v_i \in D_a} c_a^i \cdot \ln f_a^i)$$

Since we assume attribute independence, maximizing matching probability for each attribute maximizes overall probability. Assuming n_a is the size of the domain D_a;

$$\begin{aligned} &\max_{f_a}(\sum_{v_i \in D_a} c_a^i \cdot \ln f_a^i) \\ &= \max_{f_a}(c_a^1 \cdot \ln f_a^1 + \cdots + c_a^{n_a-1} \cdot \ln f_a^{n_a-1} + c_a^{n_a} \cdot \ln f_a^{n_a}) \\ &= \max_{f_a}(c_a^1 \cdot \ln f_a^1 + \cdots + c_a^{n_a-1} \cdot \ln f_a^{n_a-1} \ + c_a^{n_a} \cdot \ln(1 - f_a^1 - \cdots - f_a^{n_a-1})) \end{aligned}$$

Taking the derivatives of the last equation with respect to each parameter f_a^i and setting them to 0;

$$\begin{aligned} \frac{c_a^1}{f_a^1} - \frac{c_a^{n_a}}{1 - f_a^1 - \cdots - f_a^{n_a-1}} &= 0 \\ &\vdots \\ \frac{c_a^{n_a-1}}{f_a^{n_a-1}} - \frac{c_a^{n_a}}{1 - f_a^1 - \cdots - f_a^{n_a-1}} &= 0 \end{aligned} \tag{7}$$

$$\begin{aligned} c_a^1 \cdot \sum_{i=1}^{n_a-1} f_a^i + c_a^n f_a^1 &= c_a^1 \\ &\vdots \\ c_a^{n_a-1} \cdot \sum_{i=1}^{n_a-1} f_a^i + c_a^{n_a} f_a^{n_a-1} &= c_a^{n_a-1} \end{aligned}$$

Summing up side by side;

$$\begin{aligned} \sum_{i=1}^{n_a-1} c_a^i \cdot \sum_{i=1}^{n_a-1} f_a^i + c_a^{n_a} \sum_{i=1}^{n_a-1} f_a^i &= \sum_{i=1}^{n_a-1} c_a^i \\ \sum_{i=1}^{n_a} c_a^i \cdot \sum_{i=1}^{n_a-1} f_a^i &= \sum_{i=1}^{n_a-1} c_a^i \\ |EC| \cdot (1 - f_a^{n_a}) &= |EC| - c_a^{n_a} \\ f_a^{n_a} &= \frac{c_a^{n_a}}{|EC|} \end{aligned}$$

substituting $f_a^{n_a}$ in Eqn 7, we get, for $1 \leq i \leq n_a$;

$$f_a^i = \frac{c_a^i}{|EC|}$$

Above equality maximizes the matching probability.

Theorem 3: For an equivalence class EC, let $F^{(o)} : \bigcup_a f^{(o)}{}_a$ be the utility optimal distribution and let $F^{(1)}$ and $F^{(2)}$ be two other distribution functions with $|f^{(1)}{}_a(v_i) - f^{(o)}{}_a(v_i)| \leq |f^{(2)}{}_a(v_i) - f^{(o)}{}_a(v_i)|$ for all attribute a and for all $v_i \in D_a$ then $\mathscr{P}_{F^{(1)}} \geq \mathscr{P}_{F^{(2)}}$.

Proof By Theorem 2 and Definition 10, utility optimal distribution maximizes the matching probability. Since there is no other root that makes the derivatives in Equation 7 zero, the matching probability monotonically decreases as each f_a^i gets far away from the utility optimal distribution.

B Effects of PDFs on Rule Mining and Classification

Association rule mining is a process of finding binary rules (e.g., 'M ⇒ USA') that hold frequently in a given dataset (e.g., T). Frequency is defined in terms of minimum *support* (percentage of tuples in T that contain M and USA together, $\mathscr{P}(M \cup USA) = \frac{3}{8}$) and *confidence* (percentage of tuples in T containing M that also contain USA, $\mathscr{P}(USA \mid M) = \frac{3}{4}$). In our methodology, an anonymization is assumed to be successful in terms of rule mining, if the associated reconstruction respects exactly the same frequent rules as the original dataset does. The success is obviously correlated with the probability that the reconstruction correctly simulates the original dataset.

Let T^* be a PDF generalization of T and $b(T')$ is a boolean function that returns 1 iff dataset T' respects rule r with minimum support s and confidence c, then the probability that T^R will also respect rule r is given by

$$\begin{aligned} \mathscr{P}(b(T^R) = 1) &= \sum_{T'} Pr(T^R = T') \cdot b(T') \\ &= \sum_{T'} \prod_{i,j} T^R[i][j].f(T'[i][j]) \cdot b(T') \end{aligned}$$

Since the matching probabilities are higher for utility optimal PDF anonymizations, the expected rule mining success rate of such anonymizations should be at least as good as that of other anonymizations

Table 11 Rules holding in table T with $s \geq 0.25, c \geq 0.75$ and holding probabilities of the same rules for T_n^* and T_p^*

Rules	**NDGH:T_n^***	**PDF:T_p^***
USA ⇒ M	0.68	0.95
Italy ⇒ F	0.68	0.95
Singer ⇒ Italy	0.09	0.36
Singer ⇒ F	0.41	0.68
M ⇒ USA	0.31	0.74
F ⇒ Italy	0.31	0.74

Table 12 Probabilities that reconstructed T_n^* and T_p^* will respect rule 'Italy ⇒ >50K' for different minimum support and confidence

	$s \geq 0.25$		$s \geq 0.375$	
	$c \geq 0.66$	$c = 1$	$c \geq 0.75$	$c = 1$
T_n^*	0.52	0.32	0.12	0.06
T_p^*	0.84	0.52	0.42	0.1

(e.g., NDGH). Table 11 lists the rules holding in T with minimum support 0.25 and minimum confidence 0.75 along with the probabilities that the rules apply for the reconstructed NDGH anonymization T_n^* and PDF anonymization T_p^*. As expected, T_p^* has higher probabilities for creating the original rules.

It is also not desirable to have false rules (rules that do not hold frequently in the original dataset) in the reconstructed datasets. It is stated in [16] that only higher level rules can be mined from overly generalized single dimensional anonymizations without significant errors (e.g., '{Ca, US} ⇒ M' will be mined from T_n^* as opposed to 'US ⇒ M'). The reason is that there is no probabilistic way of distinguishing different atomic values of a given generalized value (e.g, for T_n^*, if the probability of getting rule 'US ⇒ M' is 0.68, then the probability of getting the false rule 'Canada ⇒ M' is also 0.68). This is true for anonymizations that make use of DGH, interval, or NDGH generalizations [16]. However, PDF anonymizations provide distributions to differentiate between atomic values. The same problem does not exist in such anonymizations (e.g., probability that 'Canada ⇒ M' holds for T_p^* is 0.26, whereas 'USA ⇒ M' holds with 0.95 probability).

The effects of PDFs on classification are very similar because many classification algorithms basically build models based on rules of the form $\{qi_1, \cdots, qi_n\} \Rightarrow s$ where s is a class value (e.g., salary) and qi_i are non class values (e.g., sex, job, nation). The more actual *class rules* the reconstructed data supports, the more successful it is in terms of classification. PDFs will have the same probabilistic advantage over previous generalization types with respect to classification. (in T, rule 'Italy ⇒ >50K' is a class rule holding with high confidence. Table 12 shows the probabilities that the reconstructions of T_n^* and T_p^* will respect this class rule for various levels of support and confidence. T_p^* has higher probabilities for each level.)

C The Maximum and Minimum Existence Probabilities in a given Projected Set

In this section, we prove Theorem 4. To do this, we first prove that tuples with bigger likelihood probabilities have bigger existence probabilities. This is expected, since likelihood probability for a tuple t can be thought as the share of t on the sum of existence probabilities in a given projected set (which is equal to n_1).

Theorem 7 *Given a likelihood set $P = \{p^{low}, p^{high}, p_1, \cdots, p_m\}$ and the number of present tuples n_1, if $p^{low} < p^{high}$, then $ex^{low} \leq ex^{high}$.*

Proof Difference between two existence probabilities would be

$$ex^{low} - ex^{high} = \frac{\sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p^{low} \in S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S} - \frac{\sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p^{high} \in S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S}$$

$$= \frac{\sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p^{low}, p^{high} \in S}} P_S + \sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p^{low} \in S \wedge p^{high} \notin S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S} - \frac{\sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p^{low}, p^{high} \in S}} P_S + \sum\limits_{\substack{S \subset P \wedge |S| = n_1 \wedge \\ p^{high} \in S \wedge p^{low} \notin S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S}$$

Since $\sum\limits_{p \in S \wedge |S| = n_1} P_S = p \sum\limits_{p \notin S \wedge |S| = n_1 - 1} P_S$;

$$ex^{low} - ex^{high} = \frac{p^{low} \sum\limits_{\substack{S \subset P \wedge |S| = n_1 - 1 \wedge \\ p^{low}, p^{high} \notin S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S} - \frac{p^{high} \sum\limits_{\substack{S \subset P \wedge |S| = n_1 - 1 \wedge \\ p^{low}, p^{high} \notin S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S}$$

$$= \frac{(p^{low} - p^{high}) \sum\limits_{\substack{S \subset P \wedge |S| = n_1 - 1 \wedge \\ p^{low}, p^{high} \notin S}} P_S}{\sum\limits_{S \subset P \wedge |S| = n_1} P_S}$$

First component of the numerator is negative, the second component and the denominator is non-negative. So the difference between the existence probabilities is non-positive.

Theorem 4: Given a likelihood set $P = \{p^{min}, p^{max}, p_1, \cdots, p_m\}$ and the number of present tuples n_1, let $p^{min} \leq p_i \leq p^{max}$ for $i \in [1-m]$. If $ex^{min} \geq \delta_{min}$ and $ex^{max} \leq \delta_{max}$ then $\delta_{min} \leq ex \leq \delta_{max}$ for any $ex \in EX$.

Proof By Theorem 7, $\delta_{min} \leq ex^{min} \leq ex_i \leq ex^{max} \leq \delta_{max}$.

D Finding Upper and Lower Bounds on Max and Min Existence Probabilities in a given Projected Set

In this section, we prove Theorem 5. We first show that if the likelihood probability of a tuple is increased, its existence probability also increases (or does not change) and the existence probabilities for the rest of the tuples decrease (or do not change).

Theorem 8 *Given the number of present tuples n_1, let $P^1 = \{p^{low}, p_1^1, \cdots, p_m^1\}$ and $P^2 = \{p^{high}, p_1^2, \cdots, p_m^2\}$ be two likelihood sets with $p^{low} < p^{high}$ and $p_i^1 = p_i^2$ for all $i \in [1-m]$, then we have the following relations between the existence probabilities;*

1. *$ex^{low} \leq ex^{high}$*
2. *$ex_i^1 \geq ex_i^2$ for all $i \in [1-m]$.*

Proof We first prove item 1. The difference between the existence probabilities ex^{low} and ex^{high} can be derived as follows:

$$ex^{low} - ex^{high} = \frac{\sum\limits_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \in S}} P_S}{\sum\limits_{S \subset P^1 \wedge |S| = n_1} P_S} - \frac{\sum\limits_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \in S}} P_S}{\sum\limits_{S \subset P^2 \wedge |S| = n_1} P_S}$$

$$= \frac{p^{low} \sum\limits_{\substack{S \subset P^1 \wedge |S| = n_1 - 1 \wedge \\ p^{low} \notin S}} P_S}{\sum\limits_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \in S}} P_S + \sum\limits_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \notin S}} P_S} - \frac{p^{high} \sum\limits_{\substack{S \subset P^2 \wedge |S| = n_1 - 1 \wedge \\ p^{high} \notin S}} P_S}{\sum\limits_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \in S}} P_S + \sum\limits_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \notin S}} P_S}$$

Setting

$$C_1 = \sum_{\substack{S \subset P^1 \wedge |S| = n_1 - 1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S| = n_1 - 1 \wedge \\ p^{high} \notin S}} P_S$$

$$C_2 = \sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \notin S}} P_S$$

Since C_1 and C_2 are non-negative, we have;

$$\begin{aligned} ex^{low} - ex^{high} &= \frac{p^{low} C_1}{p^{low} C_1 + C_2} - \frac{p^{high} C_1}{p^{high} C_1 + C_2} \\ &= \frac{(p^{low} - p^{high}) C_1 C_2}{(p^{low} C_1 + C_2)(p^{high} C_1 + C_2)} \\ &\leq 0 \end{aligned}$$

We now prove item 2. The difference between the existence probabilities ex_i^1 and ex_i^2 for any possible i is given by;

$$ex_i^1 - ex_i^2 = \frac{\displaystyle\sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p_i^1 \in S}} P_S}{\displaystyle\sum_{S \subset P^1 \wedge |S| = n_1} P_S} - \frac{\displaystyle\sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p_i^2 \in S}} P_S}{\displaystyle\sum_{S \subset P^2 \wedge |S| = n_1} P_S}$$

$$= \frac{\displaystyle\sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low}, p_i^1 \in S}} P_S + \sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \notin S \wedge p_i^1 \in S}} P_S}{\displaystyle\sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \in S}} P_S + \sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \notin S}} P_S} - \frac{\displaystyle\sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high}, p_i^2 \in S}} P_S + \sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \notin S \wedge p_i^2 \in S}} P_S}{\displaystyle\sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \in S}} P_S + \sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \notin S}} P_S}$$

Setting

$$C_1 = \sum_{\substack{S \subset P^1 \wedge |S| = n_1 - 1 \wedge \\ p^{low} \notin S \wedge p_i^1 \in S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S| = n_1 - 1 \wedge \\ p^{high} \notin S \wedge p_i^2 \in S}} P_S$$

$$C_2 = \sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \notin S \wedge p_i^1 \in S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \notin S \wedge p_i^2 \in S}} P_S$$

$$C_3 = \sum_{\substack{S \subset P^1 \wedge |S| = n_1 - 1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S| = n_1 - 1 \wedge \\ p^{high} \notin S}} P_S$$

$$C_4 = \sum_{\substack{S \subset P^1 \wedge |S| = n_1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S| = n_1 \wedge \\ p^{high} \notin S}} P_S$$

We have;

$$\begin{aligned} ex^{low} - ex^{high} &= \frac{p^{low} C_1 + C_2}{p^{low} C_3 + C_4} - \frac{p^{high} C_1 + C_2}{p^{high} C_3 + C_4} \\ &= \frac{(p^{high} - p^{low})(C_3 C_2 - C_1 C_4)}{(p^{low} C_3 + C_4)(p^{high} C_3 + C_4)} \end{aligned}$$

The denominator is definitely positive.The first additive component of the numerator is positive by the assumption. We now prove the second component $(C_3 C_2 - C_1 C_4)$ is also positive. Setting $P' = P^1 - p^{high} - p_i^1$, $P'' = P^1 - p^{high}$; $C_1 C_4$ and $C_2 C_3$ can be written as the summation of likelihood products;

$$C_1 C_4 = p_i^1 \cdot \sum_{\substack{\{pr_1^1, \cdots, pr_{n_1-2}^1\} \subset P', \\ \{pr_1^4, \cdots, pr_{n_1}^4\} \subset P''}} (pr_1^1 \cdots pr_{n_1-2}^1) \cdot (pr_1^4 \cdots pr_{n_1}^4)$$

$$C_2 C_3 = p_i^1 \cdot \sum_{\substack{\{pr_1^2, \cdots, pr_{n_1-1}^2\} \subset P', \\ \{pr_1^3, \cdots, pr_{n_1-1}^3\} \subset P''}} (pr_1^2 \cdots pr_{n_1-1}^2) \cdot (pr_1^3 \cdots pr_{n_1-1}^3)$$

Let, without loss of generality, in all the additive terms of $C_1 C_4$, $pr_{n_1}^4 \neq pr_j^1$ for all $j \in [1 \cdots n_1 - 2]$ and $pr_{n_1}^4 \neq p_i^1$. Any additive term $(pr_1^1 \cdots pr_{n_1-2}^1) \cdot (pr_1^4 \cdots pr_{n_1-1}^4 \cdot pr_{n_1}^4)$ of $C_1 C_4$ also exists as an additive term in $C_2 C_3$ as $(pr_1^1 \cdots pr_{n_1-2}^1 \cdot pr_{n_1}^4) \cdot (pr_1^4 \cdots pr_{n_1-1}^4)$. It can easily be proved that $C_2 C_3$ has more additive terms than $C_1 C_4$. So $C_2 C_3 - C_1 C_4$ is also non-negative.

Theorem 8 also implies that if the likelihood probability of a tuple is decreased, its existence probability also decreases (or does not change) and the existence probabilities for the rest of the tuples increase (or do not change).

Theorem 5: Given the number of present tuples n_1, likelihood sets $P, P^{\downarrow}, P^{\uparrow}$, and their corresponding existence sets $EX, EX^{\downarrow}, EX^{\uparrow}$; $\delta_{min} \leq ex \leq \delta_{max}$ for any $ex \in EX$ if $\delta_{min} \leq (ex^{\downarrow})^{min}$ and $(ex^{\uparrow})^{max} \leq \delta_{max}$.

Proof By Theorem 4, $\delta_{min} \leq ex \leq \delta_{max}$ for any $ex \in EX$; if $\delta_{min} \leq ex^{min}$ and $ex^{max} \leq \delta_{max}$. By Theorem 8 and the assumption, $\delta_{min} \leq (ex^{\downarrow})^{min} \leq ex^{min}$. Again by Theorem 8, $ex^{max} \leq (ex^{\uparrow})^{max} \leq \delta_{max}$.

E Minimality Attack on Optimal ℓ-Diverse Anatomizations

As mentioned in Section 2.2, given an existential certainty model and a fixed set of equivalence classes, releasing anatomizations instead of anonymizations preserves the QI attributes thus increases the utility of data. However, disclosing the exact QI attributes enables attackers to perform minimality attacks on optimal ℓ-diverse groupings (e.g., groups formed by similar tuples).

Table 13 T is a private table, T_σ is a copy of T on QI attributes but with different class assignments. T^a and T_σ^a are cost optimal 1.5-diverse anatomizations of T and T_σ respectively.

	Age	Class
t_1	18	c_1
t_2	18	c_1
t_3	49	c_2
t_4	51	c_1
t_5	51	c_2

T

	Age	Class
t_1	18	c_1
t_2	18	c_2
t_3	49	c_1
t_4	51	c_1
t_5	51	c_2

T_σ

	Age	Class
t_1	18	two c_1, one c_2
t_2	18	
t_3	49	
t_4	51	one c_1, one c_2
t_5	51	

T^a

	Age	Class
t_1	18	one c_1, one c_2
t_2	18	
t_3	49	two c_1, one c_2
t_4	51	
t_5	51	

T_σ^a

As an example, suppose in Table 13, T^a is released as the optimal anatomization of private table T. An adversary knowing that the anonymization algorithm tries to cluster similar tuples derives the following conclusion. t_3 is grouped with t_1 and t_2 even if it is closer to t_4 and t_5. That means t_1 and t_2 should have the same sensitive value c_1. (Otherwise, the data owner would release T_σ^a with a better grouping instead of T^a.) Thus the adversary discovers the sensitive values for tuples t_1, t_2, and t_3; even if the released dataset is 1.5-diverse.