2008

Generalizations with Probability Distributions for Data Anonymization

M. Ercan Nergiz

Suleyman Cetintas

Ferit Akova

Report Number:

08-001

GENERALIZATIONS WITH PROBABILITY DISTRIBUTIONS FOR DATA ANONYMIZATION

M. Ercan Nergiz
Suleyman Cetintas
Ferit Akova

CSD TR #08-001
January 2008

Generalizations with Probability Distributions for Data Anonymization

M. Ercan Nergiz, *Student Member, IEEE,*
Suleyman Cetintas, and
Ferit Akova

Abstract

Anonymization based privacy protection ensures that data cannot be traced to an individual. Many anonymity algorithms proposed so far made use of different value generalization techniques to satisfy different privacy constraints. This paper presents pdf-generalization method that empowers data value generalizations with probability distribution functions enabling the publisher to have better control over the trade off between privacy and utilization. We evaluate the pdf approach for k-anonymity and δ-presence privacy models and show how to use pdf generalizations to utilize datasets even further without violating privacy constraints. Paper also shows theoretically and experimentally that information gained from pdfs increases the utilization of the anonymized data w.r.t. real world applications such as classification and association rule mining.

***Index Terms*—Privacy, Security, integrity, and protection**

I. Introduction

The tension between the value of using personal data for research 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 such as age, gender ...) can still be mapped to individuals by using external knowledge [18].

Table anonymization is one method used to prevent against identification. Many different privacy notions that make use of anonymization have been introduced for different adversary models. For sensitive information protection, k-anonymity [15], ℓ-diversity [11], t-closeness [9], anatomization [19]; for protecting the existence of individuals in shared datasets, δ-presence [12] have been proposed. 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 [16], [5], [7], [2], [14], [13] requires user specified DGH structures (domain generalization hierarchies) 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 to (in Figure 1, Peru $\rightarrow$ America). Moreover, works in [3], [8] assumed a total order between the values of each attribute domain and used interval based generalizations which are more flexible (in Figure 1, Peru $\rightarrow$ [America-USA]). Later in [13], *NDGH based generalizations* 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 $\rightarrow$ {Peru,USA}). In Section II, we briefly explain the previously proposed methods and some of the privacy models that we will be referring to in future sections.

The trend in the research literature has been to get rid of restrictions on generalization and to increase the amount of information stored in data cells. 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 $\{a,b\}$ is equally likely to be a or b. 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 tradeoff between privacy and utility. In this paper, we present *PDF-generalization* method that empowers data value generalizations with probability distribution functions. 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 III, we formally define PDF generalizations.

The impact of generalization types on utilization is explicit for k-anonymity, ℓ-diversity, t-closeness, or δ-presence (where quasi-identifiers are generalized) but implicit for anatomization (where ℓ-diversity or t-closeness is used as an inner step). As for the privacy loss, the use of different

This material is based upon work supported by the National Science Foundation under Grant No. 0428168.

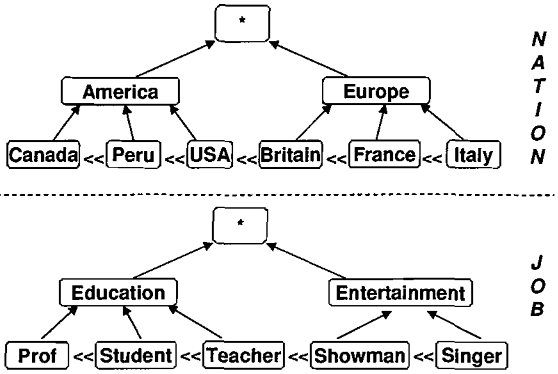


Fig. 1. DGH structures for T_d^* and total ordering for T_i^* in Table II

figure

generalization types does not introduce any privacy violation for k-anonymity, ℓ-diversity, and t-closeness (privacy models in which existence of individuals in the released datasets is known by the adversaries). This implies that in terms of privacy loss, there is no shortcoming of using a more flexible generalization type such as PDF generalizations. In such privacy models, utilization gained by PDFs can always be maximized. In Section IV, we show how to use PDFs to increase utilization by assuming k-anonymity framework and discuss theoretically how third parties can make use of the extra information provided.

For probabilistic privacy definitions such as δ-presence, when switching between generalization types, privacy loss is more observable. Thus for a better analysis of PDF generalization type in terms of utilization and privacy loss, in Section V, we use δ-presence privacy constraints. We show how to check for δ-presence constraint when non-uniform distributions are used for data cells and show how to post process output of optimal single dimensional δ-presence algorithm, SPALM [12], to make use of PDF generalizations. The final PDF algorithm, PPALM, is not optimal wrt. its domain but shows how PDFs can be used, even in a probabilistic adversary model, to increase utilization without violating privacy constraints.

Section VI evaluates the effect of the new approach on the utilization of the output dataset by presenting rule mining and classification results on real world data and shows that extra pdf information can significantly reduce rule mining and classification error on anonymized datasets without violating the privacy constraints of k-anonymity and δ-presence.

II. Background and Notation

Given a dataset (table) T, $T[c][r]$ refers to the value of column c, row r of T. $T[c]$ refers to the projection of column c on T.

While publishing person specific sensitive data, simply removing uniquely identifying information (SSN, name) from data is not sufficient to prevent identification because partially identifying information, *quasi-identifiers*, (age, gender . . .) can still be mapped to individuals by using external knowledge. E.g., in Table I, Salary attribute of private table T can be considered as *sensitive* attribute. 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 doesn't contain any uniquely identifying information [18].

In most of the privacy models, 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 adversary's knowledge on the existence of the individual:

- **Existential Certainty:**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:**Adversary doesn't know the individual is or is not in the private dataset. ($|PT| > |T|$) There are also two scenarios associated with this condition:
  - **Existential Sensitivity:**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.)
  - **Existential Identity:**Existential disclosure is not considered as a privacy violation given that sensitive information is protected according to given privacy constraints.

k-Anonymity provides (partial) privacy protection for both cases by limiting the linking of a record from a set of released records to a specific individual:

Definition 1 (k-Anonymity [17]): A given table T^* is said to satisfy k-anonymity if and only if each sequence of values in $T^*[QI_{T^*}]$ appears at least k times in T^*.

Definition 2 (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.

Table I shows an example for the privacy risk in k-anonymity framework where adversary knows PT but 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 II, respect 4-anonymity. The equivalence class of row1 in anonymized datasets is the set {row1, row2, row3, row4}. Note that by seeing one of the 4-anonymous tables, an adversary can only link, for instance, Padme into the set of salaries {>50K, $\leq$ 50K} as opposed to >50K only.

It should be noted that use of different generalization types does not violate k-anonymity definition. This makes it difficult to evaluate the privacy/utility relations for more flexible generalization types. Thus we need a probabilistic

TABLE I. k-Anonymity Framework: Public and Private Datasets. Private dataset has the same size as the Public dataset.

table

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	$\leq$ 50K
F	Showman	Italy	$>$ 50K
F	Singer	Italy	$>$ 50K
F	Singer	Italy	$>$ 50K
F	Teacher	Britain	$\leq$ 50K

TABLE II. 4-anonymous generalizations of T in Table II

table

T_d^*:DGH-anonymized Dataset

Sex	Job	Nation	Salary
M	*	America	$\leq$ 50K
M	*	America	$\leq$ 50K
M	*	America	$\leq$ 50K
M	*	America	$\leq$ 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]	$\leq$ 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}	$\leq$ 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 III. 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.

table

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

TABLE IV. 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 III. Both generalizations have the same generalization mapping.

table

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

$\Rightarrow$

T_d^*:DGH-anonymized Dataset

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

privacy notion: *δ-Presence* is defined in [12] for existential sensitivity 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 3 (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

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

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

Table III shows an example for the privacy risk in δ-presence framework where adversary knows PT and wants to identify the tuples in the private dataset T. (Attribute 'Ext' in Tables III and IV, 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 IV satisfies $(\delta_{min}, 0.8)$-presence for any $\delta_{min} \leq 0.8$. Out of 5 people {Chris, Luke, Darth, George, Obi}, 4 people is in T_d^*. So probability that Chris (or any others) is in T_d^* is 0.8. This is also true for females.

A given table can be anonymized (for k-anonymity, δ-presence, $\cdots$) by the use of generalizations:

Definition 4 (Generalization Function): Given a data value v, a generalization function ψ returns the set of all generalizations of v.

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

Definition 5 (Table Generalization): Given two tables T^1 and T^2, we say T^2 is a generalization of T^1 if and only if $|T^1| = |T^2|$ and records in T^1, T^2 can be ordered in such a way that $T^2[i][j] \in \psi(T^1[i][j])$ for every attribute $i \in QI$ and for every possible index j. We say tuple $t_1 = T_1[.][j]$ is linked to tuple $t_2 = T_2[.][j]$ and write $(t_2 \in T_2) \rightleftharpoons (t_1 \in T_1)$.

In Table II, all datasets are generalizations of table T. In each table, generalization function is defined according to generalization type being used. According to DGH structures given in Figure 1; $\psi_d(USA)$ = {USA, America, *}. T_d^* in Table II shows one DGH based anonymization of T according to the same DGH structures. According to the total ordering given in Figure 1; ψ_i(USA) $= \{[v_{min} - v_{max}] \mid v_{min} \in$ {Canada,Peru,USA} $\wedge\, v_{max} \in$ {USA,Britain,France,Italy}}. T_i^* in Table II shows one interval based anonymization of T according to the same total ordering. ψ_n(USA) $= \{S_v \mid$ {USA} $\subseteq S_v \subseteq$ {Canada, Peru, USA, Britain, France, Italy}}. NDGH based anonymizations are the most flexible anonymizations proposed so far. T_n^* in Table II shows one NDGH based anonymization of T. Tables T_d^*, T_i^*, and T_n^* use the same equivalence classes however the generalization type being used enables T_n^* to contain more specific values compared to other tables.

Work in [10] 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 evaluations in coming sections unless noted otherwise.

III. PDF Generalizations

A. Formulation

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

Definition 6 (PDF Generalization Function): A PDF generalization function ψ_p is a function, 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

TABLE V. PDF generalizations of T in Tables I and III. Tables serve as examples for both k-anonymity and δ-presence. Attribute Salary is part of the dataset in k-anonymity framework but not in δ-presence framework.

table

T_p^*:*PDF-anonymized Dataset*

Sex	Job	Nation	Salary
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	≤ 50K
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	≤ 50K
M	{Pr:0.25,St:0.75}	{Ca:0.25,US:0.75}	≤ 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}	≤ 50K

T_{p2}^*:*PDF-anonymized Dataset*

Sex	Job	Nation	Salary
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	≤ 50K
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	≤ 50K
M	{Pr:0.40,St:0.60}	{Ca:0.40,US:0.60}	≤ 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}	≤ 50K

zero probability. T_p^* and T_{p2}^* in Table V shows different PDF anonymizations of T in Table I and III. We assume for a generalized value v^* in a pdf generalization, $v^*.f$ returns the corresponding distribution function of v^* (e.g., $T_p^*[2][1].f = \{Pr:0.25, St:0.75\}$, $T_p^*[2][1].f(Pr) = 0.25$).

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}). 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.

IV. PDF and Utilization:k-Anonymity

As mentioned before, for non-probabilistic existential certainty privacy models different use of generalization types do not affect the amount of privacy provided. However this does not justify the *release* of pdf generalizations for such models. In fact, assuming total existential certainty, releasing anatomization [19] (where no QI attribute generalizations is done and a distribution for sensitive values is returned for groups of tuples) of datasets is a better approach than releasing pdf generalizations since anatomization better utilizes QI attributes without disclosing sensitive attributes. However pdf generalizations can still be used as a subprocedure to further provide utilization for anatomizations. (Anatomization makes use of generalization algorithms to form groups that contains similar tuples. Pdfs can be used to better capture similarity.)

There may be applications where k-anonymity can be classified as an existential uncertainty model. We do not defend the blind use of pdf generalizations for such scenarios since there might be privacy issues that need to be considered. We assume a k-anonymity model with existential certainty assumption in this section, because

- k-anonymity has a simple definition making it easy to understand utility aspects of different pdf generalizations. (e.g., ordering different pdf anonymizations of the same dataset in terms of utility)
- it is always possible to maximize utilization in k-anonymous datasets without violating its constraints by choosing correct distributions for pdf generalizations. This enables us to better reason about why and how extra information from pdfs can improve utilization of the data.
- when pdf utilization is maximized, it is easier to see the effects of pdfs on data-mining applications such as association rule mining and classification.

The real use of more flexible generalization types like pdfs comes into play when we assume existential uncertainty model where existence of individuals in the dataset is not a public information (and may be a sensitive information at times). In such models, use of different pdfs provide different levels of privacy. So to evaluate privacy aspects of pdfs, in Section V, we switch to δ-presence privacy model. Section V makes use of theorems on utility presented in this section.

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

A. 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 to homogeneous (leaf-level, atomic) datasets. For this purpose, we adapt the methodology proposed in [13] for pdf generalizations. Anonymized tables are first *reconstructed* before any data mining application is run.

Definition 7 (Reconstruction Function): Reconstruction function REC is a function that when given some multi-level pdf anonymized dataset T^* respecting generalization function ψ, returns an atomic data set of the same size T^R, such that

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

Informally reconstruction function converts 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 V, $T_p^R[3][1]$ will be US with 0.75 probability. For Table II, $T_d^R[3][1]$ will be US with 0.33 probability.) The reconstructed data will be suitable for all data mining applications.

B. Effects of PDF on the Reconstructed Data

Since data mining applications run on reconstructed data, effectiveness of the application application heavily depends on the similarity of the reconstructed data to the original data. Since 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. How big the *matching probability* is, depends on how much information is hidden in the anonymization. When we fix the equivalence classes EC_is in a pdf anonymization, selection of data value distributions (f functions) plays the key role in the amount of information stored in the anonymization. (e.g., T_p^* and T_{p2}^* have different matching probabilities.) Next, we derive the global optimal distribution function $GF : \{F_1, \cdots, F_\ell\}$ (where $F_i : \bigcup_{attribute\ a} f_a$) for the anonymization $T^* : \{EC_1, \cdots, EC_\ell\}$ that will maximize the matching probability.

Since each equivalence class is independent from each other, matching probability of the anonymization T^* of T is the product of matching probabilities for each equivalence class in T^*:

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

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

We now 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 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. (E.g., if we assume we have the pdf anonymization T_p^* of T in Table III and atomic value v_i is 'USA', then $c_a^i = 3$ and $f_a(v_i) = 0.75$) Then we have the following theorems:

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

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

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

PROOF. See Appendix I □

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 [6], anonymizations maximizing the negative cross-entropy minimizes KL-divergence with the original value distribution. Statistically, such an anonymization better explains the original data.

Theorem 2: The distribution function $F : \bigcup_{a,i} f_a$ defined as

$$f_a(v_i) = \frac{c_a^i}{|EC|} \qquad (2)$$

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

PROOF. See Appendix I □

Definition 8 (Utility Optimal): An anonymization T^* is utility optimal w.r.t. T if probability distribution function for every equivalence class in T^* is defined as in Eqn 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., T_p^* of Table V is utility optimal w.r.t. T of Tables III. 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 $\mathcal{P}_{F^{(1)}} \geq \mathcal{P}_{F^{(2)}}$.

PROOF. See Appendix I □

Theorem 3 gives a way to compare pdf generalizations in terms of utilization. In Tables II and V matching probability for T^*_{p2} is bigger than that of T^*_n. This due to the fact that distributions in T^*_{p2} is closer to those of 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 V, we use the 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 simulates original datasets at least as good as the other generalization types do.

As the reconstructed data becomes similar to the original data, any application run on reconstructed data increase in accuracy. Next section, we observe the effects of utility-optimal pdf generalizations on data mining applications, rule mining and classification, by looking at example datasets in Table II. Since NDGH approach is the most flexible one among previous generalization types, the comparison is carried out between datasets T^*_n and T^*_p.

C. Effects on Rule Mining and Classification

Association rule mining is a process of finding binary rules (e.g., 'M $\Rightarrow$ 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, $\mathcal{P}(M \cup USA) = \frac{3}{8}$) and *confidence* (percentage of tuples in T containing M that also contain USA, $\mathcal{P}(USA \mid M) = \frac{3}{4}$). 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 min support s and confidence c, then probability that T^R will also respect rule r is given by

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

Since matching probabilities are higher for utility-optimal pdf anonymizations, expected rule mining success rate of such anonymizations should be at least as good as that of other anonymizations. (e.g., NDGH) Table VI 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 reconstructed NDGH anonymization T^*_n and pdf anonymization T^*_p. As expected, T^*_p has higher probabilities for creating original rules.

It is also not desirable to have false rules (rules that does not hold frequently in the original dataset) in the reconstructed datasets. It is stated in [13] that only higher level rules can be mined from overly generalized single dimensional anonymizations without significant errors. (e.g., '{Ca, US} $\Rightarrow$ M' will be mined from T^*_n as opposed to 'US $\Rightarrow$ M') The reason is that there is no probabilistic way of distinguishing between different atomic values of a given generalized value. (e.g, for T^*_n, if probability of getting rule 'US $\Rightarrow$ M' is 0.68, then probability of getting false rule 'Canada $\Rightarrow$ M' is also 0.68.) This is true for anonymizations that make use of DGH, interval, or NDGH generalizations. However, pdf anonymizations provide distributions to differentiate between atomic values. The same problem does not exist in such anonymizations. (e.g., probability that 'Canada $\Rightarrow$ M' holds for T^*_p is 0.26, whereas 'USA $\Rightarrow$ M' holds with 0.95 probability.)

Effects of pdfs on classification is 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 w.r.t. classification. (in T, rule 'Italy $\Rightarrow$ >50K' is a class rule holding with high confidence. Table VII shows the probabilities that reconstructed T^*_n and T^*_p will respect this rule for different minimum support and confidence. T^*_p has higher probabilities for each level.)

In Section VI, we experiment the effect of pdf generalizations on association and class rule mining and show that use of pdf generalization increase the effectiveness of data mining applications.

V. PDF and Privacy:δ-Presence

In this section, we switch to a probabilistic existential uncertainty model, δ-presence. 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 will inject utilization into the datasets without violating the privacy constraints and next improve WPALM in terms of efficiency with a second algorithm, PPALM.

TABLE VI. 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^*

table

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

table

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

A. PDF δ-Presence Algorithms:WPALM & PPALM

In this section, we empower the previously proposed δ-presence algorithm, SPALM [12], to make use of PDF generalizations. SPALM when given a public table PT and private table T, returns an anonymization T^* of T which is δ-present wrt. PT and T. 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 section applies for both of the algorithms, so we will use the name [W,P]PALM in place of both pdf algorithms. We show experimentally in Section VI that outputs of [W,P]PALM are better utilized w.r.t. KL-cost and data mining applications.

[W,P]PALM operates on the SPALM output, which is already δ-present w.r.t. input datasets. Additionally, [W,P]PALM shifts pdfs within the output towards utility optimal distribution as long as δ-presence property is preserved. Resulting anonymization is obviously not optimal w.r.t. 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, [W,P]PALM shifts the value distributions (fs), from uniformity towards utility-optimal distribution step by step. The maximum no. of steps is set by input variable mxs. (in other words, 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 distribution function f^k being tried in step k is defined as

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

In Figure 2, $f^{(u)}$('Europe')={Italy:0.33,Britain:0.33, France:0.33}, $f^{(o)}$('Europe')={Italy:0.75,Britain:0.25, France:0}. For $mxs = 3$, f^1('Europe')= {Italy:0.47,Britain:0.30,France:0.22}, and f^2('Europe')= {Italy:0.61,Britain:0.27,France:0.11}. 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.

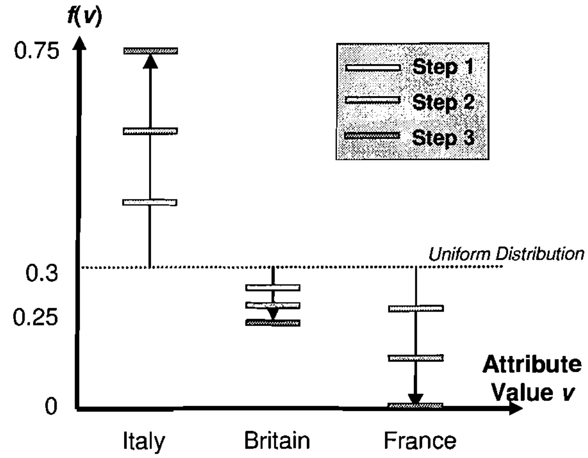


Fig. 2. Shifting of the uniform distribution (inherited in data value 'Europe') in T_d^* of Table II to the utility optimal distribution in three steps.

figure

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

Boolean function *isPresent* is called in line 8 to check for presence property. However checking for presence property for non-uniform pdfs is not as simple as in uniform pdfs.(e.g, dgh, interval, ndgh generalizations) Next two sections cover how checking process is carried out for pdf generalizations. WPALM and PPALM differs in their implementation of *isPresent*.

Algorithm 1 WPALM and PPALM

Require: public table PT; private table T, parameter δ, maximum number of shift steps mxs.
Ensure: return a pdf generalization of T respecting $(\delta_{min}, \delta_{max})$-presence with cost at most that of the optimal full domain generalization.

```
insert "Ext" attribute into PT according to T as in Table III.
run SPALM on PT, T, and δ, let PT* be the output anonymization of PT
k = 1.
while k ≤ mxs do
  for all equivalence class EC in PT* do
    for all attribute a do
      update the distribution function of values as f^k given in Eqn. 3
  if !isPresent(PT*,PT,δ_min, δ_max) then
    undo last updates.
    return
```

B. Checking for δ-Presence Property

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

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 IV). For uniform and non-overlapping generalizations, the existence probabilities can simply be calculated by working on the anonymization PT^*:

Definition 9 (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 III and IV, {Chris,Luke,Darth,George,Obi} is a projected set with j^* = <M,*,America>. In non-overlapping generalizations, 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 existence probability for any $j_i \in J$ is given by

$$\mathcal{P}(j_i \in T \mid T^*, PT) = \frac{n^1}{n^0 + n^1}$$

In other words, 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^*.) And every tuple is equally likely. Existence probabilities are the same for any tuple of the same projected set.

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. Adversary still knows n^1 tuples is selected among $|J|$ tuples but likelihood of each tuple is different due to the distribution of the outcome:

Definition 10 (Likelihood Probability): 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. $p_j^{j^*} = \mathcal{P}((j \in PT) \rightleftharpoons (j^* \in T^*)) = \prod_i j^*[i].f(j[i])$.

Given PT of Table III and T_p^* of Table V 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 11 (Likelihood Set and Existence Set): Let set of tuples $J = \{j_1, \cdots j_n\}$ be a projected set in PT w.r.t. some anonymization T^*. Likelihood set for J is defined as $P = \{p_1, \cdots, p_n\}$ where $p_i = p_{j_i}^{j^*}$. We write P_S for a set of likelihoods S for $\prod_{p \in S} p$ (product of all the likelihoods in S)
Existence set for J is defined as $EX = \{ex_1, \cdots, ex_n\}$ where $ex_i = \mathcal{P}(j_i \in T \mid T^*, PT)$.

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. Given the likelihood set and the number of existent tuples n^1, each element in the existence set can be calculated one by one. Existential probability for any tuple $j_k \in J$ takes the following conditional form:

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

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

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

Similarly, 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 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 δ.
Ensure: return true iff PT^* satisfies $(\delta_{min}, \delta_{max})$-presence.

for all projected set $J \in PT$ w.r.t. PT^* **do**
  for all tuples $j \in J$ **do**
    calculate existence probability ex for j as given in Eqn 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 Eqn 4 to check for the presence property. Basically algorithm calculates the existence probabilities for all tuples and returns true iff all existence probabilities lies 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 *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 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 Eqn 4 is infeasible even if likelihood probabilities for the tuples fits into the memory. Next subsection shows how to weaken this problem by presenting an alternative algorithm.

C. 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 δ-presence property can be speed up in two steps:

1) Existence probability of only two tuples needs to be calculated for checking.
2) Calculation of *exact* existence properties is not needed. Finding upper and lower bounds on the max and min 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 no. 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 II □

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

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

Definition 12: Given the no. 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, 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 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 lower and upper boundary likelihood sets can be used to check if the original likelihood set satisfies δ-presence. If 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 no. 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. See Appendix III. □

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 original likelihood set P (and the original projected set) satisfies $(0.31, 0.97)$-presence. Precisely P satisfies $(0.47, 0.94)$-presence.

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

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

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

Equation 5 does not require addition of many likelihood products so it is much faster to compute compared to Equation 4. However boundary sets are useful if 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.

Algorithm 3 isPresent for PPALM

Require: public table PT with attribute Ext; one anonymization of PT, PT^*; parameter δ.
Ensure: return true iff N^* satisfies $(\delta_{min}, \delta_{max})$-presence.

```
for all projected set J ∈ 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↓, P↑ of P.
    calculate ex. probability (ex↓)^min [(ex↑)^max] for the min [max] likelihood in P↓ [P↑] w.r.t. n^1
    if (ex↓)^min ≤ δ_min then
        return false
    if (ex↑)^max ≥ δ_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 algorithm creates upper and lower bound likelihood sets for the likelihood sets of each projected set in PT w.r.t. the anonymization and returns true iff bound sets satisfy *partial* presence property.

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

VI. Experiments

In this section, we experimentally evaluate pdf generalizations. We first experiment the maximum utilization we can get from pdfs by assuming k-anonymity framework and next explore the trade off between data utilization and privacy when using pdf algorithms in a δ-presence framework.

A. PDF for k-Anonymity

This section presents k-anonymity experiment to the evaluate maximum utilization one can get from pdfs. We tried "real data" experiments by adapting the Adult dataset from the UCI Machine Learning Repository [4]. The dataset was prepared the same way as in [13]. Entries with missing values are removed and the 8 attributes that are potential identifiers are used. Continuous *age* column was discretized to ten nominal values to facilitate probability distribution calculations. The dataset is k-anonymized with DGH algorithm Incognito [7] and interval algorithm Mondrian [8]. Each output is then recreated by using different generalization types but equivalence classes were preserved. (Same process as shown in Tables I,II, and V) The generalization types compared are DGH (for Incognito), interval (for Mondrian), NDGH and utility-optimal PDF. We also used two additional PDF generalizations INTER1 and INTER2 that assigns value distributions between uniform (as in NDGH) and optimal distribution. Both distributions equally partitions the euclidean distance from uniform to 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 $mxs = 3$.) 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.

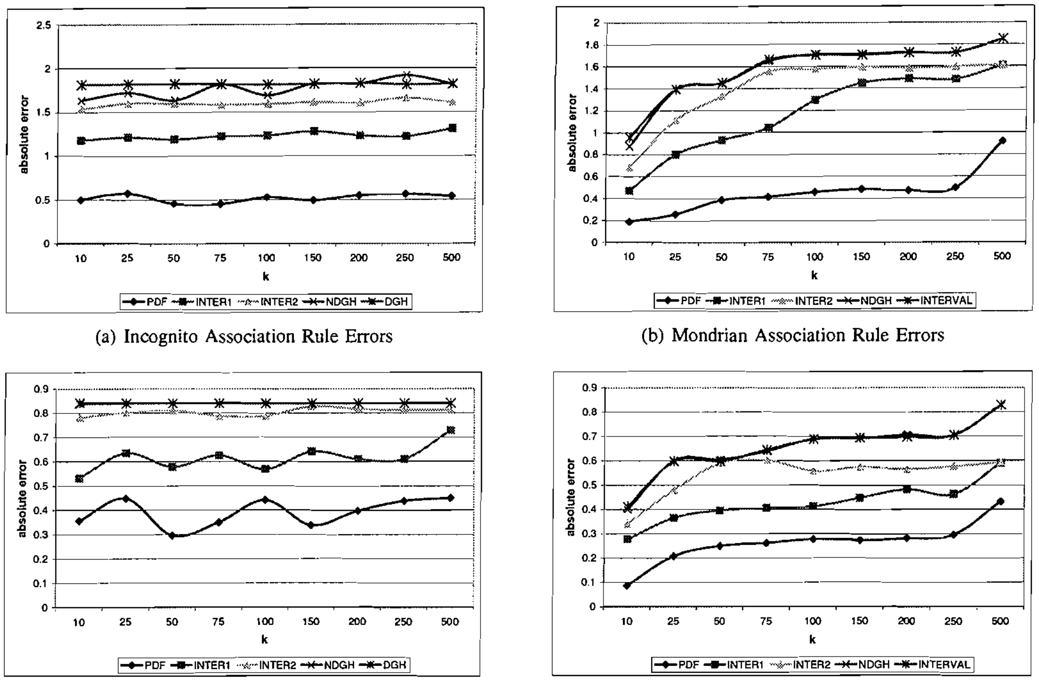
We first run association rule mining as the data mining application on the reconstructions. From each reconstruction, we extracted set of rules with confidence higher than 0.8. (0.8 was used in [13] and we observe 0.8 is a good minimum confidence level to get meaningful rules from the adult dataset.) As done in [13], before mining for association rules, to get meaningful rules, we removed the attributes *workclass*, *race*, and *native-country* since the majority of the entries in the database have the same value for these attributes. The set of rules R^o from original dataset and the set of rules R^r from reconstructed datasets created with minimum confidence c were compared with the following distance metric:

Let $C_{r,R}$ be the function that returns the confidence of rule r in R if $r \in R$, and returns 0.8 if $r \notin R$.

$$|R^o - R^r| = \sum_{r \in R^o \cup R^r} |C_{r,R^o} - C_{r,R^r}| \tag{5}$$

Informally distance metric above sums up the absolute difference of confidence levels of the same rule for two different sets of rules (assuming the minimum confidence for non-existing rules). We will name the distance between the ruleset of a particular reconstruction and the ruleset of the original distribution as the *absolute error* of the reconstruction.

Figure 3(a) and 3(b) show absolute errors of PDF, INTER1, INTER2, and NDGH reconstructions with respect to algorithms Incognito and Mondrian. As stated in Section IV, utility-optimal PDF reconstruction is much closer to the original dataset in terms of association rules supported. As



(a) Incognito Association Rule Errors

(b) Mondrian Association Rule Errors

(c) Incognito Class Rule Errors

(d) Mondrian Class Rule Errors

Fig. 3. Association and Class Rule Mining Results for k-Anonymity

figure

PDF distributions get closer to uniform distribution, the error increases for nearly all k values.

To measure classification accuracy, we conducted experiments by using decision tree classifiers. PDF reconstructions were better in terms of classification errors but not significantly. Since decision tree algorithms are very resistant to outliers, we also measured the algorithms' confidence on the created classification models by mining the class rules. Figures 3(c) and 3(d) plot the absolute errors for such rules. Similar behavior as in the case of association rule graphs suggests that utility-optimal PDF shows the relation between class values and QI attributes better than the other generalization types.

B. PDF for δ-Presence

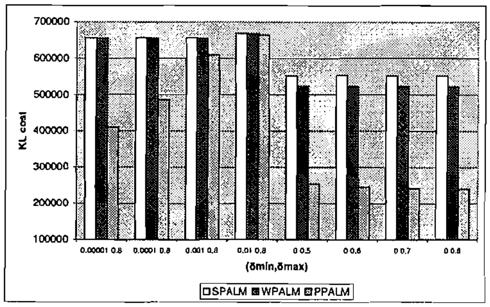
This section presents experiments regarding privacy - utility relations when using PDF generalizations in δ-presence framework. 3 different δ-presence algorithms are compared with respect to utilization of the output anonymizations and execution time: SPALM, previously proposed δ-presence algorithm [12]; PPALM, PDF δ-presence algorithm presented in Section V; and WPALM, weak version of PPALM without the speed up approach given in Section V-C. [2]

[2]WPALM is included in the experiments to show the effectiveness of the speed up process of Section V-C.

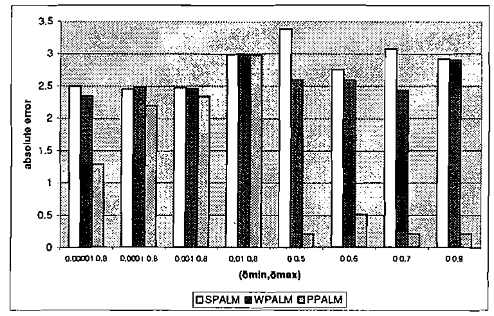
As mentioned in previous sections, both WPALM and PPALM tries to shift uniform distribution of data values given in the output of SPALM towards the utility optimal distribution without violating δ-presence. For WPALM and PPALM, we set the maximum no of steps (mxs) to 10 for the experiments. Each shift triggers a check if presence property still holds. As described in Section V-B, the checking is very costly for WPALM (time required by the checking is exponential in the size of equivalence classes, see Section V-B). 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 used the diabetes dataset prepared and used in [12] which contains a public dataset of size 45222 tuples and a private table of size 1957. ($\delta_{min} < 0.043 < \delta_{max}$ needs to hold on the constraints.) 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:

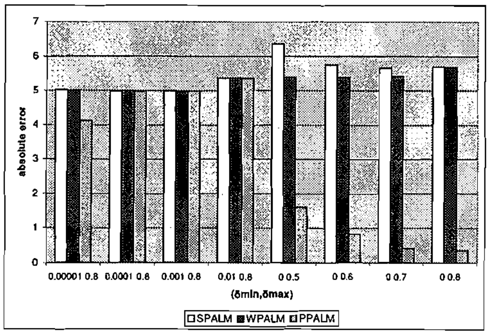
1) How effective are the proposed WPALM & PPALM algorithms compared to the SPALM algorithm in terms of data utilization?
2) How efficient are the proposed PPALM algorithm com-



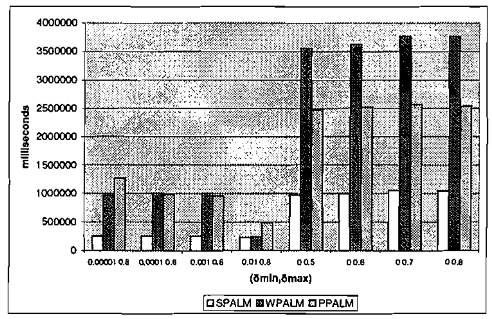
(a) KL Costs



(b) Class Rule Errors



(c) Association Rule Errors



(d) Execution Times

Fig. 4. Comparison of SPALM,WPALM, and PPALM

figure

TABLE VIII. Percentage of dataset processed by WPALM for varying δ values

table

.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%

pared to the WPALM & SPALM algorithms?

1) The Effectiveness of the WPALM & PPALM versus SPALM in terms of data utilization: We conducted experiments to compare the output utilizations of SPALM, WPALM and PPALM w.r.t. both the KL cost metric (Theorem 1) and efficiency in data mining applications, association rule mining and classification rule mining. Data mining operations on the output were carried out as described in Section VI-A.

Figure 4(a) shows the KL cost of the output anonymizations for SPALM, WPALM, and PPALM for various δ_{min} & δ_{max} intervals. WPALM improves SPALM in terms of utility however improvement introduced is not significant due to the large number of ignored ECs. On the other hand, PPALM introduces a great increase in the utilization by a factor of 3 at times. Improvement is more observable for larger δ intervals. The reason for this is that single dimensional assumption for algorithm SPALM is too strict and does not add enough information content into the output anonymization even when we lower the δ constraints. This leaves room for PPALM to inject utilization into the anonymization. Increasing δ_{max} beyond 0.5 add little utilization into pdf anonymizations. This is because anonymization mapping does not change after $\delta_{max} = 0.5$ and PPALM achieves (almost) utility optimal distribution for $\delta_{max} = 0.5$ meaning full distribution shifting occurs in all ECs. This is one more indication that lower and upper boundaries, calculated by PPALM, on exact existence probabilities are tight enough to get the maximum utilization out of pdf anonymizations.

The data mining results given in Figures 4(b) and 4(c) justify cost metric results. Error rates in finding association rules and classification rules from output anonymizations correlates with the KL costs of the anonymizations.

2) The Efficiency of the WPALM, PPALM & SPALM: We conducted a set of experiments to compare the running times of SPALM, WPALM and PPALM on a Core2duo 3GHz Linux computer with 3GB of RAM. The running times for various δ_{min} & δ_{max} configurations can be seen in Figure 4(d). As expected, SPALM is the algorithm with the shortest running time requirement, since it acts as a subroutine for the other two algorithms. PPALM requires more time than SPALM due to the post processing for shifting distribution towards utility optimal. However additional time cost is realistic and scales well with the length of the δ intervals. In most experiments, WPALM requires more execution time compared to PPALM even though it does not process most of the ECs. Table VIII shows the percentage of the database ignored by WPALM. Majority of the tuples (90+%) were ignored by WPALM. Besides as we force WPALM to process more equivalence classes, execution time for WPALM 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. To process an equivalence class of this size will require WPALM to make around 472 million multiplications. Roughly speaking WPALM will run 1345 times slower to process an additional 0.084% of the whole data.

Even though ideal WPALM acts as an upper bound for PPALM in terms of utilization, experiments in this section along with the previous section shows 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 with the size of EC groups. In reasonable settings PPALM is faster than WPALM with 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.

VII. Conclusions

We presented pdf generalizations that embed probability distributions into generalizations enabling a better control over the trade off between privacy and utility. We proposed pdf algorithms to provide δ-presence. The experiments showed that use of pdfs can increase utilization without violating privacy constraints.

There remains issues that are not addressed in this paper. First is that WPALM and PPALM algorithms are vulnerable to *reversibility* attacks by an adversary that knows the algorithm. (Such an adversary can reverse engineer the execution of the algorithm to gain more knowledge about the data.) It should be noted that such an attack is also possible (if not as easy as in here) for most algorithms proposed so far on k-anonymity and δ-presence. In [1], this problem was weakened by releasing reconstructions instead of anonymizations. Designing anonymization algorithms resistant to reversibility attacks is a nice research direction which is currently being studied by the authors. Authors also work on the evaluation of the PPALM w.r.t. varying input parameters and investigate new trade offs between the efficiency and accuracy.

References

[1] C. C. Aggarwal and P. S. Yu, "A condensation approach to privacy preserving data mining," in *EDBT'04*, Heraklion, Crete, Greece, Mar. 14 2004, pp. 183–199.

[2] K. W. B. Fung and P. Yu, "Top-down specialization for information and privacy preservation," in *Proc. of the 21st Int'l Conf. on Data Engineering*, 2005.

[3] R. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *Proc. of the 21st Int'l Conf. on Data Engineering*, 2005.

[4] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

[5] V. Iyengar, "Transforming data to satisfy privacy constraints," in *Proc., the Eigth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2002, pp. 279–288.

[6] 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.

[7] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. of the 2005 ACM SIGMOD Int'l Conf. on Management of Data*, Baltimore, MD, June 13-16 2005.

[8] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Multidimensional k-anonymity," University of Wisconsin, Madison, Tech. Rep. 1521, June 2005. http://www.cs.wisc.edu/techreports/2005/TR1521.pdf

[9] N. Li and T. Li, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proceedings of the 23nd International Conference on Data Engineering (ICDE '07)*, Istanbul, Turkey, Apr. 16-20 2007.

[10] 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.

[11] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Proc. of the 22nd IEEE Int'l Conf. on Data Engineering (ICDE 2006)*, Atlanta Georgia, Apr. 2006.

[12] 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.

[13] M. E. Nergiz and C. Clifton, "Thoughts on k-anonymization," in *ICDEW '06: Proc. of the 22nd Int'l Conf. on Data Engineering Workshops*. Washington, DC, USA: IEEE Computer Society, 2006, p. 96.

[14] M. E. Nergiz, C. Clifton, and A. E. Nergiz, "Multirelational k-anonymity," in *Proceedings of the 23nd International Conference on Data Engineering (ICDE '07)*, Istanbul, Turkey, Apr. 16-20 2007.

[15] P. Samarati, "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, no. 6, pp. 1010–1027, 2001.

[16] L. Sweeney, "Guaranteeing anonymity when sharing medical data, the datafly system," in *Proc., Journal of the American Medical Informatics Association*. Hanley & Belfus, Inc., 1997.

[17] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, vol. 10, no. 5, 2002.

[18] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[19] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proceedings of 32nd International Conference on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, Sept. 12-15 2006. http://www.vldb.org/conf/2006/p139-xiao.pdf

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

$$\begin{aligned} \mathcal{P}_F &= \Big(\prod_{v_i \in D_{A+1}} c_{A+1}^i!\Big) \cdot \Big(\prod_{a=1}^{A} \prod_{v_i \in D_a} (f_a^i)^{c_a^i}\Big) \\ &= C_1 \cdot \prod_{a=1}^{A} \prod_{v_i \in D_a} (f_a^i)^{c_a^i} \end{aligned}$$

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

$$\begin{aligned} &\ln \mathcal{P}_F \\ &= C_2 + \sum_{a=1}^{A} \sum_{v_i \in D_a} c_a^i \cdot \ln f_a^i \end{aligned}$$

This is nothing but the negative KL cost given in Eqn 1, so this proves Theorem 1. For a fixed equivalence class, the F function that maximizes Eqn 1:

$$\begin{aligned} &\max_F(\ln \mathcal{P}_F) \\ &= C_2 + \sum_{a=1}^{A} \max_{f_a}\Big(\sum_{v_i \in D_a} c_a^i \cdot \ln f_a^i\Big) \end{aligned}$$

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;

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

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

$$\begin{aligned} \frac{c_a^1}{f_a^1} - \frac{c_a^{n_a}}{1 - f_a^1 - \cdots - f_a^{n_a-1}} &= 0 \\ &\vdots \\ \frac{c_a^{n_a-1}}{f_a^{n_a-1}} - \frac{c_a^{n_a}}{1 - f_a^1 - \cdots - f_a^{n_a-1}} &= 0 \qquad (6) \\ c_a^1 \cdot \sum_{i=1}^{n_a-1} f_a^i + c_a^{n} f_a^1 &= c_a^1 \\ &\vdots \\ c_a^{n_a-1} \cdot \sum_{i=1}^{n_a-1} f_a^i + c_a^{n_a} f_a^{n_a-1} &= c_a^{n_a-1} \end{aligned}$$

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 6, we get, for $1 \leq i \leq n_a$;

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

Above equality maximizes the matching probability. □

Since there is no other root that makes the derivatives in Eqn 6 zero, matching probability monotonically decreases as each f_a^i gets far away from the utility-optimal distribution. This proves the correctness of Theorem 3.

Appendix II
. 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 6: Given a likelihood set $P = \{p^{low}, p^{high}, p_1, \cdots, p_m\}$ and the no. of present tuples n^1, if $p^{low} < p^{high}$, then $ex^{low} \le ex^{high}$.

PROOF. Difference between two existence probabilities would be

$$
\begin{aligned}
& ex^{low} - ex^{high} \\
= & \frac{\sum_{\substack{S\subset P \wedge |S|=n^1 \wedge \\ p^{low}\in S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S} - \frac{\sum_{\substack{S\subset P \wedge |S|=n^1 \wedge \\ p^{high}\in S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S} \\
= & \frac{\sum_{\substack{S\subset P \wedge |S|=n^1 \wedge \\ p^{low},p^{high}\in S}} P_S + \sum_{\substack{S\subset P \wedge |S|=n^1 \wedge \\ p^{low}\in S \wedge p^{high}\notin S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S} \\
& - \frac{\sum_{\substack{S\subset P \wedge |S|=n^1 \wedge \\ p^{low},p^{high}\in S}} P_S + \sum_{\substack{S\subset P \wedge |S|=n^1 \wedge \\ p^{high}\in S \wedge p^{low}\notin S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S}
\end{aligned}
$$

Since $\sum_{p\in S\wedge|S|=n^1} P_S = p \sum_{p\notin S\wedge|S|=n^1-1} P_S$;

$$
\begin{aligned}
= & \frac{p^{low} \sum_{\substack{S\subset P \wedge |S|=n^1-1 \wedge \\ p^{low},p^{high}\notin S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S} - \frac{p^{high} \sum_{\substack{S\subset P \wedge |S|=n^1-1 \wedge \\ p^{low},p^{high}\notin S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S} \\
= & \frac{(p^{low} - p^{high}) \sum_{\substack{S\subset P \wedge |S|=n^1-1 \wedge \\ p^{low},p^{high}\notin S}} P_S}{\sum_{S\subset P\wedge|S|=n^1} P_S}
\end{aligned}
$$

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 no. of present tuples n^1, let $p^{min} \le p_i \le p^{max}$ for $i \in [1-m]$. If $ex^{min} \ge \delta_{min}$ and $ex^{max} \le \delta_{max}$ then $\delta_{min} \le ex \le \delta_{max}$ for any $ex \in EX$.

PROOF. By Theorem 6, $\delta_{min} \le ex^{min} \le ex_i \le ex^{max} \le \delta_{max}$ for all i. □

Appendix III
. Finding Upper and Lower Bound 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 doesn't change) and existence probabilities for the rest of the tuples decrease (or do not change).

Theorem 7: Given the no. 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) $ex^{low} \le ex^{high}$
2) $ex_i^1 \ge ex_i^2$ for all $i \in [1-m]$.

PROOF. We first proof item 1. The difference between existence probabilities ex^{low} and ex^{high} is as follows:

$$\begin{aligned}
& ex^{low} - ex^{high} \\
= & \frac{\sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \in S}} P_S}{\sum_{S \subset P^1 \wedge |S|=n^1} P_S} - \frac{\sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \in S}} P_S}{\sum_{S \subset P^2 \wedge |S|=n^1} P_S} \\
= & \frac{p^{low} \sum_{\substack{S \subset P^1 \wedge |S|=n^1-1 \wedge \\ p^{low} \notin S}} P_S}{\sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \in S}} P_S + \sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \notin S}} P_S} - \frac{p^{high} \sum_{\substack{S \subset P^2 \wedge |S|=n^1-1 \wedge \\ p^{high} \notin S}} P_S}{\sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \in S}} P_S + \sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \notin S}} P_S}
\end{aligned}$$

Setting

$$\begin{aligned}
C_1 &= \sum_{\substack{S \subset P^1 \wedge |S|=n^1-1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S|=n^1-1 \wedge \\ p^{high} \notin S}} P_S \\
C_2 &= \sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \notin S}} P_S
\end{aligned}$$

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

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

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;

$$\begin{aligned}
& ex_i^1 - ex_i^2 \\
= & \frac{\sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p_i^1 \in S}} P_S}{\sum_{S \subset P^1 \wedge |S|=n^1} P_S} - \frac{\sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p_i^2 \in S}} P_S}{\sum_{S \subset P^2 \wedge |S|=n^1} P_S} \\
= & \frac{\sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low}, p_i^1 \in S}} P_S + \sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \notin S \wedge p_i^1 \in S}} P_S}{\sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \in S}} P_S + \sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \notin S}} P_S} - \frac{\sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high}, p_i^2 \in S}} P_S + \sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \notin S \wedge p_i^2 \in S}} P_S}{\sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \in S}} P_S + \sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \notin S}} P_S}
\end{aligned}$$

Setting

$$\begin{aligned}
C_1 &= \sum_{\substack{S \subset P^1 \wedge |S|=n^1-1 \wedge \\ p^{low} \notin S \wedge p_i^1 \in S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S|=n^1-1 \wedge \\ p^{high} \notin S \wedge p_i^2 \in S}} P_S \\
C_2 &= \sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \notin S \wedge p_i^1 \in S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \notin S \wedge p_i^2 \in S}} P_S \\
C_3 &= \sum_{\substack{S \subset P^1 \wedge |S|=n^1-1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S|=n^1-1 \wedge \\ p^{high} \notin S}} P_S \\
C_4 &= \sum_{\substack{S \subset P^1 \wedge |S|=n^1 \wedge \\ p^{low} \notin S}} P_S = \sum_{\substack{S \subset P^2 \wedge |S|=n^1 \wedge \\ p^{high} \notin S}} P_S
\end{aligned}$$

We have;

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

Denominator is definitely positive.The first additive component of the numerator is positive by the assumption. We now prove the second component $(C_3C_2 - C_1C_4)$ is also positive. Setting $P' = P^1 - p^{high} - p_i^1$, $P'' = P^1 - p^{high}$; C_1C_4 and C_2C_3 can be written as summation of likelihood products;

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

$$C_2 C_3$$

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

Let, without loss of generality, in all the additive terms of C_1C_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_1C_4 also exist as an additive term in C_2C_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_2C_3 has more additive terms than C_1C_4. So $C_2C_3 - C_1C_4$ is also non-negative. □

Theorem 7 also implies that if the likelihood probability of a tuple is decreased, its existence probability also decreases (or does not change) and existence probabilities for the rest of the tuples increase (or do not change).

Theorem 5: Given the no. 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 7 and the assumption, $\delta_{min} \leq (ex^{\downarrow})^{min} \leq ex^{min}$. Again by Theorem 7, $ex^{max} \leq (ex^{\uparrow})^{max} \leq \delta_{max}$. □