Computer Science Technical Reports Department of Computer Science

2008

S-Presence Without Complete World Knlowedge

M. Ercan Nergiz

Chris Clifton
Purdue University, clifton@cs.purdue.edu

Report Number:
08-016

S-Presence Without Complete World Knlowedge

M. Ercan Nergiz
Chris Clifton

CSD TR #08-016
June 2008

δ-Presence Without Complete World Knowledge

M. Ercan Nergiz, *Student Member, IEEE,*
Chris Clifton, *Senior Member, IEEE*

Abstract—Advances in information technology, and its use in research, are increasing both the need for anonymized data and the risks of poor anonymization. [1] presented a new privacy metric, δ-presence, that clearly links the quality of anonymization to the risk posed by inadequate anonymization. It was shown that existing anonymization techniques are inappropriate for situations where δ-presence is a good metric (specifically, where *knowing an individual is in the database* poses a privacy risk). This article addresses a practical problem with [1], extending to situations where the data anonymizer is not assumed to have complete world knowledge. The algorithms are evaluated in the context of a real-world scenario, demonstrating practical applicability of the approach.

Index Terms—*k*-anonymity, privacy, delta presence, medical databases

Note to reviewers: This work extends the work in [1]. Around 70% of the paper (including Sections 4,5,7,8,9, and part of 6) is new material. The material carried over from [1] has been revised and summarized, and is needed to make this submission a standalone paper.

1 Introduction

The increasing ability to collect, manage, and share information is raising every-increasing privacy concerns. This poses a challenging tradeoff between the value (both to society, and to individuals) from the knowledge available from ubiquitous, shared information, and the risk to individuals posed by disclosure and misuse of private data.

One solution to this problem is *anonymity*: ensuring that disclosed data cannot be linked to the individual whom the data is about. The European Community Directive 95/46/EC protects 'personal data':

> 'personal data' shall mean any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;

This lends credence to using anonymity to protect privacy. The United States Healthcare Information Portability and Accountability Act (HIPAA) [2] protects 'individually identifiable data', and allows disclosure of data that has been de-identified. But what does it mean to be 'de-identified'?

> Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.

How do we interpret these rules with respect to anonymity? Is it enough to say that if we cannot positively identify a record as belonging to an individual, it is suitably anonymous? What if we can identify the individual with 90% probability? The U.S. HIPAA rules do give some guidance: if someone applying generally accepted statistical and scientific principles "determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information". While this could be interpreted as data is de-identified if the recipient could not be absolutely certain a record applied to an individual, the regulations give further guidance suggesting that de-identification can be accomplished by removing not only identifying numbers/names/images, but also geographic information that limits granularity to less than 20,000 individuals or dates more specific than the year. This implies that identification with high probability, even if less than 100%, would probably not be considered suitably de-identified.

An alternative view is to look at the risk posed by disclosure of information. It is easy to see that anonymity is not enough; for example, suppose we use *k*-anonymity to protect data [3], [4]. This says that knowing identifying information about an individual, there are at least *k* records in the database that could (with equal probability) refer to that individual. However, suppose that those records also include sensitive information, e.g., if an individual is diabetic. If all *k* individuals share the same value for the sensitive information (e.g., all are diabetic), then *k*-anonymity provides no protection against disclosure of that fact. This has lead to alternate approaches, such as discernibility [5] / ℓ-diversity [6]. However, it is still difficult to answer the question, "is the data anonymous enough?"

This paper looks at a basic, and yet common and practical, problem: the risk is simply from identifying that an individual is (or is not) in an anonymized dataset. This could occur when there is a desire to publish a dataset to support research on a specific condition, but identifying individuals meeting that condition is damaging. Examples could range from counter-terrorism, publishing a database containing information about suspected terrorist groups to support research in automated support for discoverying terrorism; to medical research, such as a database of patients with a particular type of cancer. In both cases, identifying that an individual is present in the database is damaging, both to the individual, and in the terrorism example by disclosing to real terrorist groups that their "cover organization" is suspect (or not suspected).

The basic idea (introduced in [1]) is that anonymizing such a database should mean that a recipient of the database should not be able to identify any individual as being in that database

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

with certainty greater than δ. This is actually the primary value of anonymization; anonymizing to protect against linking an individual with sensitive data *in* the released dataset can be done just as effectively without anonymization [7]. As we shall see, this δ-presence measure has the nice property that it can be interpreted in terms of increased risk of disclosure. This enables a meaningful bridge between human-understandable policy and mathematically sound standards for anonymity. Another, perhaps surprising, outcome is that the k-anonymity approach is a *bad* way to meet this standard; requiring a substantial and unnecessary loss of detail in the anonymized data. Unfortunately, while k-anonymity (and related approaches) can be checked knowing only the data to be released, evaluating δ-presence requires complete knowledge of all entities, not just those in the dataset to be anonymized. This paper addresses the issue, presenting approaches that enable the δ-presence measure to be used to anonymize a dataset assuming knowledge of only global *distributions* of data.

1.1 Example: Diabetes

During the paper, we will use the "medical research dataset" problem as a running example. Diabetes is an expensive and widespread health problem, representing 11% of U.S. health care expenditures [8]. Of note is that people with diabetes have medical expenditures 2.4 times the expenditures if they did not have diabetes [9]; under the "employer pays" system used at most large U.S. companies, this would certainly be an incentive for an employer to (illegally) discriminate against hiring someone with diabetes. As we can see, it is clear that there is both great value in making data available to support research on diabetes, and a clear need to protect the individuals in such data.

Take the Diabetes dataset from the UCI machine learning repository [10] as an example. This contains data on 70 patients. What is a reasonable risk of identifying an individual as being in this dataset? At first glance, we might say that we don't want an adversary to be able to identify with certainty greater than a random guess: $70/260,000,000$ (the size of the dataset divided by the number of individuals in the U.S. in 1994), or 0.000027%. However, if we look at the larger problem, we realize that the risk is identifying that an individual *has* diabetes. As 7% of the U.S. population has diabetes [8], even without the anonymized dataset an adversary would know the probability that an individual has diabetes is much greater than 0.000027%. The real question is, how much could the anonymized database improve the adversary's estimate of the probability that an individual has diabetes? The "no better than a random guess" standard is clearly too conservative.

It is hard to address the issues stated above without assuming some knowledge about the world from which the private dataset was drawn from. The original definition of δ-Presence in [1] assumes that the data anonymizer has complete information about the world in the form of a public database (and allows the adversary to have the same knowledge.) This is perhaps unrealistic; more appropriate is to assume the anonymizer has partial knowledge about the outside world such as statistics, limited selections/projections, count queries, $\cdots$. Such information is not privacy sensitive and likely to be publicly available (or might be publicized upon request). In this paper, we revisit δ-presence when only distributions for attributes are known by the data owner and the connections between distinct attributes is not known. (E.g., the data owner knows exactly how many male-female, single-married, young-old $\cdots$ people there are in the outside world; but does not know if there is a married, young woman.[1]) To ensure privacy, we still keep the strong adversary assumption; the adversary might have access to the whole world knowledge (except the presence/absence of individuals in the private dataset) in the form of a public database. This way, we guarantee an upper bound on privacy at a user selected confidence level.

From the view of a data owner that has access to attribute distributions only, the outside world acts as a random variable. Given each possible value from the sample space, a given anonymization satisfies different levels of δ constraints. A given δ-presence constraint can only be satisfied at a certain confidence level. In Section 4, we formulate and show how to check for *c-confident* δ*-presence* from a given anonymization and attribute distributions. In Section 5, we also show the checking process is computationally expensive and present certain heuristics to speed up the operations.

We now give background and notations used in the paper, based on the original definition of δ-presence in [1]; Section 3, briefly summarizes that work. In Section 4, we reformalize δ-presence assuming the anonymizer knows only attribute distributions. Section 8 gives a set of experiments evaluating the performance of the checking process.

2 Background and Notation

Before formalizing the problem of hiding presence of individual from a given database, we give some basic notation and review the original *k-anonymity* framework.

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 and $T[.][r]$ refers to selection of row r on T (the rth tuple or record). We write $|t \in T|$ for the cardinality of tuple $t \in T$.

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

Although there are many ways to generalize a given value, in this paper, we will stick to generalizations according to DGH structures given in Figure 1. (e.g., $\psi(\text{USA}) = \{$USA, N. America, America, *$\}$) We will also write, for tuples t and t^*, $t^* \in \psi(t)$ when $t^*[i] \in \psi(t[i])$ for all possible index i.

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

In Tables 1-2, tables P_2^* and P_3^* are different generalizations of table P. (The T tables will be discussed in Section 3.2, and should be ignored for now.)

Definition 3 (Non-Overlapping Generalization):
Given a private table T and a generalization T^* of T, we say T^* is non-overlapping if and only if there does not exist $t_1^*, t_2^* \in T^*$ and any possible tuple t such that $t_1^* \neq t_2^*$ and $t_1^* \in \psi(t)$, $t_2^* \in \psi(t)$ In other words, a generalization is non-overlapping when any possible tuple can match at most one generalized tuple. In Tables

[1] unless implied by the attribute distributions

TABLE 1
Public dataset P and research subset T

P

	Publicly Known Data				
	Name	Zip	Age	Nationality	*Sen.*
a	Alice	47906	35	USA	*0*
b	Bob	47903	59	Canada	*1*
c	Christine	47906	42	USA	*1*
d	Dirk	47630	18	Brazil	*0*
e	Eunice	47630	22	Brazil	*0*
f	Frank	47633	63	Peru	*1*
g	Gail	48973	33	Spain	*0*
h	Harry	48972	47	Bulgaria	*1*
i	Iris	48970	52	France	*1*

T

	Research Subset		
	Zip	Age	Nationality
b	47903	59	Canada
c	47906	42	USA
f	47633	63	Peru
h	48972	47	Bulgaria
i	48970	52	France

$D = P - T$

	Negative Subset		
	Zip	Age	Nationality
a	47906	35	USA
d	47630	18	Brazil
e	47630	22	Brazil
g	48973	33	Spain

(Initial "key" columns for clarity only; Sen. represents sensitive data not publicly known.)

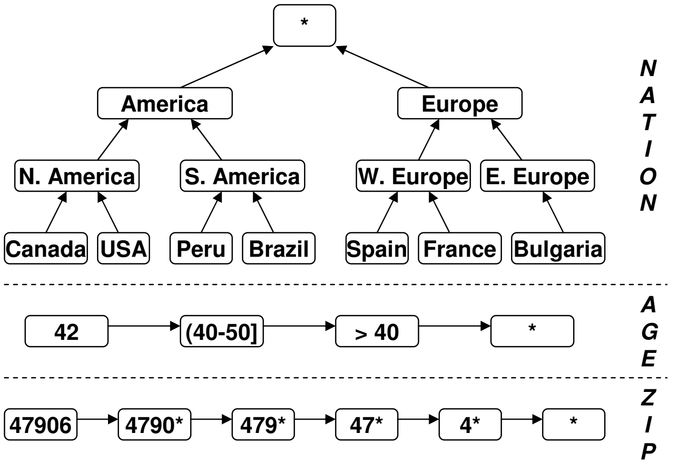


Fig. 1. DGH structures

1 and 2; datasets P_2^* and P_3^* are non-overlapping generalizations of P. Similarly T_3^* is such a generalization of T.

Definition 4 (Same Mapping Generalizations): Let T_1^* and T_2^* be two non-overlapping generalizations of tables T_1 and T_2 respectively. We say T_1^* and T_2^* have the same generalization mapping if the following holds;
for all quadruple $(t_1^* \in T_1^*) \rightleftharpoons (t_1 \in T_1)$ and $(t_2^* \in T_2^*) \rightleftharpoons (t_2 \in T_2)$, if $t_1^* \in \psi(t_2)$ then $t_2^* = t_1^*$ and similarly if $t_2^* \in \psi(t_1)$ then $t_1^* = t_2^*$.

In Table 1 and 2; generalization P_3^* of P, and T_3^* of T are same mapping generalizations.

Definition 5 (k-Anonymity): A table T^* is k-anonymous w.r.t. a set of attributes QI if each record in $T^*[QI]$ appears at least k times.

The idea behind this definition is the following; each record in the private dataset contains publicly available information in some attributes QI (*quasi-identifiers*). The values of these attributes can be exploited to (almost uniquely) link those records to records in other tables. The goal of k-anonymity is to limit an adversary's ability of linking a record from a set of released records to a specific individual. (E.g., for dataset P in Table 1, attributes Zip, Age, Nationality can be considered as QI attributes. Attribute Sen. can be considered as sensitive. Dataset P_2^* of Table 2 is a 3-anonymous generalization of P. Note that by only seeing P_2^*, an adversary can at best link a tuple <47906,35,USA>, Alice, to the tuples a, b, and c of P_2^*.)

Definition 6 (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.
In dataset P_2^*, the equivalence class for tuple a is $\{a, b, c\}$.

3 δ-Presence given Public Database

In this section, we summarize and reformalize δ-presence from [1]. Next sections will build upon the ideas given in this section.

3.1 Public Table Assumption

In this framework the adversary is presumed to have access to all publicly known data (represented in a, possibly huge, *public table P*) that links names to other set of attributes (e.g., day of birth, sex, race.) When a data holder (e.g., a medical institution) releases a (private) research subset data, the adversary can match quasi-identifiers (or common attributes) in both tables to discover unique links between records in the public and released table. Given that being linked to the research subset is a privacy risk, we instantly see that releasing the research subset T in Table 1 is not acceptable; each individual can be uniquely linked with the publicly known data. Though by releasing some generalization of T, T^*; the risk can be controlled. Thus framework can be summarized as follows:

Publisher View: P, T
Adversary View: P, T^*

The challenge is to find a suitable anonymization T^* of T that will limit the risk. [1] shows that neither k-anonymization nor enforcing ℓ-diversity solve the problem.

We now give a definition for δ-presence, a metric to evaluate the risk of identifying an individual in a table based on generalization of publicly known data.

Definition 7 (δ-Presence): Given an external public table P, and a private table T, we say that δ-*presence* holds for a generalization T^* of T, with $\delta = (\delta_{min}, \delta_{max})$ if

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

In such a dataset, we say that each tuple $t \in P$ is δ-*present* in T or we say the *existence probability* of t is within δ. Therefore, $\delta = (\delta_{min}, \delta_{max})$ is a range of acceptable probabilities for existence probability $\mathcal{P}(t \in T \mid P, T^*)$. From now on, we assume $T \subseteq P$.

TABLE 2
P_2^*: $(2,2)$ recursive diverse P; T_3^*: $(\frac{1}{2},\frac{2}{3})$ present T

P_2^*

	Public Dataset			Sen.
	Zip	Age	Nationality	
a	4790*	*	N. America	0
b	4790*	*	N. America	1
c	4790*	*	N. America	1
d	4763*	*	S. America	0
e	4763*	*	S. America	0
f	4763*	*	S. America	1
g	4897*	*	Europe	0
h	4897*	*	Europe	1
i	4897*	*	Europe	1

P_3^*

	Public Dataset			Sen.
	Zip	Age	Nationality	
a	47*	*	America	0
b	47*	*	America	1
c	47*	*	America	1
d	47*	*	America	0
e	47*	*	America	0
f	47*	*	America	1
g	48*	*	Europe	0
h	48*	*	Europe	1
i	48*	*	Europe	1

T_3^*

	Research Subset		
	Zip	Age	Nationality
b	47*	*	America
c	47*	*	America
f	47*	*	America
h	48*	*	Europe
i	48*	*	Europe

$D_3^* = P_3^* - T_3^*$

	Negative Subset		
	Zip	Age	Nationality
a	47*	*	America
d	47*	*	America
e	47*	*	America
g	48*	*	Europe

Before we show how to check for the δ-Presence property, we define two more properties:

3.2 Checking for δ-Presence

Given the public dataset, it is trivial to check a non-overlapping generalizations of the private dataset for δ-presence. Given a non-overlapping generalization T^* of T and a public dataset P, construct the generalization P^* of P with the same mappings used in T^*. For any tuple $t \in P$, let t^* be its generalization in P^* (e.g., $(t^* \in P^*) \rightleftharpoons (t \in P)$). Then the presence (or existence) probability for t takes the following form:

$$\mathcal{P}(t \in T \mid P, T^*) = \frac{|t^* \in T^*|}{|t^* \in P^*|} \qquad (1)$$

If the existence probability is within the δ parameters for all tuples $t \in P$ then the presence property holds for T^*.

In Tables 1 and 2, dataset T_3^* shows a $(\frac{1}{2},\frac{2}{3})$-present generalization of T w.r.t. public dataset P. $\mathcal{P}(\texttt{tuple}\ a \in T \mid T_3^*) = \frac{|\{b,c,f\}|}{|\{a,b,c,d,e,f\}|} = \frac{1}{2}$. The same probability holds for tuples b,c,d,e, and f. Probability for tuples g,h, and i is $\frac{|\{h,i\}|}{|\{g,h,i\}|} = \frac{2}{3}$.

The probabilities can also be calculated by making use of the attribute Sen. in P^*s. (see dataset P_3^* of Table 2) For each tuple t in an equivalence class $|t^* \in T^*|$ is the number of entries with Sen.=1, and $|t^* \in P^*|$ is the size of the equivalence class. Checking process can be done by processing each equivalence class of P^* and by using attribute Sen..

In [1], it is discussed in detail that previous anonymization approaches such as k-anonymity, ℓ-diversity, or t-closeness do not provide mechanisms to check for δ-presence. This is mainly because public dataset P is not taken into account for the anonymization of T. Even when applied on P^*, diversity enforcing privacy mechanisms (ℓ-diversity, t-closeness) can not be used to enforce constraints on Equation 1. Due to space limitations, we do not include further discussion and refer to [1].

[1] also proves the anti-monotonicity property for δ-presence:

Theorem 1: Given a public table P, private table T, a non-overlapping generalization T_1^* of T, and a non-overlapping generalization T_2^* of T_1^*. If T_2^* is not $(\delta_{min}, \delta_{max})$ present w.r.t. P and T then neither is T_1^*.

4 δ-Presence given Public Attribute Distributions

4.1 Public Attribute Distributions Assumption

Generally the party anonymizing the data does not have knowledge of the whole population, however it is reasonable to assume knowledge of statistics about the population. (Statistics over data are not individually identifiable and thus generally not considered private/protected information.) We now relax our assumption on the availability of a public table and instead assume that a set of attribute distribution functions is known by the publisher. We *keep* the assumption that adversary has access to the public table; assuming ignorance on the part of the adversary is not a good idea if we really want to make statements about privacy protection. We redefine δ-presence against such an adversary; this definition subsumes the previous definition that assumes a public table.

Definition 8 (Distribution Function): A distribution function f_A^D for a set of attributes $A = \{a_1, \cdots, a_n\}$ defined over a population D is a function that when given a set of values $\{v_1, \cdots, v_n\}$ returns the number of entities t in D with $v_i \in \psi(t[a_i])$ for $i \in [1-n]$.

If we assume that we have the population given in $D = P - T$ of Table 1 (tuples that do not exist in T), $f^D_{\text{Zip,Nationality}}(\text{‘47906’, ‘USA’}) = |\{Alice\}| = 1$ and $f^D_{\text{Nationality}}(\text{‘America’}) = |\{\text{Alice,Dirk,Eunice}\}| = 3$. Note that knowing $f^D_{\text{Zip,Age,Nationality}}$ is the same as knowing table D. Table 3 shows two other examples for sets of distribution functions; F_{2d} and F_{3d} ($f_C(c_2) = 3$).

Throughout this section, distribution functions will describe the part of the population that does not exist in the private dataset (the negative subset $D = P - T$) and will not contain superscripts (e.g., $f = f^D$). While it is more likely that distribution functions for P are known, the publisher knows T and can easily construct distribution functions of D from those of P. We also assume, without loss of generality, each function f describes only one attribute.

The new framework then looks like the following:

Publisher View: F, T
Adversary View: P, T^*

Having probabilistic information on the public dataset, the δ-presence property can now be achieved with a given confidence:

TABLE 3
Distribution Functions F_{2d} and F_{3d} for negative subset D and the corresponding probability space for D

Distribution Functions | Probability space for D (Last column shows the frequency of the corresponding tuple.)

F_{2d}

f_A	f_B
$1a_1$	$1b_1$
$3a_2$	$3b_2$

$D_1 : \frac{1}{4}$

A	B	
a_1	b_1	1
a_2	b_2	3

$D_2 : \frac{3}{4}$

A	B	
a_1	b_2	1
a_2	b_1	1
a_2	b_2	2

F_{3d}

f_A	f_B	f_C
$1a_1$	$1b_1$	$1c_1$
$3a_2$	$3b_2$	$3c_2$

$D_3 : \frac{1}{16}$

A	B	C	
a_1	b_1	c_1	1
a_2	b_2	c_2	3

$D_4 : \frac{3}{16}$

A	B	C	
a_1	b_1	c_2	1
a_2	b_2	c_1	1
a_2	b_2	c_2	2

$D_5 : \frac{3}{16}$

A	B	C	
a_1	b_2	c_1	1
a_2	b_1	c_2	1
a_2	b_2	c_2	2

$D_6 : \frac{3}{16}$

A	B	C	
a_1	b_2	c_2	1
a_2	b_1	c_1	1
a_2	b_2	c_2	2

$D_7 : \frac{6}{16}$

A	B	C	
a_1	b_2	c_2	1
a_2	b_1	c_2	1
a_2	b_2	c_1	1
a_2	b_2	c_2	1

Definition 9 (c-Confident δ*-Presence):* Given a public set of distribution functions F, a private table T, a confidence level $c \in [0-1]$, and a generalization T^* of T, let I_t be the event that tuple $t \in T$ is δ-present w.r.t. T^* and the whole population (which is unknown). In other words; I_t holds if $\delta_{min} \leq \mathcal{P}(t \in T \mid T^*) \leq \delta_{max}$. Note that I_t is a random event since public dataset P is a random variable. We say that δ-*presence* holds for T^*, with $\delta = (\delta_{min}, \delta_{max})$ and with confidence c if

$$\mathcal{P}(I_t \mid F) \geq c \qquad \forall\, t \in T \tag{2}$$

Informally, a c-confident δ-present anonymization ensures that a given tuple t is δ-present w.r.t. the current population with c probability. This definition has two important properties.

- Privacy is defined over tuples independently, meaning *each tuple* will be δ-present with c probability. An alternative definition would enforce the *anonymization* to be δ-present with c probability. The approach taken has advantages in estimating the cost of identification, especially when privacy is personalized. (E.g., some tuples with high cost of identification may require higher confidence levels.) The latter gives guaranteed for-all privacy with c probability and is harder to achieve. In this work, we stick with the former.
- Privacy is satisfied for only those tuples that are *in* the private dataset, not all tuples that are possibly in the public dataset. What this means is that while it is impossible to determine if $e \in T$ given T^* with greater than δ probability, it may be possible to determine that $e \notin T$. The reason is that possible outliers needs to be considered, particularly challenging given that the public data (and outliers) are not actually known. (E.g., if there is only one person with age > 100 in a big population, every attribute-join with age > 100 needs to be considered.) A better approach would consider the tuples in the private dataset plus *probable* tuples from the attribute distribution. In coming sections, we discuss how to convert the methodology presented in this work to handle such tuples.

As an example, assume we have the private table T given Table 4.1 and the set of distribution functions F_{3d} given in Table 3. To keep the discussion simple, the generalization we check for c-confident δ-presence is T itself. ($T^* = T$, that means also $D^* = D$.) Let us also assume $\delta_{min} = .33$, $\delta_{max} = 1$. Throughout the section, we will show that δ presence holds for T with $\frac{15}{16}$-confidence. For this, it is sufficient to show that for both tuples $t_1^* = t_1 =< a_1, b_1, c_1 >$, $t_2^* = t_2 =< a_2, b_2, c_2 >$, $\mathcal{P}(I_{t_i} \mid F_{3d}) \geq \frac{15}{16}$, $i \in [1-2]$. However we do not require, say a tuple $t_3^* = t_3 =< a_1, b_1, c_2 >$, to satisfy Equation 2 even though it is possible for t_3 to exist in the public dataset. (Actually confidence level for t_3 is intuitively zero, since $t_3 \notin T^*$, thus existence probability is zero.)

TABLE 4
Private Table for Section 4

Private Table T			
A	B	C	
a_1	b_1	c_1	1
a_2	b_2	c_2	1

We next show how to check for c-confident δ-presence for a given anonymization and distribution function.

4.2 Checking for c-Confident δ-Presence

We now show how to calculate the confidence given a δ parameter, a private table T, an anonymization T^* of T, and a set of distribution functions F (for $P - T$).

Let t be any tuple in T, then

$$\mathcal{P}(I_t \mid F) = \sum_{D} \mathcal{P}(I_t \mid D) \cdot \mathcal{P}(D \mid F) \tag{3}$$

From Section 3, we know how to calculate $\mathcal{P}(I_t \mid P)$.

$$\mathcal{P}(I_t \mid P) = \begin{cases} 1, & \delta_{min} \leq \frac{|t^* \in T^*|}{|t^* \in P^*|} \leq \delta_{max}; \\ 0, & \text{otherwise.} \end{cases}$$

Since we assume F describes $D = P - T$, we now rewrite the above for D.

$$\mathcal{P}(I_t \mid D) = \begin{cases} 1, & \delta_{min} \leq \frac{|t^* \in T^*|}{|t^* \in T^*| + |t^* \in D^*|} \leq \delta_{max}; \\ 0, & \text{otherwise.} \end{cases}$$

We are interested in only those Ds satisfying δ-presence. Setting the cardinality numbers $n_1 = |t^* \in T^*|$ and $n_0 = |t^* \in D^*|$, the requirement for Ds to make t δ-present:

$$\begin{aligned} \delta_{min} \leq &\quad \tfrac{n_1}{n_0+n_1} &\leq \delta_{max} \\ \frac{1}{\delta_{max}} \leq &\quad \tfrac{n_1+n_0}{n_1} &\leq \frac{1}{\delta_{min}} \\ \left\lceil \frac{1}{\delta_{max}} n_1 - n_1 \right\rceil \leq &\quad n_0 &\leq \left\lfloor \frac{1}{\delta_{min}} n_1 - n_1 \right\rfloor \end{aligned}$$

In other words, the cardinality of t^* in dataset D^* needs to be within some boundaries. Each boundary number can be calculated from the T^*. From now on for the sake of compactness, we use $c_{low} = \left\lceil \frac{1}{\delta_{max}} n_1 - n_1 \right\rceil$, and $c_{high} = \left\lfloor \frac{1}{\delta_{min}} n_1 - n_1 \right\rfloor$ for the boundary requirements for D^* given T^* and t.

In Table 3, we list all possible Ds given the set of distribution functions F_{3d}. Following the above example, suppose we check tuple $t^* = t =< a_2, b_2, c_2 >$ which have cardinality 1 in T^* for presence property. By the above equation, $c_{low} = \left\lceil \frac{1}{1} \cdot 1 - 1 \right\rceil = 0$ and $c_{high} = \left\lfloor \frac{1}{.33} \cdot 1 - 1 \right\rfloor = 2$. So t will be δ-present for all D with $0 \leq |t^* \in D^*| \leq 2$. Among 5 possible D, only D_4, D_5, D_6, and D_7 makes t δ-present in T^*.

Given above, we can rewrite Equation 3;

$$\mathcal{P}(I_t \mid F) = \sum_{x \in [c_{low} - c_{high}]} \mathcal{P}(|t^* \in D^*| = x \mid F) \tag{4}$$

which basically states the summation of likelihoods for each possible D (given F) satisfying the cardinality boundary condition gives us our confidence on the δ-presence of t w.r.t. T^*.

In Table 3, we also list likelihoods of all possible D given F_{3d}. Following the example above, $|t^* \in D^*|$ is 1 for D_7 and 2 for D_4, D_5, D_6, so the confidence level is $\frac{6}{16} + (\frac{3}{16} + \frac{3}{16} + \frac{3}{16}) = \frac{15}{16}$.

We now show how to calculate the likelihood of getting a D with $|t^* \in D^*| = x$. More precisely, we show how to solve for *x-cardinality likelihood*, $\mathcal{P}(|t^* \in D^*| = x \mid F)$.

The x-cardinality likelihood for T with n dimensions and a set of distributions $F = \{f_1, \cdots, f_n\}$ explaining a population of size U can be calculated recursively by starting with only one dimension and adding one dimension at a time. Let I_x^n be the event that $|t[1-n]^* \in D[1-n]^*| = x$ and let $\ell_x^n = \mathcal{P}(I_x^n \mid F)$ (ℓ_x^n is x-cardinality likelihood for T projected on the first n dimensions.), then

$$\begin{aligned} \ell_x^n &= \mathcal{P}(I_x^n \mid F) \\ &= \sum_y \mathcal{P}(I_x^n \mid I_y^{n-1}, f_n) \cdot \mathcal{P}(I_y^{n-1} \mid F) \\ &= \sum_y \mathcal{P}(I_x^n \mid I_y^{n-1}, f_n) \cdot \ell_y^{n-1} \end{aligned}$$

$\mathcal{P}(I_x^n \mid I_y^{n-1}, f_n)$ is the probability of selecting exactly x $t^*[n]$s (from the $f_n(t^*[n])$ available) next to y available $t^*[n-1]$s. This is a hypergeometric distribution[2]. More precisely;

$$\ell_x^n = \begin{cases} 1, & n = 1, x = f_1(t^*[1]); \\ 0, & n = 1, x \neq f_1(t^*[1]); \\ \sum\limits_{y \in [1-U]} \ell_y^{n-1} \cdot hyp(x; y, f_n(t^*[n]), U), & \text{otherwise.} \end{cases}$$

where hyp is the hypergeometric density function defined as

$$hyp(x; n, M, N) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, & x = 0, \cdots, n; \\ 0, & \text{otherwise.} \end{cases}$$

ℓ_x^n can be calculated by dynamic programming in $O(n \cdot U)$ hyp calculations in $O(U)$ space. In one run, we calculate all ℓ_i^n $i \in [0 - x]$. So we can check for presence property w.r.t. a single tuple t in $O(n \cdot U)$. By Definition 9, we need to check for all $t \in T$, so the worst case complexity of checking for the presence property is $O(n \cdot U \cdot |T|)$.

Following the example, suppose we want to calculate $\ell_2^3 = \mathcal{P}(|t^* \in D^*| = 2 \mid F_{3d}, T)$. We first set, for dimension A, $\ell_0^1 = \ell_1^1 = \ell_2^1 = 0$, and $\ell_3^1 = 1$ since $f_A(a_2) = 3$. Possible Ds for dimension A, B are given in Table 3. $\ell_2^2 = 0 + 1 \cdot \frac{\binom{3}{2}\binom{1}{1}}{\binom{4}{3}} = \frac{3}{4}$ (cardinality of $< a_2, b_2 >$ is 2 in D_2 only.) $\ell_3^2 = 1 \cdot \frac{\binom{3}{3}\binom{1}{0}}{\binom{4}{3}} = \frac{1}{4}$ (cardinality of $< a_2, b_2 >$ is 3 in D_3 only.) Finally $\ell_2^3 = \frac{3}{4} \cdot \frac{\binom{3}{2}\binom{1}{0}}{\binom{4}{2}} + \frac{1}{4} \cdot \frac{\binom{3}{2}\binom{1}{1}}{\binom{4}{3}} = \frac{9}{16}$. Similarly $\ell_1^3 = 0 + \frac{1}{4} \cdot \frac{\binom{3}{1}\binom{1}{2}}{\binom{4}{3}} + \frac{3}{4} \cdot \frac{\binom{3}{1}\binom{1}{1}}{\binom{4}{2}} = \frac{6}{16}$. The confidence level for T is given by $\ell_0^3 + \ell_1^3 + \ell_2^3 = 0 + \frac{6}{16} + \frac{9}{16} = \frac{15}{16}$.

Even though checking process can be done in polynomial time, calculation of hyp is very costly for large arguments. It is practically impossible to calculate all likelihoods directly for big datasets. In Section 5, we will address this issue by presenting optimizations.

4.3 Privacy for all Possible Tuples

In this section, we show how to extend the methodology given in Section 4.2, so that we achieve the desired privacy level not just for the tuples in the private dataset but for all tuples that can possibly exist in the outside world.

In Section 4.2, we derived the confidence level for tuples in the private dataset. More precisely, Equation 3 gives the confidence level for all tuples $t \in T$ given T^*, F. However this is not the exact set of tuples whose confidence levels can be computed by Equation 3. The next theorem identifies the exact set:

Definition 10 (Coverage): A given tuple t is in the coverage of a table T^*, iff there exist at least one tuple $t^* \in T^*$ with $t^* \in \psi(t)$.

For example, in Table 2, tuples <47906,Canada>, <47903,US> are covered by a table containing the generalized tuple <4790*,N. America>.

Theorem 2: Let T^* be a c-confident δ-present anonymization of private dataset T w.r.t. public distribution F. If a given tuple t is in coverage of T^*, then t is δ-present w.r.t. T^*, F with c confidence.

Proof: Since T^* is a generalization of T, there exist a tuple $t' \in T$ with $t^* \in \psi(t')$. By Definition 9, t' is δ-present w.r.t. T^*, F with c confidence. Since T^* is the only input to the adversary, t and t' are indistinguishable. t should also be c-confident δ-present. Also note that right hand side of Equation 3 depends on the generalized tuple t^*, not the atomic tuple t. □

For any other tuple t not in the coverage of the generalization, the existence probability is intuitively zero. (If there is no generalization for t in T^*, then t is definitely not in T.) That means for tuple t, when $\delta_{min} \neq 0$, confidence becomes 0; and anonymization violates privacy requirements with $c > 0$.

It is easy to check if every possible tuple that can exist (or most likely exists) in the public dataset is covered by

[2] A sample of n balls is drawn from an urn containing M white and $N-M$ black balls without replacement. hyp gives the probability of selecting exactly x white balls.

an anonymization. However, as mentioned before providing c-confident δ-presence for all possible tuples is perhaps too strong a privacy requirement. In the following sections, we stick to definition 9.

5 Speeding up the Checking Process

We now present several optimizations to reduce the number of *hyp* function calculations. All of the optimizations presented in this section stand as a trade-off between utility and efficiency. *We do not sacrifice privacy. In other words, the privacy level achieved with the optimizations is at least as good as the original privacy level.* In Section 8, we show experimentally that the loss in the utility is very small while a huge speed up can be achieved from the use of optimizations. We also show that each optimization is effective independently of the others.

5.1 Practical Anti-monotonicity

Unfortunately c-confident δ-presence does not possess the anti-monotonicity property (which we proved for the original δ-presence in Theorem 1). In other words, given a non-overlapping generalization T_1^* of T and a non-overlapping generalization T_2^* of T_1^*, even though T_2^* does not respect c-confident δ-presence (w.r.t. a public distribution), it is possible that T_1^* does. This so because of the 'per tuple' definition we adopt for c-confident δ-presence. More precisely, for a tuple $t \in T$ and a public distribution F, let $P_{T_1^*}$ be the set of all possible public tables in which t is not δ-present w.r.t. T_1^*, and let $P_{T_2^*}$ be the set containing those violating the presence property w.r.t. T_2^*. $P_{T_1^*}$ is not necessarily a subset of $P_{T_2^*}$ and thus the sum of the corresponding likelihoods may increase towards the confidence level. This does not contradict Theorem 1. Theorem 1 states given T_2^* is not δ-present, there will be at least one tuple in T_1^* violating the presence property. Tuples violating the property in each generalization do not necessarily match, there may be tuples that are present w.r.t. T_1^* although they are not w.r.t. T_2^*.

We can argue that, although it is possible, such a situation is unlikely. Given a tuple t violating the presence property with respect to a more general anonymization (e.g., T_2^*) and a fixed public dataset P, let us derive an upper bound for the probability that t is present w.r.t. a more precise randomly created anonymization (e.g., $T_1^* \mid T_2^* \in \psi(T_1^*)$) and P:

Let P_1^* and P_2^* be the generalizations of P with the same mappings used in T_1^* and T_2^* respectively. Let ec_1, ec_2 be the equivalence classes of P_1^*, P_2^* containing t respectively. Since $T_2^* \in \psi(T_1^*)$, also $P_2^* \in \psi(P_1^*)$ and also $ec_1 \subset ec_2$. Let $ep_1 \subset ec_1$ and $ep_2 \subset ec_2$ be the set of present tuples in ec_1 and ec_2 respectively. Without loss of generality, assume upper bound condition of $\delta_{min} = 0$. We are interested in the following probability:

$$\mathcal{P}\left(\frac{|ep_1|}{|ec_1|} \leq \delta_{max} \mid \frac{|ep_2|}{|ec_2|} \geq \delta_{max}\right)$$

It is complicated to derive the exact probability. However, if we restrict the random ec_2s to be sufficiently large[3], we can model each tuple in ec_2 as an independent random variable. Let X_i be a Bernoulli random number representing $t_i[sen.]$ with $\mathcal{P}(X_i = 1) = p = \frac{|ep_2|}{|ec_2|}$. Note that assuming independence of random variables is not realistic but unbiased towards what we try to prove. Most of the time, the tuples (data points) are skewed in the whole space. Present tuples are more likely to be clustered together and are relatively far from absent tuples. The presence property is more likely to fail in such skewed spaces compared to uniform spaces.

Suppose ec_1 contains n of these tuples then the probability above takes the following form:

$$\mathcal{P}(\overline{X_n} \leq \delta_{max})$$

where $\overline{X_n} = \frac{X_1 + \cdots + X_n}{n}$

We can bound the mean of Bernoulli variables using Hoeffding's Inequality:

$$\begin{aligned}
\mathcal{P}(\overline{X_n} - p < -\varepsilon) &\leq e^{-2n\varepsilon^2} \\
\mathcal{P}(\overline{X_n} < p - (p - \delta_{max})) &\leq e^{-2n(p-\delta_{max})^2} \\
\mathcal{P}(\overline{X_n} < \delta_{max}) &\leq e^{-2n(p-\delta_{max})^2}
\end{aligned}$$

Similarly it can be showed for any $\delta_{min}, \delta_{min}$;

$$\mathcal{P}(\overline{X_n} < \delta_{max}) \leq e^{-2n(p_{dist})^2}$$

where $p_{dist} = min(|p - \delta_{min}|, |p - \delta_{max}|)$.

For a small value of $n = 100$ and $p_{dist} = 0.1$, the upper bound probability is 0.13. For a moderate value of $n = 1000$ and $p_{dist} = 0.05$, the probability is 0.006. Generally when p_{dist} is more than 0.1, the bounding probability is sufficiently low for most probable n values.

Using this tool, we can have a rough estimate of the confidence level for T_2^* without calculating it (thus only calculating the final anonymization for verification). It should also be noted that there may be more than one tuple t violating the presence property which further makes an anti-monotonic behavior very probable. In Section 8, we show experimentally that p_{dist} is generally high enough to make a correct estimate, and it is safe to use apriori pruning even though anti-monotonicity does not hold theoretically.

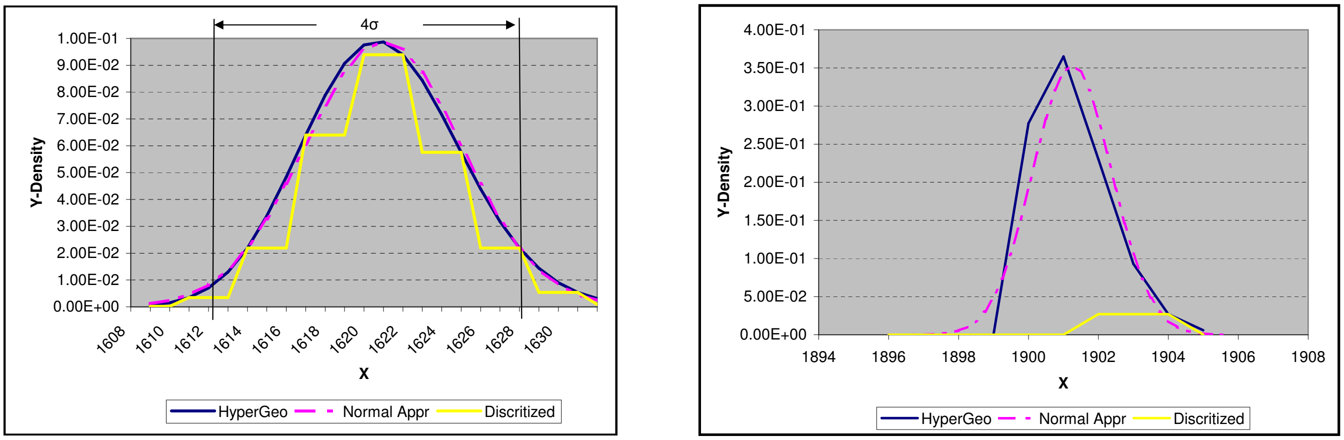
5.2 Normal Approximation to Hypergeometric Distribution

Hypergeometric calculations are costly for big datasets since they require factorization of huge numbers. Fortunately $hyp(x;n,M,N)$ can be approximated by the normal distribution $N(\mu, \sigma)$ where $\mu = x\frac{M}{N}$, $\sigma = \sqrt{N\frac{M}{N}(1-\frac{M}{N})\frac{n}{N}(1-\frac{n}{N})}$ when $\frac{M}{N}$ and $\frac{n}{N}$ are bounded away from 0 and 1 which are called *standard cases* [11]. *Non-standard cases* are also studied [12] and guaranteed error bounds have been proposed. For example, theoretical upper bound for total deviation between the normal approximation and the hypergeometric distribution (over all possible x) is 0.12 for $N = 2000$, $\frac{M}{N} = 0.9$, $\frac{n}{N} = 0.9$. In Figure 2(a), we plot the two distributions for this case. Actual total deviation is 0.037, much smaller than the theoretical upper bound. However when $\frac{M}{N} = 0.975$, $\frac{n}{N} = 0.975$ as shown in Figure 2(b), deviation increases to 0.24.

Unfortunately, we cannot use the normal approximation as an alternative to hypergeometric distribution, even in standard cases, for the following reasons:

- The sign of the deviation is not known. In other words, a normal approximation can give a higher confidence level

[3] Large datasets are more likely to have large equivalence classes.



(a)

(b)

Fig. 2. Normal Approximation and Discretization for Hyper Geometric Distribution

than the original level. Since we want to guarantee an upper bound on the privacy, this is not desirable.

- A guaranteed cdf-error bound of the approximation could be subtracted from the normal approximation to ensure that approximation is smaller than the original value. However, we are required to calculate the probability at each integer value (which is as big as the size of the population) and probabilities tend to be smaller than the error bound itself.

Instead of using the normal approximation as an alternative to the hypergeometric distribution, we use the approximation to identify those parameters that contribute little to the total confidence level. About 95% of values drawn from a normal distribution are within two standard deviations of the mean (see Figure 2(a)). For standard cases, we can assume the same for the hypergeometric distribution and avoid calculating *hyp* function on x points far from the mean and assume a zero probability for those points. So we define a new function *hyp'* as

$$hyp'(x;n,M,N) = \begin{cases} hyp(x;n,M,N), & x \in [\mu - 2\sigma \cdots \mu + 2\sigma] \\ & \quad \vee \text{ Standard Case;} \\ 0, & \text{otherwise.} \end{cases}$$

In Figure 2(a), there are 200 possible x values for which function *hyp* would return a positive probability. Only 16 of these values are within two standard deviations of the mean. So *hyp'* returns 0 instantly for 92% of the cases.

In our experiments, we considered all parameters $\frac{M}{N} < 0.99, \frac{n}{N} < 0.99$ for $N = 40000$ as standard cases. The Normal approximation caused little distortion for these parameters. It should also be noted that it is easier to compute *hyp* for non-standard cases compared to standard cases, since there are not as many possible x values (with a non-zero probability). In Figure 2(b), we have only 50 possible x values that require a hypergeometric calculation as opposed to 200.

5.3 Tuple Dominance

In Section 4.2, we calculate the confidence level for all tuples in T^*. In this section, we present a technique that will reduce the number of tuples required to be checked for confidence when either $\delta_{min} = 0$ or $\delta_{max} = 1$.

Definition 11 (Tuple Dominance): Given some public distribution $F = \cup_{A_i} f_i$, an anonymization T^* of table T, and tuples $t_1^*, t_2^* \in T^*$, we say t_1^* dominates t_2^* on the min parameter if $|t_1^* \in T^*| \le |t_2^* \in T^*|$ and $f_i(t_1^*[A_i]) \ge f_i(t_2^*[A_i])$ for all attributes A_i. Similarly we say t_1^* dominates t_2^* on the max parameter if $|t_1^* \in T^*| \ge |t_2^* \in T^*|$ and $f_i(t_1^*[A_i]) \le f_i(t_2^*[A_i])$ for all attributes A_i.

Given F_{3d} in Table 3 and $T^* = T$ in Table 4.1, tuple $t_2^* :< a_2, b_2, c_2 >$ dominates $t_1^* :< a_1, b_1, c_1 >$ on the min parameter since both have cardinality 1 and $(f_A(a_2) = f_B(b_2) = f_C(c_2) = 3) > (f_A(a_1) = f_B(b_1) = f_C(c_1) = 1)$. Similarly t_1^* dominates t_2^* on the max parameter.

The next Theorem proves that to check for the presence property when $\delta_{min} = 0$ or $\delta_{min} = 1$, it is enough to calculate the confidence level only for the dominant tuples.

Theorem 3: Following the definitions above, given that tuple t_1^* dominates tuple t_2^* on the min parameter, if t_1^* respects c confident $(\delta_{min}, 1)$ presence, so does t_2^*.

Proof:

1: We first assume $|t_1^* \in T^*| \le |t_2^* \in T^*|$ and $f_i(t_1^*[A_i]) = f_i(t_2^*[A_i])$. By equation 4, confidence level for tuple t_i^* is

$$\sum_{x \le c^i_{high}} \mathcal{P}(|t_i^* \in D^*| = x \mid F)$$

where $c^i_{high} = \left\lfloor \frac{1}{\delta_{min}} |t_i^* \in T^*| - |t_i^* \in T^*| \right\rfloor$.

By definition, $|t_1^* \in T^*| \le |t_2^* \in T^*|$, so $c^1_{high} \le c^2_{high}$. So the confidence for t_1^* will be higher (or the same) if $f_i(t_1^*[A_i]) = f_i(t_2^*[A_i])$ for all attribute A_i.

2: We now assume $|t_1^* \in T^*| \le |t_2^* \in T^*|$, $f_i(t_1^*[A_n]) = f_i(t_2^*[A_n]) + 1 = M + 1$ where A_n is the last attribute, $f_i(t_1^*[A_i]) = f_i(t_2^*[A_i])$ for any other attribute A_i. Therefore t_1^*, t_2^* differ on their last attribute cardinality by 1. Since the calculations are independent of the attribute order, the proof can be extended by induction even if any subset of attributes differ by any amount. From Section 4.2, the confidence level for tuple t_i^* is

$$\sum_{x \le c_{high}} \mathcal{P}(|t_i^* \in D^*| = x \mid F) = \sum_{x \le c_{high}} (\ell^n_x)_{t_i^*}$$
$$= \sum_{x \le c_{high}} \sum_{y \in [1-U]} (\ell^{n-1}_y)_{t_i^*} \cdot hyp(x; y, f_n(t_i^*[n]), U)$$

$$= \sum_{y \in [1-U]} (\ell_y^{n-1})_{t_i^*} \sum_{x \leq c_{high}} hyp(x; y, f_n(t_i^*[n]), U)$$

Since the first $n-1$ attributes are the same for t_1^* and t_2^*, $(\ell_y^{n-1})_{t_1^*} = (\ell_y^{n-1})_{t_2^*}$, we need to prove;

$$\sum_{x \leq c_{high}} hyp(x; y, M+1, U) \leq \sum_{x \leq c_{high}} hyp(x; y, M, U)$$

for all x. In other words, we need to prove that the cumulative for the hypergeometric distribution is non-increasing with increasing M. There is no closed form for the hypergeometric cumulative. Fortunately, however, we can work on the difference of two cumulatives. If we substitute for the *hyp* function:

$$\begin{aligned}
& hyp(x \leq a; n, M+1, N) - hyp(x \leq a; n, M, N) \\
= & \sum_{0 \leq x \leq a} \frac{\binom{M+1}{x}\binom{N-M-1}{n-x}}{\binom{N}{n}} - \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \\
= & C[\sum_{0 \leq x \leq a} (\binom{M}{x-1} + \binom{M}{x}) \binom{N-M-1}{n-x} \\
& - \sum_{0 \leq x \leq a} \binom{M}{x} (\binom{N-M-1}{n-x-1} + \binom{N-M-1}{n-x}))] \\
= & C \sum_{0 \leq x \leq a} \binom{M}{x-1} \binom{N-M-1}{n-x} - \binom{M}{x} \binom{N-M-1}{n-x-1} \\
= & C[\sum_{-1 \leq x \leq a-1} \binom{M}{x} \binom{N-M-1}{n-x-1} - \sum_{0 \leq x \leq a} \binom{M}{x} \binom{N-M-1}{n-x-1}] \\
= & C[\binom{M}{-1} \binom{N-M-1}{n} - \binom{M}{a} \binom{N-M-1}{n-a-1}] \\
= & -C \binom{M}{a} \binom{N-M-1}{n-a-1} \\
\leq & \ 0
\end{aligned}$$

for all a, where $C = \frac{1}{\binom{N}{n}}$. □

Corollary 1: Given that tuple t_1^* dominates tuple t_2^* on the max parameter, if t_1^* respects c confident $(0, \delta_{max})$ presence, so does t_2^*.

Definition 12 (Minimum Dominant Subset(MDS)): We say a set of tuples MDS is a minimum dominant subset for a set of tuples T^* if $MDS \subseteq T^*$, no two tuples in MDS dominate each other, and every tuple in T^* is dominated by some tuple in MDS.

For $\delta_{min} = 0$ or $\delta_{max} = 1$, we only need to calculate the confidence levels for the MDS of a set of tuples since the rest of the tuples will have higher confidence levels than at least one tuple in MDS. Following the example above, if we assume we have $\delta_{min} = .33$, $\delta_{max} = 1$, the MDS only contains tuple t_2^* which has a confidence level of $\frac{15}{16}$. Tuple t_1^* has a confidence of 1 which is higher than that of t_2^*.

5.4 Discretization

As mentioned above, the *hyp* function is calculated for a large range of x parameters. Fortunately, a hypergeometric distribution has a well defined behavior. As in a normal distribution, the probability density function monotonically increases for $x < \mu$, is maximized for $x = \mu$ and monotonically decreases for $x > \mu$, having a concave shape. We can exploit this property for further discretization of the hypergeometric distribution. In other words, we partition the probability space into fixed sized buckets and calculate the *hyp* function on only the boundaries of the buckets. We set the minimum of the boundary probabilities as the probability of the points inside the bucket, thus sacrificing utility for efficiency. If we have a fixed bucket size of b, the discretized function *hyp"* will take the following form:

$$hyp''(x; n, M, N) = \begin{cases} \min(hyp'(x; n, M, N), \\ \qquad hyp'(x+b; n, M, N)), & x|b; \\ hyp'(x-1; n, M, N), & \text{otherwise.} \end{cases} \quad (5)$$

Figure 2 shows 3-discretization of two hypergeometric distributions. An important point to make is that the utility loss due to the discretization is negatively correlated with the size of the probability space within 2 standard deviations from the mean. For example, in Figure 2(a) where we have 16 points 2σ close to μ, the area covered by the discretization is 79% of the total area under the hypergeometric distribution while it is 8% in Figure 2(b) with 6 points close to mean. In experiments, we used discretization only if we have at least 4 buckets within a 2σ range.

6 Limiting Harm: Selecting a good δ, c

The presence parameters δ_{min}, δ_{max} defines the level of trade-off between the utility and privacy of the anonymized dataset. As δ_{min} increases (or δ_{max} decreases), more information is hidden leading to better privacy protection but poorer dataset utility. This means that a maximal δ_{min} and minimal δ_{max} value should be selected such that privacy conditions of the application are met. In this section, we use the diabetes dataset example to demonstrate how to bound probability of disclosure in ways that correspond to real risk of misuse.

Let I_p be the event that person p has diabetes. Since the rate of diabetes in all US population is public information [8], any adversary will have a prior belief b_r on I_p given the public dataset P:

$$b_r = \mathcal{P}(I_p) = 0.07$$

The private dataset T is a subset of the set of all diabetes patients in P. Seeing some anonymization T^* of T, attacker will have a posterior belief b_o on I_p:

$$\begin{aligned}
& b_o \\
= & \mathcal{P}(I_p \mid T^*) \\
= & \mathcal{P}(I_p | p \in T) \cdot \mathcal{P}(p \in T | T^*) + \mathcal{P}(I_p | p \notin T) \cdot \mathcal{P}(p \notin T | T^*) \\
= & 1 \cdot \mathcal{P}(p \in T | T^*) + \frac{\mathcal{P}(I_p) \cdot |P| - |T|}{|P| - |T|} \cdot (1 - \mathcal{P}(p \in T | T^*)) \\
= & \mathcal{P}(p \in T \mid T^*) \cdot \frac{|P| \cdot (1 - b_r)}{|P| - |T|} + \frac{b_r \cdot |P| - |T|}{|P| - |T|}
\end{aligned}$$

We start with an acceptable cost due to misuse. Assume a hiring decision, and that a \$100 annual difference in total cost of employee is noise (difference in productivity, taking an extra sick day, salary negotiation, etc.) Thus if expected annual cost of medical treatment of diabetes based on misuse of the database is $m < \$100$, the risk of misuse is acceptably small. The total cost of diabetes per person is around $d = \$10,000$ [9]. The probabilistic acceptable misuse, a, is then $\frac{m}{d} = \frac{1}{100}$; we must ensure:

$$\begin{aligned} b_o \cdot d - b_r \cdot d &\leq m \\ b_o - b_r &\leq a \\ \mathcal{P}(p \in T \mid T^*) \cdot \frac{(1-b_r)|P|}{|P|-|T|} + \frac{b_r|P|-|T|}{|P|-|T|} - b_r &\leq a \end{aligned}$$

$$\mathcal{P}(p \in T \mid T^*) \leq \frac{a \cdot |P| + (1-a-b_r)|T|}{(1-b_r)|P|}$$

Unfortunately, this does not take into account the c parameter (possibility that we fail to maintain δ-presence for a tuple.) The risk that an individual is exposed with greater than δ_{min} probability is a different kind of risk; the possible damage to the individual could be significantly greater than a. While this could be calculated as $a = \frac{m+1/c}{d}$, it is more appropriate to separate this out as a *liability risk* that should fall on the anonymizer. $\frac{1/c}{d}$ can thus be viewed as a separate parameter, set based on the needs of the anonymizer rather than as protection for the individuals.

Letting $|T| \simeq 0.04|P|$ as in our experiments and applying the above numbers, we get:

$$\mathcal{P}(p \in T \mid T^*) \lesssim 0.05$$

This gives us the minimum δ_{max} parameter to protect against substantial misuse when hiring a single job applicant. However the upper bound does not protect against misuse when comparing two job applicant p_1, p_2. The reason is that in this setting, an anonymization that gives $b_o = 0.032$ for p_1 (this happens when $\mathcal{P}(p_1 \in T \mid T^*) \simeq 0$) and $b_o = b_r = 0.07$ for p_2 is perfectly okay, which implies p_2 is much more likely to have diabetes than p_1. We need to ensure that the company can't "cherry-pick" employees known *not* to be in the database. Thus the posterior belief should not be arbitrarily low. If we let probabilistic acceptable misuse $a = \frac{200}{10000} = 0.02$ then

$$\begin{aligned} b_r - b_o &\leq a \\ \mathcal{P}(p \in T \mid T^*) &\geq \frac{-a \cdot |P| + (1+a-b_r)|T|}{(1-b_r)|P|} \\ &\gtrsim 0.02 \end{aligned}$$

This gives us a maximum δ_{min} parameter.

7 Algorithms

Given the "world knowledge" assumption made in [1], we provided an optimal full-domain generalization algorithm SPALM; this works under the constraint that if a value is generalized, all occurrence of that value must be generalized. We also gave an algorithm MPALM that relaxed this restriction. (The difference between these approaches is analogous to the difference between the k-anonymization algorithms of [13] and [14].)

In this section, we present a c-confident δ-presence algorithm SFALM that makes use of the optimizations discussed in Section 5. We basically modify the SPALM algorithm from [1] to accept a confidence threshold and a public distribution instead of a public table. The practical anti-monotonicity claim discussed previously enables us to use this same approach to give an algorithm that is comparable in performance to SPALM; if anti-monotonicity does not hold we lose only the optimality (with respect to the amount of generalization); the privacy guarantee of c-confident δ-presence still holds. The following notations and definitions briefly recall the problem setting:

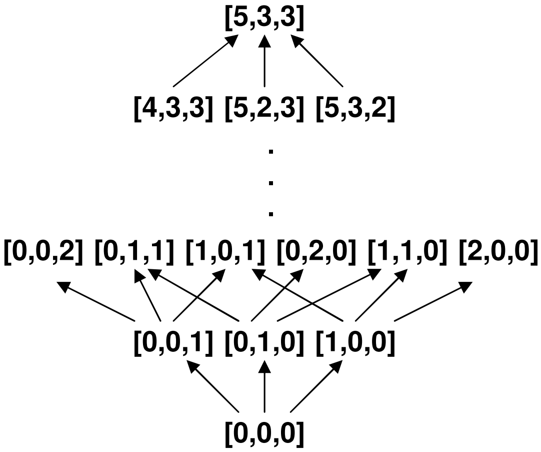


Fig. 3. Full Domain Generalization Lattice

For two values v^*, v of the same attribute A_i, we write $v^* = \Delta_i(v)$ if and only if v^* is the immediate parent of v in the domain generalization hierarchy for A_i. To express greater levels of generalization, for the n^{th} generalization of v, we write $\Delta_i^n(v) = \underbrace{\Delta_i(\cdots\Delta_i}_{n}(v)\cdots)$.

We say a table T' is a $[g'_1, \cdots, g'_n]$ full domain generalization of table T with set of attributes $\{A_1, \cdots, A_n\}$ if and only if for all pairs of tuples t, t' such that $(t \in T) \rightleftharpoons (t' \in T')$; we have $t'[A_i] = \Delta_i^{g'_i}(t[A_i])$ for all $1 \leq i \leq n$. Let T'' be a $[g''_1, \cdots, g''_n]$ full domain generalization of table T, We say T'' is a higher level generalization than T' and write $T'' \gg T'$ if and only if $T' \neq T''$ and $g'_i \leq g''_i$ for all $1 \leq i \leq n$. For cost metrics proposed so far, a high level generalization (e.g., T'') is more costly than a lower generalization (e.g., T').

In Tables 1 and 2, T_3^* is a $[3,3,2]$ full domain generalization of T.

The possible full domain generalizations of table T form a lattice on the $\gg$ relation. (see Figure 3.) To find a cost-optimal δ-present (or k-anonymous) generalization, each generalization on the lattice needs to be checked and the lowest cost δ-present dataset should be identified. However the practical anti-monotonicity property of presence can be used to prune the lattice and reduce the search space. In contrast to previous k-anonymity algorithms, we exploit only the anti-monotonicity property of presence and propose a top-down approach. (E.g., if T'' is not δ-present, neither is T'.) We observed that a top-down approach prunes much faster, especially when the data is of high dimensionality and sparsely distributed (as in the experimental data used in coming section). Notice that, for very high dimensional spaces, optimal solutions for k-anonymity (and therefore, optimal δ-presence) are subject to the curse of dimensionality, as discussed in [15].

The pseudo-code in Algorithm 1 summarizes SFALM. At line 5, the algorithm creates the Minimum Dominant Subset of the anonymization given the δ parameters are one-sided. At line 7, algorithm calls for Equation 4 (with 2σ-cut and discretization optimizations) to get confidence levels for all tuples and checks if any confidence level falls below the threshold.

Algorithm 1 SFALM

```
Require: publicly available distribution $F$; private table $T$, a cost metric COST;
Ensure: return minimum cost $c$-confident $(\delta_{min}, \delta_{max})$ present full domain generalization of $T$.
create lattice $lat$ for all possible generalization mappings for $T$. Let $n$ be the number of levels in $lat$.
for all level $i : 1 - n$ do
    for all node $m$ in level $i$ of $lat$ do
        create $T^*$; full domain generalization of $T$ according to mapping in $m$
        let set of tuples $ST$ be the MDS of $T^*$ if $\delta_{min} = 0$ or $\delta_{max} = 1$. Otherwise $ST = T^*$
        for all tuple $t^* \in ST$ do
            use Eqn 4 with function $hyp''$ to get the confidence level $c_{t^*}$ for $t^*$ given $F$.
            if $c_{t^*} < c$ then
                delete node $m$ and all children and grandchildren of $m$ from $lat$
return the least-cost generalization and the corresponding mapping among the generalizations being tested.
```

8 Experiments

As in [1], a simulated dataset is created through random selection of a 4% subset of the UCI adult dataset [10]. The entire adult dataset (specifically, the 45222 records with no unknowns) is considered the "Universe", a randomly selected subset of 1957 records is taken as the dataset of individuals whose discovery in the dataset is to be protected against (we refer to this as the *Random* dataset.) Note that for this paper, only the distribution of the Universe is used by the algorithm; the only data needed is the 1957 record subset to be anonymized.

As was done in [1], for a more realistic test we also generate a *diabetes* dataset by performing a biased selection simulating a database of individuals with diabetes; the selection is biased toward individuals with demographics matching those of actual diabetes patients (as given in [8].) Specifically, for each individual we estimate their probability of being in the diabetes subset based on independent probabilities for diabetes given age, race, and gender as shown in [8]; this gives a dataset skewed towards people with similar characteristics. (This is also the reason for 1957 records in the dataset, as this is the number obtained using these statistics to guide selection.)

We evaluate how varying δ affects the cost of anonymizing the dataset (w.r.t. utility of the dataset), as determined by the Loss Metric (LM) [16] and the Discernibility Metric (DM) [17]. We only present results on LM since the DM results were almost identical. The LM cost for a given generalized data value v^* from an atomic domain of size $|A|$ is defined as $\frac{|\psi(v^*)|-1}{|A|-1}$. For example, in Figure 1 generalizing "Canada" to "N. America" incurs a penalty of $(2-1)/(7-1) = .16$; to "America" gives a penalty of $(4-1)/(7-1) = .5$. The LM cost for a generalization T^* is the normalized cost of all data values in T^*. The higher the LM cost of the anonymization, the higher the distortion caused by the anonymization.

Fixing the private dataset and distribution functions as inputs, we evaluate the behavior of SFALM and the effectiveness of the optimizations SD (4σ cut on normal approximation), PA (Practical Anti-monotonicity), DM (minimum dominant subset), and DC (4-Discretization) presented in Section 5 with respect to efficiency and utility.

In Figure 4(a) and 4(c), we use the first 4 attributes of the diabetes dataset and random dataset respectively, fix confidence to be 0.8 [4] and plot the execution time for runs of SFALM using different subsets of 4 optimizations (note the log scale). The first 4 bars refers to SFALM with all but one optimization. (E.g., the first bar, ˜SD, shows the results when we have PA+DM+DC (all but SD).) The fifth bar is for SFALM with all optimizations. Thus the graph explains the effectiveness of a particular optimization even when all others are also in effect. We see that PA is quite effective and dominant especially for tight ranges of δ parameters. This is expected since in such levels, anonymization is very distorting and anti-monotonicity allows us to prune the majority of the lattice nodes. As we loosen the privacy constraints, SD starts to kick in and DC becomes a bit more effective, but DM optimization is not significant even when $\delta_{min} = 0$. We see that with few attributes, the time gained due to the small size of the minimum dominant subset does not compensate for the time spent on the creation of the dominant set.

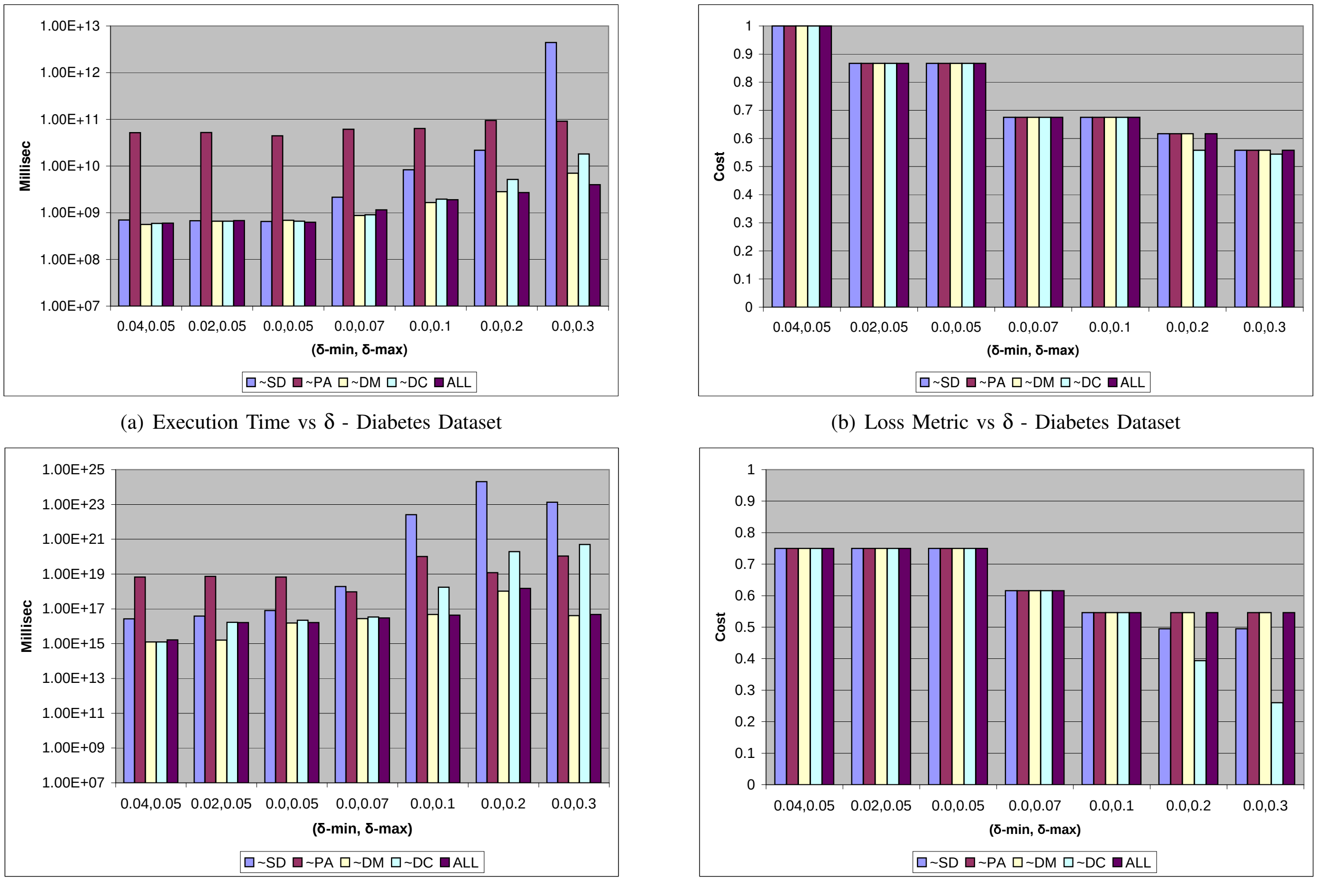
In all our experiments, SFALM with optimization PA or SD nearly always returned the same output. DM by definition is not an approximation and does not change the original output. Figure 4(b) and 4(d) plot the LM cost metric results for the above mentioned runs. DC does result in a slight increase in the loss metric, but the overall utility-time trade-off is very good. When all optimizations are in place, SFALM runs more than 1000 times faster with less than 5% worse LM cost. For the random dataset, DC results in a relatively higher utility loss but also provides a higher speed up. While the complexity of obtaining values without any optimization was prohibitive, the fact that removing any single optimization (except DC) had little impact on the loss metric gives confidence that these values would be close to the exact SFALM algorithm.

When we have more attributes, it becomes impractical to run SFALM without SD and PA. In such cases, the significance of DM and DC becomes more observable. In Figures 5(a) and 5(c), we show execution times for ˜DM˜DC, ˜DM, ˜DC, and ALL when we use the first 7 attributes of the Diabetes and Random datasets. DM and DC are both effective in speeding up SFALM especially for loose δ constraints. DC seems to be better, however, as we see in Figures 5(b) and 5(d), it introduces an 8% increase in utility cost of the output. The use of DM+DC stands as a beneficial trade-off, since SFALM runs 100 times faster when both optimizations are active. Results for the random dataset are similar except that utility cost levels were a bit lower (justifying the discussion in Section 5).

We show in Figure 6 how much pruning SD, PA, and DM do when we use all optimizations. Prune rates of 80-95% for SD and PA clearly shows why it becomes infeasible to run SFALM without them. For SD, pruning is displayed as an average for each recursive step in Equation 4. In fact, SD prunes much more than what the figures show since each cut cancels out additional calculations in the tail of the recursion.

In Figure 7 we show the behavior of SFALM with respect to decreasing confidence by setting $\delta_{min} = 0, \delta_{max} = 0.1$. From the

[4] We will show results for higher confidence levels in Figure 7, but the 0.8 confidence level provides a more effective view of the tradeoffs between optimizations



(a) Execution Time vs δ - Diabetes Dataset

(b) Loss Metric vs δ - Diabetes Dataset

(c) Execution Time vs δ - Random Dataset

(d) Loss Metric vs δ - Random Dataset

Fig. 4. Results with 4 dimensions

TABLE 5
Table shows decrease in execution time and increase in cost with varying bucket sizes for DC optimization

bs	2	4	6	8	10
x	.62	.36	.20	.13	.11
+	0	.05	.05	.09	.11

shape of the graphs, the same conclusions can be made about the effect of the various pruning techniques.

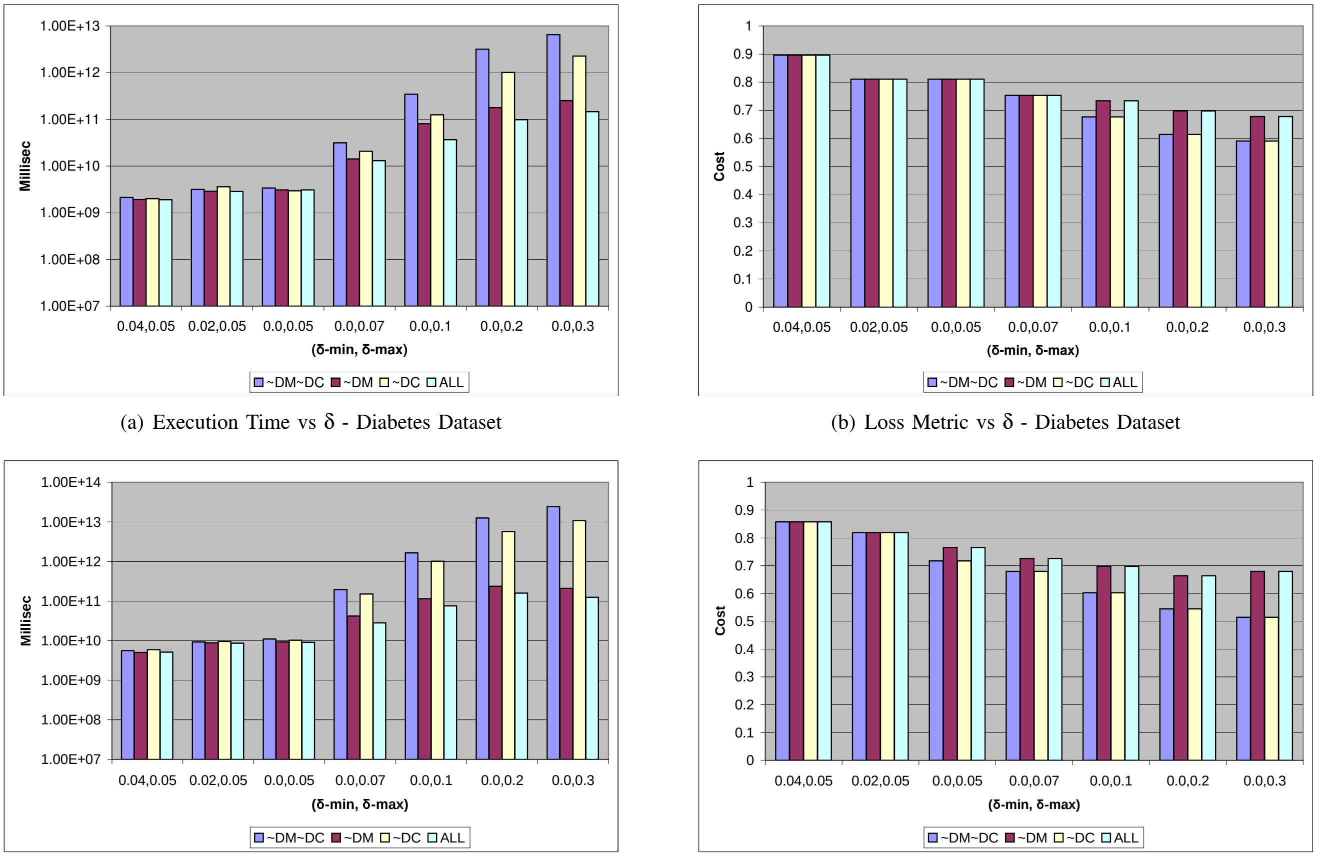
Table 5 shows how varying bucket size affects the utility-efficiency tradeoff for optimization DC. We set $c = 0.8, \delta_{min} = 0, \delta_{max} = 0.1$ for the diabetes dataset. The increase in speed starts quickly but stabilize for large bucket sizes. This is mostly due to the minimum limit we set for the number of points within 4σ.

9 Conclusions and Future Work

While *k*-anonymity and related techniques have received considerable attention, it isn't clear that they are the best way to balance privacy and data utility [7]. We have presented a problem where anonymization *is* an appropriate solution, and a metric δ-presence that correlates to the real risk/cost of a privacy violation. Datasets anonymized directly to meet the δ-presence standard distort data less than *k*-anonymization to comparable privacy levels, and provide a clear risk-based guarantee of privacy. Introducing a "liability risk" confidence metric allows the δ-presense metric to be utilized in practical scenarios.

The ability to anonymize data to meet δ-presence without complete world knowledge does come with a cost. Knowing the public dataset, the single-dimensional approach in [1] achieved LM cost under 0.6 for the scenario in Figure 7(b). They also showed that *k*-anonymity could achieve similar LM costs, but again this requires knowing the public dataset (as well as trying multiple values of *k*) as there is no direct relationship between choice of *k* and the disclosure risk measured by δ-presence, and thus no way to determine if δ-presence has been met other than by testing against the public dataset. A typical example from the experiments depicted in Figure 7(b) generalizes the tuple (39, State-gov, Bachelors, Never-married, Adm-clerical, Not-in-family, White) to (Any, Any, University, Any, Other, Not-in-family, Any) at confidence 0.8, and (Any, Any, Any, Any, Adm-clerical, Any, Any) at confidence 0.98. Note the suppression of relatively specific items (age, working for state government). Significant improvement may be possible through algorithms achieving *c*-confident δ-presence with multi-domain generalization, as done for basic δ-presence in [1].

Extending this work to linking with sensitive data in a disclosed dataset, as with ℓ-diversity at *t*-closeness, is straightforward (and could be easily accomplished using the technique of [7].) It may be possible to design δ-presence algorithms that guarantee bounds on optimality as done for *k*-anonymity in [18]. Further development of δ-presence will address a variety of real-



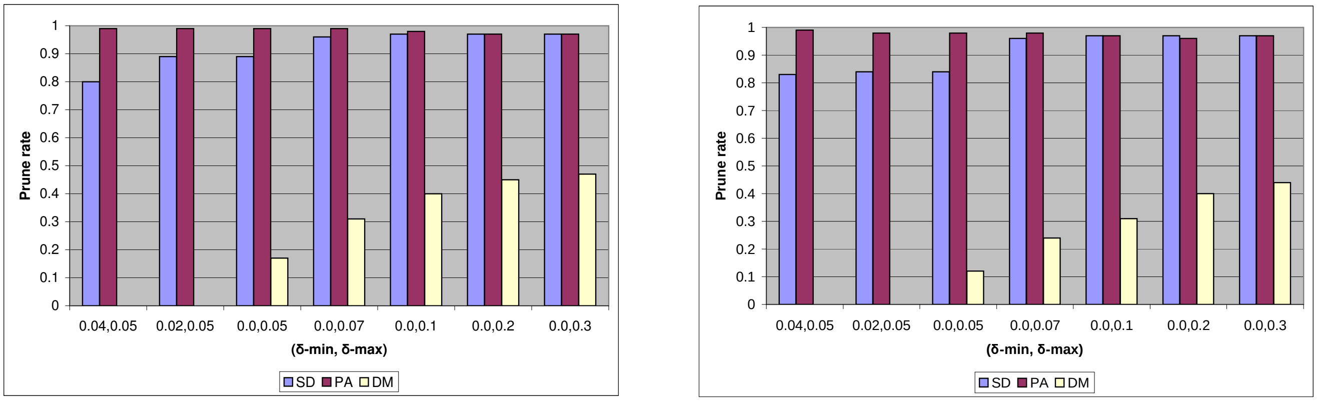
(a) Execution Time vs δ - Diabetes Dataset

(b) Loss Metric vs δ - Diabetes Dataset

(c) Execution Time vs δ - Random Dataset

(d) Loss Metric vs δ - Random Dataset

Fig. 5. Results with varying δ and 7 dimensions



(a) Diabetes Dataset

(b) Random Dataset

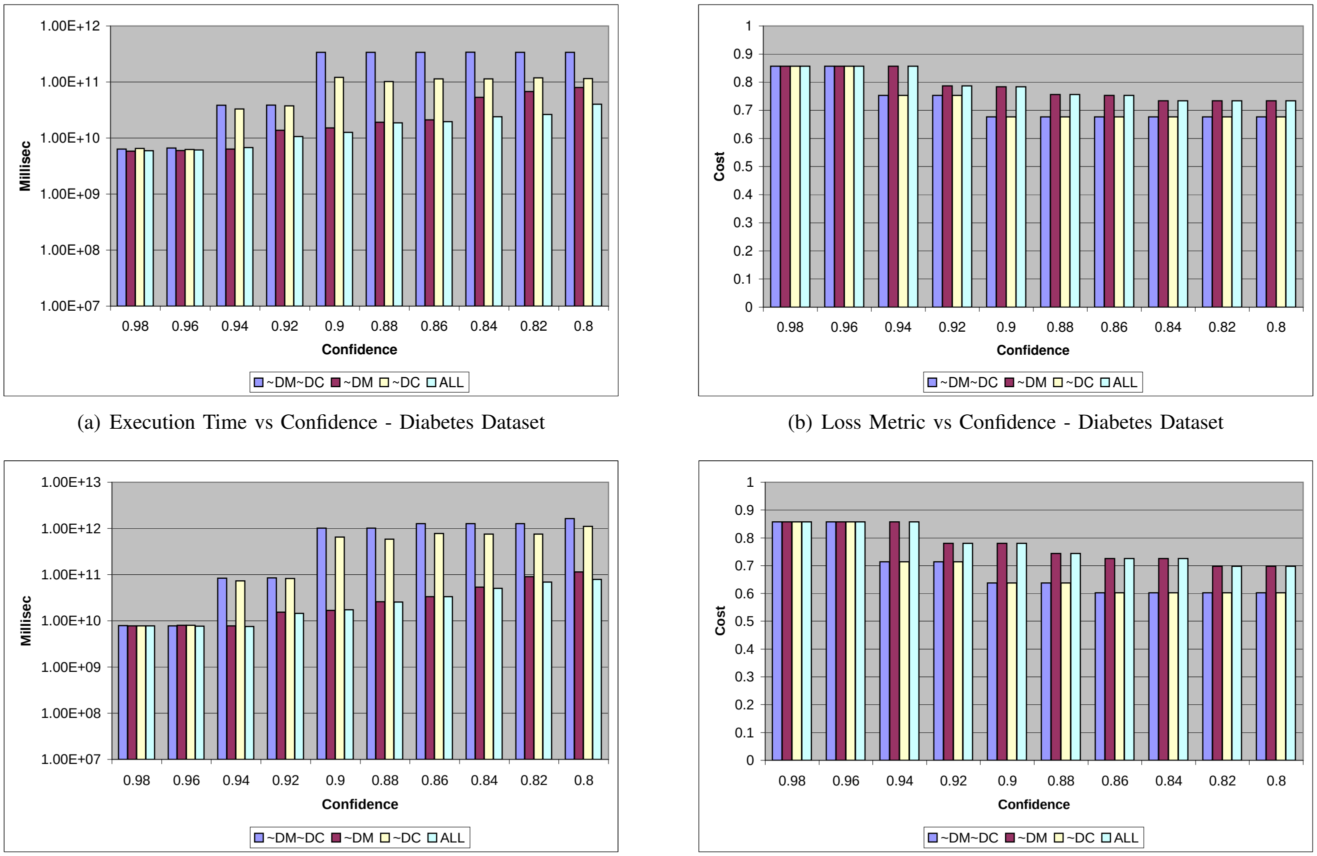
Fig. 6. Prune Rates for 7 dimensions

world privacy issues not adequately addressed by other methods.

The δ-presence definition can be revisited by assuming an adversary with varying levels of background knowledge; an adversary who knows more (e.g., the weight of an individual) gains in their ability to identify an individual, but also in their prior estimation of sensitive data. For example, knowing an individual is obese may make them easier to identify than not knowing their weight, but even without the anonymized data an adversary would have a strong reason to believe the individual was at risk for diabetes. As adversary prior knowledge increases, the probability of disclosure increases but the cost from disclosure decreases; giving a cost-utility tradeoff (instead of simply privacy-utility).

It is also possible to use randomization instead of generalizations on the private dataset to provide δ-presence. The authors are currently working on a hybrid approach where generalization is done through probability distributions. In all these cases, more advanced bayesian or statistical techniques would be required.

We have shown that the δ-presence measure first introduced in [1] not only provides a more meaningful approach to privacy than competing metrics, but with this paper we show that it can be practically achieved.



(a) Execution Time vs Confidence - Diabetes Dataset

(b) Loss Metric vs Confidence - Diabetes Dataset

(c) Execution Time vs Confidence - Random Dataset

(d) Loss Metric vs Confidence - Random Dataset

Fig. 7. Result with varying confidence with 7 dimensions

References

[1] M. Nergiz, M. Atzori, and C. Clifton, “Hiding the presence of individuals from shared databases,” in *Proc. of the 2007 ACM SIGMOD Intl. Conf. on Management of Data*, Beijing, China, Jun. 11-14 2007, pp. 665–676. [Online]. Available: http://doi.acm.org/10.1145/1247480.1247554

[2] “Standard for privacy of individually identifiable health information,” *Federal Register*, vol. 67, no. 157, pp. 53 181–53 273, Aug. 14 2002. [Online]. Available: http://www.hhs.gov/ocr/hipaa/finalreg.html

[3] P. Samarati, “Protecting respondent’s privacy in microdata release,” *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, Nov./Dec. 2001. [Online]. Available: http://dx.doi.org/10.1109/69.971193

[4] L. Sweeney, “k-anonymity: a model for protecting privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, no. 5, pp. 557–570, 2002. [Online]. Available: http://dx.doi.org/10.1142/S0218488502001648

[5] 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. [Online]. Available: http://dx.doi.org/10.1016/S0933-3657(98)00056-6

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “l-diversity: Privacy beyond k-anonymity,” in *Proc. of the 22nd IEEE Intl. Conf. on Data Engineering (ICDE 2006)*, Atlanta Georgia, Apr. 2006. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2006.1

[7] X. Xiao and Y. Tao, “Anatomy: Simple and effective privacy preservation,” in *Proc. of 32nd Intl. Conf. on Very Large Data Bases (VLDB 2006)*, Seoul, Korea, Sep. 12-15 2006. [Online]. Available: http://www.vldb.org/conf/2006/p139-xiao.pdf

[8] 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. [Online]. Available: http://diabetes.niddk.nih.gov/dm/pubs/statistics/

[9] American Diabetes Association, “Direct and indirect costs of diabetes in the United States,” http://www.diabetes.org/diabetes-statistics/cost-of-diabetes-in-us.jsp, 2006.

[10] C. Blake and C. Merz, “UCI repository of machine learning databases,” 1998. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[11] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume 1*. Wiley, 1968.

[12] S. Lahiri, A. Chatterjeea, and T. Maiti, “Normal approximation to the hypergeometric distribution in nonstandard cases and a sub-gaussian berryesseen theorem,” *Journal of Statistical Planning and Inference*, vol. 137, no. 11, pp. 3570–3590, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.jspi.2007.03.033

[13] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Incognito: Efficient full-domain k-anonymity,” in *Proc. of the 2005 ACM SIGMOD Intl. Conf. on Management of Data*, Baltimore, MD, Jun. 13-16 2005. [Online]. Available: http://doi.acm.org/10.1145/1066157.1066164

[14] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, “Mondrian multidimensional k-anonymity,” in *Proc. of the 22nd Intl. Conf. on Data Engineering (ICDE ’06)*, Atlanta, GA, Apr. 3-7 2006, pp. 25–35. [Online]. Available: http://dx.doi.org/10.1109/ICDE.2006.101

[15] C. C. Aggarwal, “On k-anonymity and the curse of dimensionality,” in *VLDB ’05: Proc. of the 31st intl. conf. on Very large data bases*. VLDB Endowment, 2005, pp. 901–909.

[16] 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. [Online]. Available: http://doi.acm.org/10.1145/775047.775089

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

[18] G. Agrawal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu, “Achieving anonymity via clustering,” in *PODS ’06: Proc. of the 25th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, Chicago, IL, USA, Jun. 26-28 2006, pp. 153–162.