Hiding the Presence of Individuals from Shared Databases

M. Ercan Nergiz [*]
CS Dept., Purdue University
305 N. University Street
West Lafayette, Indiana,
47907-2107
mnergiz@cs.purdue.edu

Maurizio Atzori [†]
KDD Laboratory, ISTI-CNR
Area della ricerca di Pisa
via G. Moruzzi 1
56124 Pisa, Italy
atzori@di.unipi.it

Christopher W. Clifton
CS Dept., Purdue University
305 N. University Street
West Lafayette, Indiana,
47907-2107
clifton@cs.purdue.edu

ABSTRACT

Advances in information technology, and its use in research, are increasing both the need for anonymized data and the risks of poor anonymization. We present a metric, δ-presence, that clearly links the quality of anonymization to the risk posed by inadequate anonymization. We show 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), and present algorithms for effectively anonymizing to meet δ-presence. The algorithms are evaluated in the context of a real-world scenario, demonstrating practical applicability of the approach.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Statistical Databases;
K.4.1 [**Public Policy Issues**]: Privacy

General Terms

Algorithms, Security, Legal Aspects

Keywords

k-anonymity, privacy, delta presence, medical databases

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) [7] 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 [15, 16]. 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

[*] Nergiz's work supported by the National Science Foundation under Grant No. 0428168.
[†] Dr. Atzori's work supported by the EU GeoPKDD project (IST-6FP-014915).

SIGMOD'07, June 12–14, 2007, Beijing, China.
Copyright 2007 ACM 978-1-59593-686-8/07/0006 ...$5.00.

protection against disclosure of that fact. This has lead to alternate approaches, such as discernibility [8] / l-diversity [12]. 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 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 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 [17]. 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. We present other approaches that meet the standard while providing much greater detail / value in the disclosed 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 [13]. Of note is that people with diabetes have medical expenditures 2.4 times the expenditures if they did not have diabetes [1]; 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 [6] 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 [13], 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.

We now give background and notations used in the paper, followed by the formal definition of δ-presence. In Section 4 we will evaluate the increase in risk posed by a given belief that an individual is or is not in the database. This increase in risk is an appropriate metric for setting policy, and can be mapped back to pick a meaningful value of δ that satisfies the policy. Section 5 gives an example of why k-anonymity and related techniques are not appropriate solutions. We give two algorithms for anonymizing a database to achieve δ-presence in Section 6; demonstrating different approaches and why each might be appropriate for different purposes. Section 7 gives a set of experiments showing how k-anonymity and the algorithms of Section 6 affect the quality of data in achieving δ-presence. We conclude with a discussion of future work in this area.

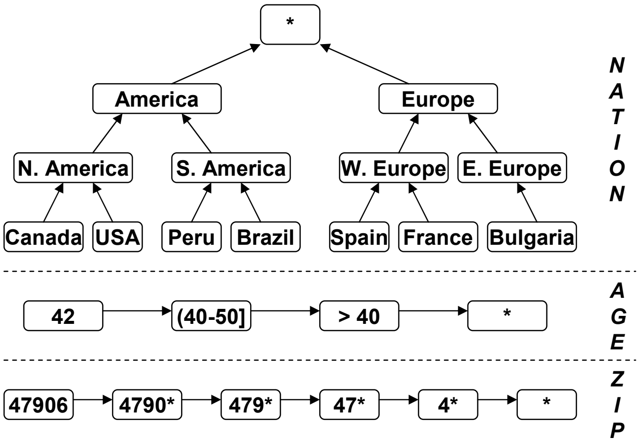


Figure 1: DGH structures

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).

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., $\psi(\text{USA}) = \{\text{USA, N. America, America, *}\}$) 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_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 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 Tables 1-3, tables P_1^*, P_2^* and P_3^* are different generalizations of table P. (The T tables will be discussed in Section 3, and should be ignored for now.)

DEFINITION 3 (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.*

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

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

Table 2: *k*-anonymization of Table 1

P_1^*

	Publicly Released Dataset			
	Zip	Age	Nationality	Sen.
a	4*	≤ 40	*	0
d	4*	≤ 40	*	0
e	4*	≤ 40	*	0
g	4*	≤ 40	*	0
b	4*	> 40	*	1
c	4*	> 40	*	1
f	4*	> 40	*	1
h	4*	> 40	*	1
i	4*	> 40	*	1

T_1^*

	Research Subset		
	Zip	Age	Nationality
b	4*	> 40	*
c	4*	> 40	*
f	4*	> 40	*
h	4*	> 40	*
i	4*	> 40	*

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_1^* of Table 2 is a 4-anonymous generalization of *P*. Note that by only seeing P_1^*, an adversary can at best link a tuple <47906,35,USA>, Alice, to the tuples a, d, e, and g of P_1^*.)

DEFINITION 4 (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_1^*, the equivalence class for tuple a is $\{a, d, e, g\}$.

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 table with sensitive information (disease attribute), the adversary can match quasi-identifiers in both tables to discover unique links between records in the public and released table. *k*-anonymity limits the linking ability of the adversary to groups of at least *k* records by the use of table generalizations.

3. PRESENCE OF INDIVIDUALS IN DATA

Given that being linked to the research subset is a privacy risk, we instantly see that releasing the research subset *T* is not acceptable; each individual can be uniquely linked with the publicly known data. *k*-anonymization does not solve the problem – even though T_1^* is anonymized to a single group, someone who knows the publicly known data in *P* can identify Bob, Christine, Frank, Harry, and Iris as being in the research subset based on their age. (This is because of the lack of ℓ-diversity with respect to the sensitive attribute that makes individuals candidates for the research subset, but as we discuss in Section 5.2 this is not the entire 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 5 (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*

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

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

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

We now briefly discuss some properties of δ-presence.

Anti-monotonicity.

DEFINITION 6 (NON-OVERLAPPING GENERALIZATION). *Given a public table P, private table T and a generalization T^* of*

T, we say T^* is non-overlapping w.r.t. P and T if and only if there does not exist $p \in P$, $t_1^*, t_2^* \in T^*$ such that $t_1^* \neq t_2^*$ and $t_1^* \in \psi(p)$, $t_2^* \in \psi(p)$

In other words, a generalization is non-overlapping when a "real" tuple can match at most one generalized tuple. In Tables 1, 2, and 3; datasets P_1^*, P_2^*, and P_3^* are non-overlapping generalizations of P. Similarly T_1^* and T_3^* are such generalizations of T.

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_1^* is $(\delta_{min}, \delta_{max})$ present w.r.t. P and T then so is T_2^*.*

PROOF. T_1^* is non-overlapping δ-present if and only if for every distinct tuple $p \in P$;

$$\delta_{min} \leq \mathsf{P}(p \in T \mid T_1^*) = \frac{C(p_1^*, T_1^*)}{\Sigma_{t \mid p_1^* \in \psi(t)} C(t, P)} \leq \delta_{max}$$

where $C(t,T)$ is the cardinality of tuple t in table T and p_1^* is the tuple in T_1^* with $p_1^* \in \psi(p)$. (There can exactly be one such distinct tuple if $\delta_{min} \neq 0$. Otherwise, if such a tuple does not exist, p_1^* is the null tuple.) Let $A(t')$ be the function for $C(t', T_1^*)$ and similarly $B(t') = \Sigma_{t \mid t' \in \psi(t)} C(t,P)$. Since T_2^* is a non-overlapping generalization of T_1^*, for every distinct tuple $p \in P$, $\mathsf{P}(p \in T \mid T_2^*)$ can be calculated in terms of A and B;

$$C(p_2^*, T_2^*) = \Sigma_{t_1^* \mid p_2^* \in \psi(t_1^*)} A(t_1^*)$$
$$\Sigma_{t \mid p_2^* \in \psi(t)} C(t,P) = \Sigma_{t_1^* \mid p_2^* \in \psi(t_1^*)} B(t_1^*)$$

where p_2^* is the tuple in T_2^* with $p_2^* \in \psi(p)$. Since $x \leq \frac{a_1}{b_1}, \cdots, \frac{a_n}{b_n} \leq y$ implies $x \leq \frac{a_1 + \cdots + a_n}{b_1 + \cdots + b_2} \leq y$;

$$\delta_{min} \leq \frac{\Sigma_{t_1^* \mid p_2^* \in \psi(t_1^*)} A(t_1^*)}{\Sigma_{t_1^* \mid p_2^* \in \psi(t_1^*)} B(t_1^*)} = \frac{C(p_2^*, T_2^*)}{\Sigma_{t \mid p_2^* \in \psi(t)} C(t,P)} \leq \delta_{max}$$

Then T_2^* is also δ present w.r.t. P, T. □

COROLLARY 1. *If T_2^* is not δ present w.r.t. P and T then neither is T_1^*.*

Ranges and Bounds. Given a dataset T^* which is a total suppression of T we have $\mathsf{P}(t \in T \mid T^*) = \frac{|T|}{|P|}$ for any $t \in P$. This means T^* respects $(\frac{|T|}{|P|}, \frac{|T|}{|P|})$-presence but does not respect $(\delta'_{min}, \delta'_{max})$-presence if $\frac{|T|}{|P|} < \delta'_{min}$ or $\delta'_{max} < \frac{|T|}{|P|}$. Since T^* is a non-overlapping generalization of any dataset T_g^* that, in turn, it is a non-overlapping generalization of T, by anti-monotonicity property we have that T_g^* does not respect $(\delta'_{min}, \delta'_{max})$-presence neither. This means any presence requirement $(\delta_{min}, \delta_{max})$ for a public dataset P and private subset T should satisfy $\delta_{min} \leq \frac{|T|}{|P|} \leq \delta_{max}$.

4. INFORMATION GAIN: SELECTING A GOOD δ

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 [13], any adversary will have a prior belief b_r on I_p given the public dataset P:

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

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:

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

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 $c <$ \$100, the risk of misuse is acceptably small. The total cost of diabetes per person is around $d = \$10,000$ [1]. The probabilistic acceptable misuse, *am*, is then $\frac{c}{d} = \frac{1}{100}$; we must ensure:

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

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

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

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

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 $\mathsf{P}(p_1 \in T \mid T^*) \backsimeq 0$) 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 $am = \frac{200}{10000} = 0.02$ then

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

This gives us a maximum δ_{min} parameter.

5. RELATED WORK

In this section, we investigate possible solutions to the δ-presence problem using previously proposed generalization-based privacy methods. Throughout the section, we assume we have public dataset P and private dataset T from Table 1, and we want to enforce a non-overlapping $(\frac{1}{2}, \frac{2}{3})$-presence.

5.1 k-anonymity

By its definition, k-anonymity assumes all public tuples reside in the private dataset so presence of a given public tuple is not an unknown. (E.g., $P(t \in T) = 1$ for any public $t \in T$.) This implies that direct optimal k-anonymization approach (an optimal k-anonymity algorithm, while providing anonymity, also minimizes a given cost metric to maximize utility) is not suitable for the δ-presence problem and finding a suitable k parameter to provide presence may be impossible. T_1^* in Table 2 shows a non-overlapping 5-anonymization of T that is optimal by the Loss (LM) and Discernibility (DM) metrics (discussed further in Section 7). T_1^* violates $(\frac{1}{2}, \frac{2}{3})$ presence (actually any presence with $\delta_{max} < 1$ is violated) since there are only 5 tuples in P and T with age '> 40' meaning every such tuple is in T. (E.g., $P(t[id = b] \in T) = 1$.) Since any optimal solution, regardless of k, maps age column to value '> 40' or a lesser granularity, it turns out that optimal solutions for any k (based on generalization) will not be $(\frac{1}{2}, \frac{2}{3})$-present for T. (k-anonymity methods can give δ-presence, as we shall see in Section 7, but there is not a direct relationship between k and δ.)

The above approach fails because public dataset P is not taken into account in the anonymization process. A simple solution would be to anonymize the public dataset P and assign the anonymization mappings to tuples in private dataset T. This parametric anonymization technique (w.r.t. a given public dataset P), called *weak k-anonymity* and proposed in [4] to reduce data distortion, still does not necessarily provide presence property since it still does not take into account which tuples are in T and which are not. P_1^* in Table 2 shows the optimal non-overlapping 4-anonymization of P w.r.t. LM and DM metrics. The mapping in P_1^*, when applied to T, creates T_1^* which is not $(\frac{1}{2}, \frac{2}{3})$-present. Dataset P_2^* in Table 3 is an optimal anonymization for $k = 2$ and $k = 3$ which is still not δ-present. The only optimal k-anonymization that gives δ-presence in this example is full suppression although there exist δ-present generalizations of higher utility for P,T.

5.2 ℓ-diversity

k-anonymization on public dataset P failed because it does not enforce mixing of *present* and *absent* tuples inside any equivalence class. One approach is to represent presence and absence information in the *Sen.* attribute of P; this could be considered a *sensitive* attribute. [12] is a related notion that extends the definition of k-anonymity to enforce diversity among sensitive values of equivalence classes. The most flexible variation of ℓ-diversity is the recursive (c, ℓ) diversity which enforce the rule; $r_1 \leq c(r_\ell + \cdots + r_m)$ in every equivalence classes *eq* where r_i is the ith frequent sensitive value in *eq* and m is the total number of distinct sensitive values in *eq*. (Original definition of recursive diversity uses $<$ in the enforced rule. Without loss of generality, we stick to $\leq$ version to ease discussion) In our case, we only have 2 different sensitive values so $m = 2$ and $\ell = 2$ with $c \geq 1$ makes sense. So given n_1 is the number of tuples with Sen.:1 and n_0 is the number of tuples with Sen.:0 in a given equivalence class *eq*, the following constraints are enforced for each *eq*;

$$
\begin{array}{lcl}
 & \frac{r_1}{r_2} & \leq c \\
(\dfrac{n_1}{n_0} \leq c \wedge n_1 \geq n_0) & \vee & (\dfrac{n_0}{n_1} \leq c \wedge n_0 > n_1) \\
(\dfrac{n_1}{n_0} \leq c \wedge n_1 \geq n_0) & \vee & (\dfrac{n_1}{n_0} \geq \dfrac{1}{c} \wedge n_0 > n_1)
\end{array}
$$

and for $c \geq 1$;

$$\frac{1}{c} \leq \quad \frac{n_1}{n_0} \quad \leq c \qquad (1)$$

For $(\frac{a_1}{b_1}, \frac{a_2}{b_2})$ presence in non-overlapping anonymizations, the following constraints should be enforced for each equivalence class:

$$
\begin{array}{rcl}
\dfrac{a_1}{b_1} \leq & \frac{n_1}{n_1+n_0} & \leq \dfrac{a_2}{b_2} \\
\dfrac{b_2}{a_2} \leq & \frac{n_1+n_0}{n_1} & \leq \dfrac{b_1}{a_1} \\
\dfrac{b_2}{a_2} - 1 \leq & \frac{n_0}{n_1} & \leq \dfrac{b_1}{a_1} - 1 \\
\dfrac{b_2 - a_2}{a_2} \leq & \frac{n_0}{n_1} & \leq \dfrac{b_1 - a_1}{a_1}
\end{array}
$$

$$\frac{a_1}{b_1 - a_1} \leq \quad \frac{n_1}{n_0} \quad \leq \frac{a_2}{b_2 - a_2} \qquad (2)$$

There are two main reasons for why ℓ-diversity is not suitable for providing (δ_{min}, δ_{max}) presence. First is that the recursive (c, ℓ) diversity only has one parameter c to express the two parameter $(\delta_{min}, \delta_{max})$-presence. Second; recursive (c, ℓ) diversity does not distinguish between the values of sensitive attributes, that is Equation 1 is symmetric (e.g., if it is okay to have m tuples with Sen.:1 and n tuples with Sen.:0 in an equivalence class then it is also okay to have n tuples with Sen.:1 and m tuples with Sen.:0.) Equation 2 is not symmetric for most values of δ_{min}, δ_{max}. This also makes it impossible to fit a recursive (c, ℓ) diversity constraint into even one of the δ constraints.

Table 3 shows an example of this. Since the number of parameters do not match, the best we can do is to match the bounds with each other one at a time. To match the upper bound constraint of presence ($\delta_{max} = \frac{2}{3}$) with the upper bound constraint of diversity, we need to set $c = \frac{a_2}{b_2 - a_2}$ where $\frac{a_2}{b_2} = \frac{2}{3}$ so $c = 2$. Table P_2^* shows an optimal recursive $(2, 2)$ diverse P. However this generalization mapping does not create a corresponding $(\frac{1}{2}, \frac{2}{3})$-present T generalization since for equivalence class of tuples d, e and f, $\frac{n_1}{n_0+n_1} = \frac{1}{3} \leq (\delta_{max} = \frac{1}{2})$. If we match the lower bounds $c = \frac{b_1 - a_1}{a_1}$ where $\frac{a_1}{b_1} = \frac{1}{2}$, we get $c = 1$. Since $n_1 > n_0$ in P, there does not exist $(2, 1)$-diverse non-overlapping generalization of P and consequently there is no generalization mapping to enforce presence on T. However dataset T has a $(\frac{1}{2}, \frac{2}{3})$ present generalization; T_3^*.

ℓ-diversity provides diversity on sensitive attributes, however we require constraining the *distribution* of sensitive attributes. A more flexible algorithm that will enforce Equation 2 will be introduced in Section 6. Although this paper focuses on non-overlapping anonymizations; checking $(\delta_{min}, \delta_{max})$ presence on anonymizations containing overlapping equivalence classes is much more difficult and complex. In that case, equivalence classes will not be independent and ℓ-diversity would be irrelevant.

It should be noted that Equation 2 is the necessary and sufficient condition for δ-presence property for non overlapping generalizations. So a direct cost-optimal approach to δ-presence that constraints non-overlapping generalizations with Equation 2 would be

Table 3: 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

an upper bound for any other approach (such as k-anonymity and ℓ-diversity) in terms of data utilization.

6. ALGORITHMS

We now introduce algorithms for achieving δ-presence. We first give an optimal full-domain generalization algorithm; this works under the constraint that if a value is generalized, all occurrence of that value must be generalized. We then relax this restriction. (The difference between these approaches is analogous to the difference between the k-anonymization algorithms of [10] and [11].)

6.1 Single-Dimensional Presence Algorithm: SPALM

In Theorem 1, we proved the monotonicity property of presence. This property can be used to create optimal (w.r.t. a precision metric) and practical presence algorithms that make use of apriori-style pruning on the universal candidate space. Such algorithms were proposed in [15, 10, 12] for k-anonymity when only *full domain generalization* of datasets are allowed. In this subsection, we present a similar algorithm SPALM that produces δ-present full-domain generalizations and at the same time maximizes a given precision metric. The following notations and definitions briefly recall the problem setting:

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 $\Delta_i^n(v) = \underbrace{\Delta_i(\cdots\Delta_i(v)\cdots)}_{n}$.

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 $1 \le i \le n$. 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 $g'_i \le g''_i$ for all $1 \le i \le n$. For cost metrics proposed so far, a high level generalization (e.g., T'') is more costly than a lower generalization (e.g., T').

The possible full domain generalizations of table T form a lattice on the $\gg$ relation. (see Figure 2.) 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 monotonicity or 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 [2].

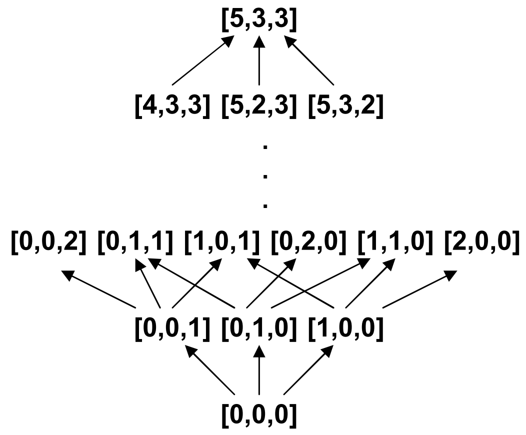


Figure 2: Full Domain Generalization Lattice

The following pseudo-code summarizes SPALM.

Algorithm 1 SPALM

```
Require: publicly available table $P$; private table $T$, a cost metric COST;
Ensure: return minimum cost $(\delta_{min}, \delta_{max})$ present full domain generalization of $T$.
insert sen. attribute into $P$ according to $T$ as in Table 1.
create lattice $lat$ for all possible generalization mappings for $N$. 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 $N^*$; full domain generalization of $N$ according to mapping in $m$
        if $N^*$ is not $(\delta_{min}, \delta_{max})$ present then
            delete node $m$ and all siblings and grandsiblings of $m$ from $lat$
return the least-cost generalization and the corresponding mapping among the generalizations being tested.
```

SPALM, in the worst case, checks for δ-presence for every generalization mapping on the lattice. Given that the fully suppressed table P^* is a $[g_1, \cdots, g_n]$ full domain generalization of the original P, the number of nodes in the lattice is $\prod_{i=1}^{n} g_i$. If no pruning can be performed, the worst-case complexity of SPALM is $O(\prod_{i=1}^{n} g_i \cdot |P|)$.

6.2 Multi-Dimensional Presence Algorithm: MPALM

While the SPALM algorithm is an optimal algorithm, it relies on single-dimensional generalizations, that is, if a value in a tuple is generalized, then all such values in the table are generalized. Clearly, if we relax this constraint, the search space is a superset that may produce better optimal solutions. Although finding optimal solutions in the multi-dimensional case may be infeasible, recent work [11] shows that sub-optimal generalizations are likely to give lower distortion than optimal single-dimension generalization. Here we consider a sub-optimal multi-dimensional generalization algorithm that provides δ-presence; we also propose a number of heuristics to be evaluated in the experiments section.

Algorithm 2 MPALM

```
Require: publicly available table $P$; private table $T$
Ensure: return a $(\delta_{min}, \delta_{max})$-present multi-dimensional generalization of $T$ w.r.t $P$
 1: Let $Q$ be an empty queue of tables;
 2: Let $R$ be an empty set of tables;
 3: $enqueue(Q, P)$;
 4: while $Q$ is not empty do
 5:     $T_i \leftarrow dequeue(Q)$;
 6:     $j \leftarrow 0$;
 7:     repeat
 8:         $j$++;
 9:         $C \leftarrow choose\_column(T_i, T, j)$;
10:         $v \leftarrow choose\_value(T_i, T, C)$;
11:     until $v \neq null$ or $j = number\_of\_columns(T)$
12:     if $v \neq null$ then {split column $C$ on value $v$}
13:         $T_i^{<v} \leftarrow \{t \in T_i \mid t[C] < v\}$
14:         $T_i^{\geq v} \leftarrow \{t \in T_i \mid t[C] \geq v\}$
15:         $enqueue(Q, T_i^{<v})$;
16:         $enqueue(Q, T_i^{\geq v})$;
17:     else {no δ-present split found}
18:         $R \leftarrow R \cup T_i$;
19: for all $T_i \in R$ do
20:     return smallest bounding box of $T_i \cap T$;
```

Algorithm 2 describes MPALM, which allows a group of tuples to be generalized while leaving other tuples with the same values in their original state. Q is a queue containing the portions of the table to be generalized further. Notice that at the beginning, it contains P, not T. This is because the algorithm, during the generalization step (lines 9–10), needs to consider tuples not in T to enforce δ-presence. In the while loop (4–18), the algorithm extracts a table T_i from Q and tries to partition the tuples into two partitions: the former (line 13) containing tuples where the value of attribute C is less than v, and the latter (line 14) with the remaining tuples. Choosing C and v can be done arbitrarily, and leads to different strategies. We considered 3 strategies for choosing the column and 3 strategies for choosing the threshold value as discussed below.

The output of MPALM is a set of smallest bounding boxes, one for each $T_i \in R$. A bounding box is a rectangle if T has two columns; a multidimensional cuboid in general, where each dimension represents a column of T. To be the smallest bounding box, the cuboid must contain all the tuples in T_i and there must not be a smaller one with the same property.

At each iteration, MPALM chooses a column to be split for tuples in T_i. Columns where no splitting leads to δ-presence are skipped, and the next one (according to the column strategy) will be chosen. The δ-presence constraint can be verified on T, and $T \cap T_i$ is the set of tuples in T_i to be disclosed. This choice can be done arbitrarily, but a good heuristic will improve data utility. The strategies we considered are the following:

next (n) gives priority to the next column (by rotating through columns);

priority (p) is a static user-defined priority. Columns with high priority are less likely to be generalized since at each loop the algorithm will try to split them. In Section 7, we used the order in the dataset (adult), i.e., age, workclass, education, marital-status, occupation, relationship, race, gender, native-country;

best (b) selects the column with the largest number of attribute values among tuples in T_i. This strategy tries to avoid generalizing values with the most diversity.

Another degree of freedom in the MPALM algorithm is how to choose the attribute value for splitting, once the column has been chosen. We looked at three heuristics:

balanced cardinality (1) selects the value that minimizes $|T_i^{<v} - T_i^{\geq v}|$;

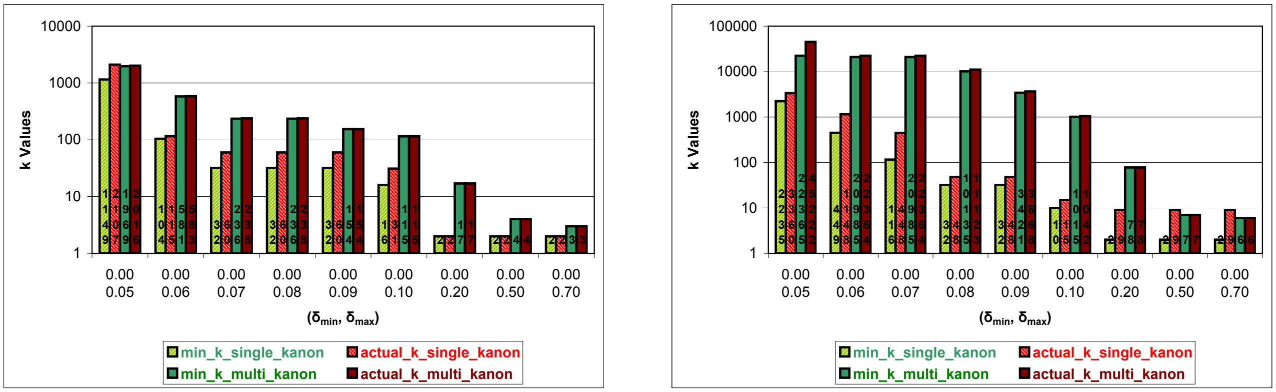
first split (2) takes the first split encountered, i.e., $0 < |T_i^{<v}|$ is minimized. This strategy clearly produces unbalanced splits;

balanced attribute-values (3) is similar to (1), but chooses a value v of a given column C s.t. $|values_C\ (T_i^{<v}) - values_C\ (T_i^{\geq v})|$ is minimized, where $values_C\ (T)$ is the number of different attribute values for column C in table T.

The most computationally expensive part of MPALM is the inner loop (repeat) of line 7–11, which is inside the outer loop (while) 4–18. For each set of tuples T_i extracted from the queue Q, the heuristics *choose_column* and *choose_value* require $|C| \cdot |T_i|$ passes in the worst case, where $|C|$ is the number of columns of table T. Thoroughly, we have $\sum_{T_i \in Q} |C| \cdot |T_i| = |C| \sum_{T_i \in Q} |T_i|$ passes. The sequence of T_i extracted from Q depends on data distribution and the heuristics used to split data (after each split, subtables are enqueued again until no further split on any column is possible.) Note that the maximum number of split is given by $|P|$. For perfectly balanced splits, we have a sequence of T_i coming out from the queue of sizes $|P|$: $\frac{|P|}{2}$, $\frac{|P|}{2}$, then $\frac{|P|}{4}$ four times, which adds up to $O(|P| \log_2 |P|)$. In this case, computational complexity of MPALM is therefore $O(|C||P| \log_2 |P|)$. Unbalanced splits give a queue Q of T_i of size $|P|, |P|-1, 1, |P|-2, 1, \ldots, |P|-(|P|-1), 1$, which leads to a worst case complexity of $O(|C||P|^2)$.

7. EXPERIMENTS

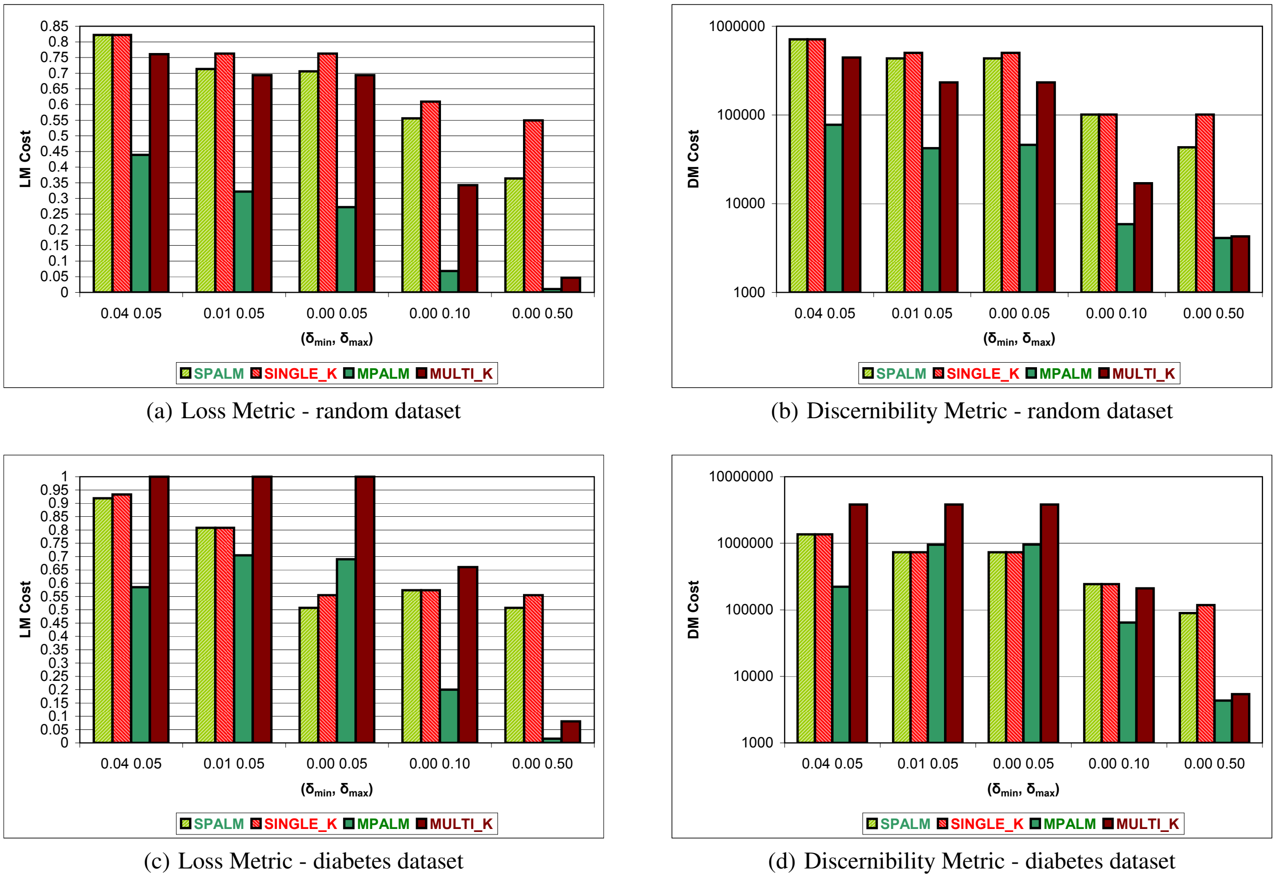
As has been learned with k-anonymization, the right way to anonymize data can be very dependent on the data and purpose for which it is to be used [14] We have presented two different approaches to enforcing δ-presence, with several variants of these approaches. We now compare those approaches on both simulated and real data.



(a) Randomly selected data

(b) Simulated diabetes patients

Figure 3: Minimum needed and actual values of *k* for achieving δ-presence



(a) Loss Metric - random dataset

(b) Discernibility Metric - random dataset

(c) Loss Metric - diabetes dataset

(d) Discernibility Metric - diabetes dataset

Figure 4: *k*-anonymity vs. δ-presence algorithms

We evaluate how varying δ affects the cost of anonymizing the dataset, as determined by the Loss Metric (LM) [9] and the Discernibility Metric (DM) [5]. The LM measures the amount of generalization as a normalized information loss. For example, in Figure 1 generalizing "Canada" to "N. America" incurs a penalty of $2/6$; to "America" gives a penalty of $4/6$. The DM measures the size of the induced equivalence classes. Each tuple is assigned a cost based on the number of identical tuples in the anonymized dataset. With k-anonymization, the penalty essentially captures how much k is exceeded. This probably unfairly penalizes k-anonymity, as it would be possible to generalize in a way that satisfies δ-presence while making each row in the anonymized dataset unique, thus giving no DM-penalty.

The simulated dataset is created through random selection of a 4% subset of the UCI adult dataset [6]. The entire adult dataset (specifically, the 45222 records with on 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.

For a more realistic test we also perform 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 [13].) 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 [13]; 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 evaluated several approaches to achieving δ-presence on these datasets. As a baseline, we start with k-anonymity. This posed several difficulties. First, what is the appropriate value of k? As discussed in Section 5.1, there is no direct mapping between δ and k, even for the relatively simple case of $\delta_{min} = 0$. It is even possible that for a given k, one anonymization may satisfy δ-presence and another may not.

Figure 3 shows the minimum value of k needed to achieve various values of δ through generalization as well as the actual size of the smallest equivalence class when the anonymization algorithm was run for that value of k (using an exhaustive search for the smallest k such that the anonymization algorithm satisfied the given δ.) This is done for two types of k-anonymization: an optimal single dimensional anonymization in the style of [10] (i.e., if a value is generalized, *all* occurrences of that value are anonymized), and a heuristic multi-dimensional k-anonymization in the style of Mondrian [11] that relaxes this restriction. In some cases, the entire dataset is anonymized (in effect suppressing all quasi-identifiers.)

For higher values of δ_{min} (e.g., $\delta = (.04, .07)$), no value of k satisfied δ-presence. The problem is that while k-anonymity prevents linking an individual to a record, it does not prevent *excluding* an individual from matching a record.

Figure 4 shows the utility impact of using k-anonymity to achieve δ-presence versus the algorithms of Section 6. (The MPALM numbers are for the n1 strategy, as this is the most analogous with k-anonymity approaches.) In general, we see that the multi-dimensional δ-presence algorithm gives significantly higher data utility as judged by both loss and discernibility metrics (although in one case the optimal single-dimensional optimization fares significantly better.) Of note is that the multidimensional approach seems to have trouble with the skewed dataset; more on this later.

We next compare several strategies for directly achieving δ-presence using the algorithms given in Section 6. Figures 5 and 6 highlight the anonymization cost; we see that the multidimensional approach achieves privacy with lower distortion of the dataset, particularly for larger values of δ_{max}. This holds even though the single-dimensional approach is finding the optimal anonymization given the constraint that if a value is generalized, all occurrences of that value are generalized. This demonstrates significant benefit to increasing the flexibility of the generalization, as is done in the multidimensional case. (Similar results were obtained on the random dataset, but are omitted due to space constraints.) In particular the n2 strategy of rotating through columns and choosing the "first fit" generalization appears effective.

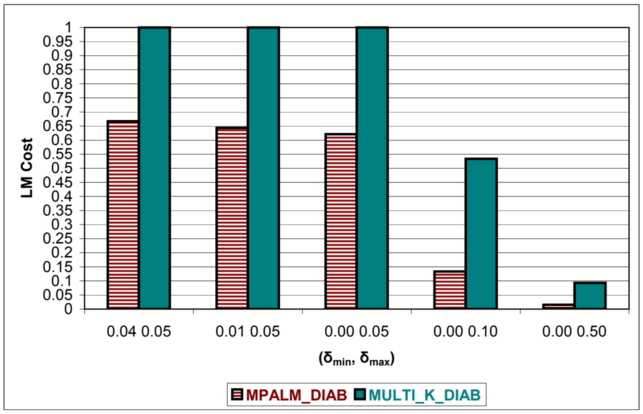


Figure 9: k-anonymity vs. MPALM for the n2 strategy on the diabetes dataset

Not only does the heuristic multi-dimensional generalization give as good or better data utility than the optimal single-dimensional generalization, it is also efficient. Figure 7 reports runtimes on a 2.16GHz Intel Core 2 Duo. It may seem surprising that as the amount of generalization required goes down, the computational cost goes up. This makes sense when we consider that the search space (number of generalizations meeting the δ-presence requirements) increases as we relax δ.

To better understand the effect of δ_{min} and δ_{max} on the distortion of the dataset, we look more deeply into the n2 strategy in Figure 8. High values of δ_{min} (i.e., preventing discovery that an individual is *not* in the dataset) comes at a high price, particularly for the biased dataset. This is not surprising; the greater homogeneity of the biased dataset makes it more likely that some individuals not in the dataset would not be similar to any of the individuals in the dataset, forcing a high level of generalization. Somewhat surprising is the outcome of low δ_{min} with low δ_{max} or high δ_{min} with high δ_{max} with the biased dataset. For a low probability that an individual is in the dataset, the algorithm is giving better utility when it is forced to ensure that individuals cannot be excluded from the dataset. Conversely, if individuals cannot be excluded from the dataset, better results are achieved if individuals cannot be determined to be in the dataset. This is an anomaly resulting from the heuristic nature of the search, as from Definition 5 it is clear that a dataset satisfying (say) $\delta = (0.02, 0.05)$ also satisfies $\delta = (0, 0.05)$ Since this anomaly doesn't appear on the randomly selected data, we feel that appropriate strategies for achieving δ-presence on skewed datasets is a good challenge for future work.

Finally, we compare the n2 strategy with Mondrian-style k-anonymity using the same strategy. Figure 9 shows that k-anonymity still is not as effective at preserving data utility while preventing disclosure that an individual is in a dataset.

8. CONCLUSIONS AND FUTURE WORK

While *k*-anonymity and related techniques have received considerable attention, it isn't clear that it is the best way to balance privacy and data utility [17]. We have presented a problem where anonymization *is* an appropriate solution, and a metric δ-presence 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.

Extending this work to linking with sensitive data in a disclosed dataset, as with ℓ-diversity, is straightforward (and could be easily accomplished using the technique of [17].) Other areas where work is needed include investigating algorithms for achieving δ-presence; the algorithms in Section 6 are given as an exploration of the space rather than a "best solution". It is also possible to design δ-presence algorithms that guarantee bounds on optimality as it is done for *k*-anonymity in [3]. Further development of δ-presence will address a variety of real-world privacy issues that are not adequately addressed by other methods.

δ-presence definition can be revisited by assuming a stronger adversary with more background knowledge. It should be noted that as adversary prior knowledge increases, the cost of disclosure decreases and a cost-utility relation could be addressed (instead of privacy-utility). In certain applications, where a public dataset is not available but some statistical properties of the public dataset (e.g., distribution of values, size, ...) is known, enforcing δ-presence property becomes even more challenging. It is also possible to use randomization instead of generalizations on the private dataset to provide δ-presence. (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.

9. ACKNOWLEDGMENTS

Dr. Atzori wishes to thank Dr. Fosca Giannotti for her support during his visit in Purdue University.

10. REFERENCES

[1] A. D. Association. Direct and indirect costs of diabetes in the United States, 2006. `http://www.diabetes.org/diabetes-statistics/cost-of-diabetes-in-us.jsp`

[2] C. C. Agrawal. On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on Very large data bases*, pp. 901–909, Trondheim, Norway, 2005

[3] G. Agrawal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas. 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, 2006.

[4] M. Atzori. Weak *k*-anonymity: A low-distortion model for protecting privacy. In *Proceedings of the 8th International Information Security Conference (ISC06)*, pages 60–71, 2006.

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

[6] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[7] Standard for privacy of individually identifiable health information. *Federal Register*, 67(157):53181–53273, Aug. 14 2002.

[8] A. Ohrn and L. Ohno-Machado. Using boolean reasoning to anonymize databases. *Artificial Intelligence in Medicine*, 15(3):235–254, Mar. 1999.

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

[10] K. LeFevre, D. DeWitt, and R. Ramakrishnan. Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, Baltimore, MD, June 13-16 2005.

[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, pages 25–35, Atlanta, GA, Apr. 3-7 2006.

[12] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. *l*-diversity: Privacy beyond *k*-anonymity. In *Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006)*, Atlanta Georgia, Apr. 2006.

[13] 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. Technical Report NIH Publication No. 06–3892, U.S. Department of Health and Human Services, National Institute of Health, Bethesda, MD, Nov. 2005.

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

[15] P. Samarati. Protecting respondent's privacy in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, Nov./Dec. 2001.

[16] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, (5):557–570, 2002.

[17] 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.

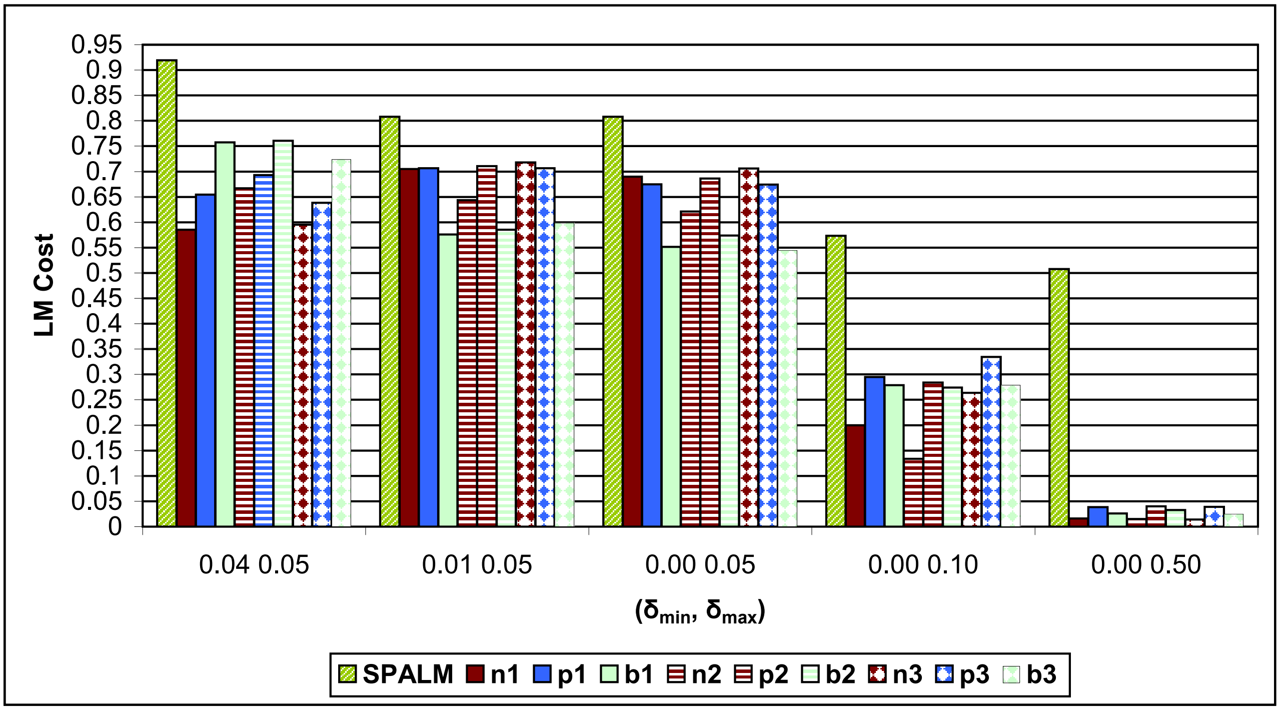


Figure 5: Loss Metric cost for a variety of strategies on the simulated diabetes patient dataset

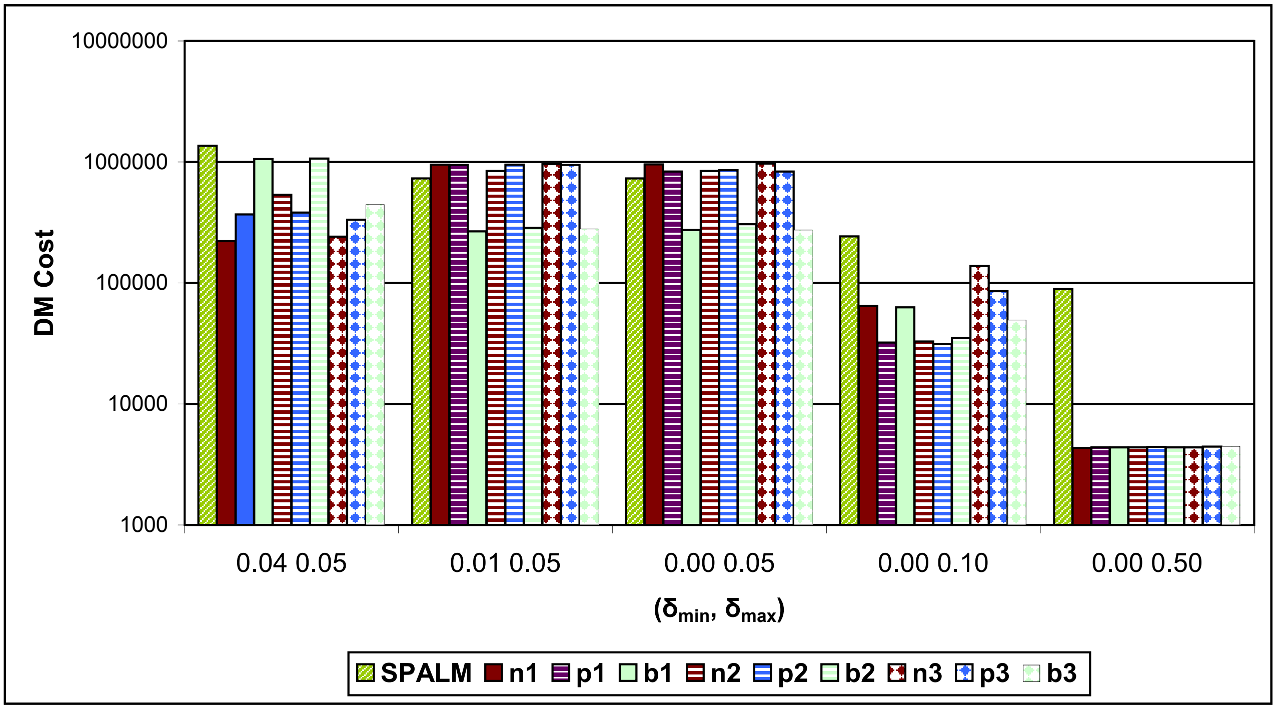


Figure 6: Discernibility Metric cost for a variety of strategies on the simulated diabetes patient dataset

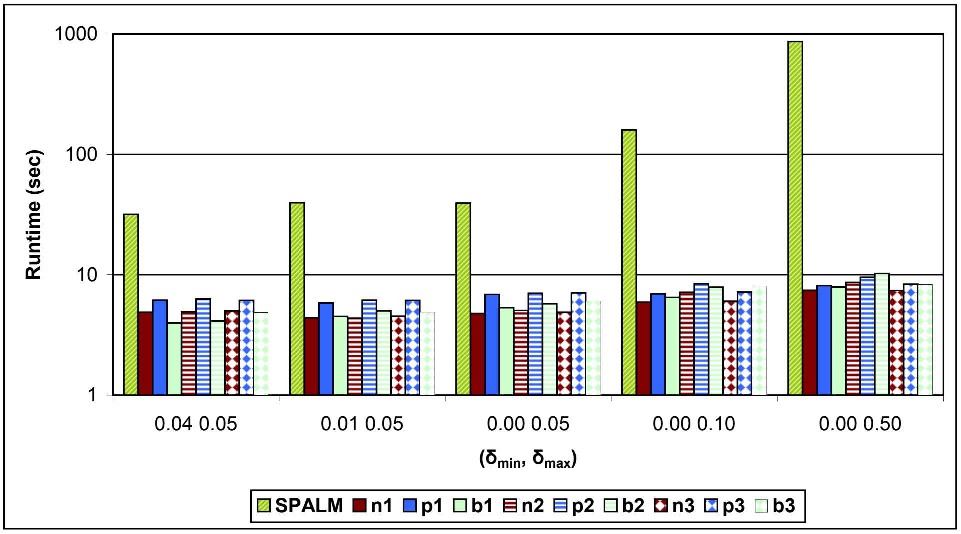
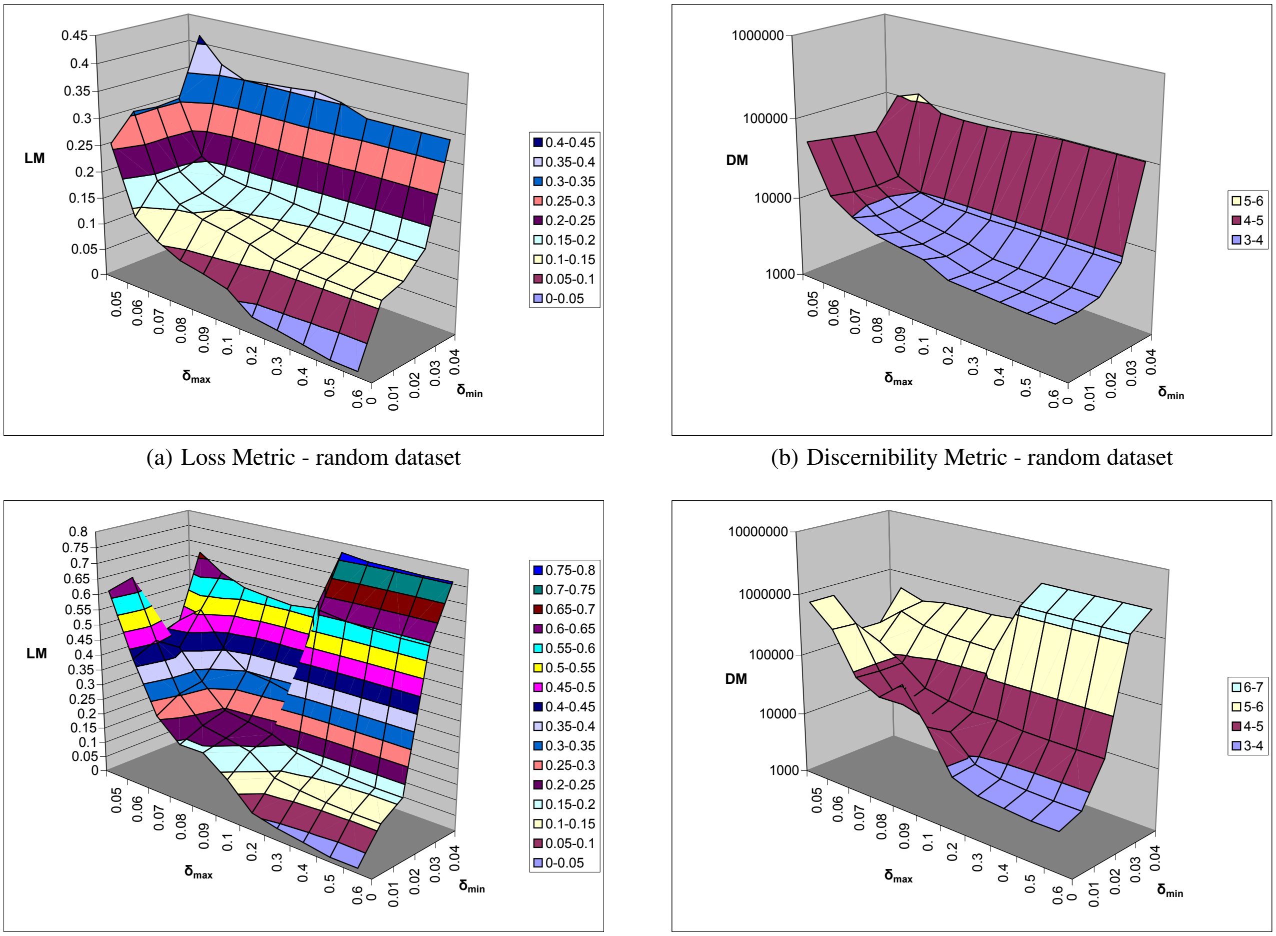


Figure 7: CPU time for a variety of strategies on the simulated diabetes patient dataset



(a) Loss Metric - random dataset

(b) Discernibility Metric - random dataset

(c) Loss Metric - diabetes dataset

(d) Discernibility Metric - diabetes dataset

Figure 8: Anonymization cost vs. various values of δ_{min} and δ_{max} for the MPALM n2 strategy