情報処理学会論文誌 Vol.56 No.5 1377–1387 (May 2015)

センシティブ属性値のランダムな追加による *l*-多様性アルゴリズムの提案

清 雄一[1,a)] 大須賀 昭彦[1]

受付日 2014年8月18日, 採録日 2015年2月4日

概要：保有している個人に関する情報データベースを他事業者と共有する場合，プライバシへの配慮が必要である．*l*-多様性等の一般的な匿名化技術では，データベースから個人を特定できる識別子を除外し，擬似識別子（QID）を一般化することで，攻撃者が各個人の属性値を推定することを防ぐ．通常，匿名化は一度のみ行われ，匿名化されたデータベースが複数のデータ利用者に共有され得る．したがって，あるデータ利用者が特に分析を行いたい QID が完全に一般化されてしまい，まったく分析ができなくなってしまう可能性がある．本研究では，QID を一般化せず，センシティブ属性にダミーの要素を追加することで，*l*-多様性を実現する．したがって各データ利用者は，好きな QID に基づいて自由に分析を行うことが可能となる．提案手法が，既存の *l*-多様性に関する手法と比べてプライバシと有効性について高いトレードオフを取れることを，実データを用いたシミュレーションによって示す．

キーワード：プライバシ，データマイニング，匿名化

An Algorithm for *l*-diversity based on Randomized Addition of Sensitive Values

Yuichi Sei[1,a)] Akihiko Ohsuga[1]

Received: August 18, 2014, Accepted: February 4, 2015

Abstract: Individual privacy needs to be studied when a data holder attempts to share databases containing personal attributes. Existing anonymization techniques remove identifiers and generalize quasi-identifiers (QIDs) from the database. By doing so, adversaries cannot specify each individual's values of the sensitive attributes. Because the database is anonymized based on one-size-fits-all measures in usual, it is possible that QIDs that a data user focuses on are all generalized, and the anonymized database has no value for the user. In this study, we propose a new technique for *l*-diversity, which keeps QIDs unchanged so that data users can analyze it based on QIDs they focus on. Through simulations of real data sets, we prove that our proposed method can result in a better tradeoff between privacy and utility of the anonymized database.

Keywords: privacy, data mining, anonymization

1. はじめに

近年，多くの組織が個人に関する情報を収集している．新しいサービス創出のために異なる事業者間で情報を共有する試みも行われている．しかし，秘匿性の高い情報を，個人を特定できるような状態のまま共有することは許されない．

このため，匿名化に関する研究が数多く提案されている [8], [9], [15]．ほとんどの研究では，**データ保有者**が以下の属性からなるデータベースを保有していることを前提としている．これらの属性は，識別子，擬似識別子（QID: Quasi-Identifiifers），センシティブ属性である．識別子は，名前や電話番号等，個人を特定できる属性である．QID は，公開情報等と組み合わせることによって個人を特定できる属性と定義され，居住地の郵便番号，年齢，職業等が当て

[1] 電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
[a)] sei@is.uec.ac.jp

表 1 患者データベース
Table 1 Patient table.

Name	Gender	Age	Address	Job	Disease
Alex	M	41	13000	Artist	Fever
Becky	F	41	17025	Artist	Sty
Carl	M	50	13021	Writer	Cancer
Diana	F	51	14053	Nurse	Pus
Edward	M	68	15000	Writer	Chill
Flora	F	69	16022	Nurse	HIV
Greg	M	72	13001	Artist	Cut
Hanna	F	77	17001	Artist	Cancer

表 2 2-多様性を満たす匿名化例
Table 2 2-anon. by existing method.

Gender	Age	Address	Job	Disease
*	41	13*-17*	Artist	Fever
*	41	13*-17*	Artist	Sty
*	50-51	13*-14*	*	Cancer
*	50-51	13*-14*	*	Pus
*	68-69	15*-16*	*	Chill
*	68-69	15*-16*	*	HIV
*	72-77	13*-17*	Artist	Cut
*	72-77	13*-17*	Artist	Cancer

はまる．センシティブ属性は，病歴や宗教等，保護対象となる属性である．

l-多様性 [11] は，k-匿名性 [8] を拡張した，プライバシ保護に関する指標である．l-多様性を実現するアルゴリズムは多数あるが，その多くは，データベースから識別子を取り除き，攻撃者が $1/l$ より高い確信度を持って個人のセンシティブ属性を特定できることのないよう，QID を一般化する[*1]．

以下に例を示す．**表 1** は，ある病院の患者に関するデータベースであり，このデータ保有者である病院はデータ分析を行う事業者（**データ利用者**）にデータベースを提供し，データマイニングを実施したいと考えている．名前が識別子，性別・年齢・住所・職業が QID，病名がセンシティブ属性であるとする．ここで，Becky がこのデータベースに含まれており，かつ Becky の QID 値をすべて知っているデータ利用者の存在を想定する．このとき，仮にこのデータベースから識別子（名前）を除外したとしても，データ利用者は Becky のセンシティブ属性が Sty であると 100%の確率で特定することができる．

表 2 は，l を 2 に設定した際における，l-多様性を実施した結果例を表している．たとえデータ利用者が，Becky が表 2 に含まれていることと，Becky の QID 値をすべて知っていたとしても，データ利用者は Becky の病気が Fever か Sty のどちらであるか特定することができない．つまり，データ利用者は Becky の病気を 50%（$= 1/2$）より高い確率で特定することができないということであり，2-多様性が満たされている．**表 3** は，l-多様性を満たす匿名化の別の実施例である．もし表 2 と表 3 の両方が公開されてしまうと，データ利用者は Becky の病気が Sty であると 100%の確率で特定できてしまう．したがって通常は，データ保有者は表 1 を一度のみ匿名化し，匿名化後のデータベースを各データ利用者に対して共通に提供する．k-匿名性や l-多様性における既存研究のほとんどがこのように，一度のみ匿名化を行うことを想定している [3], [17].

[*1] l-多様性の定義にはいくつかのバリエーションがあるが，ここではシンプルな定義を用いている．また，l-多様性の従来研究と同様に，センシティブ属性値の分布に対する攻撃者の前提知識は考慮せず，ランダムな事前分布を持つ攻撃者を想定している．

表 3 2-多様性を満たす匿名化例（2）
Table 3 2-anon. by existing method (2).

Gender	Age	Address	Job	Disease
M	41-72	13*	Artist	Fever
F	41-77	17*	Artist	Sty
M	50-68	13*-15*	Writer	Cancer
F	51-69	14*-16*	Nurse	Pus
M	50-68	13*-15*	Writer	Chill
F	51-69	14*-16*	Nurse	HIV
M	41-72	13*	Artist	Cut
F	41-77	17*	Artist	Cancer

表 4 提案手法による 2-多様性の実現例
Table 4 2-anon. by proposed method.

Gender	Age	Address	Job	Disease
M	41	13000	Artist	{Fever, Flu}
F	41	17025	Artist	{Fever, Sty}
M	50	13021	Writer	{Cold, Cancer}
F	51	14053	Nurse	{HIV, Pus}
M	68	15000	Writer	{Chill, Cut}
F	69	16022	Nurse	{Cold, HIV}
M	72	13001	Artist	{Cut, Fever}
F	77	17001	Artist	{Cancer, Flu}

あるデータ利用者 A は男女の性別による病気のかかりかたの違いを分析することを目的とし，データ利用者 B は年齢による違いを分析することを目的としていると想定する．データ利用者 B は表 2 から年齢別の分析を行うことが可能だが，データ利用者 A については，性別の値が完全に一般化されてしまっているため，目的の分析を行うことができない．

一方，データ利用者 A と B が表 3 を受け取った場合を想定する．この場合，データ利用者 A は表 3 を使って性別の差に基づく分析を行うことができるが，データ利用者 B は，年齢が大きく一般化されているために有効な分析ができない．

提案手法では QID を一般化せずにそのままの値を維持し，l-多様性を満たすためにセンシティブ属性に対して $l-1$ 個の値をランダムに追加する．このような一般化がなされた結果の例を**表 4** に示す．すべての QID がそのままの値

を維持しているため，各データ利用者は任意の QID に基づいて分析を行うことができる．本研究では，データ保有者側におけるこのような匿名化アルゴリズムだけでなく，データ利用者側においてどのような属性の人が何人いるのかを見積もるアルゴリズムもあわせて提案する．

なお，l-多様性は，k-匿名性の改良版と位置付けられ，暗黙的に k-匿名性を満たすことが多い．しかし本研究では，k-匿名性を満たすことを要求せず，l-多様性を満たすことのみを目標とする．4 章でも述べる Anatomy という手法 [16] も k-匿名性を満たさない匿名化を行っている．

本論文の構成を示す．2 章でアプリケーションモデルと攻撃モデルを記述する．3 章においてプライバシ指標を定義する．4 章で関連研究を紹介する．提案アルゴリズムを 5 章で説明し，評価結果を 6 章において示す．最後に 7 章で本論文をまとめる．

2. 想定環境

2.1 シナリオ

データ保有者は，個人に関する情報を含むデータベースを保有している．このデータベースには 1 章で述べたように，識別子，QID，センシティブ属性の各属性が含まれている．データベースはこれらの属性以外に，「QID とならない非センシティブ属性」が含まれている場合もある．この属性に関しては，一般化することなくそのまま公開することができる．

データ保有者は，このデータベースをデータ利用者に提供したいと考えている．プライバシの問題があるため，データ利用者に悪意のある人物がいたとしても，個人のセンシティブ属性を特定されることのないよう，データベースを匿名化する必要がある．

本研究では，データ利用者はどのような属性を持つ人がどのようなセンシティブ属性を持つかを分析したいと考えていると想定する．たとえば，匿名化されたデータベースが，QID（または「QID とならない非センシティブ属性」）として性別と年齢を含み，センシティブ属性として病名を含んでいる場合，年齢が高い男性は，年齢が低い女性よりも肺がん罹患率が高いことが分かる可能性がある．

なお，k-匿名性や l-多様性をプライバシ指標とする場合は，QID を保護する必要がないということに注意いただきたい．これは，k-匿名性や l-多様性等に関する研究における共通の想定である．これらの指標に基づく場合，QID の値はデータ利用者にとって既知である可能性があるという前提の下で，そのような状況においても各個人をデータベースのレコードに紐付けられないようにすることや，各個人のセンシティブ属性を特定されないようにすることを目標としている．結果として QID が一般化されることがあるが，一般化されれないこともあり，これらの指標に基づく場合は QID の保護についてはなんら保証しない．具体的には，もし QID 値がまったく同じ l 個のレコードがあり，かつそのセンシティブ属性がすべて異なっている場合は，l-多様性の観点からは，この l 個のレコードについてはまったく一般化されない．たとえば，データ保有者が**表 5** を保有していて，$l = 2$ の場合を考えると，このデータベースは最初から 2-多様性を満たしているため，匿名化されることはない．

表 5 オリジナルデータベースが l-多様性を満たしている場合に QID 値が保護されないことを示す例（$l = 2$）

Table 5 l-diversity does not protect any values of QIDs when the original database satisfies l-diversity (here, $l = 2$).

Gender	Age	Address	Job	Disease
M	31	13000	Artist	Fever
M	31	13000	Artist	Sty
M	52	13021	Writer	Cancer
M	52	13021	Writer	HIV

2.2 攻撃モデル

データ保有者は信頼できるが，データ利用者は信頼できないものと想定する．つまり，データ利用者は，匿名化されたデータベースを受け取った後，ある個人のセンシティブ属性値を推測しようとするものとする．

さらに，複数のデータ利用者は共謀する可能性があるものとする．もし，データ提供者が匿名化されたデータベースを複数のデータ利用者に提供した場合，これらのデータ利用者は受け取ったデータベースを突き合わせて個人のセンシティブ属性値をより高い精度で推測しようとする可能性があるものとする．

3. プライバシ指標

本研究では l-多様性というプライバシ指標を利用する．この指標は多くの研究で用いられている [3], [11], [17]．なお，l-多様性にはいくつかのバリエーションがあるが，本論文では以下の定義を用いる．

定義 1（QID グループ）：QID 値がすべて同じレコードの集合を QID グループと定義する．

たとえば，表 2 における 1 番目と 2 番目のレコードは，QID 値がすべて同じ（*, 41, 13*-17*, Artist）であるから，同じ QID グループに属している．

Machanavajjhala ら [10] の l-多様性の定義に従い，本論文における l-多様性を以下のように定義する．

定義 2（l-多様性）：データベースの各 QID グループにおいて，センシティブ属性値の最大出現割合がその QID グループ内で $1/l$ を超えないとき，そのデータベースは l-多様性を満たしていると定義する．

たとえば，表 2 や表 3 における各 QID グループはそれぞれ，2 つのレコードから構成されており，各レコードは異なるセンシティブ属性を持っている．したがって，これ

らのデータベースは 2-多様性を満たしている．

次に，提案手法を適用した例である表 4 を確認する．各レコードがそれぞれ 1 つの QID グループを構成している．各 QID グループを見ると，2 つの異なるセンシティブ属性から構成されている．定義 2 より，表 4 も同様に 2-多様性を満たしていると考えることができる[*2]．

4. 関連研究

データベースを匿名化する際のプライバシ指標の 1 つとして，k-匿名性が広く利用されている．データベースにおいて，すべての QID 値が同一であるレコードが少くとも k 個以上存在する場合，k-匿名性が満たされると定義される．

k-匿名性を実現するための手法として，Mondrian が広く利用されている [8]．Mondrian アルゴリズムは，まずすべてのレコードの QID 値を完全に一般化した状態からスタートして，k-匿名性が満たされなくなるまで QID 値を詳細化していくトップダウン型のアプローチである．Mondrian は本来 k-匿名化を実現するために提案されているものであるが，アルゴリズムをほとんど変えずに，l-多様性にも適用することができる．Mondrian では QID 値を詳細化していくたびに「k-匿名性を満たすかどうか」をチェックするが，この部分を「l-多様性を満たすかどうか」のチェックに変えるだけでよい．

k-匿名性は，データ利用者がある特定のユーザの QID 値をすべて知っている場合においても，匿名化後のデータベースからどのレコードが当該ユーザのレコードであるかを特定されることを防ぐことができる．しかし，当該ユーザのセンシティブ属性値が特定されることを防ぐことができない場合があることが指摘されている．

l-多様性 [11] は k-匿名性を拡張した指標に位置付けられ，前述のとおり，各個人のセンシティブ属性値を，$1/l$ より大きい確率で特定されることがないことを保証する指標である．

Xiao ら [17] は，l-多様性を行うことによる情報損失を分析し，l-多様性における他の手法と比べて，プライバシと，匿名化後のデータの有効性におけるトレードオフをより高いレベルで取ることのできる TP と呼ばれる手法を提案している．TP は，l-多様性を満たすために，1 つずつレコードを完全に一般化していく手法である．ここで，完全に一般化するとは，QID 値を “*” に置き換えることを意味する．アルゴリズムの概要を以下に示す．各 QID グループに対し，センシティブ属性値の最大出現割合がその QID グループ内で $1/l$ 以下になるまで，QID グループ内で最も出現数の多いセンシティブ属性値を持つレコードを完全に一般化する．さらに，完全に一般化されたレコード集合に対しても，センシティブ属性値の最大出現割合がその集合内で $1/l$ 以下になることが要求されるため，追加的なアルゴリズムが提案されている．QID の数が少なく，各 QID の取りうる値の値域が小さい場合は，匿名化後のデータベースからも有用な情報を抽出できると主張されている．しかし，表 1 を対象として TP を実行すると，各 QID グループにはそれぞれ 1 レコードしかないため，すべてのレコードが完全に一般化されてしまう．この場合，データ利用者は有用な分析を行うことができなくなってしまう．

このように k-匿名性や l-多様性については多くの研究がなされているが，そのほとんどが，QID を一般化して，センシティブ属性値をそのまま公開するというものである．QID を一般化することによる弊害として，1 章に述べたとおり，データ利用者が分析したい属性値の情報が大きく失われてしまう可能性があることがあげられる．

もし，データ利用者間で情報を共有することがないと信じることができるなら，各データ利用者の要望に応じた匿名化を行うことができる [9]．しかし，2 章で述べたように，データ利用者間で共謀する可能性を否定できないのであれば，別の匿名化技術が必要となる．

Xiao ら [16] や Sun ら [14] は，本研究と同じように QID の値を一般化しない，Anatomy と呼ばれる匿名化手法を提案している．Anatomy について理解するために，QID として性別および年齢，センシティブ属性として病名の 3 項目からなる，簡単なデータベース（**表 6**）を例にとり，$l = 2$ の場合の説明を記述する．

Anatomy は，**表 7** のように，オリジナルのテーブルを QID テーブルとセンシティブテーブルの 2 テーブルに分解することで匿名化を行う．QID テーブルは，QID 値をそのまま保持する．また，センシティブテーブルは，センシティブ属性値をそのまま保持するが，オリジナルのテーブルとは出現順序が異なる．さらに QID テーブルとセンシティブテーブルは共通に，「Group ID」のフィールドを持つ．この Group ID により，センシティブテーブルの各レコードを，QID テーブルのレコードと最大 $1/l$ の確率で紐付けることが可能となる．

表 6 オリジナルのデータベースの例
Table 6 An example of original database.

Gender	Age	Disease
M	25	Cancer
M	29	Cancer
M	54	Cancer
M	78	Cold
F	20	Cold
F	55	Cut
F	56	Cut
F	59	Cut

[*2] この例では各 QID グループは 1 つのレコードからのみ構成されているが，複数のレコードから構成される場合もある．提案手法を実行した結果，l-多様性を満たす匿名化データベースが生成されることは，5.2 節において証明する．

表 7 Anatomy の匿名化結果
Table 7 Anonymization result of Anatomy.

(a) QID table

Gender	Age	Group ID
M	25	2
M	29	4
M	54	1
M	78	3
F	20	4
F	55	3
F	56	1
F	59	2

(b) Sensitive table

Group ID	Disease
1	Cancer
1	Cut
2	Cancer
2	Cut
3	Cold
3	Cut
4	Cancer
4	Cold

プライバシが保護されていることが分かる例を以下に示す．データ利用者は Alex の QID 値を知っており，Alex は男性で 25 歳であるとする．表 6 より，Adam の病名は Cancer であるが，データ利用者はこの事実は知らない．表 7 (a) を見たデータ利用者は，1 番目のレコードが Adam のものであると分かる．また，Adam の Group ID が「2」であることも分かる．次に，表 7 (b) を見ると，Group ID が「2」であるレコードは 2 つあり，それぞれの病名は「Cancer」と「Cut」である．つまり，データ利用者は，Adam の病名が「Cancer」か「Cut」のどちらかであることしか分からないため，2-多様性が満たされている．

このように，Anatomy を用いて匿名化を行った場合，QID 値をそのまま維持できるため，本研究の目標と同様，データ利用者は，好きな QID に基づいて自由に分析を行うことが可能となる．

しかし，元のデータに偏りがある場合，データ利用者は高い精度での分析ができない場合がある．たとえば，表 7 において，男性と女性の差を分析したいと考える．4 人ずつ男性と女性がいることは分かるが，男性も女性もそれぞれ Group ID が 1 から 4 まで設定されている．つまり，表 7 からは男女差の分析を行うことはできない．また女性だけを対象に分析した場合，女性で Cancer ある人は本来 0 人であるにもかかわらず，3/8 の確率で Cancer である，という分析結果が得られてしまう．何故なら，QID テーブルを見ると女性 4 人はそれぞれ Group ID が 1 から 4 まで設定されており，各 Group ID に対応するセンシティブテーブルを見ると，全 8 レコード中，3 レコードに Cancer が含まれているからである．

Anatomy では上記のように，出現回数の多いセンシティブ属性値の影響を受けやすい．一方，提案手法を利用すると，各レコードで独立して匿名化を行うため，QID 値が異なるものどうしで，差がまったく出ないという事態が発生する可能性は低く，出現回数の多いセンシティブ属性値の影響を受けるということもない．提案手法は確率的な手法を用いるため，このような悪い状況が起きる確率が 0%であるというわけではないが，多くの場合はデータ利用者がより高精度に分析が可能であることは 6 章の評価によって示す．

本論文では，6 章において，提案手法，Mondrian，Anatomy，TP について比較評価を行っている．

Differential privacy（差分プライバシ）という指標が近年特にさかんに研究されている [5]．この指標は最も強力なプライバシ定義の 1 つだといわれることもあり [13]，直観的には，ある 1 人の個人がいてもいなくても匿名化の出力にほとんど差がない，ということを要求する．差分プライバシは，データベース保有者に対し，データ利用者がクエリを発行する．その回答を匿名化するというシナリオの下で利用される．つまり，データ保有者は匿名化したデータベースを他者に提供するのではなく，データベースを保有したまま，データ利用者からのクエリを逐一受け付ける必要があるということである．したがって，データ保有者のコストが高くなる，データ利用者は自由に分析を行うことができない，といったデメリットがある [4]．本研究ではこれらのデメリットを考慮し，匿名化後のデータベースを他者に提供できることが求められている環境を想定する．

5. 提案手法

提案手法は，データ保有者による匿名化と，データ利用者による分析対象レコード数の推測の 2 つから構成される．

データ保有者は，匿名化対象となるオリジナルのデータベースにおいて，各レコードのセンシティブ属性に，$l-1$ 個の属性値をランダムに追加することによって匿名化を行う．

データ利用者は，まず分析したい QID を選択する．次に，選択された QID とセンシティブ属性の各組合せにおけるレコード数を推測する．たとえば，センシティブ属性値として HIV，Fever，Cancer があるとする．また，データ利用者は，性別によってこれらの病気の罹患率が異なるかどうかを分析したいと想定する．このときデータ利用者は，（男性，HIV），（男性，Fever），（男性，Cancer），（女性，HIV），（女性，Fever），（女性，Cancer）の各組合せにおけるレコード数を推測することになる．

以下では，まず記号の定義をした後，データ保有者による匿名化アルゴリズムおよび，データ利用者による分析対象レコード数の推測アルゴリズムの詳細を説明する．

5.1 記号の定義

オリジナルのデータベースと匿名化後のデータベースをそれぞれ T と T^* で表す．T および T^* のレコード数を N とする．また，T および T^* の i 番目のレコードをそれぞれ r_i，r_i^* と表す．

センシティブ属性の取りうる値の順序付き集合を S とし，S における i 番目の要素を s_i（$i = 0, \ldots, |S|-1$）と表す．

データ利用者が分析対象とするQIDの順序付き集合を$\mathfrak{Q}$とし，$\mathfrak{Q}$におけるi番目の要素を$\mathfrak{Q}(i)$（$i=0,\ldots,|\mathfrak{Q}|-1$）と表す．

$\mathfrak{Q}(i)$に相当するQIDの取りうる値の順序付き集合を$Q(i)$とし，$Q(i)$におけるj番目の要素を$q(i)_j$と表す．

たとえば，表1においてSは{Fever, Sty, ...}である．もしデータ利用者が性別による違いを分析したい場合は，$\mathfrak{Q}$は{Gender}，$\mathfrak{Q}(0)$はGender，$Q(0)$は{M, F}，$q(0)_0$はMであり，$q(0)_1$はFとなる．

他の例として，データ利用者が性別（M, F）と年代（[0-9], [10-19],..., [90-99]）に基づいて分析を行いたい場合は，$\mathfrak{Q}$は{Gender, Age}，$Q(0)$は{M, F}，$Q(1)$は{[0-9],..., [90-99]}となる．

レコードrのセンシティブ属性値を$E(r)$と表す．たとえば，表1において$E(r_0)$はFeverを表す．

また本論文では，提案手法を用いた匿名化後のレコードにおいて，センシティブ属性に設定された値を「**センシティブ属性集合値**」と呼ぶ．一方，単に「センシティブ属性値」と呼ぶ場合は，単一のセンシティブ属性値を指す．たとえば表4において，センシティブ属性値はFever，Flu，Styといった値である．また，表4の最初のレコードにおいて，{Fever, Flu}は，センシティブ属性値を2つ含む，センシティブ属性集合値である．

匿名化されたレコードr^*のセンシティブ属性集合値は$E(r^*)$と表す．たとえば表4における$E(r_0^*)$は{Fever, Flu}である．

また，$Q(0),\ldots,Q(|\mathfrak{Q}|-1)$におけるすべての組合せを$C$と表す．つまり，$C$は以下の式で表される集合である：

$$C = Q(0) \times Q(1) \times \ldots \times Q(|\mathfrak{Q}|-1). \quad (1)$$

さらに，Cにおけるi番目の要素をc_i $(i=0,\ldots,|C|-1)$と表す．たとえば，データ利用者が性別と年代に基づいてセンシティブ属性値を分析したい場合は，c_0は(M, [0-9])，c_1は(M, [10-19]),..., $c_{|C|-1}$は(F, [90-99])となる．

5.2 匿名化アルゴリズム

データ保有者は，オリジナルのデータベースにおいて，各レコードのセンシティブ属性に，$l-1$個の値をランダムに追加する．この匿名化アルゴリズムをAlgorithm 1に示す．

Algorithm 1 Anonymization protocol for record r

Input: Domain of a sensitive attribute S, Privacy level l
Output: Set of anonymized sensitive values for record r

1: Create Set R
/*Adds original sensitive attribute*/
2: $R \Leftarrow \{E(r)\}$
/*Adds dummy sensitive attributes*/
3: $R \Leftarrow R \cup rand(l-1, S \setminus \{E(r)\})$
4: **return** R

関数$rand(b, B)$は，集合Bからランダムに異なるb個の要素を抽出する関数である．データ保有者はAlgorithm 1を，オリジナルのデータベースの各レコードに対して実行する．

ここで，以下の定理が成り立つ．

THEOREM 5.1. $|S| \geq l$である場合，データベースTに対して*Algorithm 1*を実行すると，必ずl-多様性を満たすデータベースT^*が生成される．

Proof. Algorithm 1をデータベースTに対して実行して生成される匿名化データベースをT^*とおく．T^*における各レコードrのセンシティブ属性集合値は，それぞれl個のセンシティブ属性値から構成される．データベールT^*においてある1つのQIDグループに注目する．このQIDグループを構成するレコード数をδとおく（δは1からNまでのいずれかの値を取る）．このδ個のレコードにおいて，各センシティブ属性値の出現回数の取りうる値の最大値はδである．なぜなら，1つのレコード内で同じセンシティブ属性値が2回以上出現することはないからである．

一方，そのQIDグループにおけるδ個のレコードにおいて，センシティブ属性値ののべ出現回数は$\delta \times l$回である．なぜなら，Algorithm 1を実行すると，各レコードにはl個のセンシティブ属性値が設定されるからである．

上記より，そのQIDグループにおける，センシティブ属性値の最大出現割合はそのQIDグループ内で$1/l$である．

このことはすべてのQIDグループに対して成り立つため，定義2より，定理5.1が成り立つ． □

5.3 推測アルゴリズム

匿名化後のデータベースを受け取ったデータ利用者は，各c_i（$i=0,\ldots,|C|-1$）において，各センシティブ属性値の出現回数を推測する．ここで，センシティブ属性集合値の取り得る値域の大きさは${}_{|S|}C_l$となるが，各センシティブ属性値が取り得る値域の大きさは$|S|$であることに注意いただきたい．

QID値の組合せがc_iであり，センシティブ属性値としてs_jを持つ「実際の」レコード数を$X_{i,j}$とおく．同様に，「推測された」レコード数を$\widehat{X_{i,j}}$とおく．

匿名化後のデータベースにおいて，QID値の組合せがc_iであり，センシティブ属性値としてs_jを持つレコード数を以下のようにカウントする：

$$W_{i,j} \sum_{m=0}^{N-1} H(r_m^*, c_i, s_j), \text{where } H(r^*, c, s)$$
$$= \begin{cases} 1 & (r^*\text{'s QIDs are categorized to } c \text{ and } s \in E(r^*)) \\ 0 & (\text{otherwise}) \end{cases} \quad (2)$$

シンプルな方法としては，式(2)より，

$$\widehat{X_{i,j}} = \frac{W_{i,j}}{l} \tag{3}$$

と計算することで，各 c_i（$i = 0, \ldots, |C|-1$）における，各センシティブ属性値 s_j（$j = 0, \ldots, |S|-1$）の出現回数を計算することができる．

しかし，ベイズ法 [1], [2] を利用することで，より精度良く推測を行うことを考える．

まず，匿名化後のデータベースにおいて，QID 値の組合せが c_i であるレコード数を以下のようにカウントする：

$$\begin{aligned} N_i &= \sum_{m=0}^{N-1} G(r^*_m, c_i), \text{ where } G(r^*, c) \\ &= \begin{cases} 1 & (r^{*\prime}\text{s QIDs are categorized to } c) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned} \tag{4}$$

もしレコード r がセンシティブ属性値として s_α を持っている場合，s_β が $E(r^*)$ に含まれる確率は以下の式で表される：

$$p_{\alpha,\beta} = \begin{cases} 1 & (\alpha = \beta) \\ \frac{l-1}{|S|-1} & (\text{otherwise}) \end{cases} \tag{5}$$

QID 値が c_i である，ある匿名化前のレコード r のセンシティブ属性値が確率変数 U_i によって表されるとする．また，匿名化後のセンシティブ属性集合値が確率変数 V_i によって表されるとする．このとき，$Pr(U_i = \alpha)$ は以下の式で計算することができる：

$$\begin{aligned} Pr(U_i = \alpha) &= \frac{\sum_{\beta=0}^{|S|-1} Pr(V_i \ni \beta) Pr(U_i = \alpha | V \ni \beta)}{\sum_{\beta=1}^{|S|} Pr(V_i \ni \beta)} \\ &= \frac{\sum_{\beta=0}^{|S|-1} Pr(V_i \ni \beta) Pr(U_i = \alpha | V_i \ni \beta)}{l} \end{aligned} \tag{6}$$

またベイズの定理より，

$$\begin{aligned} &Pr(U_i = \alpha | V_i \ni \beta) \\ &= \frac{Pr(V_i \ni \beta | U_i = \alpha) Pr(U_i = \alpha)}{Pr(V_i \ni \beta)} \\ &= \frac{Pr(V_i \ni \beta | U_i = \alpha) Pr(U_i = \alpha)}{\sum_{\gamma=0}^{|S|-1} Pr(V_i \ni \beta | U_i = \gamma) Pr(U_i = \gamma)} \\ &= \frac{p_{\alpha,\beta} \widehat{X_{i,\alpha}}}{\sum_{\gamma=0}^{|S|-1} p_{\gamma,\beta} \widehat{X_{i,\gamma}}} \end{aligned} \tag{7}$$

となる．

式 (6)，(7) における $Pr(U_i = \alpha)$ の値は，未知の値である $X_{i,\alpha}$ を利用して，$X_{i,\alpha}/N_i$ と表すことができる．ここでは，$X_{i,\alpha}$ の推測値を $\widehat{X_{i,\alpha}}$ と置くことにより，

$$Pr(U_i = \alpha) = \widehat{X_{i,\alpha}}/N_i \tag{8}$$

と表すことで，$\widehat{X_{i,\alpha}}$ の値を算出することを試みる．

$Pr(V_i \ni \beta)$ は，あるセンシティブ属性集合値が，あるセンシティブ属性値 β を要素に含んでいる確率を表している．センシティブ属性集合値は全部で N_i 個あり，センシティブ属性値 β はそのうちの $W_{i,\beta}$ 個に出現していることから，

$$Pr(V_i \ni \beta) = W_{i,\beta}/N_i \tag{9}$$

とおくことができる．

以上，式 (6)，(7)，(8)，(9) より，

$$\widehat{X_{i,\alpha}}^{R+1} = \sum_{\beta=0}^{|S|-1} W_{i,\beta} \frac{p_{\alpha,\beta} \widehat{X_{i,\alpha}}^{R}}{\sum_{\gamma=0}^{|S|-1} p_{\gamma,\beta} \widehat{X_{i,\gamma}}^{R}} \Big/ l \tag{10}$$

が得られる．ここで，式 (10) は再帰的に実行されるものであり，$\widehat{X_{i,j}}^{R}$ $(j = 0, \ldots, |S|-1)$ は R 回目の反復結果を表している．初期値として $\widehat{X_{i,j}}^{0}$ を $W_{i,j}$ に設定する．

あらかじめ閾値 ϵ を定義しておき，すべての j について $\widehat{X_{i,j}}^{R}$ と $\widehat{X_{i,j}}^{R+1}$ の差が ϵ 以下になるまで，式 (10) の計算を繰り返し実行する．

5.4 解析

本節では，匿名化の前提条件について議論する．

もし $l \leq |S|$ が満たされている場合，各レコードのセンシティブ属性値に $l-1$ 個の値をランダムに追加することができるため，$l \leq |S|$ であればどのようなデータベースでも l-多様性を満たす匿名化を行うことができる．逆にいうと，$l > |S|$ である場合は，l-多様性を満たす匿名化を行うことはできない．

しかし，この条件は l-多様性における既存研究においても必須の条件であるため，本提案手法のデメリットとはならないと考える．たとえば，センシティブ属性値として，HIV，Cancer，Fever の 3 つしか存在しない場合に，4-多様性を満たす匿名化を行うことは提案手法においても既存手法においてもできない[*3]．

一方，l-多様性を実現する既存手法では，eligibility requirement [17] と呼ばれる前提条件を満たすことが必要である．この条件とは，「オリジナルのデータベース T において，最も出現回数が多いセンシティブ属性値の出現回数を m とすると，m が T/l を超えないこと」である．つまり，この条件が満たされていない場合は匿名化を行うことができない．一方，提案手法では，この前提条件を満たす必要はない．

6. 評価

6.1 評価指標

データ利用者は 2 章で述べたように，各 c_i $(i =$

[*3] ただし，架空のセンシティブ属性値を追加し，かつ，データ利用者がそれを架空のものであると判断できない場合は，匿名化を行うことは可能である．これは，既存研究でも提案手法でも同じである．

表 8 OCC と SAL の各属性の取り得る値の個数
Table 8 Number of distinct values of OCC and SAL.

Age	Gender	Marital Status	Race	Birth Place	Education	Work Class	Occupation	Income
80	2	6	9	124	24	9	50	50

表 9 Adult の各属性の取り得る値の個数
Table 9 Number of distinct values of Adult data set.

Age	Work Class	Final Weight	Education	Education-num	Marital Status	Occupation	Relationship
74	7	26741	16	16	7	14	6
Race	Gender	Capital Gain	Capital Loss	Hours Per Week	Country	Salary Class	
5	2	121	97	96	41	2	

$0, \ldots, |C|-1$) における，センシティブ属性の分布を推測する．QID 値の組合せが c_i であり，センシティブ属性値として s_j を持つ「実際の」レコード数を $X_{i,j}$ とおく．同様に，「推測された」レコード数を $\widehat{X_{i,j}}$ とおく．

このとき，真の値の分布 $X_{i,j}/N_i$ と推測された値の分布 $\widehat{X_{i,j}}/N_i$ の平均二乗誤差（MSE: Mean Squared Error）を指標として利用することができる．この MSE は以下の式で計算される：

$$\sigma^2(i) = \frac{1}{|S|}\sum_{j=0}^{|S|-1}\left(\frac{X_{i,j}}{N_i} - \frac{\widehat{X_{i,j}}}{N_i}\right)^2 \qquad (11)$$

この MSE の指標は多くの既存研究でも利用されている（文献 [6], [7], [18]）．

TP や Mondrian を使った手法では，一部のレコードの QID 値が一般化される．このとき，あるセンシティブ属性値を持ったユーザ数の推測については，文献 [16] に基づいて計算した．具体的には，一般化された各レコードに対し，ユーザが分析したい値を持つ確率を計算して，それを全レコードについて足し合わせる．たとえば，表 2 に対して，「年齢が 73 歳であり，病名が Cancer である」レコード数は以下のように計算される．最初の 6 レコードについて，年齢が 73 歳である確率は 0.0 である．また，7 番目のレコードも，病名が Cancer である確率は 0.0 である．8 番目のレコードについては，年齢が 72 歳から 77 歳までの可能性がある．この中で，73 歳である確率は $1/(77-72+1) = 1/6$ であると考えられる．また，8 番目のレコードは確率 1.0 で病名が Cancer である．この結果，8 番目のレコードについて，「年齢が 73 歳であり，病名が Cancer である」確率は，$(1/6) \times (1.0) = 1/6$ と計算される．8 番目以外のレコードについては，「年齢が 73 歳であり，病名が Cancer である」確率は 0.0 である．この結果，これらの値を足し合わせることで，表 2 において，「年齢が 73 歳であり，病名が Cancer である」レコード数は 1/6 であると計算される．この推測値と，真の値を使って，式 (11) に基づいてそれぞれ MSE を計算した．

Anatomy についても，文献 [16] に基づいて計算を行っている．考え方は上記と同じであり，各レコードがクエリの対象レコードである確率を求め，それを全レコードについて足し合わせた結果を推測値としている．

6.2 評価に用いるデータ

OCC，SAL，Adult の 3 つの実データを用いて評価を行う．OCC と SAL は文献 [12] より入手したものであり，文献 [17] や文献 [16] 等多く既存研究がこのデータセットを利用している．それぞれ 60 万件のレコードからなる．OCC は，センシティブ属性として *Occupation* を持ち，QID として *Age*，*Gender*，*Marital Status*，*Race*，*Birth Place*，*Education*，*Work Class* の各属性を持つ．一方，SAL はセンシティブ属性として *Income* を持ち，OCC と同じ QID を持つ．各属性における属性値の個数を**表 8** に示す．

OCC から 7 つのデータセット OCC-1, OCC-2, ..., OCC-7 を作成した．ここで，OCC-d は，表 8 における最初の d 個の属性を QID として持ち，*Occupation* をセンシティブ属性として持つデータセットである．たとえば，OCC-3 は *Age*，*Gender*，*Marital Status* を QID として持ち，*Occupation* をセンシティブ属性として持つ．同様に SAL からも，SAL-d ($d = 1, \ldots, 7$) を生成した．

Adult は，**表 9** に示すように 15 の属性を持つデータである．欠損を含むレコードを除外すると 45,222 レコードから構成される．この Adult も多くの既存研究で利用されているものである（文献 [11], [14]）．この Adult から，15 のデータセット Adult(1), Adult(2), ..., Adult(15) を生成した．各 Adult(d) は，表 9 の d 番目の属性をセンシティブ属性として持ち，それ以外の属性を QID として持つデータセットである．たとえば Adult(2) は，*Work Class* をセンシティブ属性として持ち，*Age, Final Weight, Education, ...* を QID として持つ．

比較対象の Anatomy を提案している文献 [16] に従って，データ利用者が，g 個のランダムな QID $A_1, \ldots, A_g$ に基づく分析をするものと想定する．たとえば，SAL-4 データセットを対象とし，$g = 3$ である場合，$\{A_1, A_2, A_3\}$ は，{Age, Gender, Marital Status, Race} からランダムに選択された 3 つの要素からなる集合である．つまり $g = 3$ の場合，データ利用者は分析対象としてこの集合から 3 つの

QID を選択する．さらにデータ利用者は，$A_1, \ldots, A_g$ の各属性が取り得る値のすべての組合せを対象とするのではなく，各属性 A_i が取り得る属性値のうち，b_i 個の属性値に対応するものに注目するものと考える．たとえば，$A_1 = \text{Age}$ であり，データ利用者が 50 歳の人におけるセンシティブ属性値の分布を分析したい場合，$b_1 = 1$ である．50 歳〜59 歳の合計値に注目する場合は $b_1 = 10$ である．この b_i についても Xiao ら [16] に従って，b_i を決定付けるためのパラメータ s を導入し，

$$b_i = \lceil |A_i| \cdot s^{1/(g+1)} \rceil \tag{12}$$

の計算式で b_i の値を決定する．ここで $|A_i|$ は属性 A_i において取りうる属性値の個数である．

6.3 評価結果

本章では，提案手法を 4 章で紹介した Mondrian，TP，Anatomy と比較する．各実験は，12 GB の RAM を搭載した，Intel Xeon CPU E5-2687W v2 @ 3.40 GHz のパーソナルコンピュータ上で実施した．

また，式 (10) を再帰的に計算する際に事前に設定する閾値 ϵ は 1.0 に設定した．

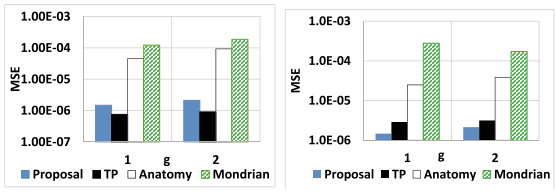
デフォルト値として，l を 10，分析対象の QID の数 g を 3，各 QID において分析対象とする属性値の個数を決定するパラメータ s を 0.07，に設定した．各シミュレーションでは，データ利用者が上述の g および s の値に基づいて 1,000 回ランダムに推測アルゴリズムを実施し，式 (11) に基づいて MSE を計算し，その平均値を計算した．

OCC と SAL を対象にした最初の実験では，QID の数と g の値が MSE に与える影響を分析するため，QID の数 d を 3 から 7 まで，g の数を 1 から $d-1$ まで変化させた．**図 1** に結果を示す．図 1 (a)，1 (a)，1 (e) はそれぞれ OCC の結果を，図 1 (b)，1 (d)，1 (f) はそれぞれ SAL の結果を表している．QID の数 d が 3 であるとき，TP と提案手法の MSE はほとんど同じ値である．しかし，他の設定においては，提案手法の MSE が最も小さいことが分かる．

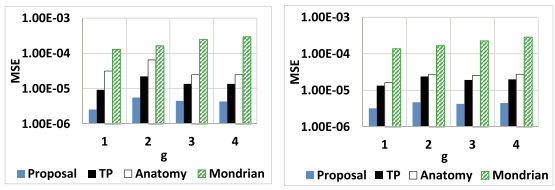
次の実験では，l の値を 5 から 15 まで変化させて MSE を計測した．結果を**図 2** に示す．l の値が大きいほど，MSE の値が増加していることが分かる．提案手法においては，各センシティブ属性にランダムに追加する値の数が増え，サーバ側での予測がより難しくなっていくためである．しかし，図 2 (a)，2 (b) から，提案手法がやはり最も小さい MSE を実現していることが読み取れる．

次に s の値を 0.04 から 0.1 まで変化させて MSE の値を計測した．結果を**図 3** に示す．s の値が大きいほど MSE の値が減少している．これは，s の値が大きいほど，正解となるレコード数も増加するため，相対的に誤差が小さくなっていくためであると考えられる．

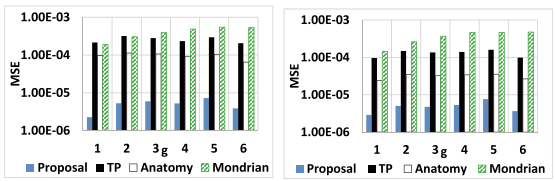
Adult(1), Adult(2), . . . , Adult(15) の各データセットを用いた実験では，l の値を 2 から 10 まで変動させて MSE を計測した．Mondrian，TP や Anatomy は，eligibility requirement（5.4 節参照）が満たされないために匿名化できない場合があった．提案手法は eligibility requirement を満たす必要はないが，$l > |S|$ の場合は匿名化を行うことが



(a) OCC-3 (b) SAL-3

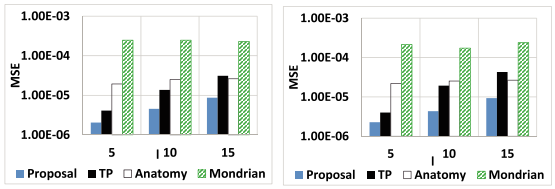


(c) OCC-5 (d) SAL-5



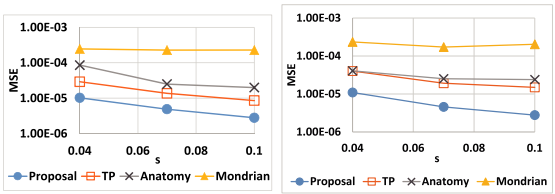
(e) OCC-7 (f) SAL-7

図 1 MSE と g の関係
Fig. 1 MSE vs. g.



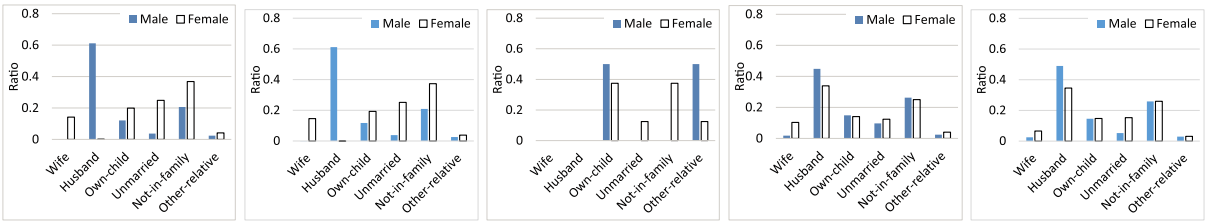
(a) OCC-5 (b) SAL-5

図 2 MSE と l の関係
Fig. 2 MSE vs. l.



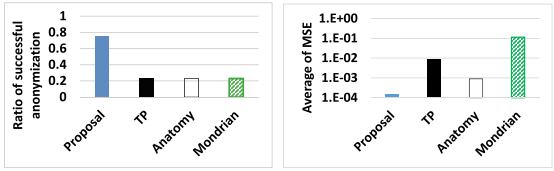
(a) OCC-5 (b) SAL-5

図 3 MSE と s の関係
Fig. 3 MSE vs. s.



(a) Original database (b) Anonymization by Proposal (c) Anonymization by TP (d) Anonymization by Anatomy (e) Anonymization by Mondrian

図 5 $l = 2$ の場合の Relationship 属性と Gender 属性における分布
Fig. 5 Relationship vs. Gender in $l = 2$.



(a) Ratio of successful anonymization (b) Average of MSE

図 4 Adult データセットを用いた結果
Fig. 4 Result overview of Adult dataset.

できない．各データセットと各 l のパラメータの組合せにおいて，匿名化が可能であった組合せの割合を**図 4** (a) に示す．提案手法の前提条件のほうが eligibility requirement よりも緩い条件であるため，提案手法における匿名化可能な組合せの割合のほうが値が高い．

図 4 (b) は，Adult(1), . . . , Adult(15) の各データセットに対して行った実験で計測された MSE の平均値を表している．提案手法の MSE が最も値が小さいことが分かる．

次の実験では，提案手法の効果を視覚的に見るために，“relationship” をセンシティブ属性とする Adult(8) のデータセットを利用し，分析対象の QID として Gender を設定して推測を行った[*4]．**図 5** (a) は，Male と Female それぞれにおけるオリジナルの分布を表している．図 5 (b)，5 (c)，5 (d)，5 (e) はそれぞれ，提案手法，TP，Anatomy，Mondrian を用いて匿名化および推測を行った結果を表している．提案手法が最も精度良くオリジナルの分布を推測できていることが図から分かる．

最後に，推測にかかる時間を Adult データセットを用いて計測した．提案手法では，平均 3.7 秒程度で推測を行うことができた．また，人工的にデータを 100 万件生成して実験を行った結果，約 58.7 秒程度で推測を行うことができた．

[*4] 結果のグラフが視覚的に見やすくなるように，ここでは属性が取り得る値の個数が少ないものを分析対象の QID およびセンシティブ属性として選定した．恣意的な選定でなく，網羅的に実験を行った結果は，図 1 から図 4 に示したとおりである．

7. おわりに

個人に関する情報を保有するデータ保有者が，データ利用者とデータを共有するための匿名化手法として，l-多様性に基づく研究が広く行われている．l-多様性に関する既存研究のほとんどは，データベース内の QID を一般化することで，匿名化後のデータベースを見たデータ利用者が，特定のユーザのセンシティブ属性を高精度に推測することができないようにしている．しかし，データ利用者が重視したい QID 値が一般化されてしまうと，匿名化後のデータベースの有用性が大きく損なわれてしまうという問題がある．

本論文では，QID をそのまま維持し，センシティブ属性にランダムな $l - 1$ 個の値を追加することで l-多様性を実現する手法を提案した．また，データ利用者側で分析したい QID に基づいて，該当するレコード数を推測するアルゴリズムをあわせて提案した．既存研究と比べ，高い有効性を実現できることを実データを用いたシミュレーションによって示した．

将来課題として，位置情報，特にユーザの位置追跡情報を対象に匿名化を行うことを考えている．

謝辞 本研究は JSPS 科研費 24300005，26330081，26870201 の助成を受けたものです．

参考文献

[1] Agrawal, R. and Srikant, R.: Privacy-Preserving Data Mining, *Proc. ACM SIGMOD*, pp.439–450 (2000).
[2] Agrawal, R., Srikant, R. and Thomas, D.: Privacy preserving OLAP, *Proc. ACM SIGMOD*, pp.251–262 (2005).
[3] Cheong, C.H.: Non-Centralized Distinct L-Diversity, *International Journal of Database Management Systems*, Vol.4, No.2, pp.1–21 (online), DOI: 10.5121/ijdms.2012.4201 (2012).
[4] Clifton, C. and Anandan, B.: Challenges and Opportunities for Security with Differential Privacy, *Information Systems Security*, Springer, pp.1–13 (2013).
[5] Dwork, C.: Differential Privacy, *Automata, Languages and Programming*, Lecture Notes in Computer Science, Vol.4052, Springer, pp.1–12 (2006).

[6] Groat, M.M., Edwards, B., Horey, J., He, W. and Forrest, S.: Enhancing privacy in participatory sensing applications with multidimensional data, *Proc. IEEE PerCom*, pp.144–152 (2012).
[7] Huang, Z. and Du, W.: OptRR: Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining, *Proc. IEEE ICDE*, pp.705–714 (2008).
[8] LeFevre, K., DeWitt, D. and Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity, *Proc. IEEE ICDE*, pp.25–25 (2006).
[9] LeFevre, K., DeWitt, D.J. and Ramakrishnan, R.: Workload-aware anonymization techniques for large-scale datasets, *ACM Trans. Database Systems*, Vol.33, No.3, pp.1–47 (2008).
[10] Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M.: l-diversity: Privacy Beyond k-Anonymity, *Proc. IEEE ICDE*, pp.24:1–24:12 (2006).
[11] Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity, *ACM TKDD*, Vol.1, No.1, pp.3–es (2007).
[12] Minnesota Population Center: IPUMS, available from ⟨https://www.ipums.org/⟩.
[13] Nikolov, A., Talwar, K. and Zhang, L.: The geometry of differential privacy: The sparse and approximate cases, *Proc. ACM STOC*, pp.351–360 (2013).
[14] Sun, X., Wang, H., Li, J. and Ross, D.: Achieving P-Sensitive K-Anonymity via Anatomy, *Proc. IEEE International Conference on e-Business Engineering* (*ICEBE*), pp.199–205 (2009).
[15] Wu, S., Wang, X., Wang, S., Zhang, Z. and Tung, A.K.: K-Anonymity for Crowdsourcing Database, *IEEE Trans. Knowledge and Data Engineering*, Vol.26, No.9, pp.2207–2221 (2014).
[16] Xiao, X. and Tao, Y.: Anatomy: Simple and effective privacy preservation, *Proc. VLDB*, pp.139–150 (2006).
[17] Xiao, X., Yi, K. and Tao, Y.: The hardness and approximation algorithms for l-diversity, *Proc. EDBT*, pp.135–146 (2010).
[18] Xie, H., Kulik, L. and Tanin, E.: Privacy-aware collection of aggregate spatial data, *Data & Knowledge Engineering*, Vol.70, No.6, pp.576–595 (2011).

清 雄一 （正会員）

1981 年生．2009 年東京大学大学院情報理工学系研究科博士後期課程修了．同年（株）三菱総合研究所入社．同社情報技術研究センター，金融ソリューション本部等に所属．2013 年より電気通信大学助教，現在に至る．分散コンピューティング，セキュリティ，プライバシ保護技術等の研究に従事．電子情報通信学会，IEEE Computer Society 各会員．

大須賀 昭彦 （正会員）

1958 年生．1981 年上智大学理工学部数学科卒業．同年（株）東芝入社．同社研究開発センター，ソフトウェア技術センター等に所属．1985〜1989 年（財）新世代コンピュータ技術開発機構（ICOT）出向．2007 年電気通信大学大学院情報システム学研究科教授．2012 年より，国立情報学研究所客員教授兼任．工学博士（早稲田大学）．主としてソフトウェアのためのフォーマルメソッド，エージェント技術の研究に従事．1986 年度情報処理学会論文賞受賞．IEEE Computer Society Japan Chapter Chair，人工知能学会理事，日本ソフトウェア科学会理事を歴任．電子情報通信学会，人工知能学会，日本ソフトウェア科学会，IEEE Computer Society 各会員．