データ流通における漏洩者を特定可能にするサンプリング方式の提案

山岡 裕司[1,a)]

概要：データ駆動型社会において，データ流通の拡大が期待されている．データ流通において，提供元は契約により，提供先に再提供を禁じることが一般的であるが，データ流通が拡大するほど，データ漏洩が発覚したとしても漏洩元の特定が難しくなるため，提供先の漏洩対策がおろそかになる恐れがある．本稿では，提供データをサンプリングデータとし，サンプリングパターンを提供先毎に変え，そのパターンを記録しておくことで，データ漏洩発覚時に漏洩元を特定できるようにする方式を提案する．提案方式は特にパーソナルデータに対し有効であることなどを考察する．

キーワード：データ流通，漏洩者特定, 電子指紋

Sampling Method for Identifying Data Leaker in Data Distribution

Yamaoka Yuji[1,a)]

Abstract: In a data driven society, expansion of data distribution is expected. In data distribution, providers generally prohibit re-provision to recipients by contract. However, as the data distribution expands, it becomes more difficult to identify the leaker when data leakage is occurred. Therefore, there is a risk that the countermeasure for leakage of recipients will be neglected. In this paper, we propose a method that makes it possible to identify the leaker when data leakage is detected by changing the sampling pattern for each recipient. We consider that the proposed method is particularly effective for personal data.

1. はじめに

データ駆動型社会において，データ流通の拡大が期待されている．特に，パーソナルデータは「インターネットの新たな石油，デジタル世界の新たな通貨」[5] といわれ，その流通拡大への期待が高い．そのため，近年ではデータ流通仲介事業者がいくつも登場し，とりわけ日本では個人の関与の下でデータ流通・活用を進める仕組みである情報銀行の社会実装を官民で協力し推進している．

情報銀行などの仲介事業者における課題の一つに，データ提供先からの再販や漏洩のリスクがある．仲介事業者は，一つのデータを複数の提供先に提供することが多い．そのため，仲介事業者が，かつて複数の提供先に提供したデータを闇市場などで発見した場合，その漏洩者を特定することは難しい．提供先は，通常，契約によって再販や漏洩を禁じられ，それを遵守するためのセキュリティ施策を講じる．しかし，セキュリティ施策にもコストがかかり，たとえ漏洩事故を起こしても漏洩者が特定されることがなければ特段不利益とならないことが想定されることから，モラルハザードに陥りセキュリティ施策を怠る恐れがある．

そのリスクを低減する技術に，電子指紋 (digital fingerprinting) がある．電子指紋は，データ提供時に，データに提供先 ID を，提供先にとって取り除くのが困難なように埋め込む技術である．また，埋め込んだ者はそのデータから提供先 ID を抽出できるようにする．仲介事業者が電子指紋を活用することで，漏洩データを発見した際にそこから提供先 ID を抽出することで漏洩者を特定できるようになるため，モラルハザードを抑止する効果を期待できる．

[1] 株式会社富士通研究所
FUJITSU LABORATORIES LTD.
[a)] yamaoka.yuji@fujitsu.com

© 2019 Information Processing Society of Japan

我々は，情報銀行での必要性が高いと考えられる，パーソナルデータかつビッグデータである，構造化データである表データを対象とする電子指紋を研究対象とした．そのようなデータは，近年発展が目覚ましい AI 技術による分析を適用し易いため，必要性が高いと考える．さらに，情報銀行では，データ流通の可否を本人が制御でき，個人全員が許諾するデータ流通は起こり難いことからも，そのパーソナルデータは標本データ（全数データではない）とみなして差し支えないと考える．

Kamran ら [3] のサーベイによると，電子指紋は電子透かしの応用であり，表データを対象とする電子透かしの研究は，画像をはじめとするメディアデータに比べて新しい分野である．Kamran らは表データを対象とする電子透かしの従来技術を次の 3 つに分類している．

- Bit-resetting techniques (BRT). 重要性の低いビットなどを改変して透かしを埋め込む.
- Data statistics-modifying techniques (DSMT). データ全体を改変して統計量に透かしを埋め込む.
- Constrained data content-modifying techniques (CDCMT). レコードの並び順などを改変して透かしを埋め込む.

しかし，それらのほとんどは，提供データに，元データからの歪みを導入するため，提供先にとっての品質低下をもたらすという問題がある．歪みを導入しない従来技術 [1], [2] もあるが，それらは脆い透かし (fragile watermark) であり，電子指紋として使用するにはロバスト性が低いという問題がある．ここで，歪みを導入する従来技術は，データの変更か挿入による方法で実現しており，削除による方法が見られないことに我々は着目した．

本稿では，サンプリングパターンによる電子指紋方式を提案する．対象として想定しているような，標本データに対しては，更なるサンプリングによる微小なレコード削除がもたらす提供先にとっての品質低下は，問題にならないくらい小さいと考える．本稿の貢献は次の通り．

- 表データに対する電子指紋としてこれまで見られなかった，サンプリングによる方式を提案．提案方式は次の特徴を持つ.
  - ロバスト性： 電子指紋は，データ正規化により取り除かれることがない上，ある程度のレコードやセル値の削除にも耐性がある.
  - データ真正性： 提供する各レコードは真正，つまり歪みがない.
- 本稿で特定能力と呼ぶ，レコード削除へのロバスト性の度合いを示すパラメーターを提案し，提供先数と特定能力を満たすサンプリング率の算出式を導出.
- 提案方式のロバスト性について，最も代表的なオープンなパーソナルデータである Adult データセットでの実験で確認.

本稿の以降の構成は次の通り．第 2 節では，本稿が対象とする問題を整理する．第 3 節では，従来方式とその問題点を説明する．第 4 節では，提案方式を説明する．第 5 節では，提案方式のロバスト性を実験評価する．第 6 節にてまとめる.

2. 問題整理

本稿が取り扱う問題を本節で整理する.

データ提供元が，各提供先に異なる電子指紋を埋め込んだデータを提供することで，漏洩データを入手した際に，漏洩者を特定できるようにする．そのことを提供先に知らせることで，モラルハザードによる漏洩を抑止する効果を期待できる．また，漏洩が起きた際には，その漏洩者への以降のデータ提供に際しては審査を厳しくすることなどで，リスク低減の効果を期待できる．なお，漏洩者に契約違反を根拠にするなどで賠償請求することも考えられるが，現時点では電子指紋が漏洩者特定の決定的な根拠とされた判例は我々が知る限りはなく，そのような使い方は難しいと考える.

データ提供元の典型例として情報銀行を想定し，提供データは次の全て—(1) パーソナルデータ，(2) ビッグデータ，(3) 構造化データ，(4) 表データ に当てはまるとする．**図 1** は情報銀行における電子指紋活用の全体像である.

また，次を前提とする.

- 同一のデータを（加工して）複数の提供先に提供する．もし一者にしか提供しない場合，漏洩者特定は容易であるため，この前提を置く．同一のデータを複数の提供先に提供する形態として，レディメイド（どの提供先にも加工せずに同じデータを提供）とカスタムメイド（各提供先の意向に合わせて加工したデータを提供）が考えられる．元データに後述するロバスト性のある電子指紋を埋め込めば，そこから多少カスタムメイドしたデータにも電子指紋が残る傾向があるため，本稿では簡明のためレディメイドを前提とする.
- データのレコード数は，提供先の数に比べて桁違いに多い．AI の訓練データの件数としては万件以上が望ましく，提供先からの需要が高いビッグデータのレコード数は万件以上であることが典型的と考える．一方，オープンではないデータの流通において，提供先の数が千以上のケースは我々が知る限りない．よって，この前提を置けると考える.
- 各レコードは識別可能である．本人同意の下でのデータ流通を仲介する情報銀行では，各人のレコードが識別可能なレベルに詳細なパーソナルデータの流通が期待されているため，この前提を置けると考える.
- 標本データとみなせる．本人同意の下でのデータ流通を仲介する情報銀行では，母集団には本人同意の取れなかった個人なども含まれると考えると，提供データ

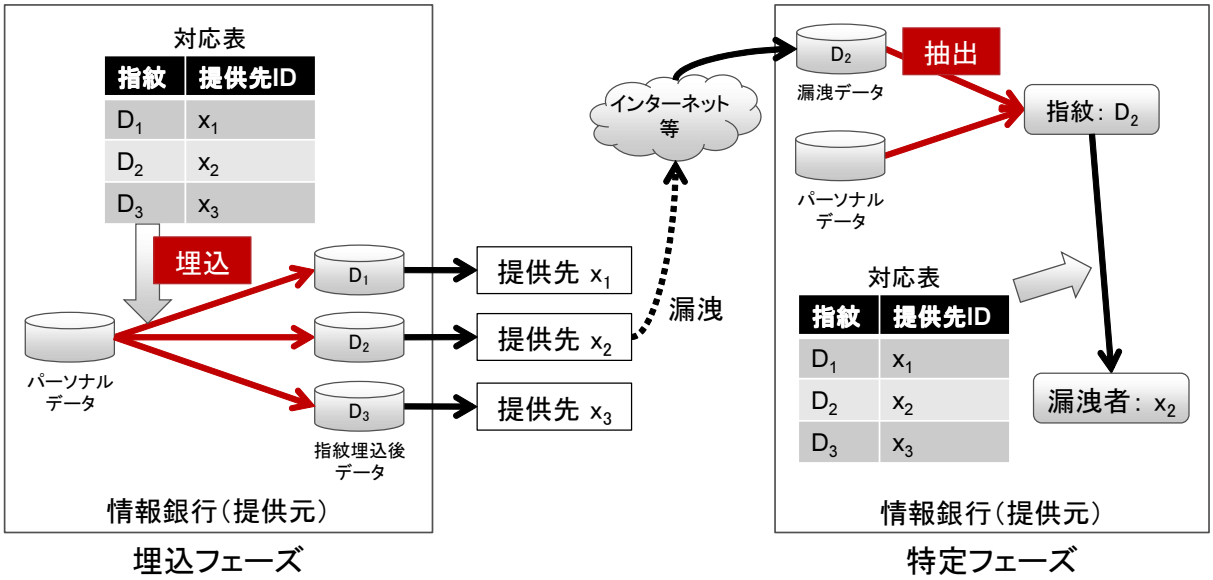

図 1: 全体像

Fig. 1 Ovewview

は標本データとみなせる．その要素は一人分のデータ，すなわちレコードとする．

解決技術となる電子指紋への要件は次の二つとする．

- ロバスト性： 電子指紋は，データ正規化により取り除かれることがない上，ある程度のレコードやセル値の削除にも耐性がある．
- データ真正性： 提供する各レコードは真正，つまり歪みがない．

ロバスト性を要件とした理由を説明する．本稿でのデータ正規化とは，提供先が自組織で採用しているデータレイアウト等に合うように変換することである．不要なデータや空白を削除したり，用語や符号を変換したり，ソートしたり，といった処理が含まれる．用語変換では，たとえば「女性」というセル値を「女」へ変換する．符号変換では，たとえば「女」というセル値を 0 へ，「男」というセル値を 1 へ，それぞれ変換する．ビッグデータ分析の際は，データを専用の DBMS に格納してからおこなわれるなど，データ正規化は一般的におこなわれる．これまでに，DBMS へのアクセス権限を持つ内部者によるデータ漏洩事例も複数起きていることを鑑みると，データ正規化により取り除かれるような電子指紋では効果が小さい．また，提供先ではレコードやセル値の削除もしばしばおこなわれる．特にレディメイドで提供されたデータに対しては，専用の DBMS に格納する際に，当面必要な値を持つレコードだけや，当面必要な属性だけを抽出することがある．カスタムメイドの場合でも，複数の部署がそれぞれの目的で分析するような場合など，当面必要な部分を抽出することがある．よって，そのような場合でも電子指紋が取り除かれにくいことが望まれる．

データ真正性を要件とした理由を説明する．ビッグデータ分析において，データ真正性を期待する事例は多い．たとえば，医療・ヘルスケア分野では僅かな歪みも嫌う傾向がある．偏ったレコード提供も歪みの一種と捉えられるが，ランダムサンプリングであれば，少なくともそのサンプリング率を明らかにすれば許容されると考える．標本データの分析に基づく多数の研究実績があるためである．

一方，従来の電子指紋でしばしば要件とされる次の事項は，本稿では要件としない．

- 提供元による自作自演の防止： ある提供先を漏洩者だとする根拠を，提供元が偽造できないこと．
- 提供先同士の結託への耐性： 提供先同士が互いのデータを見せても，電子指紋を無効化できないこと．

その理由は，前述の通り，現時点での電子指紋の主目的はリスク低減であるため，上記のような攻撃的な行為への対策がなくても十分に有用と考えるためである．

3. 従来方式

前述した Kamran ら [3] の分類での，BRT や DSMT の方式は多数提案されている．しかし，いずれも提供レコードに歪みを導入するものであり，データ真正性の要件を満たさない．

一方，レコードの並び順を改変して電子指紋を埋め込む方式（CDCMT）では，データ正規化により無効化されてしまい，データ真正性の要件を満たさない．

文章データ一般への電子指紋方式として，同義語への置換や空白位置調整等，データ真正性に配慮した方式も提案されている．しかし，これらもデータ正規化により無効化されてしまい，ロバスト性の要件を満たさない．

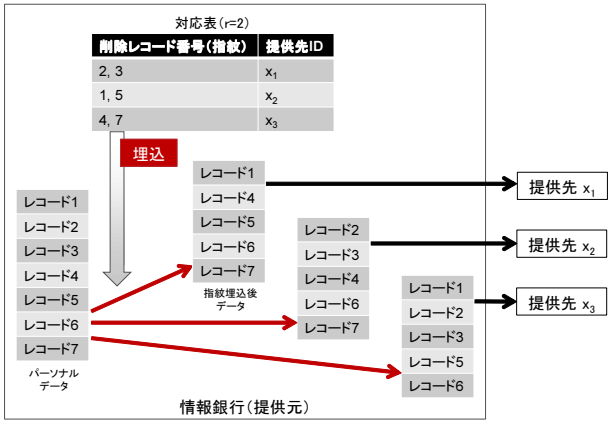

図 2: 提案方式での電子指紋の埋め込み

Fig. 2 Our method of embedding digital fingerprint

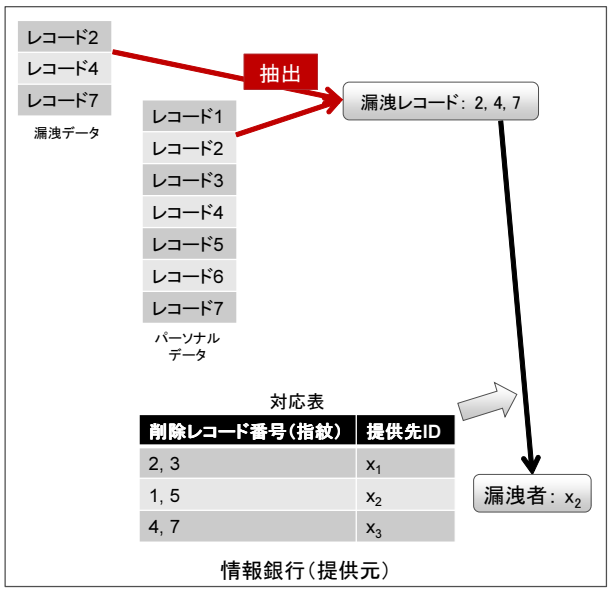

図 3: 提案方式での電子指紋の抽出

Fig. 3 Our method of extracting digital fingerprint

4. 提案方式

本節では，前述の要件を満たす方式である，我々の提案方式を説明する．

提案方式は，サンプリングベースの，つまり削除するレコード群の違いを電子指紋とする方式である．埋め込み時に，ランダムサンプリングにより削除した全てのレコード番号と，そのデータの提供先とを対応付けて対応表として記録しておく．埋め込みの様子を**図 2** に示す．漏洩データを入手した際，元データと突合せて，漏洩データに対応する元データのレコード番号を全て抽出し，対応表から漏洩可能な提供先を絞り込む．絞り込みは，漏洩データに含まれるレコードのいずれか 1 レコードでも提供を受けていない提供先を漏洩者でないとする消去法によっておこなう．抽出と漏洩者特定の様子を**図 3** に示す．

提案方式は要件を満たす．まず，データ真正性について，提供データはランダムサンプリングされただけのデータであるため，満たされる．ロバスト性について，まずデータ正規化への耐性について，提案方式による電子指紋はデータ正規化の影響を受けないのは明らかである．データ正規化されたデータが漏洩した場合，データ正規化の内容を把握できれば，元データにその正規化をおこなうことにより，漏洩データと容易に突合せできるようになる．漏洩データが提供データ由来のものであると認識できている以上，データ正規化の内容は漏洩データを見ることで把握できるはずである．次に，レコード削除への耐性について，レコードが全く削除されない場合，電子指紋は影響を受けない．その場合，漏洩データの全レコードが前述の前提通りに識別可能であれば，元データのレコードと 1 対 1 に突合せられるため，漏洩可能な提供先は一意に絞り込まれる．さらに，ある提供先が一部のレコードだけを漏洩した場合でも，漏洩可能な提供先を一意に絞り込まれる場合が多いほど望ましい．

そこで，本稿で特定能力と呼ぶ，ロバスト性のうちレコード削除への耐性の度合いを示すパラメーターを導入し，与えられた提供先数と特定能力を満たすサンプリング率を計算する．特定能力とは，提供先が漏洩可能な全てのレコード組み合わせの場合の数のうち，漏洩した場合に漏洩者を一意に絞り込むことが可能な組み合わせの場合の数の割合とする．たとえば，特定能力を 95%とした電子指紋を使い，ある提供先が漏洩可能な全てのレコード組み合わせのうちからあるレコード組み合わせを無作為に一つ選び漏洩させた場合，その漏洩データの漏洩者を確率 95%で一意に特定できる．

提案方式の詳細は次の通りである．

準備 提供元が，提供する元データ T を決めた際，本手順をおこなう．

(1) 提供元は，T から作る提供データの最大提供先数 $p \in \mathbb{N}$ と特定能力 $a \in \mathbb{R}, 0 < a \leq 1$ を決める．

(2) 削除行数 $r \in \mathbb{N}$ を次式により求める．

$$r \geq \log_2 \frac{p-1}{1-a} \tag{1}$$

本式の導出については後述する．

(3) 空の対応表 M を用意する．M は，削除するレコード番号を要素とするサイズ r の集合 $\{d_1, d_2, \ldots, d_r\}(d \in \mathbb{N}, 1 \leq d \leq |T|)$ ($|T|$ は T のレコード数) と提供先 ID x との対応付けの，集合を記録する表である．

提供 提供先 ID x に提供データ T_x^- を提供する際，本手順をおこなう．

(1) M に記録されている全レコード番号 E を取得する．たとえば，M が図 2 の場合，$E = \{1, 2, 3, 4, 5, 7\}$ となる．

(2) 削除するレコード番号の集合 D =

$\{d_1, d_2, \ldots, d_r\}(d \in \mathbb{N}, 1 \leq d \leq |T|, d \notin E)$ をランダムに生成する．

(3) 関数 s を，入力としてリストとリスト内番号集合とを受け，そのリストの内からその番号集合に対応する要素を削除したサブリストを出力する関数とすると，$T_x^- \leftarrow s(T, D)$ により T_x^- を作成する．

(4) x に対応する提供先に T_x^- を提供し，D と x を紐付けて対応表に追加する．

漏洩者特定　T から作成した提供データの一部を漏洩データ $T^{-'}$ として入手した際，本手順により漏洩者を絞り込む．

(1) T と $T^{-'}$ を突合せ，$T^{-'}$ に相当する T のレコード番号集合 $D' = \{d'_1, d'_2, \ldots\}(d' \in \mathbb{N}, 1 \leq d' \leq |T|)$ を算出する．前提通りに識別可能であれば，突合せにより $T^{-'}$ の各レコードが T のどのレコードに対応するか一意に決まることが期待できるが，一意に決まらない場合は $T^{-'}$ のそのレコードは無視する．

(2) 対応表の各要素につき，レコード番号の集合 D について $D \cap D' = \emptyset$ ならその要素の提供先 ID x を取得する．

(3) 取得した全ての提供先 ID が一つしかなかったなら，それが漏洩データを漏洩した提供先の ID とする．

最後に，削除行数 r の算出について説明する．

式 (1) は次のように導出できる．まず，データのレコード数は提供先の数に比べて桁違いに多いという前提に基づき，

$$|T| \geq rp \tag{2}$$

が成立すると仮定する．式 (2) は，各提供先で重複がないように削除するレコード番号の集合を割り当てられることを意味している．もっとも，r は導出前であるが，その値が小さいことを見越しての仮定である．実際，提案方式では，各提供先で重複がないように削除するレコード番号の集合を割り当てる．このとき，特定能力の定義より，提供先によらず，漏洩可能な全てのレコード組み合わせの場合の数を C_A，漏洩した場合に漏洩者を一意に絞り込むことが可能なレコード組み合わせの場合の数を C_1 とすると，$a \leq C_1/C_A$ となる．C_A は，各提供先は $|T| - r$ レコードの提供を受けるため，$2^{|T|-r} - 1$ となる．C_1 は，ある提供先から見て，他の提供先が提供を受けていない r レコードのいずれか一つ以上のレコードを漏洩した，ということが他の各提供先（最大 $p-1$ 者）に対しても成立する場合に限り漏洩者を一意に絞り込むことが可能なため，全提供先が提供を受けている $|T| - rp$ レコードの組み合わせの場合の数も勘案すると，$(2^r - 1)^{p-1} 2^{|T|-rp}$ となる．よって，次が成り立つ．

$$a \leq \frac{(2^r - 1)^{p-1} \cdot 2^{|T|-rp}}{2^{|T|-r} - 1} \tag{3}$$

$$= (2^r - 1)^{p-1} \cdot \frac{1}{2^{r(p-1)}} \cdot \frac{2^{|T|-r}}{2^{|T|-r} - 1} \tag{4}$$

$$\approx (2^{r(p-1)} - (p-1)2^{r(p-2)}) \cdot \frac{1}{2^{r(p-1)}} \tag{5}$$

$$= 1 - \frac{p-1}{2^r} \tag{6}$$

ここで，式 (5) の近似は，$2^r \gg 1$ とみなし $(2^r - 1)^{p-1} \approx 2^{r(p-1)} - (p-1)2^{r(p-1)}$ と，$2^{|T|-r} \gg 1$ とみなし $(2^{|T|-r})/(2^{|T|-r} - 1) \approx 1$ と，それぞれ近似したことによる．式 (6) を変形することで，式 (1) を得る．

実際に r を算出する際は，式 (1) にて近似解を求め，その r が式 (2) および式 (3) を満たしていることを確認すれば良い．

5. 実験評価

提案方式が，実データに対してもロバスト性を発揮することを確認するため，実験をおこなった．

対象データは，最も代表的なオープンなパーソナルデータである，UCI Machine Learning Repository[4] の Adult データとした．Adult データは米国の国勢調査に対する回答であり，各レコードは一人の回答内容に相当し，48,842 レコードある．実験ではそのうち，重複レコードを削除した，識別可能な 48,790 レコードを用いた．属性は「age」，「workclass」，「INCOME」など 15 属性からなる．属性「INCOME」は「<=50K」か「>50K」の 2 値のみを取り，それぞれのレコード数は 37,109, 11,681，全レコードに占める割合は 76%, 24% である．このデータを，提供元しか保有していないとの想定で，提案方式にて複数の提供先に提供する設定とした．

提供先数 $p = 1000$，特定能力 $a = 0.95$ とした．提供先数について，前述の通り千以上のケースは我々が知る限りなく，この設定で確認できれば現実的に十分と考える．

ある提供先が属性「INCOME」のマジョリティである値「<=50K」のレコードを全て漏洩するシナリオとし，各提供先のいずれかが漏洩する計 p 通りのシナリオのうち，漏洩者を一意に特定できた数を数えた．

その結果，全てのシナリオで漏洩者を一意に特定できたため，一定のロバスト性を確認できたと考える．

なお，このケースでの r は，式 (1) より $r \geq 14.286\cdots$ となるため $r = 15$ とした．これは，式 (2) および式 (3) を満たしている．また，この実験では重複レコードを削除したデータを用いたが，実用の際は識別可能な（重複のない）レコードのみに対して電子指紋を埋め込めば良い．Adult データでは大部分である 99.9% が識別可能なレコードであり，重複レコードの有無は実用上無視できると考える．前述の通り，流通が期待されるパーソナルデータの識別可能性は高く，提案方式は特にパーソナルデータ流通において

実用的と考える．

6. まとめ

本稿では，データ流通における提供先からのデータ漏洩のリスクを低減することを狙いとした，表データに対するサンプリングによる電子指紋方式を提案した．データ漏洩発覚時に漏洩元を特定できるようになるため，提供先に対し漏洩対策の動機を与えられるようになる．提案方式は，提供データをサンプリングデータとし，サンプリングパターンを提供先毎に変える．これにより，従来方式では達成できていなかった，ロバスト性とデータ真正性の両立を達成した．ここで，ロバスト性とは，提供先での正規化や，ある程度のレコードやセル値の削除にも耐性があることを指す．本稿ではさらに，本稿で特定能力と呼ぶ，レコード削除へのロバスト性の度合いを示すパラメーターを提案し，提供先数と特定能力を満たすサンプリング率の算出式を導出した．加えて，提案方式のロバスト性については，最も代表的なオープンなパーソナルデータである Adult データセットでの実験で確認した．

参考文献

[1] Bhattacharya, S. and Cortesi, A.: A Distortion Free Watermark Framework for Relational Databases, *In: ICSOFT* (2009).

[2] Hamadou, A., Sun, X., Gao, L. and Shah, S. A.: A Fragile Zero-Watermarking Technique for Authentication of Relational Databases, *International Journal of Digital Content Technology and its Applications*, Vol. 5, No. 5 (2011).

[3] Kamran, M. and Farooq, M.: A Comprehensive Survey of Watermarking Relational Databases Research, *CoRR*, Vol. abs/1801.08271 (online), available from ⟨http://arxiv.org/abs/1801.08271⟩ (2018).

[4] Lichman, M.: UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml` (2013).

[5] World Economic Forum: Personal Data: The Emergence of a New Asset Class, `https://www.weforum.org/reports/personal-data-emergence-new-asset-class`.