プライバシーを保護した名寄せプロトコル

菊池 亮　　五十嵐 大

NTT セキュアプラットフォーム研究所
180-8585 東京都武蔵野市緑町 3-9-11
kikuchi.ryo@lab.ntt.co.jp

あらまし 昨今様々なパーソナルデータが蓄積されているが，ある個人のパーソナルデータは，店舗 A と施設 B のような別の主体によって管理されている場合が考えられる．このとき，店舗 A と施設 B の持つデータを名寄せできればより多くの情報を利活用できるが，店舗 A が名寄せデータを得ると，自身のデータと突き合わせることにより名寄せデータ中の個人を識別し，相関分析等には本来不必要な情報を得ることができてしまう．本論文では，名寄せデータを作成する際に，最終的に名寄せデータを得る主体以外は何も情報を得られず，且つ名寄せしたデータを得た主体も，名寄せデータから個人が識別できないようにするプロトコルを提案する．まず基本方式として，公開鍵暗号と Pk-匿名化を組み合わせたプロトコルを提案する．次に，基本方式を拡張し，より多様なデータや状況に対応する方法，およびより安全性を高める方法について述べる．

Simple privacy-preserving join protocols for vertically partitioned data

Ryo Kikuchi　　Dai Ikarashi

NTT Corporation.
3-9-11 Midoricho, Musashinoshi, Tokyo 180-8585, Japan
kikuchi.ryo@lab.ntt.co.jp

Abstract We propose join protocols with anonymity for vertically partitioned data. We first propose a basic protocol that produces the joined table with two security notions: secrecy of tables for all parties except the party that obtains the joined table, and (probabilistic) k-anonymity for the party. In the basic protocol, we consider only a categorical attribute, the risk of distinguishing an individual, and assume that the common records among all tables is known and the adversary's knowledge is restricted. We then show how to extend the basic protocol. We give a way to treat a numerical attribute while preserving the same security level, reduce the risk of attribute estimation, manage the protocol even if the common records is unknown, and make the basic protocol be secure against an adversary with any background knowledge.

1　はじめに

位置情報や診療履歴などの個人に関する様々な情報（パーソナルデータ）を容易に取得できる環境が整い，データ分析技術の進歩と相まって，個人に関する情報の活用への期待が高まっている．その一方で，個人情報保護やプライバシ保護の観点から，パーソナルデータは慎重な取扱いが必要とされ，データの保護と活用をどう両立させるかが課題となっている．

このようなデータの保護と活用の両立を目

指した技術として，何らかのプライバシ保護指標を満たしながら，データの分析もしくは加工を行うプライバシ保護データ分析（Privacy-Preserving Data Analysis: PPDA）の研究が盛んに行われてきている．プライバシ保護指標やデータの分析・加工方法は対象とするデータによって異なるため，本論文で扱うデータは，ミクロデータと呼ばれるテーブル形式のデータを前提とする．すなわち，個人の情報は1レコードに対応し，レコードには複数の属性があり，複数のレコードを組み合わせてテーブルとなっているものとする．

1.1 k-匿名性と Pk-匿名性

パーソナルデータと特定の個人を攻撃者に関連付けられるリスクは識別リスクと呼ばれ，プライバシを保護するために考慮すべきリスクとされている [5,12]．識別リスクを定量化したプライバシ保護指標として一般によく知られている指標は Sweeny の提案した k-匿名性 [11] である．k-匿名性とは，どのレコードについても「自身と同じ値を持つレコードが $k-1$ 個以上存在する」という指標であり，直観的には「ある個人が k 個のレコードのうちどれか絞り込めない」ため識別リスクを低減している．

k-匿名性を用いることで，属性値の丸めや削除を行った際の識別リスクが評価できる．一方，データへのノイズ付加や確率的な値の変更も実際には識別リスクを低減しているはずだが，k-匿名性では識別リスクを評価できない．極端な例として，属性値を乱数に置き換えた場合，識別リスクはかなり低くなるはずだが，値域が大きいとほぼ 1-匿名性しか持たない．

ノイズ付加や確率的な値の変更に対する識別リスクの指標として，k-匿名性を確率的空間へ拡張した Pk-匿名性 [6] がある．Pk-匿名性は「攻撃者は高々確率 $1/k$ でしか個人を識別できない」ことを保証した指標であり，k-匿名性と同じ尺度で識別リスクを評価できる．Pk-匿名性を満たすテーブルを作り出す方法としては Post-RAndomization Method (PRAM) [9] が用いられている．PRAM は遷移確率行列と呼ばれる行列に従って値を確率的に変化させる手法全般を指すクラスであり，具体的な実装としては維持置換撹乱 [1] やラプラスノイズ付加 [14] がある．

1.2 パーソナルデータの名寄せ

昨今様々なパーソナルデータが蓄積されているが，ある個人のパーソナルデータは，店舗 A と施設 B のような別の主体によって管理されている場合が考えられる．このとき，店舗 A と施設 B の持つデータの名寄せできれば 2 つのデータの相関関係のような分析が可能となる．

しかし，例えば店舗 A が名寄せしたテーブルを得ると，自身のデータと突き合わせることにより名寄せデータの個人を識別し，相関関係のような分析には本来不必要な個人に関する情報が，店舗 A に渡ってしまう．この問題は，仮に暗号化しつつ名寄せをしたとしても，復号した結果から識別するため回避できない．

1.3 本研究の成果

本論文では，ある個人のパーソナルデータが複数の主体によって管理されている場合に，安全に名寄せデータを作成するプロトコルを提案する．すなわち，最終的に名寄せデータを得る主体以外は何も情報を得られず，且つ名寄せしたデータを得た主体も名寄せデータから個人を識別できないという 2 つの要件を満たすプロトコルである．

1.3.1 基本方式

本論文では，まず基本方式と呼ばれるものを提案する．n 人のパーティが互いにテーブルを持っており，それぞれのテーブルは同じ人物のレコードで構成されているが，あるパーティのテーブルには年齢，ある異なるパーティのテーブルには年収など，テーブル毎に属性が異なる．このとき，暗号化したままテーブルを名寄せし，暗号化した横に長い（年齢と年収が結合された）テーブルを作成する．その後，暗号化したまま Pk-匿名化を行い，名寄せしたテーブルを得ても個人が識別できないようにしてから復号し，名寄せされたテーブルを得るものである．

なぜ Pk-匿名化か？　k-匿名化では，一般に全体のレコードの属性値を見てどの程度一般化や削除を行うか判定するアルゴリズムが用いられる．しかしこのような操作は，属性値を秘匿しながら行うことは難しい．

属性値を秘匿しながら計算を行う際は，各サーバの挙動が属性値に依存してはいけない．一般化を属性値に依存せずに行う場合，単純には一般化する場合としない場合の両方を計算する必要がある．すると，計算量は一般化の回数に対し指数的に増加してしまう．

一方 Pk-匿名化で用いられる PRAM は，レコード毎に独立して処理を行う．また，この処理は属性の値域やレコード数にのみ依存し，属性値に依存しない．そのため秘匿しながら計算を行う場合に適している．

1.3.2　基本方式の拡張

基本方式は，扱う属性や攻撃者の背景知識，考慮するリスクが限定的であり，アプリケーションによっては，より多様な属性値や強い背景知識等を考慮する必要がある．そのような場合に対応するため，本論文では，

- 数値属性を扱う場合，
- 全テーブルの共通レコードが未知の場合，
- 強い背景知識を持つ攻撃者を想定する場合，
- 属性推定リスクを考慮する場合

について，基本方式の拡張方法を提案する．これらの拡張は排反ではなく互いに組み合わせることが可能である．

2　準備

2.1　記法

$\mathrm{abs}(x)$ は x の絶対値を指し，集合 A に対し $|A|$ は集合の要素数を表し，$\mathscr{G}$ は群を意味する．また $x \leftarrow B$ は，B が集合ならば B から一様ランダムに要素を選び x に代入することを指し，B がアルゴリズムならば B の出力を x に代入することを指す．

2.2　データ形式と記法

本論文での匿名化前後のデータは，各行が一個人のデータを表し（これをレコードと呼ぶ），各列には属性と呼ばれる年齢，年収等の値が入力されているようなテーブルであるとする．

テーブルに対する匿名化を形式的に議論するための定義を与える．

- T, τ: 匿名化前のテーブルの確率変数及びそのインスタンス
- T', τ': 匿名化テーブルの確率変数及びそのインスタンス
- $\mathbb{R}, \mathbb{R}'$: 匿名化前/後のテーブルのレコード集合
- $\mathbb{A}, \mathbb{A}'$: 匿名化前/後の属性の集合
- $\mathbb{V}_a, \mathbb{V}'_a$: 匿名化前/後の属性 $a \in \mathbb{A}$ の取りうる属性値の集合
- $\mathbb{V}, \mathbb{V}'$: 匿名化前/後のレコードが取りうる属性値の組み合わせの集合．すなわち $\prod$ で直積を表すと $\mathbb{V} = \prod_{a \in \mathbb{A}} \mathbb{V}_a$ である．
- V_a, v_a: あるレコードの属性 $a \in \mathbb{A}$ の属性値の確率変数及びそのインスタンス．
- Δ, δ: プライバシー保護処理の確率変数及びそのインスタンス
- f_X: 確率変数 X に関する確率密度関数
- Π, π: $\mathbb{R} \to \mathbb{R}'$ であるようなランダム置換の確率変数及びそのインスタンス

τ, τ', δ はそれぞれ $\tau : \mathbb{R} \to \mathbb{V}$, $\tau' : \mathbb{R}' \to \mathbb{V}'$, $\delta : (\mathbb{R} \to \mathbb{V}) \to (\mathbb{R} \to \mathbb{V}')$ であるような関数であり，π, δ, τ, τ' の間には $\delta(\tau) = \tau' \circ \pi$ が成り立つ．また，$\tau(r)^{(a)}$ を $\tau'(r')^{(a')}$ をそれぞれレコード r の属性 a の属性値，レコード r' の属性 a' の属性値とする．

2.3　Pk-匿名性

ある τ', $r \in \mathbb{R}$, $r' \in \mathbb{R}$, $\Delta(T) = T' \circ \Pi$ について，$\Pr[\Pi(r) = r' \mid T' = \tau']$ を攻撃成功の確率とし，Pk-匿名性は以下のように定義される．

定義 2.1　*[6]* 実数 $k \geq 1$ について，あるプライバシー保護処理 Δ と匿名化テーブル τ' が, 背景知識 f_T を持つ攻撃者，任意のテーブル T，レコード $r \in \mathbb{R}$，レコード $r' \in \mathbb{R}'$ について，

$$\mathcal{E}(f_T, \tau', r, r') \leq \frac{1}{k}$$

であるならば，(Δ, τ') は背景知識 f_T を持つ攻撃者に対して Pk-匿名性を満たすという．

なおこの定義は文献 [6] の定義から，弱い背景知識を持つ攻撃者も想定できるように拡張した定義となっている．

2.4 Pk-匿名化

Pk-匿名化とは，Pk-匿名性を満たすための処理であり，カテゴリ属性に対する維持置換撹乱 [1,6]，数値属性に対するラプラスノイズ付加 [14] および有界ノイズ付加 [15] が提案されている．

2.4.1 維持置換撹乱

Pk-匿名化の一種として，属性値を一定の確率で維持し，それ以外ではランダムな値に置換する手法を維持置換撹乱と呼ぶ．

遷移確率行列 A を，$A^{(a)}$ は属性 a の値が匿名化によってどのように変化するかを表した $|\mathbb{V}_a| \times |\mathbb{V}'_a|$ の行列とする．$A^{(a)}$ の各要素 $A^{(a)}_{v_a,v'_a}$ は，匿名化によって属性値 v_a が v'_a に遷移する確率となっている．維持置換撹乱とは，任意の $a \in \mathbb{A}$, $v_a \in \mathbb{V}_a$, $0 \leq \rho_a \leq 1$ について

$$A^{(a)}_{v_a,v'_a} = \begin{cases} \rho_a + \frac{(1-\rho_a)}{|\mathbb{V}_a|} & \text{if } v_a = v'_a \\ \frac{(1-\rho_a)}{|\mathbb{V}_a|} & \text{otherwise} \end{cases}.$$

を満たすものである．遷移確率行列 A とは，$\prod$ をクロネッカー積としたとき $A = \prod_{a \in \mathbb{A}} A^{(a)}$ である．このとき，$A_{v,v'}$ は $A^{(a)}_{v_a,v'_a}$ と同様に，レコードの属性値が v のとき，匿名化によって属性値が v' に変わる確率となっている．

維持置換撹乱は，$\alpha := \left(\frac{k-1}{|\mathbb{R}|-1}\right)^{\frac{1}{|\mathbb{A}|}}$ としたとき，

$$\rho_a = \frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}(|\mathbb{V}_a|-1)} \text{ for } a \in \mathbb{A} \quad (1)$$

とすることで，Pk-匿名性を満たせることが知られている [6]．

2.5 有界ノイズ付加

維持置換撹乱では属性値の近さを考えず，維持されない場合一様ランダムに変換しているが，年収やポイント等の数値属性では，元の値に近い値に遷移する確率を高めた方が有用性が高まると期待できる．

そのような Pk-匿名化として，有界ノイズ付加が提案されている．有界ノイズ付加では，ある値域に区切られたノイズを加算することにより，Pk-匿名性を満たす方法である．

本論文では確率密度関数

$$f_{\mathrm{lap},b}(x) = \frac{1}{2b}\exp(-\frac{\mathrm{abs}(x)}{b}).$$

に従うラプラスノイズを基とした有界ノイズを用いる．ある数値属性 $a \in \mathbb{A}$ の値域が $[v_{\min}, v_{\max}]$ のとき，$f_{\mathrm{lap},b}(x)$ のパラメータ b を

$$b := -\sqrt{\frac{\inf_{u,v \in \mathbb{V}_{a_i}}(\|u-v\|_1)}{\ln \alpha}} \quad (2)$$

と設定し，

$$\alpha_v = \int_{v_{\min}-v}^{v_{\max}-v} f_{\mathrm{lap},b}(x)dx \quad (3)$$

としたとき $v \in \mathbb{V}_a$ に対して確率密度関数

$$f'_v(x) = \begin{cases} f_{\mathrm{lap},b}(x)/\alpha_v & \text{if } x \in [v_{\max}-v, v_{\min}-v] \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

に従うノイズを加えることで，Pk-匿名性を満たすことができる [15]．

2.6 再乱数化可能な公開鍵暗号

本論文では，暗号文の中身を変えず，乱数を変えることができるような公開鍵暗号を再乱数化可能な公開鍵暗号と呼ぶことにする．なお，準同型性を持てば再乱数化可能である．また，パラメータ $params$ は Setup 以外のアルゴリズム全てに必要なため省略する．

- $\mathsf{Setup}(1^\lambda)$: セキュリティパラメータ 1^λ を入力として，パラメータ $params$ を出力する．
- $\mathsf{KGen}()$: $params$ を入力として，公開鍵と秘密鍵のペア (pk, sk) を出力する．
- $\mathsf{Enc}(pk, m)$ pk とメッセージ m を入力として，暗号文 c を出力する．
- $\mathsf{Dec}(sk, c)$ sk と暗号文 c を入力として，m を出力する．

- $\mathsf{ReRand}(pk, c)$ pk と c を入力として，再乱数化された暗号文 c' を出力する．

正当性として，任意の $params \leftarrow \mathsf{Setup}(1^\lambda)$, 任意の $(pk, sk) \leftarrow \mathsf{KGen}(params)$ について $m = \mathsf{Dec}(sk, \mathsf{ReRand}(pk, \mathsf{Enc}(pk, m)))$ かつ $m = \mathsf{Dec}(sk, \mathsf{Enc}(pk, m))$ が成り立つ．

本論文では，再乱数化可能な公開鍵暗号は IND-CPA 安全性 [3] を持つと仮定する．すなわち，任意の多項式時間で動くチューリング機械 $\mathcal{A}$ に対し，

$$\Pr\left[\begin{array}{l} params \leftarrow \mathsf{Setup}(1^\lambda); \\ (pk, sk) \leftarrow \mathsf{KGen}(); \\ (st, (m_0, m_1)) \leftarrow \mathcal{A}_0(pk); \ : b' = b \\ b \leftarrow \{0,1\}; \\ b' \leftarrow \mathcal{A}(st, \mathsf{Enc}(pk, m_b)) \end{array}\right] \leq \mathsf{negl}(\lambda)$$

が成り立つ．

3 基本方式: 再乱数化可能な暗号と維持置換撹乱を用いた名寄せ

本章では，まず基本となる名寄せプロトコル（基本方式）を提案する．

基本方式では，以下のステップで名寄せテーブルを作成している．

1. $\mathcal{P}_0$ はあらかじめ再乱数化可能な公開鍵暗号の鍵生成を行い，公開鍵を配布する．
2. それぞれの $\mathcal{P}_i$ は自身の持つテーブルの属性値を全て暗号化し，（名寄せできるように順序を明らかにして）$\mathcal{P}_{n-1}$ に渡す．
3. $\mathcal{P}_{n-1}$ は受け取った暗号文を並べることで，暗号化された名寄せテーブルを得る．
4. $\mathcal{P}_{n-1}$ は $\mathcal{P}_0$ が持つ属性に維持置換撹乱を行う．
    1. 式 1 を用いて各属性の維持確率 ρ を計算する．このとき必要な情報は所望の k，レコード数，$\mathcal{P}_0$ が持つ属性の数と値域であるため，$\mathcal{P}_{n-1}$ は計算可能である．
    2. $\mathcal{P}_0$ のテーブルに含まれる全ての属性の暗号文に対し，維持確率に応じて撹乱を行う．維持する場合には ReRand を用いて再暗号化し，置換する場合には乱数を暗号化したもので元の暗号文を上書きする．
    3. レコードをシャッフルする．
4. $\mathcal{P}_{n-1}$ は全ての暗号文を $\mathcal{P}_0$ に渡し，$\mathcal{P}_0$ は復号することで名寄せされたテーブルを得ることができる．

では，具体的にプロトコルを記述する．パーティ $\mathcal{P}_0, \ldots, \mathcal{P}_{n-1}$ はそれぞれ $\tau_0, \ldots, \tau_{n-1}$ を所持し，τ_i は $\mathbb{A} = \bigcup_{i<n} \mathbb{A}_i$ であるような $\mathbb{A}_i$ を属性として持ち，$\mathbb{V}_i = \prod_{a \in \mathbb{A}_i} \mathbb{V}_a$ としたとき $\tau_i : \mathbb{R} \to \mathbb{V}_i$ とする．ここで，$\mathbb{R}$ は i によらず一定であるということは，ある ID となる属性が存在し，その属性値がどのテーブルでも同一だということである．なお，$\mathcal{P}_0$ は再乱数化可能な公開鍵暗号の鍵生成をあらかじめ行い，公開鍵は配布済であるとする．このとき，基本方式は Protocol 1 で表される．

Protocol 1 基本名寄せプロトコル

Input: (sk, τ_0) for $\mathcal{P}_0$, (pk, τ_i) for $\mathcal{P}_i$ where $1 \leq i < n$
Output: τ^* for $\mathcal{P}_0$

each $\mathcal{P}_i$ for $0 \leq i < n$ **do**
  $c_{\ell,a} = \mathsf{Enc}(pk, \tau(\ell)^{(a)})$ for $\ell \in \mathbb{R}$ and $a \in \mathbb{A}_i$
  Send all $c_{\ell,a}$ to $\mathcal{P}_{n-1}$
$\mathcal{P}_{n-1}$ **do**
  Set $\rho_a = \frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}(|\mathbb{V}_a|-1)}$ with $\alpha = \left(\frac{k-1}{|\mathbb{R}|-1}\right)^{\frac{1}{|\mathbb{A}_0|}}$ for $a \in \mathbb{A}$
  for $\ell \in \mathbb{R}$ and $a \in \mathbb{A}_0$ **do**
    $\rho'_{\ell,a} \leftarrow (0, 1]$
    if $\rho_a \leq \rho'_{\ell,a}$ **then**
      $r \leftarrow \mathbb{V}_a$
      $c_{\ell,a} := \mathsf{Enc}(pk, r)$
    else
      $c_{\ell,a} \leftarrow \mathsf{ReRand}(pk, c_{\ell,a})$
  Choose a random perm. $\pi : \mathbb{R} \to \mathbb{R}$
  Send $c_{\pi(\ell),a}$ to $\mathcal{P}_0$ for $\ell \in \mathbb{R}$ and $a \in \mathbb{A}$
$\mathcal{P}_0$ **do**
  $v_{\ell,a} \leftarrow \mathsf{Dec}(sk, c_{\ell,a})$ for $\ell \in \mathbb{R}$ and $a \in \mathbb{A}$
  Set $\tau^* : \ell \mapsto \{v_{\ell,a}\}_{a \in \mathbb{V}}$ for $\ell \in \mathbb{R}$

3.1 基本方式の安全性

紙面の都合上，安全性の直観のみを与え，詳しい安全性の定義および証明は別の機会に譲る．

定理 3.1 $a \notin \mathbb{V}_0$ について f_{V_a} が一様であるような背景知識を持つ $\mathcal{P}_0$ に対して，*Protocol 1* で生成された τ^* は Pk-匿名性を満たす．

[**証明のスケッチ**] $\mathcal{P}_0$ に与えられるテーブル τ^* は，属性 $a \in \mathbb{A}_0$ に対し Pk-匿名化が為されて

いる．そのため，$\{\tau^*(\ell)^{(a)}\}_{\ell\in\mathbb{R},a\in\mathbb{A}_0}$ は $\mathcal{P}_0$ に対し Pk-匿名性を満たす．また，$V_a \notin \mathbb{V}_0$ が一様であるような背景知識であれば，属性 $a \notin \mathbb{A}_0$ は個人の識別に影響しないので，τ^* は $\mathcal{P}_0$ に対して Pk-匿名性を満たす．

定理 3.2 再乱数化可能な公開鍵暗号が *IND-CPA* 安全であるとき，$\{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{n-1}\}$ の任意の結託に対し，*Protocol 1* は結託していないパーティのテーブルを秘匿する．

[**証明のスケッチ**] $1 \leq i < n$ について $\mathcal{P}_i$ は再乱数化可能な公開鍵暗号の暗号文のみしか得られないことから明らか．

3.2 基本方式の効率

基本方式は非常にシンプルな構成であり，高速に動作可能である．再乱数化可能な公開鍵暗号として楕円 ElGamal 暗号を用い，パーティ数が $n = 2$ で，各パーティ1つずつ属性を持っているとき，暗号文サイズ（圧縮無しの楕円点 ×2）が 640bit，暗号化および再乱数化（楕円スカラー倍 ×2）に 1ms，復号（楕円スカラー倍と逆元演算 1 回ずつ）に 1ms かかるとしても，2 パーティの場合 1 万レコードの名寄せに計算時間が 40 秒，通信量が往復で 1.6MB と見積もれる．暗号文を圧縮した場合（楕円点の x 座標 ×2）は，元の座標に戻すために 0.5ms かかるとすると，計算時間が 50 秒，通信量が 0.4MB と見積もれる．さらに暗号化と復号，再乱数化は全て要素毎に独立な処理のため，コア数 m の CPU を用いた場合，計算時間は $1/m$ 倍となることが期待できる．具体的な実装実験は今後の課題である．

4 基本方式の拡張

基本方式では，属性がカテゴリ属性で，各テーブルのレコードの共通集合が既知である場合に，自身の持つ属性以外の属性についての知識がない攻撃者に対し，識別リスクを低減している．

本章では，上記以外の場合に，基本方式をどのように拡張すればよいかについて記述する．なお，これらの拡張は排反ではなく，組み合わせることも可能である．

4.1 数値属性を扱う場合

基本方式は属性値がカテゴリであることを前提としているが，身長などの数値を扱いたい場合は，Pk-匿名化として有界ノイズ付加が適切と考えらえる．そのような場合は，再乱数化可能な公開鍵暗号の代わりに，平文空間が固定小数点であるような準同型暗号（例えば [2,4,7,8]）を選ぶことにより，ラプラスノイズの加算が可能となり，有界ノイズ付加が利用できる．

再乱数化可能な公開鍵暗号として青野らの方式 [2] を利用した場合を例に，数値属性を扱えるよう基本方式を拡張する方法を示す．

数値属性を扱う際は，$\mathcal{P}_{n-1}$ は基本方式のステップ 5 から 12 を以下に変更する．

1. 属性の値域と式 2 から b を計算する
2. 確率密度関数 $f_{\mathrm{lap},b}$ に従い乱数 $\rho'_{\ell,a}$ を生成し，$\rho'_{\ell,a} \notin \mathbb{V}_a$ である場合は生成をやり直す．
3. $\rho'_{\ell,a}$ を適切な値にエンコード [1] したものを $\tilde{\rho}'_{\ell,a} \in \mathbb{Z}_p^{1\times\ell}$ とすると，$c_{\ell,a} = (c_{\ell,a,1}, c_{\ell,a,2})$ に対し $c_{\ell,a} = (c_{\ell,a,1}, c_{\ell,a,2}+\tilde{\rho}'_{\ell,a}) \in \mathbb{Z}_q^{1\times n} \times \mathbb{Z}_q^{1\times\ell}$ とする．

ここでの p, q, ℓ, n は公開鍵暗号の平文空間や暗号文空間を規定するパラメータであり，セキュリティパラメータに対して適切に選ぶ必要がある．

拡張を行った場合，加算されるノイズは元々の値によらず

$$\alpha_v = \int_{v_{\min}}^{v_{\max}} f_{\mathrm{lap},b}(x)dx$$

とした際の $\frac{1}{\alpha_v} f_{\mathrm{lap},b}(x)$ に従うものであり，式 4 とは異なるものの，拡張方式の安全性は基本方式と同等である．詳細な証明は別の機会とするが，これは，Pk-匿名化のノイズの大きさは，ある値 $u \in \mathbb{V}_0$ から別の値 $v \in \mathbb{V}_0$ に移る確率とそのままである確率の比の最大値/最小値で決定されており，元の値が何であっても，平文空間で剰余がとられる場合は上記の値が一定となるからである．

一方で，拡張方式の効率は自明ではない．拡張方式では，平文空間およびエンコードの方法により，剰余を取ることにより本来非常に遠い

[1] 暗号化の際に，実数はそのままの値ではなく適切な形にエンコードされているため，定数加算の場合もそのエンコードに合わせる必要がある．詳しいエンコード方法は文献 [2] の 4 章を参照のこと．

はずの値に遷移してしまう．これを防ぐためにはある程度大きい平文空間を取る必要があり，その具体的なパラメータとその際の暗号文サイズの見積りは今後の課題である．

4.2 全テーブルの共通レコード集合が未知の場合

基本方式では，全ての $0 \leq i < n$ について，τ_i の i 番目のレコードは全て同一人物のものであることを前提としている．これは，見方を変えれば，どのテーブルにも含まれるようなレコードの集合（共通レコード集合）が，全てのパーティに既知であることを意味している．

しかし，アプリケーションによっては，$\mathcal{P}_i$ が $\mathcal{P}_j$ の持つレコードを知らない場合，すなわち $0 \leq i < n$ について互いに異なる $\mathbb{R}_i$ が存在することも想定されうる．本節では，そのような場合の拡張方法を提案する．

まず，各 $\mathbb{R}_i$ を知られても良いのであれば，各パーティは互いにレコード集合を公開し，その共通集合を得れば良い．

各 $\mathbb{R}_i$ は公開したくないが，$\mathbb{R}_0$ は公開しても良い場合は，$\mathcal{P}_0$ は自身のレコード集合の暗号文を生成するのに加えて，$\mathbb{R}_0$ をブロードキャストする．$1 \leq i < n$ について $\mathcal{P}_i$ は，$\ell \in \mathbb{R}_i$ について，$\ell \in \mathbb{R}_0$ であれば $\tau_i(\ell)$ を，そうでなければ平文空間のうち特別な値 ι を暗号化して $\mathcal{P}_{n-1}$ に集める．

この拡張を行っても，安全性は基本方式と同等である．計算および通信効率は，共通レコード集合が既知の場合と比較し，暗号化の回数と送信する暗号文の数が $\sum_{0 \leq i < n} \left(|\mathbb{R}_i| - \left| \bigcup_{0 \leq i < n} \mathbb{R}_i \right| \right)$ 程度増加する．

もし上の式が大きい場合は，基本方式を実行する前に Private Set Intersection (PSI) と呼ばれる共通レコード集合を計算するプロトコルを行うことで，全体の計算時間を削減できる可能性がある．文献 [10] によれば，2 者間，1Gbp-sLAN 環境において 2^{18} 件のレコードに対し 13 秒程度で実行可能であるため，計算量・通信量の増加が大きい場合は PSI を用いた方が効率的となる．PSI を用いた場合は，能動的な攻撃者に対して安全な PSI を用いれば，基本方式と同等の安全性を持つ．

4.3 $\mathcal{P}_0$ がより強い背景知識を持つ場合

基本方式では，攻撃者 $\mathcal{P}_0$ の背景知識は全ての $a \notin \mathbb{A}_0$ について f_{V_a} が一様ランダムであることを仮定している．これはすなわち，$\mathcal{P}_0$ は自身のテーブルに含まれない属性について，何も背景知識を持っていないことを意味する．しかし $\mathbb{A}_0$ に学歴，$\mathbb{A}_1$ に年収が含まれていれば，$\mathcal{P}_0$ はある程度年収を推測可能できるなど，背景知識を持つと見なす場合がある．

このような攻撃者の背景知識を想定する場合は，全ての属性について Pk-匿名化を行えばよい．PRAM は属性ごとに独立して匿名化を行うため比較的容易に実現可能である．具体的には，以下の様に基本方式を変更することで，Pk-匿名性を保証することができる．

基本方式のステップ 1 から 5 を以下に変更する．$\mathcal{P}_0$ の動作は変更なしとする．

1. $1 \leq i < n$ について $\mathcal{P}_i$ は，$\alpha = \left(\frac{k-1}{|\mathbb{R}|-1} \right)^{\frac{1}{|\mathbb{A}|}}$ としたとき，$a \in \mathbb{A}_i$ について
$$\rho_a = \frac{1 - \sqrt{\alpha}}{1 + \sqrt{\alpha}(|\mathbb{V}_a| - 1)}$$
を計算する．
2. $\ell \in \mathbb{R}$ および $a \in \mathbb{A}_i$ について，乱数 $\rho'_{\ell,a} \leftarrow (0, 1]$ を生成し，$\rho_a \leq \rho'_{\ell,a}$ ならば $c_{\ell,a} := \mathsf{Enc}(pk, r)$，そうでなければ $c_{\ell,a} \leftarrow \mathsf{Enc}(pk, \tau(\ell)^{(a)})$ とする．
3. 全ての $c_{\ell,a}$ を $\mathcal{P}_{n-1}$ に送る．
4. $\mathcal{P}_{n-1}$ はステップ 5 において $\alpha = \left(\frac{k-1}{|\mathbb{R}|-1} \right)^{\frac{1}{|\mathbb{A}|}}$ として計算する．

このとき，以下が成り立つ．

定理 4.1 任意の f_T を背景知識として持つ能動的な $\mathcal{P}_0$ に対して，*Protocol 1* で生成された τ^* は Pk-匿名性を満たす．

証明は省略する．

4.4 属性推定リスクを考慮する場合

基本方式で考慮しているのは個人の識別リスクのみであるが，他の代表的なプライバシーリスクとして，属性推定リスクが知られている．属性推定リスクとは，ある個人の属性値が攻撃

者に暴露されるリスクであり，識別リスクとは区別して考えられている [5,12].

識別リスクに加えて属性推定リスクを考慮する場合は，$P(c,\ell)$-多様性を利用することで，属性推定リスクを低減することができる．詳細な説明は [13] を参照されたいが，攻撃者が推定しようとしている属性をターゲット属性 $\mathbb{V}^*$ と呼ぶとき，全てのターゲット属性以外の属性値とターゲット属性単独の分布が既知であり，攻撃対象の個人のターゲット属性の値が，$\ell-2$ 個以内の属性値について確率が ϵ より低いという背景知識持っている攻撃者に対して，攻撃対象の個人のターゲット属性値の推定確率について

$$\Pr[S=s \mid T'=\tau'] \leq \frac{1}{1+\frac{\epsilon}{c(1-\epsilon(|\mathbb{V}^*|-\ell-1))}}$$

が成り立つとき，$P(c,\ell)$-多様性を持つという．維持置換撹乱において，Ω_Σ をターゲット属性の母集団の分布とし，Ω_Σ が $P(c,\ell)$-多様性を満たしているとき，維持置換撹乱において

$$\rho = \frac{c'-c}{c'+(m-1)c} \tag{5}$$

とすれば $P(c',\ell)$-多様性を満たす．

以上から，識別リスクと属性推定リスク両方を満たすように拡張する場合は，Pk-匿名性と $P(c,\ell)$-多様性を満たすように維持置換撹乱のパラメータ ρ を設定すればよい．ターゲット属性は $\mathbb{A}_0$ に含まれていない属性となるので，前節と同様に $1 \leq i < n$ について $\mathcal{P}_i$ は最初に暗号化する前に式 5 に従い撹乱すればよい．

5 まとめ

本論文では，プライバシー保護とテーブルの名寄せを両立するためのプロトコルを提案した．まず基本方式として，複数のパーティが，同じレコード集合かつ違う属性であるようなテーブルを持つときに，暗号化して名寄せを行いつつ，名寄せテーブルから個人が識別できないように，暗号化したまま Pk-匿名化を行う方法を示した．次に，基本方式を拡張し，数値属性を扱う場合，全テーブルの共通レコード集合が未知の場合，より強い背景知識を持つ攻撃者を想定する場合，属性推定リスクを考慮する場合について，拡張方法の構成法を示した．

参考文献

[1] R. Agrawal, R. Srikant, and D. Thomas. Privacy preserving OLAP. In *SIGMOD Conference*, pp. 251–262, 2005.

[2] Y. Aono, T. Hayashi, L. T. Phong, and L. Wang. Fast and secure linear regression and biometric authentication with security update. *IACR Cryptology ePrint Archive*, 2015:692, 2015.

[3] S. Goldwasser and S. Micali. Probabilistic encryption. *J. Comput. Syst. Sci.*, 28(2):270–299, 1984.

[4] T. Graepel, K. E. Lauter, and M. Naehrig. ML confidential: Machine learning on encrypted data. In *Information Security and Cryptology - ICISC 2012*, pp. 1–21, 2012.

[5] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. S. Nordholt, G. Seri, and P.-P. D. Wolf. *Statistical Disclosure Control*. Wiley Series in Survey Methodology, 2012.

[6] D. Ikarashi, R. Kikuchi, K. Chida, and K. Takahashi. *k*-anonymous microdata release via post randomisation method. In *IWSEC 2015 (to appear)*, 2015.

[7] M. Naehrig, K. E. Lauter, and V. Vaikuntanathan. Can homomorphic encryption be practical? In *Proceedings of the 3rd ACM Cloud Computing Security Workshop, CCSW*, pp. 113–124, 2011.

[8] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft. Privacy-preserving ridge regression on hundreds of millions of records. In *2013 IEEE Symposium on Security and Privacy, SP 2013*, pp. 334–348, 2013.

[9] K. Peter, W. Leon, and G. Jose. *PRAM: a Method for Disclosure Limitation of Microdata*. Research paper. CBS, 1997.

[10] B. Pinkas, T. Schneider, and M. Zohner. Faster private set intersection based on OT extension. In *Proceedings of the 23rd USENIX Security Symposium*, pp. 797–812, 2014.

[11] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[12] M. Templ, B. Meindl, and A. Kowarik. Introduction to statistical disclosure control (SDC), 2015.

[13] 五十嵐, 千田, 高橋. $P\ell$-多様性: 属性推定に対する再構築法のプライバシーの定量化. In *CSS*, 2010.

[14] 五十嵐, 千田, 高橋. 数値属性における, k-匿名性を満たすランダム化手法. In *CSS*, 2011.

[15] 五十嵐, 長谷川, 納, 菊池, 千田. 数値属性に適用可能な, ランダム化により k-匿名性を保証するプライバシー保護クロス集計. In *CSS*, 2012.