情報処理学会論文誌 Vol.58 No.9 1483–1500 (Sep. 2017)

匿名化個票開示への差分プライバシの適用

寺田 雅之[1,a)] 山口 高康[1] 本郷 節之[2]

受付日 2016年11月25日, 採録日 2017年6月6日

概要：個人に関わるデータの公開・提供にあたっては，開示されたデータから個人のプライバシが漏洩することを防ぐ必要がある．本稿では，それらのデータを匿名化された個票データ（microdata）である匿名化個票として開示する際において，強い数学的な安全性が示されている差分プライバシ基準を充足しつつ，データの有用性を高く保つ方式を提案する．提案方式は，属性空間を完全に分割した分割表と個票データは多重集合の意味で等価であることに着目し，完全分割表に対する Laplace メカニズムの適用と，ベクトル空間における最近傍探索に基づく非負制約，整数制約，総数制約の充足により匿名化個票を得る．また，提案方式の有用性を評価するために，売上履歴を模したロングテイル性を持つ擬似的な個票データを用い，L_2 距離，KS-距離を指標として従来方式と定量的に比較評価し，従来方式と比べて元データの性質をより強く保持する匿名化個票が得られることを示す．

キーワード：差分プライバシ，匿名化個票，Laplace メカニズム，最近傍探索

On Releasing Anonymized Microdata with Differential Privacy

Masayuki Terada[1,a)] Takayasu Yamaguchi[1] Sadayuki Hongo[2]

Received: November 25, 2016, Accepted: June 6, 2017

Abstract: It has become strongly required to protect privacy when releasing a dataset related to individuals. This paper proposes a novel method for generating anonymized microdata under differential privacy, which provides an ad omnia privacy guarantee based on solid mathematical foundations. Utilizing the equivalence relation of microdata and its completely-divided contingency table, the proposed method generates differentially private microdata by applying the Laplace mechanism to the contingency table equivalent to the original microdata, which is followed by a new efficient method to find the optimal contingency table that has equivalent microdata. The evaluation results in terms of L_2 distance and Kolmogorov Smirnov distance show that the output of the proposed method retains the nature of its original data much better than those of the previous methods for generating microdata with differential privacy.

Keywords: differential privacy, anonymization, microdata, Laplace mechanism, nearest neighbor search

1. はじめに

個人に関わるデータの利活用にあたっては，プライバシ保護への十分な配慮が必要となる．本稿では，それらのデータを匿名化された個票データ（匿名化個票）の形態で開示する際において，開示データの有用性をなるべく高く保ちながら，強い数学的な安全性が示されている差分プライバシ基準を充足する方式について考察する．

あるシステムで収集・作成されたデータベースを，第三者に公開もしくは提供することを考える．これをデータの開示という．データ開示の形態は，大きく「個票データ（microdata）」の開示と「集計データ（aggregated data）」の開示に大別される．個票データとは，個人を単位とした情報（レコードと呼ばれる）の集合として開示されるデータである．集計データとは，それらのデータをなんらかの条件で集約し，その個数を数えた数値データ*1（セルと呼ばれる）の集合などとして開示されるデータである [7], [46].

[1] 株式会社 NTT ドコモ先進技術研究所
Research Laboratories, NTT DOCOMO, Inc., Yokosuka, Kanagawa 239–8536, Japan
[2] 北海道科学大学工学部
Faculty of Engineering, Hokkaido University of Science, Sapporo, Hokkaido 006–8585, Japan
[a)] teradam@nttdocomo.com

*1 一般には，データの平均や総計など，他の統計量の場合もあるが，本稿では個数であるとする（詳細は 2 章で定義する）．

これらのデータの開示は，開示されるデータに含まれる個々人のプライバシを保護しつつ行われなければならない．しかし，個票データの開示においては，その有用性を保ちながらプライバシを十分に保護することは容易ではない [12]．これは，個人を単位として表現されるデータから個人のプライバシが暴露されることを防ぐという，そもそも原理的な困難が含まれるためである．

データ開示にあたってプライバシを保護するための指標としては，k-匿名性（k-anonymity）基準 [37] や差分プライバシ（differential privacy）基準 [8] などがあげられる．特に差分プライバシは，その安全性が数学的に保証可能であるという特長を持ち，米国などにおけるプライバシ保護研究の分野で大きく注目を集めている．

しかし，差分プライバシは集計データのプライバシ保護に有用である一方で，個票データへの効果的な適用法は明らかではない．たとえば，差分プライバシを満たすために用いられる最も一般的な手法である Laplace メカニズム [10], [13] は，集計データの各セル値に対して Laplace ノイズを加算するのみという簡単な処理によって，加算したノイズ強度に従った強度の差分プライバシを保証できるが，これをそのまま個票データのプライバシ保護に適用することはできない．

差分プライバシが保証された匿名化個票を作成する手段として，PRAM（post randomization）[20] や合成データ（synthetic data）生成 [15], [29], [36] など，統計分野で用いられる統計的開示制御手法（Statistical Disclosure Control, SDC）[23] の一部（およびその派生）は差分プライバシを満たすことが示されている [1], [21], [24], [31], [32], [40]．しかし，実際にはこれらの手法で得られるプライバシ保護とデータの有用性のトレードオフの関係はあまり芳しいものではない．いい換えると，十分な強度でプライバシを保護しようとすると，著しいデータの有用性の低下を招き，実用的なプライバシ保護の手段とはいいにくくなる[*2]．

本稿の目的は，上記の従来方式の課題を解決し，差分プライバシが保証された有用性が高い匿名化個票を作成する手段を与えることである．そこで，上記の従来方式の課題について具体的な数値を用いて定量的に議論するとともに，差分プライバシを備えた匿名化個票を作成する新しい方式として，属性を完全分割した高次元集計データに対する Laplace メカニズムの適用と多次元空間における最近傍探索に基づく手法を提案する．

提案方式は，個票データの各属性の値域の直積をセル集合とした集計データ（完全分割表）は元の個票データと数学的に等価であることに着目し，個票データに対してそのままプライバシ保護のための処理を施すのではなく，まず完全分割表を作成したうえで，そこに Laplace メカニズムを適用することにより差分プライバシを満たす匿名化集計データを作成する．しかし，そのように作成された匿名化集計データは，そのままでは個票データに対応づけられない（個票データに戻すことができない）．そこで，ある集計データを元と同じサイズの個票データに「戻す」ためには，集計データが非負のセル値のみから構成されること（非負制約），各セル値が整数であること（整数制約），セル値の合計が個票データのレコード数と等しいこと（総数制約），の 3 つの制約を充足すればよいことを示し，完全分割表に含まれるセルの値域の直積から構成されるベクトル空間において，上記の制約を充足する匿名化集計データからの最近傍点を探索することにより，差分プライバシを満たす匿名化個票を生成する．

以下，2 章では準備として個票データ，集計データ，完全分割表など提案方式で扱うデータ形式に対して数学的な定義を与えるとともに，差分プライバシの定義とその実現手法（メカニズム）の 1 つである Laplace メカニズムについて簡単に説明する．3 章では差分プライバシを備えた匿名化個票を作成するための従来技術を紹介するとともに，その問題点を具体的な数値に基づき議論する．4 章では，匿名化個票を得るために集計データが満たすべき条件を示すとともに，Laplace ノイズが加えられた（そのままでは等価な個票データを持たない）匿名化集計データから，前記の条件を制約とした最近傍探索により差分プライバシを満たす匿名化個票を得る方式を提案する．5 章では提案方式の有用性を従来方式との比較評価を通じて検証し，6 章で評価結果に対する考察を加える．最後に 7 章において，制約条件に基づく事後処理による差分プライバシの有用性向上という観点で関連研究との比較検討を行う．

2. 準備

本章では，議論の準備として，個票データおよび集計データの定義を与えるとともに，差分プライバシ [8] の定義と，差分プライバシを実現するための代表的な手段として知られている Laplace メカニズムについて説明する．

2.1 個票データ

個票データとは，それぞれが個人に対応づけられた 1 つ以上のレコードから構成されたデータベースであり，各レコードは 1 つ以上の属性値を持つ．これは，各レコードを元とした多重集合（multiset）[*3] として定義される．

ある個人 i に対応づけられたレコードを x_i とする．x_i は，その個人に関する何らかの情報を表す，d 個の属性値 x_{ij}（$1 \leq j \leq d$）の組（順序対）から構成される．任意のレコードにおける，j 番目の属性値は集合 A_j（$1 \leq j \leq d$）に属する（$\forall i, x_{ij} \in A_j$）．ここで，$A_j$ を属性と呼び，すべて

[*2] 詳細は 3 章で議論する．

[*3] 同値を持つ元が重複して存在することを許す（順序なし）集合．

の属性の直積 $A = A_1 \times A_2 \times \cdots \times A_d$ を属性空間と呼ぶ．

このとき，n 個のレコードから構成される個票データ D は，属性空間 A を台集合（underlying set）[*4]とする，以下の n 元の多重集合として表される．

$$\begin{aligned} &D = \{\{x_1, x_2, \ldots, x_n\}\}, \\ &x_i = (x_{i1}, x_{i2}, \ldots, x_{id}) \quad (\in A), \\ &A = A_1 \times A_2 \times \cdots \times A_d. \end{aligned} \tag{1}$$

ここで，D を（集合や順序対ではなく）多重集合として定義する理由は，個票データには同一の属性値の組合せを持つレコードが複数存在しうること，レコードの並び順のみが異なる個票データは本質的に等価であることによる．

$x_i \in A$ より，各レコードがとりうる属性値の組合せの数は，属性空間 A の濃度（cardinality）$|A|$ に等しい．これは，各属性 A_j の濃度 $|A_j|$ の総積である．すなわち，

$$|A| = \prod_{j=1}^{d} |A_j|. \tag{2}$$

一般には，属性 A_j は有限集合（カテゴリ属性や上限/下限を持った自然数をとる数値属性など）もしくは無限集合（実数を値としてとる数値属性など）のいずれもとりうる．本稿では，以降の議論において A_j は有限集合であるとする．すなわち，（実数値をとる）数値属性は，階級化などの処理により，有限なカテゴリとして表現されているものとする．このとき，$|A_j|$ は（有限な）自然数となることから，$|A|$ もまた自然数となる．

2.2 集計データ

集計データは，個票データ D において，ある定められた条件を満たす属性値（もしくは属性値の組合せ）を持つレコードの個数を数えあげた値の集合である．

A を D の属性空間とするとき，A の部分空間 C_k（$\subseteq A$）に属するレコードの個数を $Count(D, C_k) = |\{x \in D \mid x \in C_k\}|$ とする．これを計数問合せ（count query）と呼ぶ．このとき，任意の C_k からなる順序対である集計条件 $C = (C_1, C_2, \ldots, C_p)$ に対して，集計データ V は，C の各元に対応する計数問合せ $Count(D, C_k)$ からなる順序対として与えられる．すなわち，

$$V = (v_1, v_2, \ldots, v_p), \quad v_k = Count(D, C_k). \tag{3}$$

集計データ V の作成において，一般的には各属性の値域 A_j の互いに素な部分集合の直積が集計条件 C として用いられる．このとき，集計データは分割表（contingency table）と呼ばれる．分割表の各要素 v_k を，セル（cell）もしくはセル値と呼ぶ．

[*4] 多重集合の元が属する集合．

2.3 完全分割表

伊藤 [43] は，集計データは集計条件を細かくしていくと個票データとの意味的な違いが実質的になくなっていくことを指摘している．実際に，ある条件において集計データと個票データは相互に変換可能，すなわち数学的に等価になる．以下において，多重集合の概念を用いて集計データと個票データが等価となる条件を与える．

$A = \{a_1, a_2, \ldots, a_p\}$ を属性空間とする集計データ V において，$C_k = \{a_k\}$（$1 \le k \le p$），すなわち集計条件の各要素 C_k は，属性空間 A のいずれかの元のみを含む集合であり，あらかじめ定められた規則（a_k の辞書順など）に従って重複なくすべての A の元に対応づけられる（全単射 $f : A \to \{C_k\}$ が存在する）とする．この集計データは分割表であり，これ以上に集計条件を細かくした分割表は作れないことから，本稿ではこれを完全分割表[*5]と呼ぶ．完全分割表 V のセル数 p は，A の濃度 $|A|$ と等しい（$p = |C| = |A| = \prod_{j=1}^{d} |A_j|$）．たとえば，$A = A_1 \times A_2$，$A_1 = \{\text{男性}, \text{女性}\}$，$A_2 = \{\text{大人}, \text{子供}\}$ という属性空間を持つデータにおいて，集計条件 $C = ((\text{男性}, \text{大人}), (\text{男性}, \text{子供}), (\text{女性}, \text{大人}), (\text{女性}, \text{子供}))$ に対応する集計データ V は完全分割表であり，そのセル数 p は $|A_1| \times |A_2| = 4$ となる．

完全分割表 V における各要素 v_k は，（多重集合である）個票データ D における，（その台集合である）属性空間 A の元 a_k の多重度（multiplicity）$m_D(a_k)$ に他ならない．たとえば $a_k = (\text{男性}, \text{大人})$，$v_k = 3$ のとき，これは個票データ D が $(\text{男性}, \text{大人})$ というレコードを 3 個含むことを意味する．また，$v_k = 0$ のときは，D はそのようなレコードを含まないことになる．**図 1** に完全分割表の例を示す．ここで示された 2 つの個票と完全分割表は多重集合として互いに等価である．

任意の多重集合は，台集合と，その各元の多重度により一意に定義されるため，(V, A, f) の組が与えられれば D は一意に定まる．また，その逆に (D, A, f) の組が与えられれば V が定まることは V の定義から明らかである．すな

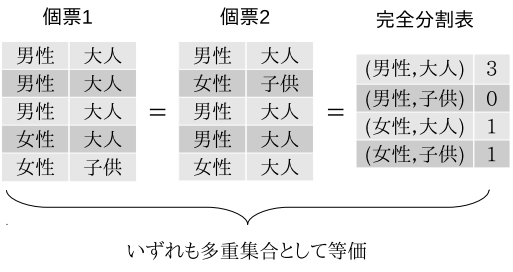

図 1 完全分割表の例

Fig. 1 An example of a completely-divided contingency table.

[*5] 文献 [43] では，個票データと等しくなるまで集計条件を細かくした集計データを「超高次元クロス集計表」と呼んでいるが，本稿では曖昧さを避けるためこの呼称を用いる．

わち，(A, f) が定められた条件において D と V は等価であり，以下の定理が成立する．

定理 1. 属性空間 A を持つ個票データ D から完全分割表 V を生成する写像を $\mathcal{F}_A$ とする（$V = \mathcal{F}_A(D)$）．このとき，$\mathcal{F}_A$ は $D = \mathcal{F}_A^{-1}(V)$ となる逆写像 $\mathcal{F}_A^{-1}$ を持つ．

2.4 差分プライバシ

差分プライバシ [8], [9] は，識別不能性に基づくプライバシ基準の一種であり，パラメータ ϵ を用いて以下のように定義される．

定義 1. 任意の隣接した[*6]データベース D_1 および D_2（$D_1, D_2 \in \mathcal{D}$）に対し，ランダム化関数（randomized function）$\mathcal{K} : \mathcal{D} \to \mathcal{R}$ が下式を満たすとき，$\mathcal{K}$ は ϵ-差分プライバシを満たす．ただし，ここで S は $\mathcal{K}$ の出力空間 $\mathcal{R}$ の任意の部分空間である（$S \subseteq \mathcal{R}$）．

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{K}(D_2) \in S]. \tag{4}$$

このとき，上記のランダム化関数 $\mathcal{K}$ は「メカニズム（mechanism）」と呼ばれる．

差分プライバシの特徴として，その安全性定義がデータの性質や攻撃者の能力（攻撃手段や攻撃者の背景知識）に依存しないことがあげられる．すなわち，データベースに異常値が混入していても安全性が損なわれることがなく，また任意の背景知識を持つ攻撃者や未知の攻撃に対して安全である．これは，差分プライバシ基準を正しく満たしたデータは，データ作成時には未知であった新たな攻撃手法が発見されたり，もしくは想定外の背景知識を持つ攻撃者が現れたりしたとしても，その安全性が損なわれないということを意味する．

2.5 隣接の定義について

定義 1 において，隣接するデータベースという概念が現れる．差分プライバシの文脈において，データベース D と D' が隣接するとは，1 レコードの削除もしくは追加（$D' = D \backslash x$ or $D \uplus y$）を意味する場合 [8], [35] と，1 レコードの置換（$D' = D \backslash x \uplus y$）を意味する場合 [11], [33] がある[*7]．これらを区別して議論するとき [25], [38]，前者は拘束なし隣接（unbounded neighbors），後者は拘束つき隣接（bounded neighbors）と呼ばれる．

拘束なし隣接と拘束つき隣接の違いは，攻撃者のゴールと最強の攻撃者が持つ背景知識の違いをモデル化したものと考えることができる．すなわち，D に含まれる（攻撃対象以外の）すべてのレコードを攻撃者は知るものとして，

- さらに攻撃者は攻撃対象が持つ属性を完全に知りえるが，データベースのサイズ $|D|$ は未知であるとし，攻撃対象が D に含まれるか否かを決定することを攻撃者のゴールと置く[*8]（拘束なし隣接）か，
- 攻撃対象が D に含まれていることと $|D|$ のいずれも攻撃者は知りえるが，攻撃対象の属性の一部は未知であるとし，これを推定することを攻撃者のゴールと置く（拘束つき隣接）か，

の違いをそれぞれ表すと考えることができる．$|D|$（$= n$）が攻撃者にとって既知か未知かの違いに着目し，拘束つき隣接を「公開 n 条件（public n regime）」，拘束なし隣接を「秘匿 n 条件（private n regime）」と呼ぶこともある [5]．

拘束つき隣接と拘束なし隣接は互いにシミュレーション可能だが，得られる安全性が異なる．拘束つき隣接の条件で ϵ-差分プライバシを満たす方式は，拘束なし隣接の条件でも ϵ-差分プライバシを満たすが，逆の場合は ϵ への加算（安全性の低下）が発生しうる [5]．具体的な加算の量は問合せやメカニズムに依存する．

これは，問合せの種類によっては，隣接の定義の違いが差分プライバシを保証するためのノイズの大きさに影響しうることを意味する．たとえば Laplace メカニズムによる計数問合せ（count query）では，いずれの定義を用いてもノイズの大きさは変わらないが，分割表の作成（histogram query [11] に相当する）においては，拘束つき隣接の条件下では拘束なし隣接の条件下に比べて 2 倍のスケールのノイズが必要となる（詳細は次節で説明する）．

本稿では，匿名化個票の作成にあたっては個票データのレコード数を保存するものとし，これは攻撃者にとって既知であるものとする．すなわち，n が公開された，拘束つき隣接の条件下にあるものとして議論を進める．

2.6 Laplace メカニズム

差分プライバシを実現するためには，定義 1 を満たすメカニズム $\mathcal{K}$ が必要となる．代表的なメカニズムとしては Laplace メカニズムがあげられる．なお，計数問合せに対しては Laplace メカニズムは最適ではない（より分散が小さいノイズで同じ安全性強度の差分プライバシを与えるメカニズムが存在する）ことが知られており，具体的には幾何メカニズム（geometric mechanism）[19] や staircase メカニズム [16], [17] が最適となることが示されている．ただし，ϵ が小さい（安全性が高い）条件下ではほぼ差がないとされる [17] ため，本稿では主に Laplace メカニズムに基づいて議論する．

Laplace メカニズムは，0 を平均とした Laplace 分布に従う乱数である Laplace ノイズを問合せ結果に加算することにより実現される．Laplace 分布の確率密度 $\ell(x)$ は，平均 μ とスケール λ を用いて下式で与えられる．

[*6] 「隣接」の厳密な定義は次節で議論する．

[*7] x, y（$x \neq y$，$x \subseteq D$）は D と共通の台集合を持つ濃度 1 の多重集合であり，$\uplus$ は多重集合の直和（multiset sum）を表す．

[*8] 攻撃対象の属性と $|D|$ のいずれも既知だと攻撃が自動的に成功する（守るべきプライバシが存在しなくなる）ことに注意．

$$\ell(x;\mu,\lambda)=\frac{1}{2\lambda}e^{(-|x-\mu|/\lambda)}. \quad (5)$$

以降，平均 0，スケール λ の Laplace 分布に従って発生させた Laplace ノイズを Lap(λ) とし，q 個の互いに独立した Lap(λ) からなるベクトル列を Lap(λ)^q と記載する．

Laplace メカニズムにおけるノイズのスケール λ は，定義 1 における安全性パラメータ ϵ と，問合せの種類ごとに定まる感度（sensitivity）によって与えられる．具体的には，S_f を問合せ $f:\mathcal{D}\to\mathbb{R}^q$ の感度としたとき，f に対応するメカニズム $\mathcal{K}_f$ は下式で定義される．

$$\mathcal{K}_f(X)=f(X)+\mathrm{Lap}(S_f/\epsilon)^q, \quad (6)$$

$$S_f=\max_{D_1,D_2}|f(D_1)-f(D_2)|_1. \quad (7)$$

ここで，D_1（$\in\mathcal{D}$）および D_2（$\in\mathcal{D}$）は任意の隣接したデータベース（定義 1 参照）のペアである．

Laplace メカニズムを用いることにより，差分プライバシを満たす分割表 V^*（$|V^*|=p$）を簡単に作成することができるが，前節で議論したとおり隣接の定義の違いにより必要となるノイズの大きさが異なることに注意が必要となる．

拘束なし隣接の条件下では，1 レコードの違いは分割表における 1 セルの（計数問合せ結果の）違いにしか相当しない．計数問合せの感度 S_{count} は 1 であることから，差分プライバシの並列合成則（parallel composition theorem）[34] により，

$$V^*=V+\mathrm{Lap}(1/\epsilon)^p \quad (8)$$

により ϵ-差分プライバシを満たす集計データ V^* を得ることができる．

その一方，拘束つき条件下では，1 レコードの違いは分割表における 2 つのセルの違いを生む [9], [11]．したがって分割表全体としての感度は 2 となることから，

$$V^*=V+\mathrm{Lap}(2/\epsilon)^p \quad (9)$$

とすることが必要となる．

拘束つき条件下において分割表（完全分割表を含む）の感度が 2 となることについて，図 1 の例を用いて簡単に説明する．たとえば，同図の個票において，ある「男性，大人」のレコードが「女性，子供」に書き換えられたとする．書き換え前の個票を D，書き換え後の個票を D' とおくと，D と D' との間には 1 レコードの差（置換）しかないため，これら 2 つの個票は（拘束つき条件下において）隣接する．このとき，対応する完全分割表は (男性, 大人) の値が 1 減じられ，(女性, 子供) の値が 1 増加することになるため，$V=(3,0,1,1)\to V'=(2,0,1,2)$ に変化する．これら 2 つの分割表 V，V' の距離（1 次ノルム）は，

$$|V-V'|_1=|(-1,0,0,+1)|_1=2 \quad (10)$$

となる．書き換え対象のレコードや書き換え後の値の組合せを任意に変更しても，また元々の D の内容が異なっていたとしても，これは同様に成立する．すなわち，隣接する 2 つの分割表の距離は（拘束つき条件下において）いずれも 2 を超えないことから，分割表の感度は 2 として与えられる．

3. 従来技術と課題

前述のとおり，集計データに対しては，Laplace メカニズムを用いることにより差分プライバシを容易に保証することができ，その有用性を向上させるための改善方式も近年さかんに議論されている [2], [18], [28], [39], [41], [42], [44]．しかし，Laplace メカニズムの適用には「問合せ」と，それにより定まる「感度」が定義できることが必要となるが，個票データにはそれらは存在せず，直接 Laplace メカニズムを適用することはできない．これは幾何メカニズムや Staircase メカニズムでも同様である．

差分プライバシを保証した匿名化個票を作成するための従来技術としては，大域的再符号化や，PRAM，合成データ（synthetic data）生成などの他のプライバシ保護手法を適用し，それが差分プライバシを保証することを証明する手法があげられる．以下，それぞれについて説明する．

3.1 大域的再符号化に基づく手法

Li ら [27] は，個票データからのランダムサンプリング結果に対して大域的再符号化と（k を閾値とした）セル秘匿 [23] によって k-匿名性を実現することにより，差分プライバシを満たす匿名化個票を作成する手法を示している．ここで，k-匿名性の実現手段として（データの内容に依存しない）大域的再符号化を用いる必要があることに留意が必要である．一般的な k-匿名化手法では，有用性の劣化を抑制するために，データの内容に応じて適応的に k 個のレコードを集約する（clustering and local recoding，CLR）方法がとられるが，CLR は差分プライバシを満たすことができないとされる．そのため，この手法では一般的な k-匿名化手法と比較して有用性が低くなる．また，大域的再符号化をともなうことにより，出力される匿名化個票の属性空間 A' は，元の属性空間 A に対し，より「丸められた」ものとなる（$|A'|<|A|$）．

3.2 PRAM に基づく手法

Dwork ら [13] は，コイントスの例を用いて，あるデータを開示する際に「ある確率でランダムな値を答える」ことにより差分プライバシが得られるとしている．この手法はランダム化回答（randomized response）と呼ばれ，Google 社の RAPPOR [14] において利用者からの情報収集の際のプライバシ保護手段として実用化されている．

ランダム化回答は個々の問合せに対する回答においてプライバシを確保するための手段であるが，これが差分プラ

イバシを満たすことは，類似した考えに基づく統計的開示制御手法の 1 つである PRAM（Post Randomization）[20] が差分プライバシのメカニズムとして利用可能であることを示唆している．

PRAM とは，個票データの各レコードに含まれる属性値を，ある定められた確率（遷移確率）に従ってランダムに書き換えることによってプライバシを保護する手法である．特に，この遷移確率が属性ごとに一様であるもの，すなわち，それぞれの属性値 x_{ij} について確率 ρ_j で属性値を維持し，確率 $1-\rho_j$ で属性値を $x'_{ij} \in A_j$ へと一様ランダムに置換するものを維持置換攪乱（retention-replacement perturbation）と呼ぶ．ここで ρ_j は維持確率（retention probability）と呼ばれ，加工後のデータの精度と安全性とのトレードオフの関係を定めるパラメータとなる．

Lin ら [31] と Ikarashi ら [24] は，いずれも PRAM が差分プライバシを満たすことを定量的に示している．文献 [24] では，維持置換攪乱における維持確率 ρ_j と差分プライバシの安全性 ϵ の関係を直接的に計算可能とする式を与えている．具体的には，d 次の属性を持ち n レコードからなる個票データに対し，属性 A_j に対応する維持確率を ρ_j とする維持置換攪乱は，以下の ϵ-差分プライバシを与える．

$$\epsilon = \sum_{j=1}^{d} \ln \frac{1 + (|A_j| - 1)\rho_j}{1 - \rho_j}. \tag{11}$$

3.3 合成データに基づく手法

公的統計の分野において，「誰もが自由に使えるような」高い安全性を持つような匿名化個票を作成する手段として，合成データ（synthetic data）に基づく手法が注目を集めている．合成データとは，元となる個票データの統計的性質を保つ分布から標本を抽出することにより作成（synthesize）されるデータであり，Liew ら [29] により提案された．Rubin [36] らにより公的統計分野への適用が議論されて以降，海外の公的統計では主に一般公開可能な匿名データ（Public Use File，PUF）を作成する手段として数多く利用されている [45]．

合成データの作成手法には，標本を抽出する際に用いる分布の作成方法により様々な派生があり，それらの一部は差分プライバシを備えることが示されている．これらは大別すると，平滑化に基づく手法 [32], [40] と攪乱に基づく手法 [1], [21], [40] に分けられる．

Machanavajjhala ら [32] [*9]は，個票データの頻度分布に基づいて平滑化されたディリクレ分布を作成し，これを事前分布とする多項サンプリングにより差分プライバシを持つ合成データが生成できることを示している．ここで，頻度分布として完全分割表を用いれば，得られる合成データは元の個票データと同一の属性空間を持つ匿名化個票となる．具体的には，個票データから得られる完全分割表を V ($|V| = p$)，平滑化パラメータを $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ とするとき，ディリクレ分布 $Dir(\alpha + V)$ からの確率ベクトルの抽出と，抽出された確率ベクトルに基づく A からのサンプリングを繰り返すことにより匿名化個票を得る．この手法により作成された匿名化個票の安全性は $\min \alpha$ とサンプリング回数（作成する匿名化個票のレコード数）k に依存し，

$$\forall \alpha_j, \quad \alpha_j \geq \frac{k}{e^{\epsilon} - 1} \tag{12}$$

のとき，得られた合成データは ϵ-差分プライバシを満たす．

また，Wasserman ら [40] は，ディリクレ事前分布を用いることなく，平滑化された頻度分布を正規化することにより得られる確率ベクトルからの直接サンプリングによっても，差分プライバシを満たす合成データが生成できることを示している．この手法では，個票データ D の頻度分布 $f_i = v_i/n$ を平滑化パラメータ δ で平滑化したヒストグラム $f_i^* = (1-\delta)f_i + \delta$ を作成し，そのヒストグラムを正規化した確率ベクトルを生成分布としたサンプリングにより合成データを得る．ここで，δ は得られた合成データの安全性を制御するパラメータとして機能する．試行の回数を k とするとき，この合成データは，

$$\epsilon = k \ln \left(\frac{(1-\delta)p}{n\delta} + 1 \right) \tag{13}$$

の ϵ-差分プライバシを満たす．

Abowd ら [1] は，元データから得られるヒストグラムを Laplace メカニズムを用いて攪乱し，そこから得られる（攪乱された）確率ベクトルからのサンプリングにより合成データを得る方法を示している[*10]．なお，Laplace メカニズムにより攪乱されたヒストグラムは負の値をとりうることから，負値は 0 とみなすように補正する．すなわち，合成データの生成分布となる確率ベクトル $Q = (q_1, q_2, \ldots, q_p)$ は以下のように計算される．

$$v_i^* = \max(v_i + \mathrm{Lap}(2/\epsilon), 0), \tag{14}$$

$$q_i = v_i^* / \sum_s v_s^*. \tag{15}$$

なお，ここで Laplace メカニズムを適用する対象となるヒストグラムは分割表であり，集計区間に重複がない（差分プライバシの並列合成則が適用可能である）ことを前提としている．もし集計区間に重複がある場合は（並列合成則でなく）直列合成則の適用となり，重複の度合いに応じた強度の Laplace ノイズが必要となる．

Hardt ら [21] は，差分プライバシのメカニズムの一種である指数メカニズム（exponential mechanism）[33] を用

[*9] 文献 [1] においても平滑化されたディリクレ分布に基づく合成データが差分プライバシを満たしうることが示唆されているが，その条件などは定式化されていない．

[*10] Wasserman ら [40] も同様な手法を示しており，より厳密に定義と議論をしている．式 (14) は Wasserman らの定義に基づく．

いた乗算的重み更新法（multiplicative weights update）[3]により（攪乱的に）作成した確率ベクトルに基づいて，差分プライバシを満たす合成データを作成する手法であるMWEM（multiplicative weights exponential mechanism）を提案している．MWEMは，乗算的重み更新法における更新対象重みの選択を指数メカニズムにより行い，重みの更新量の攪乱をLaplaceメカニズムにより行うことを繰り返すことによって，差分プライバシを満たした確率ベクトルを作成し，これを生成分布としたサンプリングにより合成データを得る．MWEMにおいて，重み更新の繰返し回数をT，指数メカニズムで用いる安全性パラメータをϵ_1，Laplaceメカニズムの安全性パラメータをϵ_2とするとき，最終的に得られる合成データは$\epsilon = (\epsilon_1 + \epsilon_2)T$の$\epsilon$-差分プライバシを満たす．

また，MWEMは確率的に誤差の上界を持つ．Qを保存したい統計量に対応する問合せの集合とし，D^+を生成された合成データとするとき，少なくとも$1 - 2T/|Q|$の確率で，

$$\max_{q \in Q} |q(D^+) - q(D)| \leq 2n\sqrt{\frac{\ln |A|}{T}} + \frac{10T \ln |Q|}{\epsilon} \quad (16)$$

が成立する．（差分プライバシを持たない）乗算的重み更新法の誤差の上界は（決定的に）$2n\sqrt{\ln |A|/T}$である[21]ため，右辺第二項の存在と，上界が確率的にしか得られなくなる点が差分プライバシによる安全性の代償に相当する．

3.4 従来技術の課題

以上で議論したとおり，PRAMや合成データなどの統計的開示制御手法を用いることにより，差分プライバシを備える匿名化個票を得ることができる．しかし，これらの手法が与える差分プライバシの安全性強度はあまり高くない．いい換えると，十分な安全性を与えるようにこれらの手法を適用すると，有用な匿名化個票を出力できるとはいえなくなる．

以下，具体例を用いて定量的に議論する．たとえば，$A = A_1 \times A_2$，$|A_1| = |A_2| = 10$という属性空間を持つ，レコード数100の（簡単な）個票データを考える．この個票データに対し，集計データに対する差分プライバシの議論で（十分な安全性を持つとして）しばしば用いられる$\epsilon = 0.1$の差分プライバシを満たす，同数のレコードを持つ匿名化個票を得るために必要となるパラメータを具体的に計算する．

まず，PRAMに基づく手法について議論する．$\epsilon = \sum_{j=1}^{d} \epsilon_j$とするとき，式(11)の変形により維持確率$\rho_j$について以下の式を得る．

$$\rho_j = \frac{e^{\epsilon_j} - 1}{|A_j| + e^{\epsilon_j} - 1} \quad \text{s.t.} \sum_j \epsilon_j = \epsilon \quad (17)$$

このとき，$\epsilon_1 = \epsilon_2 = 0.05$とするならば，式(17)より維持確率$\rho_1$，$\rho_2$は以下の値をとる．

$$\rho_1 = \rho_2 = \frac{e^{0.05} - 1}{10 + e^{0.05} - 1} \simeq 0.005. \quad (18)$$

すなわち，上記の個票データに対して維持置換攪乱により0.1-差分プライバシを満たそうとすると，約0.5%の確率でしか元の個票データの属性値は維持されず，約99.5%の確率で属性値がランダムに置き替わることを意味する．

この悲観的な結果は，PRAMが差分プライバシを満たすための手法としてはあまり適していないことを示唆している．3.2節で示したように，PRAMは個票データに含まれるそれぞれのレコードについて，（他のレコードを参照することなく）それぞれ独立に攪乱を適用している．その結果，もし他に同じ属性値の組合せを持つレコードが複数個あった場合，その属性値の組合せに対応する完全分割表の値には，ほぼそのレコード数に比例した分散のノイズが付与される．すなわち，同じ属性値の組合せを持つレコードが多ければ多いほど（いい換えると「よくある」レコードであるほど）対応する完全分割表の値には大きなノイズが加算されることになる．

これは，「よくある」レコードほど（大きなノイズで）強固に守り，逆に珍しい属性値の組合せを持つレコードほど相対的に（小さなノイズで）弱く守るというアンバランスな結果をもたらす．差分プライバシの安全性パラメータϵは，定義1が示すように最も弱いレコードの安全性強度により定められるので，これはPRAMの結果にはϵを保証するために必要となる以上のノイズが含まれることを意味する．

なお，これはあくまでPRAMが差分プライバシにおいて考慮されるべき攻撃に対してバランスが良くないノイズを付与することを意味しており，それらの攻撃を考慮しない指標に対しては良好な結果を与える場合もある．たとえば，PRAMは（k-匿名性と等価な安全性を持つとされる）Pk-匿名性を満たすことが示されている[24]．同文献で与えられるPRAMとPk-匿名性の関係式から前述の条件における維持確率ρ_jを導出すると，たとえば$k = 3$のときに$\rho_1 = \rho_2 \simeq 0.14$を得る．これは式(17)で得られた値（約0.005）と比べると現実的といえる．ただし，k-匿名性やPk-匿名性は，（差分プライバシで考慮されている）属性値の開示（属性開示攻撃）に対する安全性や差分開示が生じうる状況における安全性に関して保証を持たない点に十分な留意が必要である．

次に合成データに基づくアプローチについて議論する．Machanavajjhalaら[32]による平滑化ディリクレ分布に基づく手法では，式(12)に上記のパラメータを代入することにより$\alpha_j \geq 100/(e^{0.1} - 1) \simeq 951$を得る．この数字の意味は直感的には分かりにくいが，これはDから完全分割表を作成したときに，すべてのセル値に951（以上の値）を足したものに基づいて合成データの生成モデルを作成するこ

とに相当する．この値は D の全レコード数すら大きく上回るものであり，元データの性質をほとんど反映しなくなることを意味する．

Wasserman ら [40] による平滑化ヒストグラムに基づく手法でも状況は似たものになる．式 (13) の変形により，該手法の安全性パラメータ δ は以下のように導出される．

$$\delta = \frac{p}{(e^{\epsilon/k}-1)n+p} = \frac{100}{(e^{0.1/100}-1)100+100} \simeq 0.999. \tag{19}$$

これを f_i^* の定義に代入すると，匿名化個票を得るためのサンプリング元となる分布は，その 99.9%が定数部分となることが分かる．すなわち，ほとんど元の個票データの情報は残らず，一様分布からサンプリングされたようなデータが匿名化個票として出力される．

これらの平滑化に基づく合成データ生成手法で芳しくない結果が得られる理由の 1 つとして，サンプリングにより得られる安全性は抽出率に依存することがあげられる．たとえば式 (13) によれば，安全性パラメータ ϵ と抽出するサンプル数（出力するレコード数）k は比例の関係にあり，k を小さくすれば（抽出率を小さくすれば）それに比例して ϵ が小さくなる（安全になる）という関係にあることが分かる．これは，これらの手法が（本稿が目的とする）すべての個票データを匿名化して出力するという応用より，大規模なデータから少数のサンプルを（匿名化して）抽出するという応用に適していることを示唆している．

Laplace メカニズムによる攪乱に基づく合成データ生成手法 [1] では，これまでに説明した手法より良い結果が得られる．式 (14) で加えられる Laplace ノイズの標準偏差は $\sqrt{2\sigma^2} = \sqrt{2}(2/0.1) \simeq 28.3$ であり，いままで議論した手法に比べて大幅に精度が改善されることが期待できる．これは，前述の Laplace メカニズムの準最適性，すなわち ϵ が小さい条件下においては Laplace メカニズムがほぼ最適なメカニズムとされることからも説明できる．匿名化個票の生成分布となる確率ベクトルのノイズが最小であれば，そこから生成される匿名化個票のノイズもまた最小化される．

ただし，この手法では，Laplace ノイズを加えた結果，ヒストグラムの値が負の値になる場合に，単純に 0 に切り上げるという操作を加えることにより，元の個票データにおいて出現頻度が低い属性組合せを持つレコードの個数は匿名化個票において増える傾向（正のバイアス）を持ち，出現頻度が高い属性組合せは減る傾向（負のバイアス）を持つことに注意が必要である．

Laplace ノイズ自体は平均が 0 の（バイアスを持たない）ノイズであるが，式 (14) による切り上げの可能性により v_i^* の期待値は v_i より大きいものとなる（正のバイアスが発生する）．ここで生じるバイアスの大きさは，v_i が小さいほど（出現頻度が低いほど）大きい．その一方，式 (14) における正規化により，このような「切り上げ」が発生しないような頻度を持つ属性組合せに対しては，逆に負のバイアスが発生することになる．

これは，属性空間の濃度に比べて十分な数のレコード数があり（$p \ll n$），かつデータの分布に偏りがない場合には，切り上げが発生する可能性が小さいため大きな問題とはならないが，元データがロングテイル性を持つ場合など，そうではない（完全分割表が小さい値もしくは 0 値を持つセルを多く含む）条件下では，生成された匿名化個票の分布を大きく歪める（分布のファットテイル化[*11]を引き起こす）原因となる．

最後に，MWEM について検討する．前述のとおり MWEM は確率的に誤差の上界を与えるが，実用的にはほぼ意味をなさない．式 (16) によれば，乗算的重み更新法の性質に従って，右辺第一項による誤差は繰返し回数 T の平方根に反比例する形で低減していくが，その一方で右辺第二項による誤差は T に比例して増加する．すなわち，全体として発散する．また，T の値が小さい領域でも実際には良い結果を与えない．たとえば，具体的に $Q = A$，$T = 10$ として式 (16) の右辺の値を計算すると，

$$2n\sqrt{\frac{\ln|A|}{T}} + \frac{10T\ln|Q|}{\epsilon} = 200\sqrt{\frac{\ln 100}{10}} + \frac{100\ln 100}{0.1} \simeq 136 + 4{,}605 = 4{,}741 \tag{20}$$

を得るが，これは上記の設定における誤差の上界として意味を持つ数字ではない．

また，MWEM は Q により指定される統計量を保存することに主眼を置いた方式であることに注意が必要である．できるだけ多くの種類の統計量を保存しようとして Q を増やしすぎると，式 (16) が示すとおり精度が劣化する．具体的な応用があらかじめ定められている場合には，その応用に必要となる統計量に限定して Q を指定すればよいが，一般的に匿名化個票が必要となるのはその後の使い方を限定したくない，もしくは使い方が定まっていない場合である．このとき，たとえば上記の計算例のように $Q = A$ とするなど Q の次元を大きくせざるを得ないが，これは MWEM の適用には適した条件ではない．

4. 提案方式

前章で示したように，従来手法では個票データを差分プライバシを満たすように加工すると，その有用性が大きく劣化する．そこで本稿では，個票データを直接処理するのではなく，完全分割表を経由して，Laplace メカニズムと最近傍探索により差分プライバシを満たす匿名化個票を得る方式を提案する．

[*11] ロングテイル性を持つ分布において，相対的に度数が小さい「テイル」部分が全体的に持ち上がり，その分だけ度数が大きい「ヘッド」部分が頭打ちになる．

提案方式は，個票データと等価な集計データである完全分割表に対して差分プライバシを満たすためのノイズ付与を行う．個票データに対して直接差分プライバシを与える方式に比べ，(Laplace メカニズムを代表とする）集計データに差分プライバシを与える方式は，データの有用性を効率的に保つ特長を持つ．ただし，ノイズが付与された完全分割表は，個票データの形にはそのまま変換できない．そこで，個票データの形式に変換可能な完全分割表の条件と，その導出手段を与える．

提案方式は，個票データから得られた集計データに Laplace メカニズムを適用する点において，Laplace メカニズムによる攪乱に基づく合成データ生成方式 [1], [40] と類似する．ただし，これらの方式が（攪乱により発生する）負値の単純な切り上げによって，生成される匿名化個票の分布を歪めてしまうことに対し，提案方式では Laplace メカニズムの出力から，匿名化個票を得ることができる最適な分布を最近傍探索により発見し，そこから（サンプリングを経ることなく）直接的に匿名化個票を得る点が異なる．これにより，特に大規模なデータ（ロングテイル性を持つ疎なデータであることが多い）において，元の個票データの性質をより良く残す匿名化個票を得られることが期待される．

4.1 処理の流れ

本方式は，以下に示す流れにより，個票データ D から差分プライバシを満たす匿名化個票 D^+ を作成する．

(1) 個票データ D に対応する完全分割表 V を作成する．
(2) V に対して（Laplace メカニズムなど）集計データに対して差分プライバシを与えるメカニズムを適用し，完全分割表 V^* を作成する．
(3) V^* から個票データに対応づけられる完全分割表 V^+ を導出する．
(4) 最後に V^+ を対応する匿名化個票 D^+ に変換する．

提案方式において差分プライバシを与えるためのメカニズムは Laplace メカニズムに限らない（たとえば幾何メカニズム [19] を用いてもよい）が，以降では Laplace メカニズムの適用を例として議論を進める．

上記手順において，手順 (1) および (2) は，それぞれ式 (3) および式 (9) により容易に実現することができる．しかし，式 (9) により得られた V^* は，そのままでは個票データに対応づけられない．これは，Laplace メカニズムによりノイズが加算されたセル値は，負の値や実数をとりうるため，V^* に対応する個票データが存在するとは限らないためである．いい換えると，V^* が f_A^{-1} の定義域に属するとは限らず，このとき $f_A^{-1}(V^*)$ は値をとりえない．

そこで，手順 (3) として，上記の V^* を入力として，そこから最も「近い」f_A^{-1} の定義域に属する（個票データと等価な）完全分割表 V^+ を導出した後に，手順 (4) で $D^+ = f_A^{-1}(V^+)$ により差分プライバシが保証された匿名化個票 D^+ を生成する．

4.2 匿名化個票を得るために満たすべき条件

個票データの形式に変換可能な完全分割表 V^+ の導出にあたり，まず最初に V^+ が満たす必要がある条件を議論する．

2 章で議論したように，完全分割表 V のセル値 v_i $(1 \leq i \leq p)$ は，個票データ D の台集合 A における各元の多重度 $m_D(a_i)$ $(a_i \in A)$ である．ここで，任意の多重集合において多重度の値域は非負整数であり，かつ，(台集合と）多重度を与えることにより対応する多重集合が一意に定まることから，V^+ の要素 v_i^+ について，

$$\forall v_i^+ \in V^+, v_i^+ \in \mathbb{Z}, v_i^+ \geq 0 \tag{21}$$

が成立すれば，V^+ は対応する多重集合を持つ（A をあわせて与えることにより個票データの形に戻すことができる）．すなわち，V^+ のすべての要素は非負の値を持ち（非負制約），かつ整数でなくてはならない（整数制約）[*12]．以降，V^+ により与えられる個票データを D^+ と表記する．

また，個票データ D のレコード数 n は，D の多重度の総和である（$n = \sum m_D(a_i) = \sum v_i$）．したがって，$D^+$ のレコード数を保存する（D と同じレコード数にする）ためには，

$$\sum v_i^+ = n \tag{22}$$

という束縛条件をあわせて満たす必要がある（総数制約）．

これら 3 つの制約条件を合わせると，以下の定理を得る．

定理 2. A を属性空間とする完全分割表 V^+ は，その要素 v_i^+ について非負制約，整数制約，総数制約をいずれも満たすとき，対応するレコード数 n の個票データ D^+ を持つ．すなわち，

$$\exists D^+ \mid \{D^+ = f_A^{-1}(V^+),\ |D^+| = n\},\\ \forall v_i^+ \in V^+, v_i^+ \in \mathbb{Z}, v_i^+ \geq 0 \quad \text{s.t. } \sum v_i^+ = n. \tag{23}$$

4.3 最近傍点の探索

定理 2 により，手順 (3) は V^* から最近傍にある（式 (23) の条件を満たす）V^+ を探索する問題に帰着される．ここで最近傍とは，(V^*, V^+) 間の距離を p 次元の実数ベクトル空間 $\mathbb{R}^p$ におけるユークリッド距離（L_2 距離）とし，これを最小にすることを意味する．すなわち，以下の問題を解くことにより V^+ を得る．

$$V^+ = \underset{\Phi \in \mathbb{N}_0^p}{\arg\min} |V^* - \Phi|_2 \quad \text{s.t. } |\Phi|_1 = n. \tag{24}$$

幾何学的には，これは p 次元の実数ベクトル空間 $\mathbb{R}^p$ にお

[*12] 2 つの制約条件を合わせて「各要素は（0 を含む）自然数である $(\forall v_i^+, v_i^+ \in \mathbb{N}_0)$」と一言で表すこともできるが，以降の議論のため非負制約と整数制約を分けて表現する．

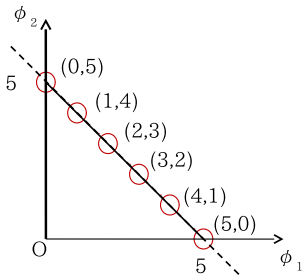

図 2 $p=2$, $n=5$ における式 (24) の解の候補点
Fig. 2 Feasible solutions of Eq. (24) for $p=2$, $n=5$.

いて，$p-1$ 次元の超平面 $Z:\sum_{\varphi_i\in\Phi}\varphi_i=n$ 上に配置された $\mathbb{N}_0^p$ 上の格子点を候補点とし，V^* から最近傍に存在する候補点を抽出する問題に相当する．たとえば，$p=2$, $n=5$ の場合における候補点は，座標 $(0,5)$ と $(5,0)$ を結ぶ線分上の（端点を含む）6 個の整数格子点となる（**図 2**）．

n や p が十分に小さいときは，これはすべての候補点に対する距離計算や，kd-木などを用いた探索など，一般的な最近傍探索手法により求めることができる．しかし，n もしくは p が大きくなると候補点の数は階乗オーダで増加するため，計算量の観点からあまり実用的ではない．

式 (24) における候補点の数は，D のレコード数 n と V^* のセル数 p（$=|A|$）によって定まる．具体的には，p 元からなる集合から（重複を許した）n 元を選ぶ組合せ（重複組合せ）の数 ${}_pH_n$ となるため，

$$ {}_pH_n = {}_{p+n-1}C_n = \frac{(p+n-1)!}{n!(p-1)!}. \tag{25}$$

すなわち，$O((p+n)!)$ の計算量オーダとなる．kd-木を用いれば，時間計算量はその対数となるが，それでも現実的とはいえない．

ただし，候補点が超平面上の非負格子点として規則的に並んでいることを利用すれば，下記の手順により，より効率的に V^+ を探索することができる．

(1) V^* から最近傍にある，超平面 Z 上の点を求める．これは，V^* から Z への垂線の足に相当するため，Z の法線ベクトル $\mathbf{z}=(1,1,\ldots,1)$ とパラメータ s を用い，直線 $V^*+s\mathbf{z}$ と Z との交点となる．これを s について解くことにより下式を得る．

$$ \Phi_1 = V^* - \frac{(|V^*|_1-n)}{p}\,\mathbf{z}. \tag{26}$$

ここで Φ_1 は Z 上にあるため，総数制約を満たす．

(2) Φ_1 から最近傍にある，超平面 Z 上の非負領域内の点 Φ'（$\forall\varphi'_i\in Z$, $\varphi'_i\geq 0$）を求める．具体的な手順は後述する．Φ' は超平面 Z 上の非負領域に存在するため，総数制約と非負制約をいずれも満たす．

(3) 最後に，Φ' の最近傍にある Z 上の整数格子点を探索する．これは，Φ' の各要素について，小数部分が大きいものは小数点以下を切り上げ，小さいものは同じく切り下げたものになる．具体的には，Φ' の各要素の小数点以下を切り下げたものを $\lceil\Phi'\rceil$ とするとき，$|\Phi'-\lceil\Phi'\rceil|_1$ 個の要素を（小数部分が大きい順に）切り上げ，それ以外を切り下げる．これを V^+ とする．

以上の手順により，総数制約，非負制約，整数制約のいずれも満たす，V^* の最近傍点 V^+ を得る．これにより，定理 1 が示すように，f_A^{-1} を用いて等価な匿名化個票 $D^+=f_A^{-1}(V^+)$ を得ることができる．

4.4 総数制約と非負制約を満たす最近傍点の探索手順

前節の探索手順 (2) における，超平面 Z 上にある非負領域上の Φ_1 からの最近傍点の探索手順について具体的に議論する．これは，Φ_1 に含まれる要素の最小値を φ_i として，$\varphi_i<0$ であるならば Z と $\varphi_i=0$ とが交わる（$|Z|-1$ 次元の）超平面上にある Φ_1 との最近傍点 Φ_2 を求め，得られた Φ_2 について同様の処理を（負値がなくなるまで）再帰的に繰り返して $\Phi_3,\Phi_4,\ldots,\Phi_k$ と求めていくことにより，$\Phi'=\Phi_k$ として得ることができる．

これをナイーブに実装すると，その計算量は $p\simeq m'$ のときに $O(p)$，そうでないときに $O(p^2)$ となる．ここで，m' は Φ' において非 0 値をとる要素の個数である．Φ_j を得るための計算に，$p-j+1$ 次元のベクトルに対する最小値の探索と点の移動をともなうことから，第 j ステップの計算量は $O(p-j+1)$ となり，これを k 回繰り返すと全体の計算量は $O(\sum_{j=1}^{k}(p-j+1))$ となる．ここで，繰返しの回数 k は，$k=p-m'$ で表されることから $O(p(p-m'))$ を得る．これは $p\simeq m'$ のときに $O(p)$，そうでないときに $O(p^2)$ となる．

ただし，下記の手順のとおり複数の負値をまとめて処理する工夫を導入することにより，これはより小さい計算量で求めることができる．なお，下記の手順において，$j=1$ を初期値とする．

(1) Φ_j の要素を，値の正負により $\Phi_j^+=\{\varphi_i\}\mid\varphi_i>0$ と $\Phi_j^-=\{\varphi_i\}\mid\varphi_i\leq 0$ に分割する．

(2) Φ_j^+ の各要素に対し，Φ_j^- の総和を Φ_j^+ の要素数で除した値 $-|\Phi_j^-|_1/|\Phi_j^+|$ を加算する．

(3) Φ_j^- に対応する Φ' の要素をすべて 0 にする．

(4) もし Φ_j^+ が負値を含むのであれば，$\Phi_{j+1}=\Phi_j^+$ とし，j に 1 を加えたうえで本手順を再実行する．含まなければ，Φ_j^+ に対応する Φ' の要素に Φ_j^+ の値を代入し，処理を終了する．

この計算量は，$p\simeq m'$ のとき $O(p)$，それ以外で $O(p\log p)$ となる．その導出方法は若干複雑であるため付録 A.1 に示す．定量的な実行時間については次章で議論する．

5. 評価

提案方式の有用性を検証するため，提案方式により得られる匿名化個票 D^+ がどのくらい元の個票データ D と近

いものであるかについて，従来方式との比較を通じて評価する．以下，評価に用いる評価データ，評価手順，評価結果の順に説明する．

5.1 評価データ

評価のための個票データ（評価データ）としては，商品の売上履歴を擬似的に模した，下記の 3 属性からなるロングテイル性を持つデータを用いる．ある商品が，どのような顧客に売れたかを記録した売上履歴を考える．ここで，商品種別は r 種類あり，それぞれの商品を購入した顧客について性別（男性/女性）と年齢層（20 代から 60 代まで 10 歳刻み）が記録されているとする．すなわち，評価データ D は，以下の属性空間 A を台集合とする，n 個の要素から構成される多重集合となる．

$$\begin{aligned} &A = A_1 \times A_2 \times A_3, \\ &A_1 = \text{商品種別} = \{h_1, \ldots, h_r\}, \\ &A_2 = \text{性別} = \{\text{男性}, \text{女性}\}, \\ &A_3 = \text{年齢層} = \{20\text{代}, \ldots, 60\text{代}\}. \end{aligned} \tag{27}$$

このとき，属性空間 A の濃度 p は，$p = |A| = r \times 2 \times 5 = 10r$ となる[*13]．すなわち，D に対応する完全分割表 V は $10r$ 次元のベクトルとして構成される．

パレートの法則（Pareto principle）が示唆するように，商品売上など，自然現象や社会現象による事象から得られる高次元データの頻度分布はロングテイル性を持つ（べき乗則もしくは近い性質が成り立つ）ことが多いとされる．そこで，評価データにおいて商品 h_k が売れる確率 $\Pr[x_{i1} = h_k]$ は，Zipf の法則に従うものとした．すなわち，

$$\forall i, \Pr[x_{i1} = h_k] = \frac{k^{-1}}{\Sigma_{j=1}^{r} j^{-1}}. \tag{28}$$

また，男女による嗜好差を表すものとして，k が奇数の商品 h_k は男性が女性の 2 倍多く買う（2/3 の確率で男性が購入，1/3 で女性が購入）とし，k が偶数の場合はその逆とした．年齢層による差は特に導入していない．購入した商品種別や性別にかかわらず，顧客の年齢層は 20 代から 60 代まで一様に分布しているものとする．

本章の評価は，この評価データについて商品種別数を 100，1,000，10,000 とし，売上件数を 10,000，100,000，1,000,000 としてそれぞれ生成した 3 種類の規模のデータセットを用いて実施した．前述のとおり，属性空間の大きさは商品種別数の 10 倍（$p = 10r$）であるため，これらの評価データの規模はそれぞれ ($p = 1{,}000,\ n = 10{,}000$)，($p = 10{,}000,\ n = 100{,}000$)，($p = 100{,}000,\ n = 1{,}000{,}000$) となる．

[*13] A_1 は単一の商品を元とすることに注意．複数商品の組合せ購入を 1 つのレコードとして扱いたい場合は，A_1 の元は r 種類の商品からの組合せとなり，たとえば k 種類以内の商品の組合せを扱うとき $|A_1| = \sum_{i=1}^{k} {}_nC_i$ となる．

5.2 評価手順

前述の評価データを用い，差分プライバシの安全性パラメータ ϵ を変化させながら，提案方式を PRAM の一種である維持置換攪乱に基づく手法 [24], [31]（以降, PRAM と記載する）および合成データに基づく手法と比較することによりその性質を議論する．3 章で議論したとおり，合成データに基づく手法はいくつかの派生があるが，本評価では代表的なものとして，平滑化されたディリクレ分布に基づく方式 [32]（以降，Synth-MD と記載する）と Laplace メカニズムを用いた攪乱分布に基づく方式 [1], [40]（以降，Synth-Lap と記載する）を比較対象として採用する．ただし，Synth-MD は他の方式と比べて試行に時間がかかることから，最も小さい規模のデータセット ($p = 1{,}000,\ n = 10{,}000$) のみを評価対象とした．

PRAM における維持確率 ρ_j の値は，対応する属性 A_j の値の保護のために（$\sum_j \epsilon_j = \epsilon$ の制約の下で）どれだけの大きさの ϵ_j を配分するかに依存する（式 (17) 参照）が，本実験ではすべての ϵ_j は等しい（$\epsilon_{\text{商品種別}} = \epsilon_{\text{性別}} = \epsilon_{\text{年齢層}} = \epsilon/3$）とした．また，Synth-MD における平滑化パラメータ α_j（式 (12) 参照）は，すべて等しく ϵ により定まる最小値をとるものとした．

また，それぞれの手法の出力が元の個票データの性質をどのくらい維持しているかを示す評価指標は，文献 [40] にならい，一般的な誤差指標として L_2 距離（$|V - V^+|_2$）を，分布の歪みを示す指標として KS-距離（Kolmogorov Smirnov distance）を用いる．KS-距離とは，ある 2 つの標本群があったときに，それらが同じ確率分布に従うか否かを検査する検定（KS-検定）で用いられるノンパラメトリックな統計量であり，それぞれの累積分布（経験分布）が最大でどのくらい異なるかを距離とする．

いずれの統計量も値が小さいほど優れていることを示しており，完全に元データと一致すると 0 になる．たとえば個々のデータの出現頻度に誤差があっても分布全体の形状は保たれているような場合，L_2 距離は大きくなるが KS-距離は小さく抑えられる．その一方，出現頻度に薄く広くバイアスがかかるような形で分布が歪められた場合（たとえばロングテイル分布のファットテイル化など），L_2 距離は小さく抑えられるが KS-距離は大きな値を持つ．

それぞれの指標値について，100 回の独立試行による平均値を得る．その際，評価データについても，1 回の試行ごとに前節で示した分布に基づいて作り直すものとした．

なお，本評価は intel Core i5-6200U（2.3 GHz, 2 core）を CPU とする一般的なノート型 PC 上で，Python 2.7.12 + NumPy 1.11.1 を用いて実施した．評価のために作成したプログラムは（NumPy をライブラリとして用いた）純粋な Python プログラムであり，Cython などの事前コンパイル拡張は適用していない．評価にあたり，各方式の実行時間についてもあわせて測定した．

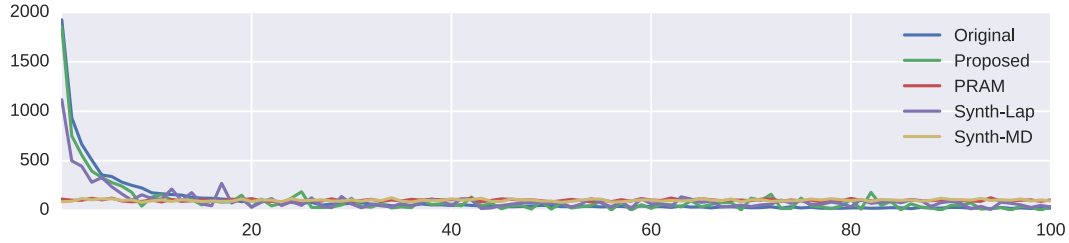

図 3 度数分布の比較（$\epsilon = 0.1$，$p = 1{,}000$，$n = 10{,}000$）

Fig. 3 Comparison of frequency distributions ($\epsilon = 0.1$, $p = 1{,}000$, $n = 10{,}000$).

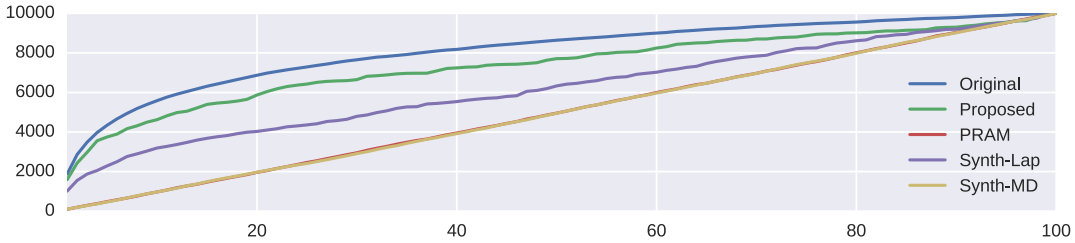

図 4 累積度数分布の比較（$\epsilon = 0.1$，$p = 1{,}000$，$n = 10{,}000$）

Fig. 4 Comparison of the cumulative frequency distributions ($\epsilon = 0.1$, $p = 1{,}000$, $n = 10{,}000$).

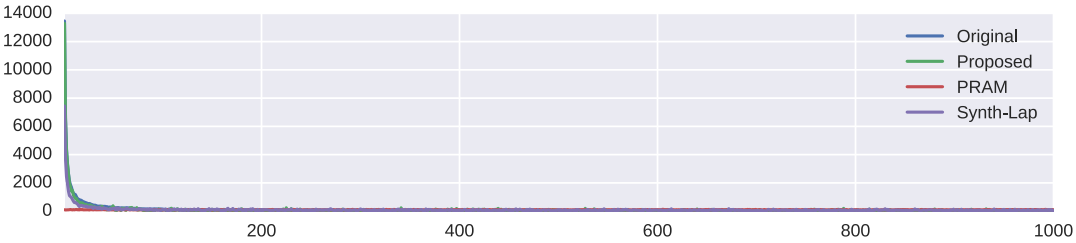

図 5 度数分布の比較（$\epsilon = 0.1$，$p = 10{,}000$，$n = 100{,}000$）

Fig. 5 Comparison of frequency distributions ($\epsilon = 0.1$, $p = 10{,}000$, $n = 100{,}000$).

5.3 評価結果

まず，直感的な理解のために，$\epsilon = 0.1$ の条件下のある試行における，評価データ（“Original”），提案方式の出力（“Proposed”），PRAM による出力（“PRAM”），Laplace メカニズムにより攪乱した分布に基づく合成データ（“Synth-Lap”），平滑化されたディリクレ分布に基づく合成データ（“Synth-MD”）のそれぞれについて，商品種別を x 軸とした度数分布と累積度数分布をデータセットの規模ごとに示す（**図 3**，**図 4**，**図 5**，**図 6**，**図 7**，**図 8**）．なお，本評価で評価指標の 1 つとして用いる KS-距離は，累積度数分布のグラフにおける，“Original” との間の最大距離にほぼ比例する（これを標本数で除して正規化した値で近似できる）．評価データは商品種別に関して Zipf 分布に基づいて生成しているため，いずれのグラフからもデータのロングテイル性が確認できる．

次に，差分プライバシの安全性パラメータである ϵ を変化させながら，各手法の出力における L_2 距離と KS-距離の変化を定量的に評価した結果について，L_2 距離の評価結果を**表 1**，**表 2**，**表 3** に，KS-距離の評価結果を**表 4**，**表 5**，**表 6** に，データセットの規模ごとにそれぞれ示す．ここで，ϵ を変化させる範囲は，$\epsilon \in \{0.1, 0.2, \ln 2, \ln 3, 10, 100\}$ とした[*14]．便宜上，以降では $\epsilon \in \{0.1, 0.2\}$ を高プライバシ条件，$\epsilon \in \{\ln 2, \ln 3\}$ を中プライバシ条件，$\epsilon \in \{10, 100\}$ を低プライバシ条件と呼ぶ．

いずれの表も，横軸は ϵ の値を，縦軸は適用した手法をそれぞれ表す．各手法のうち，最も良い（同一の ϵ において，最も値が小さい）値を太字で表現している．

各手法の実行時間は Python の Timeit ライブラリによりデータセットの規模ごとに測定した．対象データセットの規模が小さい順（$p = 1{,}000$，$10{,}000$，$100{,}000$）に，提案方式の 1 回あたりの実行時間は 0.21 ms，1.32 ms，12.6 ms であった．また，PRAM は 5.63 ms，19.4 ms，164 ms であり，Synth-Lap が 1.71 ms，20.2 ms，271 ms という結果をそれぞれ得た．なお，Synth-MD は $p = 1{,}000$ のデータセットに対する実行時間は 1.21 s であったが，$p = 10{,}000$，$100{,}000$ ではメモリ不足で動作しなかった．

[*14] なお，直感的には，$\epsilon = \ln k$ とは yes/no などの二値で回答する問合せに対し，「k 回に一度の割合でランダムな答えを返す（もしくは，$k+1$ 回に一度の割合で嘘の答えを返す）」としたときに得られるプライバシレベルに相当する．

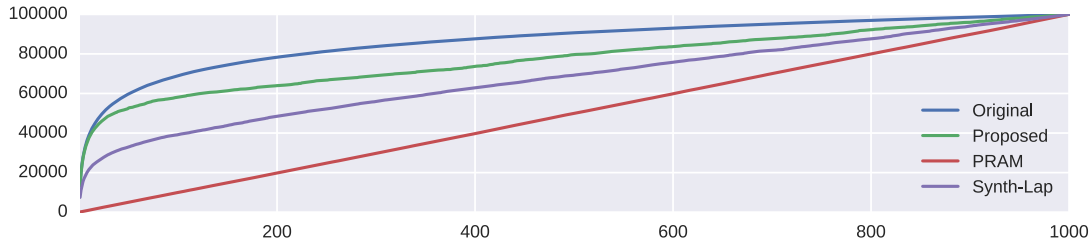

図 6 累積度数分布の比較（$\epsilon = 0.1$, $p = 10{,}000$, $n = 100{,}000$）

Fig. 6 Comparison of the cumulative frequency distributions ($\epsilon = 0.1$, $p = 10{,}000$, $n = 100{,}000$).

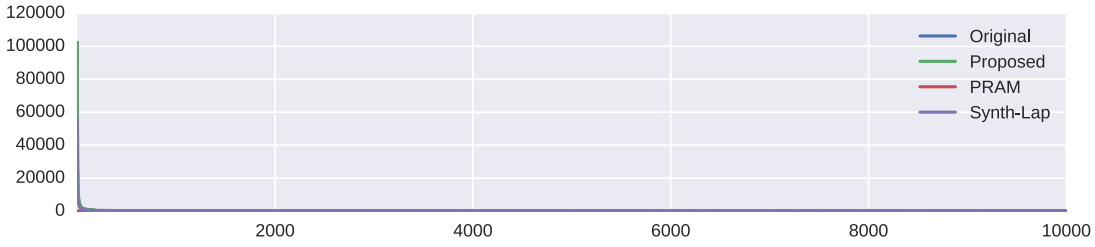

図 7 度数分布の比較（$\epsilon = 0.1$, $p = 100{,}000$, $n = 1{,}000{,}000$）

Fig. 7 Comparison of frequency distributions ($\epsilon = 0.1$, $p = 100{,}000$, $n = 1{,}000{,}000$).

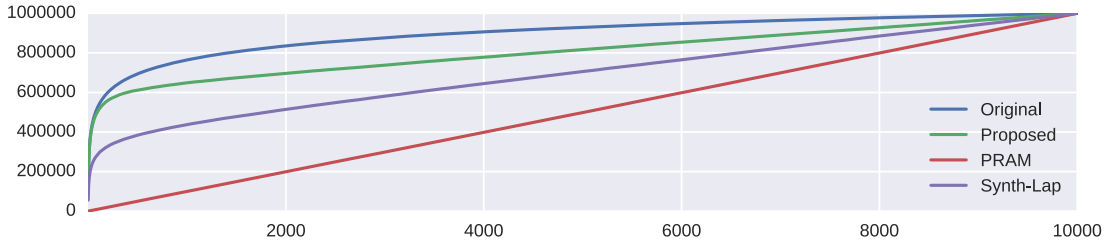

図 8 累積度数分布の比較（$\epsilon = 0.1$, $p = 100{,}000$, $n = 1{,}000{,}000$）

Fig. 8 Comparison of the cumulative frequency distributions ($\epsilon = 0.1$, $p = 100{,}000$, $n = 1{,}000{,}000$).

表 1 L_2 距離の比較（$p = 1{,}000$, $n = 10{,}000$）

Table 1 L_2 distance ($p = 1{,}000$, $n = 10{,}000$).

$\epsilon =$	0.1	0.2	ln 2	ln 3	10	100
提案方式	**504.0**	**296.6**	**107.7**	**72.6**	**9.0**	**0.0**
PRAM	770.4	771.8	771.1	769.1	57.7	**0.0**
Synth-Lap	512.0	333.2	149.5	122.6	99.3	99.8
Synth-MD	771.2	772.2	770.1	769.2	104.8	99.6

表 2 L_2 距離の比較（$p = 10{,}000$, $n = 100{,}000$）

Table 2 L_2 distance ($p = 10{,}000$, $n = 100{,}000$).

$\epsilon =$	0.1	0.2	ln 2	ln 3	10	100
提案方式	**1,470**	**874.5**	**322.1**	**218.3**	**28.1**	**0.0**
PRAM	5,644	5,639	5,639	5,640	1,945	**0.0**
Synth-Lap	2,773	1,670	547.0	416.5	312.3	315.8

表 3 L_2 距離の比較（$p = 100{,}000$, $n = 1{,}000{,}000$）

Table 3 L_2 distance ($p = 100{,}000$, $n = 1{,}000{,}000$).

$\epsilon =$	0.1	0.2	ln 2	ln 3	10	100
提案方式	**4,330**	**2,603**	**974.1**	**664.0**	**87.4**	**0.0**
PRAM	43,588	43,597	43,587	43,588	36,420	**0.0**
Synth-Lap	20,141	11,994	3,162	1,898	1,005	993.6

表 4 KS-距離の比較（%, $p = 1{,}000$, $n = 10{,}000$）

Table 4 KS-distance (%, $p = 1{,}000$, $n = 10{,}000$).

$\epsilon =$	0.1	0.2	ln 2	ln 3	10	100
提案方式	**16.6**	**8.3**	**1.9**	**1.0**	**0.1**	**0.0**
PRAM	49.5	49.5	49.6	49.5	2.2	**0.0**
Synth-Lap	26.0	15.1	3.7	1.9	0.9	0.9
Synth-MD	49.6	49.5	49.6	49.6	2.5	0.8

表 5 KS-距離の比較（%, $p = 10{,}000$, $n = 100{,}000$）

Table 5 KS-distance (%, $p = 10{,}000$, $n = 100{,}000$).

$\epsilon =$	0.1	0.2	ln 2	ln 3	10	100
提案方式	**15.2**	**8.1**	**1.8**	**1.0**	**0.0**	**0.0**
PRAM	59.8	59.8	59.8	59.8	18.7	**0.0**
Synth-Lap	30.0	18.1	4.3	2.2	0.3	0.3

表 6 KS-距離の比較（%, $p = 100{,}000$, $n = 1{,}000{,}000$）

Table 6 KS-distance (%, $p = 100{,}000$, $n = 1{,}000{,}000$).

$\epsilon =$	0.1	0.2	ln 2	ln 3	10	100
提案方式	**14.0**	**7.9**	**2.0**	**1.1**	**0.0**	**0.0**
PRAM	66.5	66.5	66.5	66.5	54.5	**0.0**
Synth-Lap	33.0	20.1	5.2	2.8	0.2	0.1

6. 考察

5 章の評価結果に基づき，提案方式の有用性に関して従来手法と比較して議論する．

まず最初に，$p = 1{,}000$，$n = 10{,}000$ における度数分布および累積度数分布のグラフ（図 3，図 4）について議論する．いずれのグラフにおいても，PRAM の出力と Synth-MD の出力[*15]は元データの性質をほとんど残しておらず，ほぼ一様分布と等価な出力となっていることを示している．これは，たとえば PRAM の場合，式 (17) に基づいて ρ_j を具体的に計算した結果と符合する．本評価の条件下で式 (17) により商品種別の維持確率 ρ_1 を計算すると $\rho_1 \simeq 3.4 \times 10^{-4}$ となる．すなわち，確率的には 3,000 回に 1 回未満しか商品種別が維持されず，それ以外は一様ランダムに 100 種類の中から商品が選択されていることになる．さらに他の属性を含めて同時確率 $\rho = \prod_j \rho_j$ を計算すると，$\rho \simeq 1.7 \times 10^{-5}$ となり，10 万回に 1〜2 回しかすべての属性値が維持されることはない．詳細は割愛するが，Synth-MD でも同様に，平滑化パラメータ α_j が著しく大きい値となることにより，合成データのサンプリング元となる生成分布がほぼ一様分布になってしまっている．

それに比べると，提案方式と Synth-Lap はいずれも元データの性質を残しているといえる．度数分布（図 3）では差異は一見して明らかではないが，累積度数分布（図 4）で見ると，提案方式のほうがより良好に元データの性質を残している．3 章で議論したとおり，Synth-Lap はノイズの付与後に負値を単純に切り捨ててしまうことにより，分布がファットテイル化する影響を持つことを反映した結果となっている．提案方式もファットテイル化の兆候は見られるが，Synth-Lap と比べてその影響は軽減されている．なお，これらの傾向はデータセットの規模が違ってもほとんど変わっていない（図 5，図 6，図 7，図 8）．

次に，評価指標である L_2 距離および KS-距離の傾向を，ϵ の変化による影響とデータセットの規模による影響を含めて考察する．

まず $p = 1{,}000$，$n = 10{,}000$ における状況（表 1，表 4）に着目し，ϵ の違いによる変化について議論する．高プライバシ条件（$\epsilon \in \{0.1, 0.2\}$）と中プライバシ条件（$\epsilon \in \{\ln 2, \ln 3\}$）では，いずれも相対的な状況は大きく変わらない．提案方式と Synth-Lap では ϵ の増加に対して誤差が低減していく様子が見られるが，いずれの条件においても提案方式が安定して Synth-Lap を上回る精度を持つ．特に KS-距離を指標として見た場合，提案方式は Synth-Lap の約半分強の誤差となっている．PRAM と Synth-MD は，この領域では ϵ を増やしても精度が改善しない．$\epsilon = \ln 3$ などの中プライバシ条件においても，ほぼ出力が一様分布のままである．

低プライバシ条件（$\epsilon \in \{10, 100\}$）においては状況が変化する．提案方式はこの領域でも ϵ の増加に従って精度が改善され，$\epsilon = 100$ の条件下では元の評価データとの差が観測不能となった（100 回程度の試行では入力データと差が生じることがなかった）．その一方，Synth-Lap は L_2 距離で 100 前後，KS-距離で 0.9%程度で頭打ちとなった．PRAM と Synth-MD は，この領域では大きな改善を見せた．特に，PRAM は $\epsilon = 100$ の条件下では提案方式と同様に評価データとの差が観測不能となった．Synth-MD は Synth-Lap と同様の「頭打ち」傾向が見られた．これは，合成データに基づく手法では，匿名化個票を得るために分布からの無作為抽出をともなうことから，ϵ が大きい条件下ではそれによるノイズの影響が無視できなくなるためと考えられる．

次に，データセットの規模の違いによる変化について議論する．実験結果から分かるように，データセットの規模が大きくなっても上記で議論した傾向は定性的にはほぼ変わらない．提案方式はいずれの条件下においても安定的に高い精度を保ち，$\epsilon = 10$ において PRAM と Synth-Lap の順位が入れ替わっていることを除き，方式ごとの優劣にも差は見られない．

ただし，提案方式は他の 2 方式に比べて，データセットの規模の増加に対する精度の悪化が抑えられた結果になった．特に，KS-距離についてはほとんど変化がない．Synth-Lap も中低プライバシ条件では比較的精度の悪化は抑えられているが，高プライバシ条件での L_2 距離の悪化が顕著である．PRAM は安定的に精度が悪化している．

提案方式や Synth-Lap が PRAM と比べてデータセット規模の影響を受けにくい理由は，プライバシ保護のために加えられる摂動の大きさが属性空間に依存しないことが理由として考えられる．3.4 節の式 (17) が示すように，PRAM における維持確率の計算式には分母に属性の大きさが含まれている．すなわち，データセットの規模が大きくなればなるほど維持確率は減少する（摂動が増える）ことになる．それに対し，提案方式や Synth-Lap は Laplace メカニズムによるノイズが加わるが，そのノイズ強度は属性空間に依存しない．

なお，高プライバシ条件で Synth-Lap の精度悪化が顕著であったことについては，3.4 節で課題として述べた「負値の切り上げ」による悪影響がデータのロングテイル部に多く現れるようになったためと推測される．低プライバシ条件では付与されるノイズ強度が小さく，ロングテイル部であっても負値の切り上げが発生する可能性やその影響が小さく抑えられる一方で，高プライバシ条件では，より高頻度で切り上げが発生する．Zipf 分布の性質より，大規模なデータセットであるほどロングテイル性が高くなることから，その影響が高プライバシ条件において強く現れたことが考えられる．

[*15] 図 3，図 4 のいずれにおいても，PRAM と Synth-MD のグラフはほぼ重なっている．

最後に，実行時間について考察する．性能測定の結果，本実験の条件下では（メモリ不足が発生したSynth-MDを除き）各方式とも実用上の課題となるような処理時間は見られなかったが，その中でも提案方式が最も高速であった．ただし，より大きなデータセットに対しては計算量の観点から留意が必要となる．

それぞれの方式の計算量について簡単に議論する．PRAMの計算量は，レコードあたりの属性値の個数をdとすれば明らかに$O(dn)$であり，Synth-Lapは Laplaceメカニズムによる攪乱に$O(p)$，そこからのnレコードの抽出に$O(n)$をそれぞれ要することから，全体としての計算量は$O(p+n)$となる．Synth-MDは，p次元のディリクレ乱数の生成をn回繰り返し行うことが計算量的には支配的である．p次元のディリクレ乱数はp個のガンマ乱数の正規化により生成されるため，Synth-MDの計算量は全体で$O(np)$となる．それに対し，提案方式は4.2節で示した近傍探索が支配的であり，その計算量は$O(p)$ないし$O(p\log p)$である[*16]．

すなわち，PRAM以外の手法は，提案方式も含めてpに依存する計算量を持つ．これは，地理空間データなどのようにきわめて大きなpを持ちうるデータを扱う際に，深刻な速度低下を引き起こす懸念がある[6], [44]．その改善は今後の課題である．

7. 関連研究

本稿では差分プライバシを満たした匿名化個票を得るにあたり，(a) 完全分割表と個票データの等価性に着目し，(b) ノイズが付与された完全分割表から匿名化個票を得るための条件を定式化（総数制約，非負制約，整数制約）したうえで，(c) Laplaceメカニズムを適用した完全分割表への事後処理（post processing）により前記条件を（L_2距離の意味で）最適かつ効率良く満たす方式を提案した．

すなわち，提案方式は制約条件に基づく事後処理により，差分プライバシの有用性を向上させる試みの1つとして考えることもできる．本章では，その観点における関連研究と提案方式との関係や違いについて議論する．

Barakら[4]は，集計データに対して離散Fourier変換を適用し，Fourier空間においてノイズ付与をすることにより，各Fourier級数に対応する周辺度数（marginal）の精度を向上させるとともに，周辺度数と個々のセル値との整合性を確保する方式を提案している．しかし，その結果として得られる集計データは（周辺度数の精度を向上させる代償として）個々のセル値については精度が悪いものとなり，さらに等価な個票データを持つための制約条件をいずれも満たさない．

ここで，非負制約と整数制約については，得られた（ノイズ付き）Fourier級数に対して事後的に線形計画法（LP）などを適用することにより解決できることが示唆されている．しかし，上記のとおりFourier変換に基づくノイズ付与は（周辺度数については良好な精度を持ちながらも）個々のセル値には最適な精度を与えない（Laplaceメカニズムより悪くなる）．すなわち，完全分割表にこの方式を適用した出力から作成された匿名化個票の頻度分布は，元の個票データの頻度分布と大きく異なることになる．さらに，この事後処理で用いられる線形計画問題の計算量は明らかにされておらず，そのスケーラビリティは不明である．

なお，同文献では線形計画問題を用いずに非負制約を与える簡便な手法として，ノイズ付与後のFourier級数の直流成分に対し，非負制約を満たすために十分な量として計算された値を加算する方法についても触れられている．この場合の計算量は線形計画法の適用と比較して十分に小さいものとなるが，Fourier級数の直流成分への加算は（逆変換後の）すべての出力値に正のバイアスを加えることに相当し，総数制約を明らかに逸脱させる．

Hayら[22]は，unattributed histogram（本稿における分割表から，属性空間Aとの対応を失わせたもの）とuniversal histogram（本稿における一般的な集計データに相当し，分割表とは限らない）の2種類のクラスの集計データについて，事後処理により精度向上を図る手法を提案している．

これらのうち，unattributed histogramは属性との対応が失われていることから，匿名化個票の作成に用いることは原理的に不可能である[*17]．universal histogramに対する事後処理は，集計データを半分ずつに分割していったときに現れる部分和に対してLaplaceノイズを加えたうえで[*18]，包含関係にある部分和間の整合性を制約条件としたL_2最適化により出力を得る．この最適化は2回の部分和全体の走査で実現され，その計算量は$O(p)$である．

この手法は，Barakらの手法（線形計画問題を除いても$O(p\log p)$を要する）より高速であるが，その一方で非負制約などの個票データとの等価性に関する制約条件は考慮されない．したがって，この手法を完全分割表に適用したとしても，その出力が匿名化個票を持つことは保証されない．

寺田ら[44]は，地理空間データの分析において重要となる部分和の精度向上と非負制約の逸脱の解決を目的として，集計データに対して局所性保存写像（Morton符号化）とHaar Wavelet変換を適用したうえでノイズを付与し，Wavelet変換の性質を用いて非負制約を満たした分割表を効率的に得る手法を提案している．その具体的な構成

[*16] 厳密には，入出力において個票データと完全集計表との間の形式変換が必要な場合，その処理のために$O(n)$を要する．

[*17] なお，unattributed histogramに対する精度向上のための事後処理としてはLinら[30]がより洗練された手法を提案しているが，同様に匿名化個票の作成に用いることはできない．

[*18] それぞれの部分和は互いに$\log_2 p$個の重複する部分和を持つことになるため，ここで加えられるLaplaceノイズの強度は（直列合成則により）それに比例して強くなることに注意．

法として（構成が容易だが $O(p)$ の計算量を持つ）直列構成法と（複雑だが計算量が $O(m \log p)$ に抑えられる）並列構成法の 2 種類の等価な出力分布を持つ構成法が示されている．そのうち直列構成法は Xiao らの Privlet [41] に対する事後処理の一種と考えることができる．

この手法は計算量 $O(p)$ もしくは $O(m \log p)$ と高速に非負制約を満たした分割表を出力する（$p = 2^{18} = 262{,}144$ のデータを約 20〜30 ms で処理するとしている）が，それ以外の制約条件を直接には満たさない．また，Barak らの手法 [4] と同様に，部分和の精度向上の代償として個々のセル値の精度は良くない（個々のセル値や小領域の部分和では単純な Laplace メカニズムよりノイズの分散が大きくなる）とされる．したがって，この手法の出力は（なんらかの形で総数制約や整数制約をさらに与えたとしても）最適な匿名化個票を与えない．

Lee ら [26] は，文献 [22] の universal histogram に対する事後処理と同様に，部分和間の整合性確保を事後処理として行うことにより集計データの精度を向上させる方式を提案している．この手法では，整合性確保における最適化の目的関数として（L_2 距離だけでなく）L_1 距離をあわせて用い[*19]，数値最適化法の一種である ADMM（alternating direction method of multipliers）を用いてこれを解決するアルゴリズムを与えている．

しかし，この最適化にあたっては，文献 [44] と同様に非負制約を考慮しているが他の制約条件は考慮しておらず，そのままでは匿名化個票の作成に用いることができない．また，ADMM は汎用性が高い数値最適化法として知られているが，その計算量や実行性能は前述の文献 [44] や提案方式に比べて大きく劣る．たとえば文献 [26] で示された実験によれば，Opteron 2216（2core，2.4 GHz）を用いた $p = 2^{12} = 4{,}096$ のヒストグラムの最適化に 21 秒を要している．計算量が明示されていないため該手法のスケーラビリティは不明であるが，$p \times p$ 次元の行列演算を含む処理を許容誤差が得られるまで繰り返し行うことから，$O(p^2)$ もしくはそれ以上の計算量を必要とすると考えられる．したがって，より大きな属性空間を持つデータに対しては実用的な実行速度が得られないことが懸念される．

これに対し，提案方式では $p = 100{,}000$ の完全分割表に対して 12.6 ms で処理を完了させている（5.3 節）．問題のサイズは 25 倍であるにもかかわらず実行時間は約 1/1,700 に抑えられており，計算量もより小さい．評価に用いられた環境が異なることなどから定量的に厳密な意味を持つ比較ではないが，実行性能の違いは明らかである．

[*19] 文献 [26] では最尤推定の観点から L_2 最適化より L_1 最適化が優れていると主張しているが，L_1 のみによる最適化はユニークな解を持つとは限らないことから実際には L_2 最適化を併用している．

8. まとめ

本稿では，差分プライバシを満たす匿名化個票の作成方法について検討し，集計データの一種である完全分割表に対する Laplace メカニズムの適用と，ベクトル空間における最近傍探索に基づく非負制約，整数制約，総数制約の充足によりこれが得られることを示した．

提案方式の有用性を評価するために，売上履歴を模したロングテイル性を持つ擬似的な個票データを用いて評価を行い，従来方式として PRAM の一種である維持置換攪乱方式 [24], [31]，Laplace ノイズにより攪乱された分布からの合成データ生成に基づく方式 [1], [40]，平滑化されたディリクレ分布からの合成データ生成に基づく方式 [32] と比較した．その結果，提案方式は L_2 ノルム，KS-距離のいずれの評価指標からも，従来方式と比べて元となる個票データの性質をより強く保持することを示した．

また，その実行性能について実装に基づく比較評価と計算量に基づく考察を行った．本稿での評価実験の範囲では，提案方式が最も高速に処理を終えた．ただし，提案方式は属性空間の濃度 p に依存する計算量を持つことから，地理空間データなどのようにきわめて大きな p を持ちうるデータを扱う際には留意が必要であり，その改善は今後の課題である．

参考文献

[1] Abowd, J.M. and Vilhuber, L.: How protective are synthetic data?, *Privacy in Statistical Databases*, LNCS, Vol.5262, pp.239–246 (2008).

[2] Acs, G., Castelluccia, C. and Chen, R.: Differentially private histogram publishing through lossy compression, *Proc. 2012 IEEE 12th Intl. Conf. Data Mining* (*ICDM*), pp.1–10 (2012).

[3] Arora, S., Hazan, E. and Kale, S.: The Multiplicative Weights Update Method: A Meta-Algorithm and Applications, *Theory of Computing*, Vol.8, No.6, pp.121–164 (2012).

[4] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F. and Talwar, K.: Privacy, accuracy, and consistency too: A holistic solution to contingency table release, *Proc. 26th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems - PODS '07*, pp.273–282, ACM Press (2007).

[5] Blum, A., Ligett, K. and Roth, A.: A Learning Theory Approach to Noninteractive Database Privacy, *J. ACM*, Vol.60, No.2, pp.1–25 (2013).

[6] Cormode, G., Procopiuc, M., Srivastava, D. and Tran, T.: Differentially Private Publication of Sparse Data, *Proc. Intl. Conf. Database Theory* (*ICDT2012*) (2012).

[7] Department of Economic and Social Affairs, United Nations: *Multilingual Demographic Dictionary* (1982).

[8] Dwork, C.: Differential Privacy, *Proc. 33rd Intl. Conf. Automata, Languages and Programming - Volume Part II*, LNCS, Vol.4052, pp.1–12, Springer (2006).

[9] Dwork, C.: Differential privacy: A survey of results, *Proc. 5th Intl. Conf. Theory and Applications of Models of Computation*, pp.1–19, Springer (2008).

[10] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. and Naor, M.: Our Data, Ourselves: Privacy via Distributed Noise Generation, *Proc. EUROCRYPT 2006*, pp.486–503 (2006).
[11] Dwork, C., McSherry, F., Nissim, K. and Smith, A.: Calibrating noise to sensitivity in private data analysis, *Proc. 3rd Conf. Theory of Cryptography*, LNCS, Vol.3876, pp.265–284, Springer (2006).
[12] Dwork, C., Naor, M., Reingold, O., Rothblum, G.N. and Vadhan, S.: On the Complexity of Differentially Private Data Release: Efficient Algorithms and Hardness Results, *Proc. 41st Annual ACM Symp. Theory of Computing* (*STOC '09*), pp.381–390 (2009).
[13] Dwork, C. and Roth, A.: The Algorithmic Foundation of Diferential Privacy, *Foundations and Trends in Theoretical Computer Science*, Vol.9, No.3-4, pp.211–407 (2014).
[14] Erlingsson, Ú., Pihur, V. and Korolova, A.: RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, *Proc. 2014 ACM SIGSAC Conf. Computer and Communications Security* (*CCS '14*), pp.1054–1067 (2014).
[15] Fienberg, S.E.: A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality, Technical Report, Carnegie Mellon University (1994).
[16] Geng, Q., Kairouz, P., Oh, S. and Viswanath, P.: The Staircase Mechanism in Differential Privacy, *IEEE J. Selected Topics in Signal Processing*, Vol.9, No.7, pp.1176–1184 (2015).
[17] Geng, Q. and Viswanath, P.: The Optimal Mechanism in Differential Privacy, Technical Report (2012).
[18] Geng, Q. and Viswanath, P.: Optimal Noise Adding Mechanisms for Approximate Differential Privacy, *IEEE Trans. Information Theory*, Vol.62, No.2, pp.952–969 (2016).
[19] Ghosh, A., Roughgarden, T. and Sundararajan, M.: Universally Utility-maximizing Privacy Mechanisms, *SIAM J. Computing*, Vol.41, No.6, pp.1673–1693 (2012).
[20] Gouweleeuw, J., Kooiman, P., Willenborg, L. and de Wolf, P.-P.: The Post Randomisation Method for Protecting Microdata, *Quaderns d'Estadística i Investigació Operativa* (*QÜESTIIÓ*), Vol.22, No.1, pp.145–156 (1998).
[21] Hardt, M., Ligett, K. and McSherry, F.: A Simple and Practical Algorithm for Differentially Private Data Release, *Proc. 26th Annual Conf. Neural Information Processing Systems* (*NIPS 2012*), pp.2339–2347 (2012).
[22] Hay, M., Rastogi, V., Miklau, G. and Suciu, D.: Boosting the accuracy of differentially private histograms through consistency, *Proc. VLDB Endowment*, Vol.3, No.1-2, pp.1021–1032, VLDB Endowment (2010).
[23] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. and de Wolf, P.-P.: *Statistical Disclosure Control*, John Wiley & Sons (2012).
[24] Ikarashi, D., Kikuchi, R., Chida, K. and Takahashi, K.: k-anonymous Microdata Release via Post Randomisation Method, *eprint arXiv*, Vol.1504.05353, pp.1–22 (2015).
[25] Kifer, D. and Machanavajjhala, A.: No free lunch in data privacy, *Proc. 2011 Intl. Conf. Management of Data* (*SIGMOD '11*), pp.193–204, ACM (2011).
[26] Lee, J., Wang, Y. and Kifer, D.: Maximum Likelihood Postprocessing for Differential Privacy under Consistency Constraints, *Proc. 21st ACM SIGKDD Intl. Conf. Knowledge Discovery and Data Mining* (*KDD '15*), pp.635–644, ACM Press (2015).
[27] Li, N., Qardaji, W. and Su, D.: On Sampling, Anonymization, and Differential Privacy: Or, k-Anonymization Meets Differential Privacy, *Proc. 7th ACM Symp. Information, Computer and Communications Security* (*ASIACCS '12*), pp.32–33, ACM (2012).
[28] Li, Y.D., Zhang, Z., Winslett, M. and Yang, Y.: Compressive Mechanism: Utilizing Sparse Representation in Differential Privacy, *Proc. 10th Annual ACM Workshop on Privacy in the Electronic Society* (*WPES '11*), pp.177–182, ACM Press (2011).
[29] Liew, C.K., Choi, U.J. and Liew, C.J.: A Data Distortion by Probability Distribution, *ACM Trans. Database Systems*, Vol.10, No.3, pp.395–411 (1985).
[30] Lin, B.-R. and Kifer, D.: Information preservation in statistical privacy and Bayesian estimation of unattributed histograms, *Proc. 2013 Intl. Conf. Management of Data* (*SIGMOD '13*), pp.677–688 (2013).
[31] Lin, B.-R., Wang, Y. and Rane, S.: A Framework for Privacy Preserving Statistical Analysis on Distributed Databases, *2012 IEEE Intl. Workshop on Information Forensics and Security* (*WIFS*), pp.61–66 (2012).
[32] Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L.: Privacy: Theory meets Practice on the Map, *24th Intl. Conf. Data Engineering*, pp.277–286, IEEE (2008).
[33] McSherry, F. and Talwar, K.: Mechanism Design via Differential Privacy, *Proc. 48th Annual IEEE Symp. Foundations of Computer Science* (*FOCS '07*), pp.94–103 (2007).
[34] McSherry, F.D.: Privacy integrated queries: An extensible platform for privacy-preserving data analysis, *Proc. 35th SIGMOD Intl. Conf. Management of Data* (*SIGMOD '09*), pp.19–30 (2009).
[35] Miklau, G.: The Matrix Mechanism: Optimizing Linear Counting Queries Under Differential Privacy (2011).
[36] Rubin, D.B.: Discussion—Statistical Disclosure Limitation, *J. Official Statistics*, Vol.9, No.2, pp.461–468 (1993).
[37] Sweeney, L.: k-anonimity: A model for protecting privacy, *Intl. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol.10, No.5, pp.557–570 (2002).
[38] Tramèr, F., Huang, Z. and Ayday, E.: Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies Categories and Subject Descriptors, *Proc. 22nd ACM Conf. Computer and Communications Security* (*CCS '15*), pp.1286–1297, ACM (2015).
[39] Wang, J., Liu, S. and Li, Y.: A Review of Differential Privacy in Individual Data Release, *Intl. J. Distributed Sensor Networks*, Vol.2015, No.1, pp.1–18 (2015).
[40] Wasserman, L. and Zhou, S.: A Statistical Framework for Differential Privacy, *J. American Statistical Association*, Vol.105, No.489, pp.375–389 (2010).
[41] Xiao, X., Wang, G., Gehrke, J. and Jefferson, T.: Differential Privacy via Wavelet Transforms, *IEEE Trans. Knowledge and Data Engineering*, Vol.23, No.8, pp.1200–1214 (2011).
[42] Yuan, G., Zhang, Z., Winslett, M., Xiao, X., Yang, Y. and Hao, Z.: Low-rank mechanism: Optimizing batch queries under differential privacy, *Proc. VLDB Endowment*, Vol.5, No.11, pp.1352–1363 (2012).
[43] 伊藤伸介：政府統計データにおける匿名化について，中央大学経済研究所年報，No.46, pp.457–478 (2015).

[44] 寺田雅之，鈴木亮平，山口高康，本郷節之：大規模集計データへの差分プライバシの適用，情報処理学会論文誌，Vol.56, No.9, pp.1801–1816 (2015).
[45] 星野伸明：公的統計ミクロデータ提供制度の課題，日本統計学会誌，Vol.40, No.1, pp.23–45 (2010).
[46] 統計センター：統計データ開示抑制に関する用語集（改訂版），製表関連国際用語集，No.2 (2005).

付　　録

A.1 探索手順（4.4 節）の計算量について

4.4 節において，非負制約と総数制約を満たす最近傍点の（効率化された）探索手順の計算量は，$p \simeq m'$ のとき $O(p)$ であり，それ以外のとき $O(p \log p)$ であるとした．以下，その導出方法を簡単に説明する．

本手順において，j 回目の処理の計算量は

$$O\left(\sum_{s=j}^{k} |\Phi_s^-| + |\Phi_k^+|\right) \tag{A.1}$$

となることに着目する．すなわち，全体の計算量は下式で表される．

$$O\left(\sum_{j=1}^{k}\left(\sum_{s=j}^{k} |\Phi_s^-| + |\Phi_k^+|\right)\right) = O\left(\sum_{j=1}^{k}\sum_{s=j}^{k} |\Phi_s^-| + k|\Phi_k^+|\right). \tag{A.2}$$

ここで，$\sum_{s=1}^{k} |\Phi_s^-| + |\Phi_k^+| = p$，$|\Phi_k^+| = m'$ より，

$$\sum_{s=1}^{k} |\Phi_s^-| = p - m' \tag{A.3}$$

である．これを用いると，

$$\sum_{j=1}^{k}\sum_{s=j}^{k} |\Phi_s^-| < k\sum_{s=1}^{k} |\Phi_s^-| = k(p - m') \tag{A.4}$$

と表される．したがって，式 (A.2) は，

$$O(k(p - m') + km') = O(kp) \tag{A.5}$$

となる．繰返し回数 k について $k = O(\log(p - m'))$ となることから，

$$O(kp) = O(p \log(p - m')). \tag{A.6}$$

すなわち，$p \simeq m'$ のとき $O(p)$，それ以外のとき $O(p \log p)$ を得る．

寺田 雅之 （正会員）

1995 年神戸大学大学院工学研究科修士課程修了．同年日本電信電話（株）入社．同社情報通信研究所，情報流通プラットフォーム研究所を経て，2003 年（株）NTT ドコモへ転籍．2008 年電気通信大学大学院電気通信研究科博士後期課程修了．博士（工学）．2009 年より現職．情報セキュリティ技術，プライバシ保護技術，大規模統計処理技術の研究開発に従事．2015 年度本学会論文賞，2015 年度本学会山下記念研究賞受賞．電子情報通信学会会員．

山口 高康 （正会員）

2001 年電気通信大学大学院電気通信学研究科博士前期課程修了．同年株式会社 NTT ドコモ入社．以後，携帯端末での撮影対象判別技術，権利価値流通技術，コンテンツ検索技術，統計情報作成技術，プライバシ保護技術の研究開発に従事．2015 年度本学会論文賞受賞．

本郷 節之 （正会員）

1984 年岩手大学大学院工学研究科修士課程修了．同年日本電信電話公社入社．1987 年 ATR 視聴覚機構研究所へ出向．1991 年 NTT ヒューマンインタフェース研究所へ復帰．この間，視覚情報処理モデルの研究に従事．著書『脳・神経システムの数理モデル』（共著）ほか．工学博士．1999 年 NTT ドコモマルチメディア研究所へ転籍．2001 年セキュリティ方式研究室長．2010 年北海道工業大学（現北海道科学大学）教授に着任，現在に至る．モバイルセキュリティならびにプライバシ保護技術の研究開発に従事．2015 年度本学会論文賞受賞．電子情報通信学会，IEEE 各会員．