Computer Security Symposium 2014
22 - 24 October 2014

大規模集計データへの差分プライバシーの適用

寺田 雅之[1]　鈴木 亮平[1]　山口 高康[1]　本郷 節之[2]

概要：データの有効な活用による社会・産業の発展への期待が高まる中，プライバシーを保護した上でデータを利用するための技術が注目を集めている．その中で，Dwork らによる差分プライバシーは，その高い安全性から大きな期待が寄せられているが，データの有用性や処理効率の観点から実用上の課題を持つ．本論文では，地理空間データなどの大規模な集計データに差分プライバシーを適用する上での課題を示すとともに，これを解決する手法について安全性証明と実データに基づく評価を与える．本手法は，集計データの非負制約に着目し，その逸脱を補正する過程を導入することにより有用性と処理効率の向上を実現する．

On Publishing Large Tabular Data with Differential Privacy

Masayuki Terada[1]　Ryohei Suzuki[1]　Takayasu Yamaguchi[1]　Sadayuki Hongo[2]

Abstract: Big data become widely expected to enhance the quality and efficiency of our daily life. On the other hand, facile utilization of such data can easily derive serious privacy breach; data must be utilized or published with preserving privacy, while it isn't an easy task. Differential privacy is a promising paradigms to achieve proven privacy, but previous methods to assure differential privacy have several drawbacks on data utility and scalability in practice, in particular when applied to publishing large tabular data such as geospatial data. This paper introduces a novel differential private method, which simultaneously solves the utility and scalability problems through correcting the deviation of its output from the non-negative restriction. According to the evaluation results using Japanese census data in 2010, the output data from proposed method has much superior precision (i.e. lower noises) to those of the previous methods, the Laplace mechanism and Xiao's Privelet.

1. はじめに

人々に関係するデータベースから作成された集計データを公開するにあたっては，プライバシー保護への十分な配慮が必要となる．ここで，集計データとは，元のデータベースに含まれる個々のデータ群 (個票，あるいは生データ (raw data) とも呼ばれる) から作成した，「ある条件を満たす」データの個数を数えあげた数値データ (セル) の集まりである．特に，本研究では，広範囲の空間分布を表す集計データ (たとえば人口分布や交通量分布) などの，大規模な地理空間に関する集計データ (地理空間データ) を主な検討の対象とする．

本研究では，集計データのプライバシーを保護するための基準として，Dwork らにより 2006 年に提案された差分プライバシー (differential privacy)[3] に着目する．これは，「ある人が (加工データを作成する上での元データとなる) データベースに含まれるか否かの，加工データからの判別困難性」を安全性の根拠とするプライバシー保護基準である．差分プライバシーは，k-匿名性 (k-anonymity) 基準 [10] などのプライバシー保護基準と異なり，任意の背景知識を持つ攻撃者や未知の攻撃に対して数学的な安全性が与えられているという優れた性質を持つ．

差分プライバシー基準を実現する代表的な手段としては，集計データの各セルに対して，平均 0 の Laplace 分布に従う独立した乱数 (Laplace ノイズ) を付与する手法が挙げられる．この手法は Laplace メカニズムと呼ばれる．

しかし，Laplace メカニズムをそのまま実際の集計データのプライバシー保護に適用することは，特に大規模な集計データにおいて実用上の困難を伴う．その理由として，Laplace メカニズムが適用された集計データは (実際の集計データではありえない) 負数を多く含みうるため，その

[1] (株)NTT ドコモ 先進技術研究所
Research Laboratories, NTT DOCOMO, Inc.
[2] 北海道科学大学 工学部
Faculty of Engineering, Hokkaido University of Science

後の利用に困難を伴うこと (**非負制約の逸脱**)，複数セルの部分和を取った際の誤差が大きく有用性が劣化すること (**部分和精度の劣化**)，集計データの密度 (非 0 値の割合) を大きく増大させてしまい，大規模な集計データに適用した際に計算量や出力データ量が現実的ではなくなること (**計算量の増大**)，の三点の課題が挙げられる．

これらの課題に対して，いくつかの部分的な改善方式が提案されている [1], [2], [7], [8], [11], [12]．しかし，いずれの方式においても，前述の三点の課題を同時に解決することはできず，またこれらの方式を単純に組み合わせることも困難である．

そこで，これらの問題を解決するため，Wavelet 変換と Top-down 精緻化と呼ぶ手法を組み合わせた，差分プライバシー基準を満たすプライバシー保護方式を提案する [*1] とともに，二次元データである地理空間データを提案方式に適用するための適用方式を与え，提案方式による改善効果を検証する．

本提案方式は，部分和精度の劣化を抑えるためのアプローチとして Xiao らによる手法 (Privelet)[11], [12] と同じ戦略を採る．すなわち，集計データにそのまま Laplace ノイズを付与するのではなく，集計データに離散 Wavelet 変換の一種である Haar Wavelet 変換 (HWT) を適用し，得られた Wavelet 係数群に対して Laplace ノイズを付与することにより差分プライバシーを達成する．

ただし，Xiao らの手法では，Wavelet 係数群への Laplace ノイズの付与後に，そのまま (HWT の逆変換である) 逆 HWT により出力となる集計データを得ていることにより，非負制約の逸脱や計算量の増大を招いている．それに対し，提案方式は Wavelet 係数に対応する部分和が非負制約を満たすよう，再帰降下的に各係数に補正を加えながら逆 HWT を適用する．また，非負制約を明らかに逸脱させる Laplace ノイズは，安全性を損なうことなく除去できる (発生させる必要がない) ことに着目する．これにより，Xiao らの手法では解決できなかった非負制約の充足と計算量の抑制を併せて達成する．さらに，提案方式は，プライバシーを犠牲にすることなしに，Xiao らの手法より優れた部分和精度を達成する．その効果は，特に疎な集計データへの適用時に顕著となる．

また，入出力が一次元ベクトルである提案方式を (二次元データである) 地理空間データに適用するにあたり，地理空間分析で多用される正方ブロック領域における部分和精度の向上に着目した次元変換方式を示す．この方式を用い，国勢調査のメッシュ人口を対象として提案方式の精度を Laplace メカニズムおよび Xiao らの手法と比較することにより，実際の地理空間データにおける提案方式の改善効果を定量的に評価する．

[*1] なお，本方式の原型は [13] で示している．本稿で示す方式は，[13] に対し，ノイズスケールを効率化する改善を加えるとともに，アルゴリズムの整理や誤記修正を施し，安全性および有用性に関する証明を与えている．

2. 従来技術と課題

2.1 集計データ

集計データとは，1 以上の属性を持つレコードの集合から構成されるデータベースにおいて，ある属性 (もしくは属性の組み合わせ) に該当するレコードの個数を数えあげた値の集合である．集計データは，様々な統計分析における基礎データとして広く使われている．

集計データは以下のように定義できる．l 個のレコードから構成されるデータベース $D = \{x_1, x_2, \ldots, x_l\}$ を考える．ここで，各レコード x_i は d 次の属性空間 $A = A_1 \times A_2 \times \cdots \times A_d$ に属するベクトル値を持つ $(x_i \in A)$．集計データとは，ある与えられた A の部分空間の集合 $C = (C_1, C_2, \ldots, C_n)$ $(C_j \subseteq A)$ に対する，D 内の各部分空間に属するレコードの個数の集合 $V = (v_1, v_2, \ldots, v_n)$ である．ここで，$v_j = Count(D, C_j)$ であり，これは C_j に属する D 中のレコードの個数 $|x|(x \in D, x \in C_j)$ を意味する．

一般的には，各属性の定義域 A_k の互いに素な部分空間集合の直積が C として用いられる．この時の集計データは分割表 (contingency table) と呼ばれる．分割表におけるそれぞれの値 v_j を，セル (cell) もしくはセル値と呼ぶ．以降，本稿では断りがない限り集計データは分割表の形式をとるものとする．

実世界に基づく大規模な集計データは，0 値のセルを多数含む，疎 (sparse) なデータになることが多い．すなわち，論理的なセルの総数を $n(= |C|)$，そのうち非 0 値を持つセル数を m とすると，$m \ll n$ となる．たとえば，ある日時における日本全国の属性別人口分布を，500m メッシュ単位に，5 歳区分の年齢層別，男女別，居住市区町村別に集計したとする．日本の国土にかかる 500m メッシュの数は約 150 万 [14]，年齢層の数は約 20，市区町村数は約 2,000 であるため，論理的なセルの総数 $n = |C|$ はおよそ $1.5 \cdot 10^6 \times 20 \times 2 \times 2 \cdot 10^3 = 1.2 \cdot 10^{11}$ となる．これは日本の総人口である約 $1.2 \cdot 10^8$ を 1,000 倍ほど上回る数字であり，そのまま計算機上で扱うには極めて効率が悪い．

そのため，実際の集計データは，実装上は非 0 値を持つ m 個のセルのみについての，(j, v_j) の組のリストとして表現されることが多い．これは COO 形式 (coordinate format) とも呼ばれる [9]．

2.2 差分プライバシー

差分プライバシー [3], [5] は，識別不能性に基づくプライバシー基準の一種である．直感的には，「ある個人のデータを含むデータベースに対する問い合わせ結果が，その個人のデータを含まないデータベースへの問い合わせ結果と

区別できないなら，その問い合わせは安全である (個人に関するプライバシーを開示しない)」という考え方によりプライバシーを規定する．

たとえば，個人に関するデータの集合から構成されるデータベース D と，データベース問い合わせ $f(\cdot)$ を考える．なお，$f(\cdot)$ は (出力に摂動を加えるなど) 確率的な出力を持つ関数 (randomized function) である．このとき，データベース D への問い合わせ $f(D)$ と，ある個人 i に関するデータ $x_i(\in D)$ を D から取り除いたデータベース $D'(= D \setminus x_i)$ への問い合わせ結果 $f(D')$ が区別できないなら，$f(D)$ から x_i に関して意味がある情報を抽出することはできない．すなわち，個人 i のプライバシーは保護される．

より厳密には，差分プライバシーはパラメータ 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 (互いにたかだか 1 要素しか異ならない) データセット 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-差分プライバシーを満たす．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 $(S \subseteq \mathcal{R})$.

$$\Pr[\mathcal{K}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{K}(D_2) \in S]. \qquad (1)$$

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．

差分プライバシーの特徴として，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挙げられる．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，差分プライバシー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われたとしても，その安全性が損なわれないということを意味する．Dwork は，差分プライバシーが持つこの性質について，差分プライバシーは (“ad hoc” ではなく) “ad omnia” なプライバシー保証を与える，としている [4].

2.3 Laplace メカニズム

差分プライバシー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．差分プライバシーを実現する代表的なメカニズムとしては Laplace メカニズム [*2] が挙げられる．

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 ノイズを問い合わせ結果に加算することにより実現できる．Laplace 分布の確率密度は平均 μ とスケール λ を用いて下式で与えられる．

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

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 $\mathrm{Lap}(\lambda)$ とし，k 個の互いに独立した $\mathrm{Lap}(\lambda)$ からなるベクトル列を $\mathrm{Lap}(\lambda)^k$ と記載する．

Laplace メカニズムにおける Laplace ノイズのスケール λ は，定義 1 におけるパラメータ ϵ と，問い合わせの種類ごとに定まる「(大域的) 感度 (global sensitivity, GS)」によって与えられる．具体的には，GS_f を問い合わせ $f : \mathcal{D} \to R$ 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．

$$\mathcal{K}_f(X) = f(X) + \mathrm{Lap}(GS_f/\epsilon), \qquad (3)$$

$$GS_f = \max_{D_1, D_2} ||f(D_1) - f(D_2)||_1. \qquad (4)$$

ここで，D_1 および D_2 は任意の隣接したデータセット (定義 1 参照) のペアである．

2.4 集計データへの Laplace メカニズムの適用

理論的には，Laplace メカニズムを用いることにより差分プライバシーが保証された集計データを簡単に作成することができる．

前述の通り，集計データは計数問い合わせ結果 $v_j = Count(D, C_j)$ の集合である．ここで，C を構成する各部分集合は互いに素 $(\forall i, j (i \neq j), C_i \cap C_j = \phi)$ であるとする．計数問い合わせの感度 GS_{count} は 1 であることが知られているため，集計データのセル v_i にスケール $1/\epsilon$ の Laplace ノイズを加えた値，

$$v_i^* = v_i + \mathrm{Lap}(1/\epsilon) \qquad (5)$$

は ϵ-差分プライバシーを満たす．

v_i は互いに素な集合から生成されていることから，差分プライバシーの並列合成則により，v_i^* の集合 $V^* = \{v_1^*, v_2^*, \ldots, v_n^*\}$ もまた ϵ-差分プライバシーを満たす．すなわち，$V^* = V + \mathrm{Lap}(1/\epsilon)^{|V|}$ により ϵ-差分プライバシーが保証された集計データを得られる．

しかし，実際の集計データにこの方法を適用することは，しばしば現実的ではない．その理由として，以下の 3 点が挙げられる．

第 1 の問題は，非負制約の逸脱である．集計データは計数値の集合であるため，その定義から各セルの値は非負でなくてはならない．しかし，Laplace メカニズムを適用したデータは (実際の集計データではありえない) 負数を多く含みうる．これは，データの利用者にとって不自然に感じられるだけでなく，分析プログラムの予期せぬ異常動作を引き起こす可能性をもたらすなど，データの利用に著し

[*2] 計数問い合わせに対しては，幾何分布に従う乱数を用いた幾何メカニズム (geometric mechanism) のほうが適しているとされる ([2], [6]). しかし，ϵ が小さいときには両者の出力はほぼ変わらないことから，本稿では代表して Laplace メカニズムを扱う．

い困難を生じさせる．

第 2 の問題は，部分和精度の劣化である．集計データを利用する際には，個々のセルの値だけではなく，複数のセルの値を合算した部分和が用いられることも多い．たとえば，500m メッシュを単位とした人口分布から，2x2 個のメッシュ人口を合算して 1km メッシュ人口として利用することなどは一般的である．しかし，Laplace メカニズムを適用した集計データにおける部分和には，和をとる対象のセル数と等しい数の Laplace ノイズが重畳して加算される．たとえば，上記の例では，1km メッシュには 4 個分のノイズが，2km メッシュであれば 16 個分のノイズが重畳することになる．すなわち，部分和の対象範囲が広ければ広いほどノイズによる真値からの偏差が大きくなり，その有用性が大きく劣化する．

第 3 の問題は，計算量の増大である．前述の通り，実世界から得られた大規模な集計データは疎であることが多い．すなわち，論理的なセルの総数を n とし，そのうち 0 以外の値をとるセルの個数を m とすると，$m \ll n$ となる．Laplace メカニズムによるノイズの付与は，(0 値のセルを含めた) n 個のセルに対して行なう必要がある [*3]．すなわち，$O(m)$ のデータ量で表現された集計データに対し，$O(n)$ の計算量による Laplace ノイズの付与により $O(n)$ のデータ量を持つ集計データを出力することになる．これは $m \ll n$ の場合に非効率であるだけでなく，そもそも前述の日本全国の属性別人口の例のように n が非常に大きくなる場合には現実的ではない．

2.5 関連研究

これらの課題に対し，これまでにいくつかの部分的改善手法が提案されている．

Cormode ら [2] は，「ある閾値」を越える値を持つセルの値だけがあれば良いという応用を前提とした上で，計算量の増大を回避する方式を提案している．この方式では，Laplace ノイズの付与により 0 値のセルが「閾値」を超える (= 出力の対象となる) 確率をあらかじめ計算しておき，その確率に従った個数のセルをランダムに抽出する．そして，以降の処理は，ここで抽出されたセルと非 0 値を持つセルのみを対象とする．閾値が十分に大きければ計算量・データ量ともに大きく削減されるため，計算量の問題は回避される．また，閾値は正の値をとることから，非負制約の問題も生じない．

その一方，Cormode らの手法は「閾値」を下回る値を持つセルは無視されてしまうことから，部分和にノイズだけでなく (過少方向の) バイアスを生じさせる．そのため，単に Laplace メカニズムを適用したデータ以上に部分和の利用は困難なものとなる．

Barak ら [1] は，離散 Fourier 変換の導入と周波数領域における Laplace メカニズムの適用により部分和精度を改善し，さらに線形計画法に基づいて非負制約の逸脱を解消する方法を提案している．具体的には，集計データに離散 Fourier 変換を適用した上で，各周波数成分 (部分和をとる範囲に相当する) に対応する Fourier 係数にそれぞれ Laplace メカニズムを適用することにより，部分和の精度を改善する．その後に，(元の集計データを参照することなく) 適用後データのみを参照して線形計画法を適用することにより，差分プライバシーを保ちつつ非負制約の逸脱を解消する．しかし，計算量の増大への対処はなされておらず，また線形計画法の計算負荷が大きいことから，大規模な集計データへの実用的な適用は困難である．

Xiao ら [11], [12] は，部分和精度の改善に離散 Wavelet 変換を用いる手法を提案している．この手法は “Privelet” と名付けられている．Barak らが Fourier 変換を導入したのに対し, Privelet では Haar 基底に基づく離散 Wavelet 変換 (HWT) を導入し，その Wavelet 係数に対して Laplace メカニズムを適用する．HWT は概念や実装が単純であり，Fourier 変換に基づく手法では自明ではない階層的な名義尺度への適用を，比較的簡単な拡張で可能としている．しかし，その一方で非負制約の逸脱を解決する手段は与えられていない．また，Barak らの手法と同様に計算量の増大についても解決されない．

このように，いずれの手法も前述の 3 つの問題の全てを同時に解決しない．また，これらの手法を単純に組み合わせて問題を解決することも困難である．

3. 提案方式

前節で上げた 3 点の課題を解決する新たなプライバシー保護方式を提案する．

提案方式は, 部分和精度の劣化を抑制するために, Xiao らの手法 (Privelet 法) と同様に, 集計データにそのままノイズを加えるのではなく, 集計データ $V = (v_1, v_2, \ldots, v_n), \forall v_i \geq 0$ に対して Haar Wavelet 変換 (HWT) $\mathcal{H}$ を適用した Wavelet 係数系列 $W = \mathcal{H}(V)$ に対してノイズを加えるアプローチを採る．簡単のため，$n = 2^k$ $(k \in \mathbb{N})$ であるとする．

Privelet 法では，W へのノイズ付与後に，単に逆 HWT $\mathcal{H}^{-1}$ を適用することにより出力となる集計データを得る．しかし，この方法で得られた集計データは，ノイズの影響により非負制約を逸脱し，さらに計算量は $O(n)$ となり計算量の問題も解決しない．

この問題を解決するために, 提案方式は, Top-down 精緻化と呼ぶ処理過程を導入する．Top-down 精緻化は，HWT により得られた W の各要素に対し，Laplace ノイズの付与と非負精緻化，および逆 Haar Wavelet 変換を再帰降下

[*3] 厳密には，構造的ゼロ (structual zero)，すなわち「0 以外の値をとることが論理的にありえない」セルに対してはノイズの付与は不要である．

法で適用していく．これにより，差分プライバシー基準を満たす出力 V^+ を得るとともに，非負制約からの逸脱への対処と計算量の増大を併せて解決する．

3.1 Haar Wavelet 変換

Haar Wavelet 変換 (HWT) $\mathcal{H} : \mathbb{R}^n \to \mathbb{R}^n$ は，階段関数の一種である Haar 関数を母 Wavelet とした離散 Wavelet 変換であり，長さ $n = 2^k$ $(k \in \mathbb{N})$ のベクトル列 $V = (v_1, v_2, \ldots, v_n)$ を，同じ長さを持つベクトル列 $W = (w_1, w_2, \ldots, w_n)$ に変換する．$\mathcal{H}$ は逆変換関数 $\mathcal{H}^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ を持ち，任意の $V \in \mathbb{R}^n$ について $V = \mathcal{H}^{-1}(\mathcal{H}(V))$ が成立する．

$\mathcal{H}$ は Haar 分解 $\mathcal{H}_1 : \mathbb{R}^n \to \mathbb{R}^{n/2} \times \mathbb{R}^{n/2}$ を再帰的に k 回適用することにより構成できる．Haar 分解 $\mathcal{H}_1$ は，長さ 2^l のベクトル列 $Y = (y_1, y_2, \ldots, y_{2^l})$ を，長さ 2^{l-1} のベクトル列 cA, cD に分解する．

$$\mathcal{H}_1(Y) = cA \mid cD, \tag{6}$$

$$cA = (\frac{y_1 + y_2}{2}, \frac{y_3 + y_4}{2}, \ldots, \frac{y_{2^l-1} + y_{2^l}}{2}), \tag{7}$$

$$cD = (\frac{y_1 - y_2}{2}, \frac{y_3 - y_4}{2}, \ldots, \frac{y_{2^l-1} - y_{2^l}}{2}). \tag{8}$$

cA と cD はそれぞれ，Y において隣り合う 2 つの値の平均のベクトルと差分のベクトルである．cA を近似係数ベクトル，cD を詳細係数ベクトルと呼ぶ．

Haar 分解により生成された近似係数ベクトル cA を入力として，再び Haar 分解をほどこすと，長さ 2^{l-2} の近似係数ベクトルと詳細係数ベクトルの組が得られる．V を初期入力として，この分解を再帰的に k 回繰り返すと，最終的には k 個の詳細係数ベクトルと 1 個の近似係数ベクトルが得られる．これらの連接が HWT の出力 W となる．すなわち，$W = \mathcal{H}(V)$ は Haar 分解 $\mathcal{H}_1$ を用いて以下の式により定義される．

$$cA_0 = V, \tag{9}$$

$$cA_i \mid cD_i = \mathcal{H}_1(cA_{i-1}), \quad (i \in \{1..k\}) \tag{10}$$

$$W = (cA_k \mid cD_k \mid cD_{k-1} \mid \ldots \mid cD_1). \tag{11}$$

式 (9) の手順は，そのままナイーブに実装すると $O(n)$ の計算量となる．しかし，非 0 値を持つセルのみに着目するアルゴリズムを用いることにより，これを $O(km) = O(m \log n)$ に削減することができる．その具体的な構成法を附録 A.1 に示す．

3.2 Top-down 精緻化

HWT により得られた Wavelet 係数系列 W から，ϵ-差分プライバシーを満たし，かつ非負制約を充足する集計データ V^+ を得る．説明のため，まず $O(n)$ の計算量を持つアルゴリズム $\mathcal{T}_1$ を示し，その後に計算量を $O(m^+ \log n)$ に効率化した，$\mathcal{T}_1$ と出力等価なアルゴリズム $\mathcal{T}_2$ を示す．ここで，m^+ は出力 V^+ における非 0 値の個数である．

3.2.1 $\mathcal{T}_1 : O(n)$ の構成法

HWT において，近似係数ベクトル cA_i は，V における 2^i 個のセルの平均，すなわち部分和を 2^i で除したものであることに着目する．すなわち，cA_i の x 番目の要素を $cA_{i,x}$ としたとき，$\forall x \in \{1..2^i\}$ について，$cA_{i,x} \geq 0$ でなくてはならない．これは，HWT の性質により，

$$cA_{i+1,\lceil x/2 \rceil} \geq |cD_{i+1,\lceil x/2 \rceil}| \tag{12}$$

のとき成立し，$W = \mathcal{H}(V)$ は上式を満たす．しかし，W の各要素に Laplace ノイズを付加した系列 W^* では明らかにこの性質が維持されることは保証されない．これが，W^* にそのまま逆 HWT $\mathcal{H}^{-1}$ を適用する Privelet 法の出力が非負制約を逸脱する理由となる．逆に言えば，式 (12) を満たすよう W^* を精緻化することができれば，その逆 HWT 後の出力は非負制約を満たすことになる．

この性質を利用し，アルゴリズム $\mathcal{T}_1$ は W へのノイズ付加後に，式 (12) を満たすよう各係数を精緻化した後に逆 HWT を適用する．具体的には，W を入力として，以下の三段階の手順により差分プライバシーと非負制約を満たす出力 V^+ を得る．なお，下記手順において，$g(\cdot)$ は整数値を入力として以下の値をとる符号関数である．

$$g(x) = \begin{cases} +1 & (x = 1 \pmod 2), \\ -1 & (x = 0 \pmod 2). \end{cases} \tag{13}$$

(1) W の各要素に Laplace ノイズを付加することにより，ϵ-差分プライバシーを満たす係数系列 W^* を得る．これは，$\lambda = (1 + \log_2 n)/\epsilon$ として，下式で導出される．

$$\begin{aligned} cA^*_k &= cA_k + \mathrm{Lap}(\lambda/2^k), \\ cD^*_i &= cD_i + \mathrm{Lap}(\lambda/2^i)^{2^{k-i}} \quad (i \in \{1..k\}), \\ W^* &= (cA^*_k \mid cD^*_k \mid cD^*_{k-1} \mid \ldots \mid cD^*_1). \end{aligned}$$

(2) W^* において，各 Wavelet 係数に対応する部分和が非負制約を逸脱しないよう係数値を補正 (非負精緻化) し，精緻化済み Wavelet 係数系列 W^+ を得る．

$$cA^+_{i,x} = \begin{cases} \max(cA^*_{i,x}, 0) & (i = k), \\ cA^+_{i+1,\lceil x/2 \rceil} + g(x) \cdot cD^+_{i+1,\lceil x/2 \rceil} & (\text{otherwise}), \end{cases}$$

$$cD^+_{i,x} = \begin{cases} -cA^+_{i,x} & (cD^*_{i,x} < -cA^+_{i,x}), \\ cA^+_{i,x} & (cD^*_{i,x} > cA^+_{i,x}), \\ cD^*_{i,x} & (\text{otherwise}). \end{cases}$$

$$W^+ = (cA^+_k \mid cD^+_k \mid cD^+_{k-1} \mid \ldots \mid cD^+_1).$$

(3) W^+ に対して逆 HWT $\mathcal{H}^{-1}$ を適用し，出力 V^+ を得る．

$$\begin{aligned} cA^+_{i,x} &= cA^+_{i+1,\lceil x/2 \rceil} + g(x) \cdot cD^+_{i+1,\lceil x/2 \rceil}, \\ V^+ &= cA_0. \end{aligned}$$

3.2.2 $\mathcal{T}_2 : O(m^+ \log n)$ の構成法

$\mathcal{T}_1$ の計算量は $O(n)$ となる．これは，$\mathcal{T}_1$ を構成する 3 つの手順の計算量がいずれも $O(n)$ である [*4] ことによる．この計算量は Barak らの手法 [1] (n の多項式時間) より優れており，Xiao らの Privelet と同等であるが，前述の通り $m \ll n$ となる大規模な集計データでは実用的とは言えない．そこで，$\mathcal{T}_1$ と等価な出力を得つつ計算量を削減するアルゴリズム $\mathcal{T}_2$ を構成する．

$\mathcal{T}_2$ の構成にあたり，V が疎であるときには，ほとんどのノイズは手順 2 (非負精緻化) の過程で「捨てられる」ことに着目する．すなわち，$cD^*_{i,x}$ に精緻化が適用される ($cD^+_{i,x} \neq cD^*_{i,x}$ となる) とき，$cA^+_{i-1,2x-1}$ か $cA^+_{i-1,2x}$ のいずれかは必ず 0 となる．このとき，0 値をとったほうの部分木に含まれる，2^{i-1} 個の Laplace ノイズが出力値に影響する可能性はなく，したがって安全性にも寄与しない．

そこで，非 0 値をとる $cA^+_{i,x}$ のみを対象として，Laplace メカニズムの適用と非負精緻化を再帰降下により同時に実施する．これにより，出力に寄与しない無駄な Laplace ノイズを発生させることなしに，出力 V^+ の差分プライバシーを満たす．

(1) まず，最上位の Wavelet 係数である cA_k, cD_k に関し，それぞれ対応する cA^+_k, cD^+_k を計算する．

$$
\begin{aligned}
cA^*_k &= cA_k + \mathrm{Lap}(\lambda/2^k),\\
cD^*_k &= cD_k + \mathrm{Lap}(\lambda/2^k),\\
cA^+_{k,1} &= \max\{cA^*_{k,1}, 0\},\\
cD^+_{k,1} &= \begin{cases} -cA^+_{k,1} & (cD^*_{k,1} < -cA^+_{k,1}),\\ cA^+_{k,1} & (cD^*_{k,1} > cA^+_{k,1}),\\ cD^*_{k,1} & (\text{otherwise}). \end{cases}
\end{aligned}
$$

(2) $i = \{k, ..., 2\}$ について，$\forall(x \mid cA^+_{i,x} \neq 0)$ に対して下記を実行することにより，再帰的に cA^+_{i-1} と cD^+_{i-1} が得られ，最終的には cA^+_1 と cD^+_1 を得る．

$$
\begin{aligned}
cA^+_{i-1,2x-1} &= cA^+_{i,x} + cD^+_{i,x},\\
cA^+_{i-1,2x} &= cA^+_{i,x} - cD^+_{i,x},\\
p^* &= cD_{i-1,2x-1} + \mathrm{Lap}(\lambda/2^i),\\
q^* &= cD_{i-1,2x} + \mathrm{Lap}(\lambda/2^i),\\
cD^+_{i-1,2x-1} &= \begin{cases} -cA^+_{i-1,2x-1} & (p^* < -cA^+_{i-1,2x-1}),\\ cA^+_{i-1,2x-1} & (p^* > cA^+_{i-1,2x-1}),\\ p^* & (\text{otherwise}), \end{cases}\\
cD^+_{i-1,2x} &= \begin{cases} -cA^+_{i-1,2x} & (q^* < -cA^+_{i-1,2x}),\\ cA^+_{i-1,2x} & (q^* > cA^+_{i-1,2x}),\\ q^* & (\text{otherwise}). \end{cases}
\end{aligned}
$$

(3) $\forall(x \mid cA^+_{1,x} \neq 0)$ に対して下記を実行することにより，$V^+ = (v^+_0, v^+_1, ..., v^+_n)$ を得る．

$$v^+_{2x-1} = cA^+_{1,x} + cD^+_{1,x}, \tag{14}$$

$$v^+_{2x} = cA^+_{1,x} - cD^+_{1,x}. \tag{15}$$

3.2.3 $\mathcal{T}_1$ と $\mathcal{T}_2$ の等価性について

$\mathcal{T}_1$ と $\mathcal{T}_2$ は出力に関して等価である．すなわち，以下の補題が成立する．この証明は付録 A.2 で与える．

補題 1. 入力が同一であるとき，$\mathcal{T}_1$ と $\mathcal{T}_2$ の出力分布は等しい．すなわち，$\mathcal{T}_1, \mathcal{T}_2$ の出力空間を $\mathcal{R}$ としたとき，その任意の部分空間 $S(\subseteq \mathcal{R})$ について，下記が成立する．

$$\Pr[\mathcal{T}_1(\mathcal{H}(V)) \in S] = \Pr[\mathcal{T}_2(\mathcal{H}(V)) \in S]. \tag{16}$$

4. 提案方式の安全性と有用性

第 3 章で提案した方式 $\mathcal{T}_1, \mathcal{T}_2$ により得られた出力 $V^+ = (v^+_1, v^+_2, \ldots, v^+_n)$ は，安全性と有用性に関する以下の性質を満たす．

- (安全性) V^+ は ϵ-差分プライバシーを満たす．
- (非負制約の充足) V^+ は非負制約 $\forall v^+_i \geq 0$ を満たす．
- (部分和劣化の抑制) V^+ を 2^l ごとのブロックに分割したとき，その部分和に含まれるノイズの分散は，$\frac{2}{3}\lambda^2$ より小さい．すなわち，上記ブロックの部分和のノイズの大きさの上限は，ブロック長にかかわらず等しい．

さらに，構成法 $\mathcal{T}_2$ を用いた場合，提案方式は以下の性質を満たす．

- (計算量の抑制) m^+ を出力 V^+ における非 0 値の個数としたとき，$\mathcal{T}_2$ の計算量は $O(m^+ \log n)$ である．

以下，上記のそれぞれの性質について証明する．なお，定理 1〜3 については，$\mathcal{T}_1$ の出力に対する証明のみを示すが，補題 1 により $\mathcal{T}_2$ の出力に対してもそれぞれ成立する．

4.1 安全性

定理 1. V^+ は ϵ-差分プライバシーを満たす．

証明. $W = \mathcal{H}(V)$ は V の一次変換であるため，HWT を表す行列 H を用いた行列積により $W = HV$ と表せる．ここで，W, V は列ベクトルとして表現されているとする．すなわち，W^* は，n 次の対角行列 $E_W = \{e_{ij}\}$ を用いて Laplace ノイズをスケールさせることにより，

$$W^* = HV + E_W \mathrm{Lap}(\lambda)^n \tag{17}$$

と表せる．なお，$\{e_{ij}\}$ は以下の値をとる．

$$e_{ij} = \begin{cases} 0 & (i \neq j),\\ 2^{-k} & (i = j = 1),\\ 2^{-(k - \lceil \log_2 i \rceil) + 1} & (\text{otherwise}). \end{cases} \tag{18}$$

ここで両辺に E^{-1} を左から乗ずると，

[*4] 正確には，手順 3 の $\mathcal{H}^{-1}$ の計算は，附録 A.1 と同様の工夫により容易に $O(m^+ \log n)$ とすることができる．

$$E_W^{-1}W^* = E_W^{-1}HV + \mathrm{Lap}(\lambda)^n \qquad (19)$$

となり，これは Matrix メカニズム [7], [8] ($z = Ax + \mathrm{Lap}(|A|_1/\epsilon)$) の形をとる．

[7] によれば，$|E_W^{-1}H|_1 = 1 + \log_2 n$ となることから，$\lambda = (1+\log_2 n)/\epsilon$ とすることにより，$E_W^{-1}W^*$ は ϵ-差分プライバシーを満たす．ここで，E_W は V の情報を含まないことから，差分プライバシーの事後処理則 により，これに左側から E_W を乗じた $W^* = E_W \cdot E_W^{-1}W^*$ も，ϵ-差分プライバシーが保証される．

また，$\mathcal{T}_1$ において，W^* を生成した後の手順，すなわち W^* から W^+ および V^+ を導出する過程において，(ドメイン知識である非負制約を除いて) V の具体的な値に関する知識は用いられていない．すなわち，差分プライバシーの事後処理則の適用条件を満たす．したがって，$\mathcal{T}_1$ の出力である V^+ も ϵ-差分プライバシーが保証される． □

4.2 非負制約の充足

定理 2. V^+ は非負制約 $\forall v_i^+ \geq 0$ を満たす．

証明. $\mathcal{T}_1$ において，$i \in \{0..k-1\}$ のとき，

$$cA_{i,x}^+ = cA_{i+1,\lceil x/2 \rceil}^+ + g(x) \cdot cD_{i+1,\lceil x/2 \rceil}^+ \qquad (20)$$

であることから，

$$|cD_{i+1,\lceil x/2 \rceil}^+| \leq cA_{i+1,\lceil x/2 \rceil}^+ \qquad (21)$$

が満たされるならば，$cA_{i,x}^+ \geq 0$ が成立する．$V^+ = cA_0^+$ であるため，これは $\forall v_i^+ \geq 0$ の十分条件である．

すなわち，$cA_{k,1}^+ \geq 0$ であり，$i \in \{1..k\}$ において，任意の $cA_{i,x}^+$ と $cD_{i,x}^+$ の組に対し，$|cD_{i,x}^+| \leq cA_{i,x}^+$ が成立すれば良い．これらは $\mathcal{T}_1$ の構成法における手順 (2) により明らかに満たされる． □

4.3 部分和劣化の抑制

定理 3. V^+ を 2^l ごとのブロックに分割したとき，その部分和に含まれるノイズの分散は，$\frac{2}{3}\lambda^2$ より小さい．

証明. V^+ を 2^l ごとの q $(= 2^{k-l})$ 個のブロックに分割したときの部分和を $p_1^l, p_2^l, \ldots, p_q^l$ とする．HWT の性質により，$p_j^l = 2^l \cdot cA_{l,j}^+$ となる．$\mathcal{T}_1$ において付与されるノイズは互いに独立であるため，$cA_{l,j}^+$ のノイズの分散は，$cA_k^+, cD_k^+, cD_{k-1}^+, \ldots, cD_{l+1}^+$ に与えられるノイズの分散の和になる．

ここで，$cD_{i,x}^+$ に与えられるノイズの分布は，$\mathrm{Lap}(\lambda/2^i)$ の分布の両端 (具体的には $\pm cA_{i,x}^+$ の外側) を非負精緻化により「カット」した分布となる．したがって，その分散は $\mathrm{Var}(\mathrm{Lap}(\lambda/2^i)) = 2(\lambda/2^i)^2 = 2\lambda^2/4^i$ よりも小さい．

すなわち，$cA_{l,j}^+$ のノイズ分散の上限は，

$$\mathrm{Var}(\mathrm{Lap}(\lambda/2^k)) + \sum_{i=l+1}^{k} \mathrm{Var}(\mathrm{Lap}(\lambda/2^i)) \simeq \frac{2}{3}\lambda^2/2^l \quad (22)$$

で与えられる．$p_j = 2^l \cdot cA_{l,j}^+$ であるため，任意の l, j に対し，p_j^l のノイズの分散は $\frac{2}{3}\lambda^2$ を下回る． □

4.4 計算量の抑制

定理 4. $\mathcal{T}_2$ の計算量は $O(m^+ \log n)$ である．

証明. まず $\mathcal{T}_2$ における手順 (3) に着目する．$cA_{1,x}^+ \neq 0$ のとき，v_{2x-1}^+, v_{2x}^+ の少なくともいずれか片方は非 0 値をとる．そのため，cA_1^+ に含まれる非 0 値の個数は高々 m^+ 個である．したがって，手順 (3) の計算量は $O(m^+)$ となる．

次に，手順 (2) に着目する．同様に，$cA_{i,x}^+ \neq 0$ のときに $cA_{i+1,2x-1}^+, cA_{i+1,2x}^+$ の少なくともいずれかは非 0 値をとることから，cA_i^+ に含まれる非 0 値の個数は，高々 cA_{i+1}^+ に含まれる非 0 値の個数となり，m^+ を超えることはない．したがって，手順 (2) は $O(m^+)$ の処理を $k-1$ 回実行することになるため，その計算量は $O(m^+k) = O(m^+ \log n)$ となる．

また，手順 (1) の計算量は明らかに $O(1)$ のため，$\mathcal{T}_2$ の計算量は，$O(m^+) + O(m^+ \log n) + O(1) = O(m^+ \log n)$ となる． □

5. 地理空間データへの適用と評価

提案方式の実問題への適用例として，地理空間データ (geospatial data) への適用を考える．地理空間データとは，人口分布や降雨量分布，交通量分布など，地理的な場所に応じて変化する「量」を表すデータである．

地理空間データは一般的に二次元に配列されたデータであるため，提案方式への適用にあたっては，提案方式を二次元の入力に対応できるよう拡張するか，二次元のデータを一次元のベクトルに次元変換する必要がある．本稿では，データを一次元に次元変換する適用方式について述べる．

変換にあたり，地理空間データの分析の際に頻繁に用いられる，正方ブロック領域に関する部分和の精度に着目する．本稿で示す方式は，地理空間上の (一辺が 2^k となる) 正方ブロック領域が，一次元の連続した領域に再配置されることを保証することにより，正方ブロック領域の部分和精度を向上させる．

また，上記の次元変換方式を用い，提案方式から得られる部分和の精度について評価した結果を示す．評価対象データとして，実世界に基づく地理空間データの一つである H22 国勢調査による地域メッシュ人口を用い，Privelet 法および Laplace メカニズムとの比較を通じて提案方式の改善効果を定量的に評価する．

5.1 正方ブロックの連続領域化を保証する次元変換

地理空間データなどの二次元空間上のデータを，提案方式に適用するために一次元ベクトルへ次元変換する方式について述べる．

単純な方法としては，二次元データを縦方向 (緯度方向) もしくは横方向 (経度方向) に順に走査していくことにより一次元に再配列し，これを提案方式の入力 V として用いることが考えられる．この方法では，V における連続した領域が，二次元空間上の「縦一列 (もしくは横一列)」の領域に対応することになる．

しかし，実際の地理空間データの応用においては，一般にこのような部分和が用いられることはほとんどない．500m メッシュ人口における 2×2 の 4 セルの部分和を 1km メッシュ人口として用いるなど，多くの場合は正方メッシュ領域の部分和が用いられる．したがって，このような単純な変換方法では，実際に応用分野で利用される部分和の精度が悪化する．

そこで，地理空間データの一次元ベクトルへの次元変換にあたり，二次元配列上における $2^K \times 2^K$ $(K \in \mathbb{N})$ の正方ブロックを，一次元ベクトル上の長さ 2^{2K} の連続領域に写像する再配列関数 $\mathcal{F} : \mathbb{R}^{2^n} \times \mathbb{R}^{2^n} \to \mathbb{R}^{2^{2n}}$ $(n \in \mathbb{N})$ を用いる方法を提案する．

以下，$\mathcal{F}$ の具体的な構成方法を示す．$V = \mathcal{F}(M)$ において，再配列関数 $\mathcal{F}$ は，2^n 行 2^n 列の行列からなる入力 M の i 行 j 列の要素 m_{ij} を，長さ 2^{2n} の出力 V における k 番目の要素 v_k に写像する (簡単のため，i, j, k はいずれも 0 から始まるとする)．このとき，k は以下の二進数表現により得られる．

$$k = (i_n \; j_n \; i_{n-1} \; j_{n-1} \; \cdots \; i_1 \; j_1)_2. \tag{23}$$

ここで，i_x, j_x はそれぞれ，i, j の二進数表現における下位から x bit 目の値を表す．すなわち，i, j はそれぞれ以下のように表される．

$$i = (i_n \; i_{n-1} \; \cdots \; i_1)_2, \quad j = (j_n \; j_{n-1} \; \cdots \; j_1)_2.$$

なお，提案方式の出力 V^+ を二次元空間に戻すには，再配列関数 $\mathcal{F}$ の逆関数 $\mathcal{F}^{-1} : \mathbb{R}^{2^{2n}} \to \mathbb{R}^{2^n} \times \mathbb{R}^{2^n}$ を用いる．$\mathcal{F}^{-1}$ は，式 (23) を逆方向に適用することにより容易に得ることができる．

5.2 評価

本評価では，地理空間データの例として，H22 国勢調査に基づく地域メッシュ統計 [14] の 500m メッシュ人口 (1/2 地域メッシュ人口) から，首都圏周辺の 256km 四方 $(n = 512 \times 512 = 2^{18})$ を抽出したものを評価対象データとして用いる (以下，対象データと呼ぶ)．

対象データに再配列関数 $\mathcal{F}$ を適用して得られたデータに対し，Privelet 法と提案方式をそれぞれ適用し，出力におけるノイズの大きさ (対象データに対する誤差) を評価する．評価にあたり，部分和の範囲の大きさを変動させ，誤差との関係を見る．評価指標としては MAE (mean absolute error) および RMSE (rooted mean squared error) を用いる．また，比較のため Laplace メカニズム $V + \mathrm{Lap}(1/\epsilon)^n$ における RMSE の理論値を併せて示す．なお，それぞれにおいて，$\epsilon = 0.1$ と設定した．

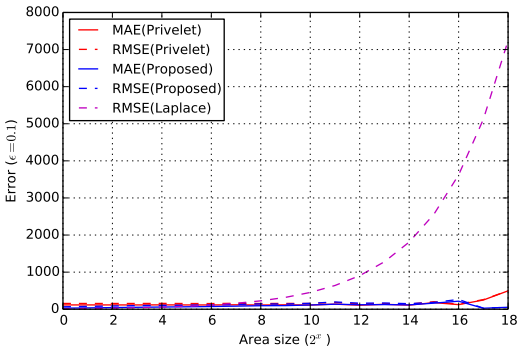

図 1 部分和の領域サイズと誤差の関係

定理 3 によれば，提案方式の有用性は Xiao らの Privelet 法と同等か，条件によってはそれ以上に改善されることが期待される．特に，人口分布などのいわゆる「自然な」集計データ，すなわちロングテイル性を持ち，0 値や小さい値を持つ v_i が多数現れるようなデータにおいて，非負精緻化によるノイズ分布の「カット効果」は大きくなる．そのため，これらのデータにおいてはノイズがより小さくなる傾向を持つことが期待される．

図 1 に評価の結果を示す．x 軸は部分和の範囲の大きさ (セル数) を 2 の対数で示し，y 軸はそのときの誤差の大きさを示している．たとえば，グラフ上の $x = 6$ の点は，$2^6 = 64$ セルの部分和，すなわち 8×8 メッシュ (4km メッシュ相当) の正方領域に含まれる人口の合算値の誤差を表す．

同図により，Laplace メカニズムは領域サイズの二乗根に比例して RMSE が増大している (部分和精度の劣化が発生している) のに対し，提案方式と Privelet では部分和の領域サイズが大きくなっても誤差の増大が抑えこまれていることが確認できる．

提案方式と Privelet の比較のため，**図 2** に図 1 のうち $y \leq 600$ の部分を拡大したものを示す．$x \leq 10$，すなわち $2^{10} = 32 \times 32$ メッシュ (16km メッシュ相当) 以下の大きさの部分和において，提案方式は Privelet に対して精度が大きく改善しており，部分和サイズが小さいほど改善効果が大きいことが確認できる．

6. まとめ

本稿では，地理空間データなどの大規模な集計データのプライバシーを差分プライバシー基準に基づいて保護する

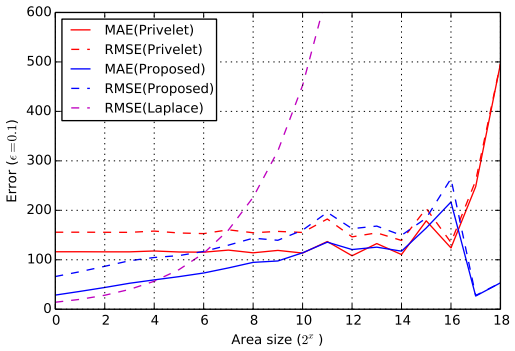

図 2 部分和の領域サイズと誤差の関係 ($y \leq 600$ 部分を拡大)

上で，データの統計的正確性と計算効率の向上に着目した手法を提案した．差分プライバシーは，数学的な安全性保証が与えられているという優れた特徴を持つ一方で，大規模な集計データの公開におけるプライバシー保護に適用するためには，(1) 非負制約の逸脱，(2) 部分和精度の劣化，(3) 計算量の増大，という 3 つの課題を解決する必要があることを示した．

上記問題を解決するために，Wavelet 変換と Top-down 精緻化と呼ぶ手法を導入する方式を提案した．提案方式は，差分プライバシーを満たしつつ，上記 3 点の課題を同時に解決することを証明した．具体的には，部分和精度に関し，提案方式は集計データを 2^l 個のブロックに分割した部分和において，l の大きさに関わらず (部分和の範囲の大きさにかかわらず) その精度を一定以上に保つことが保証され，さらに集計データが疎であるほどその精度は向上する．また，計算量に関し，Laplace メカニズムや Privelet 法の計算量が $O(n)$ であるのに対し，提案方式の計算量は $O(m^+ \log n)$ である．ここで，n はセル空間全体の大きさ，m^+ は非 0 値を持つ出力セルの数である．したがって，$m^+ \ll n$ となるような疎な集計データにおいて，提案方式は計算効率に優れる．

さらに，提案方式を二次元の地理空間データに適用するための次元変換方式を示した．また，国勢調査によるメッシュ人口に基づくサンプルデータセットへの適用を通じた比較評価により，提案方式を適用した集計データは，Privelet と同等以上に部分和の精度を向上させ，特に小範囲の部分和について Privelet と比較して大きく精度に優れることが明らかとなった．

参考文献

[1] Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., Talwar, K. and Berkeley, U. C.: 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*, ACM Press, pp. 273–282 (2007).

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

[3] Dwork, C.: Differential Privacy, *Proc. 33rd intl. conf. Automata, Languages and Programming - Volume Part II* (Bugliesi, M., Preneel, B., Sassone, V. and Wegener, I., eds.), Lecture Notes in Computer Science, Vol. 4052, Springer, pp. 1–12 (2006).

[4] Dwork, C.: An ad omnia approach to defining and achieving private data analysis, *Proc. 1st ACM SIGKDD intl. conf. Privacy, security, and trust in KDD*, Springer-Verlag, pp. 1–13 (2007).

[5] Dwork, C.: Differential privacy: a survey of results, *Proc. 5th intl. conf. Theory and applications of models of computation*, Springer-Verlag, pp. 1–19 (2008).

[6] Ghosh, A., Roughgarden, T. and Sundararajan, M.: Universally Utility-maximizing Privacy Mechanisms, *SIAM J. Computing*, Vol. 41, No. 6, pp. 1673–1693 (2012).

[7] 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, VLDB Endowment, pp. 1021–1032 (2010).

[8] Li, C., Hay, M., Rastogi, V., Miklau, G. and McGregor, A.: Optimizing linear counting queries under differential privacy, *Proc. 29th ACM SIGMOD-SIGACT-SIGART symp. Principles of database systems of data (PODS '10)*, New York, New York, USA, ACM Press, pp. 123–134 (2010).

[9] Stanimirovic, I. P. and Tasic, M. B.: Performance comparison of storage formats for sparse matrices, *Ser. Mathematics and Informatics*, Vol. 24, No. 1, pp. 39–51 (2009).

[10] Sweeney, L.: k-anonimity: a model for protecting privacy, *Intl. J. Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570 (2002).

[11] Xiao, X., Wang, G. and Gehrke, J.: Differential privacy via wavelet transforms, *Proc. 26th intl. conf. Data Engineering (ICDE 2010)*, IEEE, pp. 225–236 (2010).

[12] 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).

[13] 寺田 雅之，竹内 大二朗，齊藤 克哉，本郷節之：差分プライバシー基準に基づく情報秘匿手法の一考察，マルチメディア，分散，協調とモバイル (DICOMO2014) シンポジウム論文集，pp. 224–233 (2014).

[14] 総務省 統計局：地域メッシュ統計の特質・沿革.

付　　録

A.1 Haar Wavelet 変換の計算量削減

第 3.1 節の式 (9) の計算量を $O(km) = O(m \log n)$ に削減する構成法を示す．

V は COO 形式などの疎データ形式で入力されるとする．$\forall v_i (v_i \neq 0)$ に対し，順番に，

$$cA_{1,\lceil i/2 \rceil} := cA_{1,\lceil i/2 \rceil} + v_i, \tag{A.1}$$

$$cD_{1,\lceil i/2 \rceil} := cA_{1,\lceil i/2 \rceil} + g(i) \cdot v_i. \tag{A.2}$$

を計算することにより Haar 分解 $\mathcal{H}_1$ を実現する．ここで，$g(\cdot)$ は式 (13) で与えられる符号関数である．なお，cA_1, cD_1 もそれぞれ疎データ形式で保持するものとし，その初期値はいずれも $cA_1 = cD_1 = \{0\}^{n/2}$ とする．

この手順による Haar 分解の計算量は明らかに $O(m)$ である．これを再帰的に cA_k, cD_k まで繰り返せば，$k = log_2 n$ 回の Haar 分解により W を得ることができる．すなわち，この構成法に基づく HWT 全体の計算量は $O(m \log n)$ となる．

A.2 補題 1 の証明

$\mathcal{T}_1, \mathcal{T}_2$ において，cA_1^+ および cD_1^+ の分布がそれぞれ等しいなら，それぞれの出力分布も等しくなり，補題 1 が成立する．そこで，$\forall i \in \{1..k\}$ において，cA_i^+ および cD_i^+ が $\mathcal{T}_1, \mathcal{T}_2$ のいずれにおいても等しい分布を持つことを数学的帰納法で示す．

証明. まず，$i = k$ のとき，すなわち cA_k^+ と cD_k^+ については，それぞれの定義から明らかに $\mathcal{T}_1, \mathcal{T}_2$ で同一である．

次に，$1 \leq i < k$ のときを考える．$\mathcal{T}_1$ の手順 2 における $cA_{i,x}^+$ の導出式において，$j = i + 1, y = \lceil x/2 \rceil$ と置換すると，これは $\mathcal{T}_2$ の手順 2 における導出式と等価となる．すなわち，cA_{i+1}^+ および cD_{i+1}^+ が $\mathcal{T}_1, \mathcal{T}_2$ でそれぞれ等しい分布をとるならば，cA_i^+ の分布も等しい．

また，cD_i について，$cA_{i,x}^+ = 0$ のときは，$\mathcal{T}_1, \mathcal{T}_2$ のいずれにおいても $cD_{i,x}^+ = 0$ となる．$cA_{i,x}^+ \neq 0$ のとき，$\mathcal{T}_2$ の手順 2 における $cD_{i-1,2x-1}^+$ の導出において，$j = i-1, y = 2x-1$ と置換すると，$\mathcal{T}_1$ の手順 2 における $cD_{i,x}^+$ の導出と等価な式となる．これは $cD_{i-1,2x}^+$ の導出においても同様である．したがって，cA_i^+ の分布が $\mathcal{T}_1, \mathcal{T}_2$ で等しいならば，cD_i^+ がとる分布も等しい．

すなわち，$\mathcal{T}_1, \mathcal{T}_2$ において，(1) $i = k$ のとき，cA_i^+ および cD_i^+ の分布はそれぞれ等価である．(2) $1 \leq i < k$ のとき，cA_{i+1}^+ および cD_{i+1}^+ の分布が等価であるならば，cA_i^+ の分布も等価であり，(3) cA_i^+ の分布が等価であるならば，cD_i^+ の分布も等価である．したがって，cA_1^+ および cD_1^+ は $\mathcal{T}_1, \mathcal{T}_2$ で等しい分布を持つ． □