7G-06

集計データへの差分プライバシ適用における特性の一考察 IV

本郷 節之† 大加瀬 稔† 寺田 雅之‡ 稲垣 潤† 鈴木 昭弘†
北海道科学大学† 株式会社ＮＴＴドコモ‡

1 はじめに

プライバシ保護とデータの有効活用の両立を可能にする技術が注目を集めている．とりわけ，Dwork らが提案した差分プライバシ基準[1] は，高い安全性を保障する．しかし，データの有用性や処理効率において実用性に課題を有している．Xiao らが提案した Privelet 法[2] は，有用性の一要素である“部分和精度の劣化”の改善を実現した．この手法では，データに対して Wavelet 変換を施し，得られた Wavelet 係数に対して Laplace 摂動[1]を加えることで，差分プライバシ基準に従う秘匿を実現している．けれども，この Privelet 法を用いても，例えば人口の空間分布のような，非負の，かつ，疎な（ゼロ値が多い）データ分布に対しては，“非負制約の逸脱”，“スパース（疎）なデータのデンス（密）化”といった課題が残されていた．

我々は上記ふたつの問題を解決するために，この Privelet 法に対して非負精緻化[3] と呼ばれる処理を導入した．非負精緻化を伴う Privelet 法は，“秘匿処理後のデータに負値が存在しない”，“スパースなデータがデンスになる事態を抑制できる”という特徴をもつことが，理論的に証明されている[3]．

我々は以前，この非負精緻化を伴う Privelet 法に“枝刈り”と呼ばれる処理を適用することにより，著しい演算量の削減を実現できることを実証した[4]．しかしこの演算量削減効果が，この枝刈り処理の発生メカニズムとどのような関係にあるのかは，未だ明らかになっていない．

本稿では，枝刈り処理の発生位置と，演算量削減効果との関係に対する仮説を立てて，評価用に作成したデータを用いた検証を行う．さらに，複数のエリアにおける人口分布データを対象に，実データにおいても，枝刈り処理が同様の効果を有することを確認する．

A Study on Characteristics of Publishing Tabular Data with Differential Privacy IV
† Hokkaido University of Science
‡ NTT DOCOMO Inc.

2 非負精緻化を伴う Privelet 法の枝刈り

図１に枝刈りによる演算量削減の様子を示す．ここでは二分木構造をもつ Haar Wavelet を対象に考える．Privelet 法[2]では，秘匿対象である長さ2^Hの一次元データベクトルに対して Haar 分解を再帰的に施し，ひとつの近似係数$cA_{H,x}$と2^H-1個の詳細係数$cD_{h,x}$を求め（順変換）た上で摂動を加え，摂動が付加された詳細係数$cD^*_{h,x}$と一段上層の演算により求められた近似係数$cA^*_{h,x}$とから，一段下層の近似係数$cA^*_{h-1,2x}$および$cA^*_{h-1,2x+1}$を求める演算を繰り返す（逆変換）ことで，秘匿された一次元データベクトルを生成する．ここでcDは詳細係数を，cAは近似係数を，Hは二分木の階層数を，$0<h\leq$Hは階層番号を，$0\leq x<2^{H-h}$は階層内ノード番号を，そして，$*$は摂動が付加された値であることをそれぞれ表している．

さらに，非負精緻化を伴う Privelet 法[3]の場合には，逆変換の過程で算出された一段下層の近似係数$cA^*_{h-1,x}$の値が負値であった場合に，詳細係数$cD^*_{h,x}$の値を調整し，$cA^*_{h-1,x}$の値を非負値へと精緻化する．このとき，算出されたふたつの近似係数$cA^*_{h-1,2x}$または$cA^*_{h-1,2x+1}$のいずれかの値は0となる．さらに，近似係数の値が0となった側のノードの配下にあるノードの近似係数の値は全て0となることから，この部分の演算を全て省略してしまう（枝刈り）ことができる．これが，非負精緻化を伴う Privelet 法の演算量削減手法[4]である．

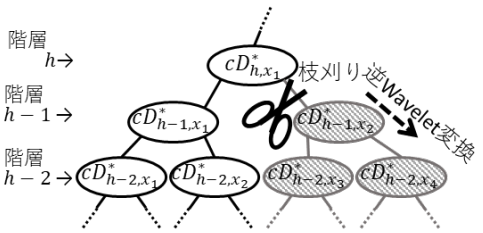

図１　枝刈りによる演算量削減の様子

3 仮説

上記の枝刈り手法を前提とすると，①枝刈りが起こったノード(h,x)では左右いずれかひとつの近似係数演算が省略され，②以下一階層下る

ごとに2倍の演算量が削減されることが想定される．すると，階層hを起点とする枝刈りの発生数をs_hとするとき，枝刈りにより省略される近似係数演算の総数(ここでは“重み付き枝刈り発生回数”と呼ぶ)の値S_wは次式によって求められる．

$$S_w = \sum_{h=1}^{H}\left(s_h \cdot \sum_{i=1}^{h} 2^{h-i}\right) \quad (式1)$$

4 検証

図2に検証実験に使用した評価用データを示す．256×256の二次元データを16のブロックに分けて，0値エリアの比率が異なる(c)～(g)のような5種類のパターンを用意し，ラスター方式，ソート方式，Morton 方式，ランダム方式[4]により変換して得られた一次元データベクトルを用いた．各方式とも摂動値を変えて3回ずつ計測を行った．評価結果では，ゼロ値比率の高いパターンほど重み付き枝刈り発生回数が多くなる．評価には Intel Core i7-875K CPU (2.93GHz)，実装メモリ 4GB のデスクトップ PC を使用した．

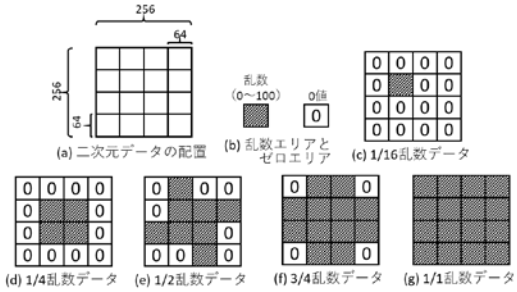

図2　評価用データの説明図

図3に評価結果を示す．評価結果から，(式1)で求めた“重み付き枝刈り発生回数”に比例して演算時間が短縮されていることがわかる．これにより，各階層における近似係数の演算省略によって演算時間の短縮が実現されていることが実証された．

5 考察

平成22 年度国勢調査に基づく地域メッシュ人口（1km メッシュ）データから，北海道，四国，関東（ゼロ値比率の高い準）エリアを切り出したデータに対しても同様の特性が見られるかについて評価を試みた．

図4に評価結果を示す．人口統計データにおいても，人工的に作成した評価用データの場合と同様に，“重み付き枝刈り発生回数”に比例して演算時間が短縮されている．

6 おわりに

各階層における近似係数の演算省略によって演算時間が短縮されることが実証された．今後は次元変換部分の効率化が課題と考える．

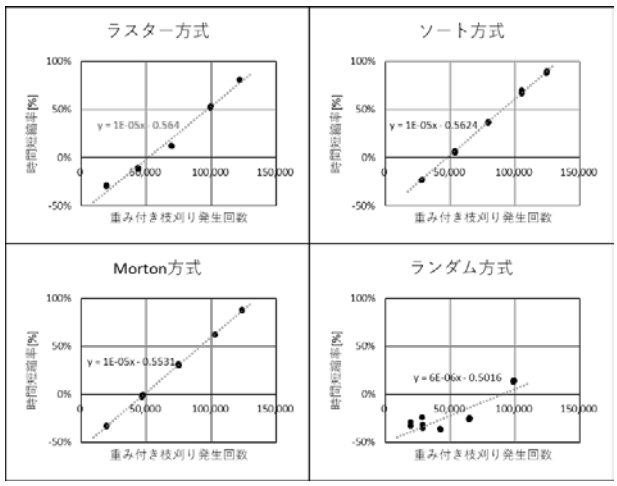

図3　評価用データによる評価結果

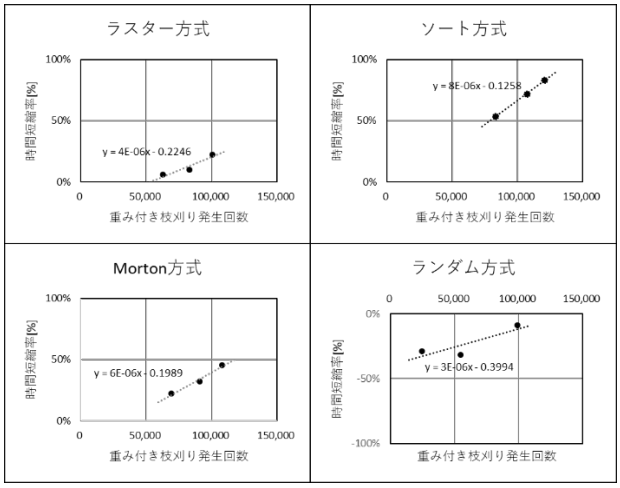

図4　人口統計データによる評価結果

謝辞

本研究の一部は日本学術振興会科学研究費補助金基盤研究(C)（課題番号：15K00190）による補助を受けて行なわれた．

参考文献

[1] Barak, B., et. al.: Privacy, accuracy, and consistency too: a holistic solution to contingency table release, Proc. PODS ’07, ACM Press, 273–282 (2007).
[2] Xiao, X., et. al.: Differential Privacy via Wavelet Transforms, IEEE Trans. Knowledge and Data Engineering, 23, 8, 1200–1214 (2011).
[3] 寺田，鈴木，山口，本郷: 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, 56, 9,1801–1816 (2015).
[4] 本郷，手塚，寺田，稲垣：Top-down 精緻化を伴う Privelet 法における演算効率化手法の検討, DICOMO2018，pp. 460-466 (2018).