Computer Security Symposium 2018
22 - 25 October 2018

Locally Private Continual Countingにおける1-Shot Reportingメカニズムの有用性解析

小野 元[1]　福地 一斗[2]　秋本 洋平[1]　佐久間 淳[1,2,3]

概要：オンラインサービス提供者は，利用者の情報をリアルタイムに解析することで迅速なシステム最適化が可能になる．しかし，情報の収集は利用者のプイラバシーを侵害する恐れがある．そこで我々は局所差分プライバシーによる利用者のプライバシー保護を行う．プイラバシーを保護したリアルタイム情報集約問題を locally private continual counting として定式化する．この問題ではラウンドが 1 から T まであり，ユーザー N 人と収集者がいる．ユーザーらはそれぞれ毎ラウンド 0 または 1 の状態を持ち，ユーザーらは毎ラウンドプライバシー保護下で状態を送信する．収集者は毎ラウンドの 1 を持つユーザーの割合を推測する．この問題の解法として我々は Glance を提案し，適用可能な提案法との比較を行った．その結果，特定のケースにおいて提案法の誤差の上界のレートが既存法の誤差の下界のレートよりも優位であることを示した．

キーワード：局所差分プライバシー，時系列データ

1. Introduction

Continual Couting とは，リアルタイム情報集約を定式化した以下のような問題である．ラウンドが 1 から T まであり，N 人のユーザーと 1 人のデータ収集者がいる．各ユーザー n は各ラウンドにおいて 0 または 1 の状態をもち，収集者に送信する．収集者は各ラウンドにおいて状態が 1 であるユーザーの割合をリアルタイムで推定する．ここでいうリアルタイムで推定するとは，あるラウンド t における割合を，ラウンド $t+1$ の報告を観測する前に推定する，ということである．この問題において推定の誤差は，全てのラウンドの中での推定値と実際の比の差の最大値であると定義される．この誤差は容易に 0 にできることは明らかである．Continual Counting の例として次のような状況が考えられる：電力会社が電力需要の予測のために，ある地域の毎時刻の在宅率の調査を行う．オンラインサービス提供者がサーバの起動数を決定するために，毎分のユーザの接続率を計測する．

Continual Counting を含め，ユーザーから情報を収集することは，サービスの向上に役立つが，ユーザ・収集者双方にリスクが伴う．ユーザーは情報の提供によってプライバシーを暴かれる危険性がある．また，収集者もユーザーから集めたデータが外部に流出させてしまった場合，社会からの批判に晒されることになる．ユーザーの情報をプライバシーに配慮せずに 収集することは，ユーザー・収集者双方にとって好ましくない．

データ収集シナリオにおいてデータ提供者のプライバシーがどの程度保護されているかを測る指標として局所差分プライバシー (Local Differential Privacy，LDP)[11] が近年注目されている．LDP はデータを収集してそこから統計的な情報を得るというシナリオにおいて，情報提供者のプライバシーが情報収集者にどのくらい確信されないかを定量的に表現する．LDP におけるプライバシー保護は提供するデータをランダマイズしてから送信することによって達成される．ランダマイズを行うという性質上，プライバシー保護と収集されるデータの有用性の間にはトレードオフの関係が生じる．そのため，プライバシー保護と有用性のよりよいトレードオフをいかにして達成するかが問題になる．

LDP ではデータ提供者のプライバシーがどの程度保護されているかを，プライバシーパラメータ $\epsilon > 0$ で表現し，ϵ が小さいほど強力なプライバシー保護を意味する．ϵ が小さいほどデータに強力なランダマイズを行うことになる．ϵ は問題ごとに決定される値であり，プライバシー保護・データの有用性を考慮して決定される．

本研究ではプライバシーを保護しながらリアルタイムで

1　筑波大学システム情報工学研究科
2　理化学研究所　革新知能統合研究センター
3　JST CREST

データを効率よく集計するために，Locally Private Continual Counting(LPCC) を考える．LPCC はユーザーに対して ϵ-LDP が保証された Continual Counting である．LDP によるユーザーのプライバシー保護を行うため，プライバシー保護と推定の正確性の間にトレードオフの関係が生じる．LPCC では Continual Couting と違い，アルゴリズムにランダム性が求められるため，決定的に誤差を 0 にすることは不可能である．そのため，LPCC のアルゴリズムは高い確率で誤差の小さい予測をすることが目的になる．

LPCC の誤差はプライバシー保護の度合い (ϵ の大きさ) によってコントロールされるはずであり，我々は誤差と ϵ の大きさの関係に注目してアルゴリズムの誤差の解析を行う．$\epsilon \to +\infty$ の場合，LPCC は Continual Counting と等しくなり，誤差を決定的に 0 にできるはずである．したがって，直感的には，ϵ が十分に大きければ T や N といった他のパラメータによらず，誤差を小さくすることが可能であるように思われる．LDP 研究では，(local でない)DP 研究に比べて，大きな ϵ が想定される場合が多いため，ϵ と誤差の関係はより重要である．

プラバシー保護の度合いは LDP アルゴリズムの研究や実践では，DP アルゴリズムでの ϵ より大きいことが多く，我々もプライバシーパラメータ ϵ が比較的大きい ($\epsilon > 1$) 領域に着目する．理論的な (局所でない) 差分プライバシーの研究では $\epsilon < 0.1$ を想定する場合が多い一方で，実践的な局所差分プライバシーのアルゴリズム研究などでは $\epsilon > 1$ の設定も珍しくない．実際に運用されている LDP メカニズムの例として，iOS ユーザーは毎日 $\epsilon = 16$ で Apple 社に端末の利用情報を送信している [1], [14]．また，BLENDER[2] の実験では，AOL データ (519,371 ユーザー，4,811,646 クエリ) から最頻 50 検索語のリストを作成するために，$\epsilon = 3$ を必要としている．

我々の目標は緩いプライバシー保護下 ($1 \leq \epsilon \leq 10$ 程度) で，よりよいトレードオフを達成するメカニズムを開発すること，あるいは開発したメカニズムが他の手法と比較して，よりよいトレードオフを達成する条件を明らかにすることである．また，一般には誤差を 0 にできなくとも，特殊な条件を仮定することで誤差を改善できるかどうかを理論的に解析し，できる場合にはどの程度改善できるか解析する．

1.1 Related Work

Private heavy hitters という問題がある．N 人のユーザーと 1 人のデータ収集者がおり，それぞれのユーザー $n \in \{1, 2, \ldots, N\}$ はアイテム集合 $\mathcal{V}$ の中からいずれかのアイテムをもつ．このとき，収集者はユーザーらが持つアイテムのうち出現頻度の高い k 種類を特定し，それらの出現頻度を推定することことを目的とする．一方，ユーザーは局所差分プライバシー保障下で自分のアイテムの情報を収集者に提供することを目的とする．この問題はアイテムインデックスをラウンドと読みかえることで LPCC と同じ問題に帰着できるが，この問題を解くためのアルゴリズムの多くは有用性を高めるために Bloom filter[8], [13] や Johnson-Lindenstrauss 変換 [3], [9], [13], [15] を用いて入力を圧縮するため，LPCC のようにリアルタイムでの推定ができない．

Basic One-time RAPPOR (BORAPPOR) [8] は入力を圧縮しないため LPCC に適用可能な private heavy hitters アルゴリズムである．それぞれのユーザーはプライバシーバジェットを ϵ/T ずつに分割して，それぞれのラウンドで Randomized Response[16] でそのラウンドにおける自分の状態を送信する．LPCC における BORAPPOR の誤差はラウンド数 T に対して $O(T\sqrt{\log T})$ であり，後に導入する Harmony や提案法と比較すると，ラウンド数 T に対して，誤差のオーダーが大きい．

Harmony[12] は private heavy hitters を拡張し，収集したデータを頻出アイテムの推定や頻度の推定だけでなく他の用途にも使えるように，データを圧縮しないため，LPCC にも適用可能なアルゴリズムである．LPCC に適用した場合，それぞれのユーザーは T ラウンドから一様ランダムに 1 ラウンドを選択し，そのラウンドにおける自分の状態を Randomized Response[16] で収集者に送信する．LPCC における Harmony の誤差はラウンド数 T に対して $O(\sqrt{T \log T})$ であり，BORAPPOR と比較して，オーダーが小さい．プライバシーパラメータ ϵ に関しては，$\epsilon \to +0$ では $O(1/\epsilon)$ である．著者らは [17] において $\epsilon \to +\infty$ で $O(1)$ であることが指摘し，十分大きな ϵ が利用可能でも誤差を小さくできない可能性を示唆した．

差分プライバシー保護下での Continual Observation[4], [7] は LPCC とは異なる問題である．この問題におけるプライバシー定義は (局所ではない) 差分プライバシーである．また，この問題では 0, 1 で構成される一本の時系列データに対して，各ラウンド t までに何個の 1 が出現したか推定する問題である．対して，LPCC は N 人のユーザーがそれぞれ時系列データを持ち，各ラウンド t において 1 を持つユーザーの割合を推定する問題である．

メモ化 [5], [8] は同じデータを繰り返し送信してもプライバシーを保護できる手法だが，LPCC において一つの記録は 1 度しか収集されないため本研究では用いない．メモ化が想定するデータ収集シナリオでは同じ情報が何度も送信されうる．この状況においては何度も同じ情報を送信させることで高い確率で真の情報がわかってしまう．メモ化ではそれを防ぐために，真の情報をランダマイズしてから保存し，保存したその情報をさらにランダマイズして収集者に送信する．こうすることによって，何度も同じ情報を送信してもランダマイズされた値しか収集者にはわからない．

表 1: 誤差の ϵ への依存性

手法	ケース	$\epsilon \to +0$	$\epsilon \to +\infty$
提案法	一般	$O(1/\epsilon)$	$O(1)$
	密または疎	$O(1/\epsilon)$	$O(1/\sqrt{\exp(\epsilon)})$
Harmony[12]	一般	$O(1/\epsilon)$	$\Theta(1)$
	密または疎	$O(1/\epsilon)$	$\Theta(1)$

1.2 Our Contribution

我々は LPCC のアルゴリズムとして，Randomized Response[16] とサンプリングを用いる Glance を提案する．Randomized Response はユーザーのプライバシーを保護するために採用する．サンプリングは誤差を小さくするために採用する．サンプリングを用いた誤差の改善は LDP メカニズムとしてはよく用いられている ([3], [12] など)．

μ_t に特に仮定を置かない場合を**一般ケース**として，一般ケースでの Glance の誤差と Harmony の誤差のオーダーの比較を行った．その結果，一般ケースでは Glance の誤差と Harmony の誤差の上界のオーダーは同じであることがわかった．また，Harmony の誤差の下界のオーダーも解析し，log 項を無視すれば Glance の誤差の上界のオーダーと一致することを明らかにした．すなわち，一般ケースにおいても Glance の誤差は Harmony の誤差と高々同程度でしかない．

また，我々は Glance の誤差の上界が Harmony の誤差の下界よりも小さくなる場合があることを発見し，その場合を**密または疎ケース**と名付けた．このケースにおいては，Glance と Harmony は ϵ に対してまったく異なる挙動を示す．他のパラメータを固定したとき，Harmony の誤差はどんなに大きな ϵ を与えられても正の定数に漸近する．一方，Glance の誤差は，ϵ が十分に大きいとき，ϵ の増加に対して指数的に小さくなる．

表 1 は誤差の ϵ への依存性の解析結果のまとめである．一般ケースにおいて，提案法と Harmony の誤差の ϵ への依存性は一致している．すなわち，一般ケースにおいて保証できる誤差への ϵ の影響は 2 つのアルゴリズムで大きな差がないことを意味する．一方，密または疎ケースにおいては提案法の ϵ への依存性は $\epsilon \to +\infty$ で $O(1/\sqrt{e^\epsilon})$ であり，Harmony の $\Theta(1)$ よりもオーダーが指数的に小さい．

1.3 Notation

本論文で用いる主な記法を表 2 にまとめる．特に，簡単のために 1 から始まる連続した自然数の集合 $\{1, ..., K\}(K$ はある自然数) を $[K]$ と略記することに注意する．

2. Preliminaries

この章では局所差分プライバシーと Randomized Response を導入する．

表 2: 記法

記法	意味
$[K]$	$\{1, 2, \ldots, K\}(K$ は自然数)
$N \in \mathbb{N}$	ユーザーの数
$T \in \mathbb{N}$	ラウンド数
$v_{n,t} \in \{0, 1\}$	ラウンド t におけるユーザー n の状態
$V \in \{0, 1\}^{N \times T}$	$\{v_{n,t}\}_{n \in [N], t \in [T]}$
$V_{:,t} \in \{0, 1\}^N$	$\{v_{n,t}\}_{n \in [N]}$
$z_{n,t} \in \mathbb{R} \cup \{\text{NULL}\}$	ランダマイズされた $v_{n,t}$
$Z \in (\mathbb{R} \cup \{\text{NULL}\})^{N \times T}$	ランダマイズされた V
$Z_{:,t} \in \mathbb{R}^N$	$\{z_{n,t}\}_{n \in [N]}$
$\mu_t \in [0, 1]$	$\mu_t = \lvert\{n : v_{n,t} = 1\}\rvert / N$

2.1 Local Differential Privacy

LDP では次のようなデータ収集シナリオにおいて，データ提供者のプライバシーの保護を考える：それぞれが秘密のデータを持っている複数人のデータ提供者と一人のデータ収集者がいる．各データ提供者はプライバシーを守りながら，個別に情報を提供することを目的とする．データ提供者はプライバシーを保護するためにメカニズム $\mathcal{Q}$ を通してランダマイズしたデータを提供する．データ収集者は提供されたデータから統計値を得ることを目的する．

形式的には，プライバシー保護の度合いを表すパラメータ ϵ を用いて，次のように定義される：

Definition 1 (ϵ-LDP [11])．集合 $\mathcal{V}$ の要素を入力とし，集合 $\mathcal{Z}$ の要素を出力するメカニズムを $\mathcal{Q}$ とする．$\epsilon > 0$ に対して，メカニズム $\mathcal{Q}$ が *ϵ-LDP* であるとは，任意の入力対 $v, v' \in \mathcal{V}$ と任意の出力集合 $\mathcal{S} \subset \mathcal{Z}$ に対して次の条件を満たすことをいう：

$$\frac{\Pr[\mathcal{Q}(v) \in \mathcal{S}]}{\Pr[\mathcal{Q}(v') \in \mathcal{S}]} \leq e^\epsilon.$$

この定義は，いかなる入力 v に対しても，値 $s \in \mathcal{S}$ を出力する確率と他の値 v' が入力された場合に $s \in \mathcal{S}$ を出力する確率の比が高々 e^ϵ でしかないことを要求している．プライバシーパラメータ ϵ が小さいほど強力なプライバシー保護，大きいほど弱い保護を意味する．決定的な出力をもつメカニズムはいかなる有限の ϵ に対しても ϵ-LDP を満たさないことに注意されたい．

同じ提供者から繰り返し情報を収集することはプライバシー侵害のリスクを高める．これは LDP の文脈においては直列合成定理として表現される．

Theorem 1 (直列合成定理 [6])．メカニズム列 $\{\mathcal{Q}_i\}_{i=1}^K$ があり，それぞれのメカニズムが ϵ_i-*LDP* を保証するとする．このとき，メカニズム列は $(\sum_{i=1}^K \epsilon_i)$-*LDP* を保証する．

メカニズム列全体として ϵ-LDP を保証するためには，各メカニズムのプライバシーパラメータ ϵ_i は ϵ より小さくなければならない．これは，メカニズム列が長ければ長いほど各メカニズムにおいてより強固なプライバシー保護が要

求されることを意味する．

2.2 Randomized Response

Randomized Response [16] は 1-bit の情報を送信する際に ϵ-LDP を保証するメカニズムである．Randomized Response は LDP が登場する前からプライバシーを保護しながら統計情報を得るための方法として知られており，LDP との関係も議論されている．Randomized Response を用いた LDP アルゴリズムとしては [3], [8], [12] などがあり非常に多くの応用がある．

形式的には Randomized Response Q_{Bin} は次のように表現できる．入力 $v \in \{0,1\}$ に対して，

$$Q_{\mathrm{Bin}}(v;\epsilon) = \begin{cases} v & \text{w.p. } \frac{e^\epsilon}{e^\epsilon+1}, \\ 1-v & \text{w.p. } \frac{1}{e^\epsilon+1}. \end{cases}$$

確率 $e^\epsilon/(e^\epsilon+1)$ で入力をそのまま出力し，確率 $1/(e^\epsilon+1)$ で入力を反転して出力する．一方がもう一方の e^ϵ 倍の確率で出力されることから，このアルゴリズムが ϵ-LDP を満たすことは容易に分かる．

Theorem 2 ([11])**.** *Randomized Response Q_{Bin} は ϵ-LDP* である．

また，Randomized Response Q_{Bin} は $[0,1]$ 上の 2 つの分布間の f-divergence を最大化する [10] ことが知られている．これは 2 つの入力を見分ける能力が，他の $[0,1]$ を入力とする LDP メカニズムよりも高いことを意味する．

3. Problem Formulation

この章では本論文で扱う問題であるプライバシーを保護しながらのリアルタイム集計問題を Locally Private Continual Counting(LPCC) として定式化する．まずは，プライベートでない，continual counting を定義する．その後に，continual counting のプライバシーを保護する場合として，locally private continual counting を定義する．

3.1 Continual Counting

N 人のデータ提供者 (ユーザー) と 1 人のデータ収集者がおり，1 から T までのラウンドがある．各ユーザー n は各ラウンド t において状態 $v_{n,t} \in \{0,1\}$ をもち (これは各ラウンドで変化しうる)，状態を毎ラウンド収集に送信する．全てのユーザー，ラウンドにおける状態をまとめてデータストリーム V と表す．また，$V_{:,t}$ はラウンド t における全てのユーザーの状態を表す．ラウンド t において状態が 1 であるユーザーを (ラウンド t における) **アクティブユーザー**と呼ぶこととする．

このとき，ユーザーらの目的は自分の状態を収集者に提供することであり，収集者の目的はリアルタイムで各ラウンドにおけるアクティブユーザーの比 μ_t を推定することである．ただし，ここでいう「リアルタイムで推定する」とは，ラウンド t の μ_t の推定をラウンド $t+1$ のユーザーらの情報を観測する前に行うということである．また，アクティブユーザーの比 μ_t は，ユーザーの状態 $v_{n,t}$ を用いて次のように表現できる:

$$\mu_t = \frac{|\{n : v_{n,t} = 1\}|}{N} = \frac{\sum_{n\in[N]} v_{n,t}}{N}.$$

3.2 Locally Private Continual Couting

我々が対象とする問題は continual counting に，データ提供者のプライバシーが ϵ-LDP によって保護されている，という制約を加えた問題である．ユーザーには局所差分プライバシーが保証されていなければならないとし，ユーザーは LDP メカニズムを通じて情報を収集者に提供する．ユーザーらはプライバシーを保護するために，ϵ-LDP メカニズム Q を通じて状態をランダマイズして送信する．メカニズム Q は公開情報であるとする．Q は $[0,1]$ を入力とするなんらかのメカニズムである．LDP メカニズムによってランダマイズされた $v_{n,t}$ を $z_{n,t}$ とする．$z_{n,t}$ はランダマイズされているため，収集者は $z_{n,t}$ を観測しても $v_{n,t}$ を決定的に知ることはできない．プライベート化されたデータストリームを Z と表記する．また，$Z_{:,t}$ はラウンド t における全てのユーザーのランダム化された状態を表す．

収集者は毎ラウンドユーザーらから送信されてきた $z_{n,t}$ から，そのラウンドにおけるアクティブユーザーの比 μ_t の推定値 $\hat{\mu}_t$ を計算する．推定量としての $\hat{\mu}_t$ は LDP メカニズムのランダムネスを考慮して，実際のアクティブユーザーの比を推定できるように設計されるべきである．したがって，μ_t は使用する LDP メカニズムに依存する．また，収集者はリアルタイムに推定を行うので，ラウンド t における推定値 $\hat{\mu}_t$ はラウンド $t+1$ に送信される情報を観測する前に出力しなければならない．

我々は LPCC における，アルゴリズムの誤差を次のように定義する．アルゴリズムが各ラウンド t において，推定量 μ_t を出力するとき，確率 $1-\beta$ で

$$\max_{t\in[T]} |\hat{\mu}_t - \mu_t| = \alpha$$

が成り立つときこのアルゴリズムの誤差は α であると定義する．我々は，より誤差が小さいアルゴリズムをより優れたアルゴリズムであるとする．

4. 1-Shot Reporting Mechanisms

この章では LPCC に対する我々の提案法 (Glance) と Harmony[12] を説明する．Glance と Harmony は Randomized Response[16] とサンプリングを基本的なアイディアとするアルゴリズムである．Glance，Harmony ともに，サンプリングによって T ラウンドからただ一つのラウンドを選択するため，これらをまとめて 1-Shot Reporting メカニズムとよぶこととする．

Algorithm 1: Glance

Input: data stream V, privacy parameter ϵ

1 **for** $n = 1$ *to* N **do**
2 　$\tau_n \sim$ uniform random from $[T]$;
3 **end**
4 **for** $t = 1$ *to* T **do**
5 　**for** $n = 1$ *to* N **do**
6 　　**if** $t = \tau_n$ **then**
7 　　　$z_{n,t} = Q_{\mathrm{Bin}}(v_{n,t}; \epsilon)$;
8 　　**else**
9 　　　$z_{n,t} = \mathrm{NULL}$;
10 　　**end**
11 　**end**
　　Output: $\hat{\mu}_t$
12 **end**

この章の内容は次のようにまとめられる．4.1 章では提案法のアルゴリズムを具体的に説明する．4.2 章では提案法がユーザーのプライバシーを保護することを示す．

4.1 Algorithm

Algorithm 1 は Glance の疑似コードである．各ユーザーはそれぞれあらかじめ Randomized Response を通じて状態をレポートするラウンド τ_n を一様ランダムに決めておく (line 2)．各ユーザーはラウンド τ_n になったら，そのラウンドにおける自分の状態 $v_{n,t}$ をプライバシーバジェットを ϵ 消費して Randomized Response を通じて $z_{n,t}$ として収集者に送信する (line 7)．それ以外のラウンドでは NULL を送信する (line 9)．各ラウンドにおいて収集者はそのラウンドにおけるアクティブユーザーの比 μ_t の推定値 $\hat{\mu}_t$ を計算する．Algorithm 1 は推定値 $\hat{\mu}_t$ を次のように計算する：

$$\hat{\mu}_t = \frac{\frac{1}{|\eta_t|}\sum_{n\in\eta_t} z_{n,t} - q}{p - q} \text{ where}$$
$$p = \Pr[z_{n,t} = 1 | n = \eta_t, v_{n,t} = 1] = \frac{e^\epsilon}{e^\epsilon + 1},$$
$$q = \Pr[z_{n,t} = 1 | n = \eta_t, v_{n,t} = 0] = \frac{1}{e^\epsilon + 1},$$
$$\eta_t = \{n : t \in \tau_n\}.$$

η_t はラウンド t において Randomized Response を用いて状態をレポートするユーザーの集合を表している．

Harmony と Glance の違いは，Randomized Response を用いた状態の報告を行わないラウンドでのユーザーの収集者への報告である (line 9)：Glance では NULL を送信するが，Harmony ではユーザーは 0 を送信する．また，μ_t の推定量 $\hat{\mu}_t$ の設計も，この違いに合わせて，少し異なる．このわずかなアルゴリズムの違いが，誤差に大きな違いを生む．

4.2 Privacy

以下の定理は Glance が ϵ-LDP であることを保証する．

Theorem 3. *Alogorithm 1* は ϵ-*LDP* である．

ラウンド τ_n の選択のランダムネスがユーザーの状態に依存していないこと，Randomized Response のランダムネスより明らかである．

5. Utility Analysis

ここでは 2 つのケースにおける 1-Shot Reporing メカニズムの誤差を解析する．我々は提案法の誤差の確率的上界を導出し，Harmony の誤差の確率的上下界を導出する．また，解析した Glance の誤差のレートと Harmony の誤差のレートの比較を行う．我々は特にそれぞれの誤差の ϵ への依存性に興味がある．

5.1 General Case Analysis

ここでは，N と T の関係以外は，特にデータストリームに対して仮定が置けない場合において，提案法と Harmony[12] の誤差を解析する．Glance，Harmony それぞれについて，任意のデータストリームの中で最大の誤差を持つ場合の誤差の上界を解析する．また，Harmony に関しては誤差の下界についても解析する．

5.1.1 Harmony

まずは Harmony の誤差の上界について論じる．我々は誤差に関して確率的不等式を導出し，次の系を導いた．

Corollary 1. *Harmony* において，$\epsilon > 0, \beta > 0$ に対して，確率 $1 - \beta$ で

$$\max_{t\in[T]} |\hat{\mu}_t - \mu_t| = \begin{cases} O\left(\frac{1}{\epsilon}\sqrt{\frac{T\log(T/\beta)}{N}}\right) & as\ \epsilon \to +0, \\ O\left(\sqrt{\frac{T\log(T/\beta)}{N}}\right) & as\ \epsilon \to +\infty \end{cases}$$
$$as\ T \to +\infty, N \to +\infty.$$

証明は省略する．Corollary 1 の $\epsilon \to +0$ での誤差のレートは [12] での解析に一致する．

また，Harmony は誤差の下界についても解析した．この解析は [12] では行われていない．上界の解析と同様に，確率的不等式を導出し次の系を導いた．

Corollary 2. *Harmony* において，あるデータストリームが存在して，$\epsilon > 0, 0 < \beta < 1/(N+1)^2$ に対して，確率 $1 - \beta$ で

$$\max_{t\in[T]} |\mu_t - \hat{\mu}_t| = \begin{cases} \Omega\left(\frac{1}{\epsilon}\sqrt{\frac{T}{N}\log\frac{1}{(N+1)^2\beta}}\right) & as\ \epsilon \to +0, \\ \Omega\left(\sqrt{\frac{T}{N}\log\frac{1}{(N+1)^2\beta}}\right) & as\ \epsilon \to +\infty \end{cases}$$
$$as\ N \to \infty, T \to +\infty.$$

証明は省略する．これらの結果は，T, N, β を固定したとき，十分に大きな ϵ を与えても誤差をある正の定数よりも小さくできないことを意味している．

5.1.2 Glance

次に提案法の誤差の上界について論じる．我々は Glance の誤差に関する確率的不等式を導出し，その不等式から次の系を導いた．

Corollary 3. *Glance* において，$\epsilon > 0, \beta > 0$ に対して，少なくとも確率 $1-\beta$ で

$$\max_{t\in[T]} |\hat{\mu}_t - \mu_t| = \begin{cases} O\left(\frac{1}{\epsilon}\sqrt{\frac{T\log(T/\beta)}{N}}\right) & as\ \epsilon \to +0, \\ O\left(\sqrt{\frac{T\log(T/\beta)}{N}}\right) & as\ \epsilon \to +\infty \end{cases}$$
$$as\ T \to +\infty, N \in \omega(T\log(T/\beta)).$$

証明は省略する，

以下で証明のアイディアを説明する．誤差を次のように 2 つに分解し，それぞれに対して上界を導出する：

$$\max_{t\in[T]} |\hat{\mu}_t - \mu_t| \le \max_{t\in[T]} |\hat{\mu}_t - \tilde{\mu}_t| + \max_{t\in[T]} |\tilde{\mu}_t - \mu_t| \text{ where}$$
$$\tilde{\mu}_t = \frac{1}{|\eta_t|}\sum_{n\in\eta_t} v_{n,t}, \eta_t = \{n : t = \tau_n\}.$$

$|\hat{\mu}_t - \tilde{\mu}_t|$ は Randomized Response に起因する誤差，$|\tilde{\mu}_t - \mu_t|$ はサンプリングでの情報の喪失による誤差と解釈できる．この分解により，本来の誤差よりも大きい値を見積もってしまうが，もとの誤差の高々 2 倍であるのでオーダー解析には影響がない．

Glance における Randomized Response に起因する誤差の上界は次の Claim で表される．

Claim 1. *Glance* において，確率 $1-\beta/2$ で，$\epsilon > 0, \beta > 0, N > T\log(4T/\beta)$ に対して，

$$\max_{t\in[T]} |\hat{\mu}_t - \tilde{\mu}_t| < \theta \ where$$
$$\theta = \frac{2e^\epsilon}{3(e^\epsilon - 1)} \log\left(\frac{1}{1 - \frac{T}{N}\log\frac{4T}{\beta}}\right) + \frac{3}{e^\epsilon - 1}\left[\left(\frac{2e^\epsilon}{9}\log\left(\frac{1}{1 - \frac{T}{N}\log\frac{4T}{\beta}}\right)\right)^2 + 4\frac{2e^\epsilon}{9}\log\left(\frac{1}{1 - \frac{T}{N}\log\frac{4T}{\beta}}\right)\right]^{1/2}. \quad (1)$$

証明は省略する．Randomized Response は ϵ が小さいほど強力なランダマイズを行うため，ϵ が小さいほど誤差を大きくする．したがって，ϵ が小さくなる方向には Randomized Response のランダムネスに起因する誤差が支配的である．この誤差は ϵ が大きければ大きいほど小さくなり，0 に近づく．そのため，ϵ が十分に大きければではこの誤差 $|\hat{\mu}_t - \tilde{\mu}_t|$ は誤差全体 $|\hat{\mu}_t - \mu_t|$ にほとんど影響を与えない．

Glace におけるサンプリングでの情報喪失による誤差の上界は次の Claim で表される．

Claim 2. *Glance* において，確率 $1-\beta/2$ で，$\epsilon > 0, \beta > 0, N > T\log(4T/\beta)$ に対して，

$$\max_{t\in T} |\tilde{\mu}_t - \mu_t| < \sqrt{\frac{1}{2}\log\frac{1}{1 - \frac{T}{N}\log\frac{4T}{\beta}}}.$$

証明は省略する．提案法は各ユーザーがラウンドをサンプルし，情報を送信しないラウンドがあり，送信が行われなかったラウンドの状態に関する情報は収集者には全くわからない．サンプリングは ϵ の大きさに関係が無いため，サンプリングにより失われた情報によって生じる誤差は ϵ の大きさに関係なく生じる．そのため，ϵ を大きくしてもこの誤差は 0 にできない．したがって，ϵ が大きい領域ではサンプリングにより情報が失われることに起因する誤差が支配的である．

5.1.3 Discussion

プライバシーパラメータ ϵ に関しては，提案法の誤差と Harmony の誤差は同じオーダーであり，これらは ϵ が小さくなっていくときと大きくなっていくときで挙動が変わる．$\epsilon \to +0$ では，Harmony の誤差の上界は $O(1/\epsilon)$ で，$\epsilon \to +\infty$ では $\Theta(1)$ である．提案法は下界は証明していないが，同様の挙動をすると考えられる．これは，十分に大きな ϵ を与えても，T, N, β を固定すると，最悪の場合，提案法，Harmony ともに誤差がある正の定数よりも小さくできないことを示唆している．

5.2 Dense or Sparse Case Analysis

一般ケースにおける誤差の解析から，ϵ が大きい領域ではサンプリングによるデータの損失に起因する誤差が支配的であることがわかる．そのため，サンプリングによるデータの損失に起因する誤差が無視できるほど小さければ，Glance の誤差はより小さくできることになる．我々は，一般ケースの誤差上界の証明過程から，次の条件が成り立つとき，サンプリングによるデータの損失に起因する誤差が Randomized Response のランダムネスに起因する誤差よりも小さくなることを発見した：

$$\min\{D(\mu_t - \theta||\mu_t), D(\mu_t + \theta||\mu_t)\} \ge \frac{\frac{(e^\epsilon-1)^2}{2e^\epsilon}\theta^2}{1 + (e^\epsilon - 1)\theta/3} \quad (2)$$
$$\text{where } D(a||b) = a\log(a/b) + (1-a)\log((1-a)/(1-b)).$$

θ は式 (1) で定義されている．$D(a||b)$ は，$0 \le a, b \le 1$ をそれぞれ平均とするベルヌーイ分布間の KL-divergence である．また，$a \ge 1, a \le 0, b = 0, 1$ のいずれかのときは，$D(a||b) = +\infty$ とみなすことで整合性がとれる．

ここでこの条件の解釈を説明する．図 1 は，$T = 10, N = 100,000, \beta = 0.1$ としたときに条件 (2) が成り立つ領域を表したものである．色が付いている領域で条件 (2) が成り立つ．直感的には ϵ がある程度大きい領域では，データス

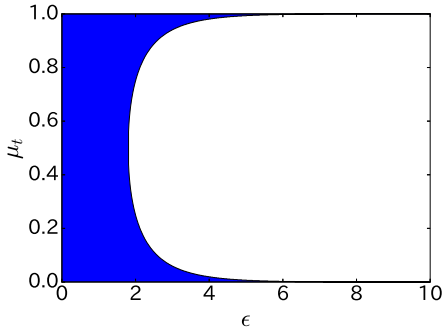

図 1: 縦軸は μ_t，横軸は ϵ を表す．色をつけた領域は $N = 10,0000, T = 10, \beta = 0.1$ において条件 (2) を満たす．

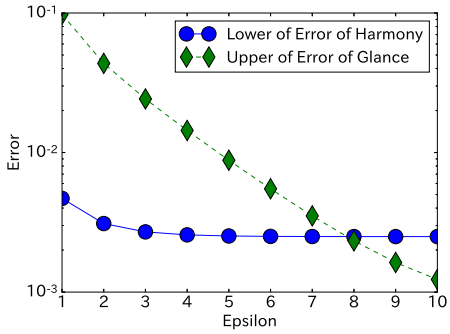

図 2: 密または疎ケースでの，$N = 1,000,000, T = 10, \beta = \frac{1}{2(N+1)^2}$ での Harmony の誤差の下界と Glance の誤差の上界の比較．$\epsilon = 8$ で逆転している．

トリームが非常に密または疎であるときに，条件 (2) が成り立つと言える．

ここでは条件 (2) 下において，提案法と Harmony[12] の誤差を比較する．我々は提案法は誤差の確率的上界を導出する．また，Harmony の誤差の確率的下界を導出する．その後，Harmony の誤差の確率的下界と提案法の誤差の確率的上界の比較を行う．

5.2.1 Harmony

密または疎ケースにおける Harmony の誤差の下界を以下の定理で示す．

Theorem 4. *Harmony* において，条件 *(2)* が成り立つとき，あるデータストリームが存在して，$\epsilon > 0, 0 < \beta < 1/(N+1)^2$ に対して，少なくとも確率 $1-\beta$ で，

$$\max_{t\in[T]} |\mu_t - \hat{\mu}_t| > \sqrt{\frac{T}{N}\frac{e^\epsilon(e^\epsilon+1)}{(e^\epsilon-1)^2}\left(1-\frac{1}{T}\frac{e^\epsilon}{e^\epsilon+1}\right)\log\frac{1}{(N+1)^2\beta}}.$$

証明は省略する．この定理から次の系が容易に導かれる．

Corollary 4. *Harmony* において，条件 *(2)* が成り立つとき，あるデータストリームが存在して，$\epsilon > 0, 0 < \beta < 1/(N+1)^2$ に対して，少なくとも確率 $1-\beta$ で

$$\max_{t\in[T]} |\mu_t - \hat{\mu}_t| = \begin{cases} \Omega\left(\frac{1}{\epsilon}\sqrt{\frac{T}{N}\log\frac{1}{(N+1)^2\beta}}\right) & as\ \epsilon \to +0, \\ \Omega\left(\sqrt{\frac{T}{N}\log\frac{1}{(N+1)^2\beta}}\right) & as\ \epsilon \to +\infty \end{cases} \quad as\ T \to +\infty.$$

証明は省略する．この定理 Harmony が十分に大きな ϵ が利用可能であっても，T, N, β が固定されていれば，誤差がある正の定数よりも小さくできないことを意味している．

5.2.2 Glance

提案法の誤差の上界について論じる．提案法の密または疎ケースにおける誤差の上界は次の定理で表現される．

Theorem 5. *Glance* において，条件 *(2)* が成り立つとき，$\epsilon > 0, \beta > 0, N > T\log(4T/\beta)$ に対して，確率 $1-\beta$ で，

$$\max_{t\in[T]} |\hat{\mu}_t - \tilde{\mu}_t| < 2\theta. \tag{3}$$

ただし，θ は式 *(1)* で定義されている．

証明は省略する．この定理から次の系が容易に導かれる．

Corollary 5. 条件 *(2)* が成り立つとき，$\epsilon > 0, \beta > 0$ に対して

$$\max_{t\in[T]} |\hat{\mu}_t - \tilde{\mu}_t| = \begin{cases} O\left(\frac{1}{\epsilon}\sqrt{\frac{T\log(T/\beta)}{N}}\right) & as\ \epsilon \to +0, \\ O\left(\sqrt{\frac{T\log(T/\beta)}{\exp(\epsilon)N}} \vee \frac{T\log(T/\beta)}{N}\right) & as\ \epsilon \to +\infty \end{cases}$$
$$as\ T \to +\infty, N \in \omega(T\log(T/\beta)).$$

証明は省略する．提案法の誤差の ϵ に対する依存性は，$\epsilon \to +0$ では $O(1/\epsilon)$ であり，一般ケースにおける誤差の上界の依存性と同じであるが，$\epsilon \to +\infty$ では $O(1/\sqrt{e^\epsilon})$ である．ある程度大きなプライバシーパラメータ ϵ に対しては，提案法は指数的に誤差を小さくできる．

5.2.3 Discussion

ϵ に注目すると，十分に大きな ϵ が与えられたとき，提案法の誤差の上界が Harmony の誤差の下界を下回る．このケースにおいては提案法の誤差の上界は，ϵ が大きければ大きいほど，小さくなる．一方，Harmony の誤差の下界は，どんなに大きな ϵ が与えられても正の定数に漸近するため，ある値よりも小さくできない．これはこのケースにおいて，ϵ に関して，提案法の誤差が Harmony の誤差よりも優位であることを示している．また，log 項を無視すれば，T, N に関しても提案法の誤差の上界は Harmony の誤差の下界とオーダーが等しく，ϵ にのみ関して優位であるわけではないことにも注意する．

6. Numerical Experiment

ここでは，人工データを用いた数値実験によって次の 2 点を確かめる: (i) 条件 (2) が成り立つとき，Glance の誤差が ϵ の増加に対して，Harmony の誤差よりも，早く小さくなる．(ii)Glance の誤差は条件 (2) が成り立つときは，条件 (2) が成り立たないときよりも，ϵ の増加に対して早く

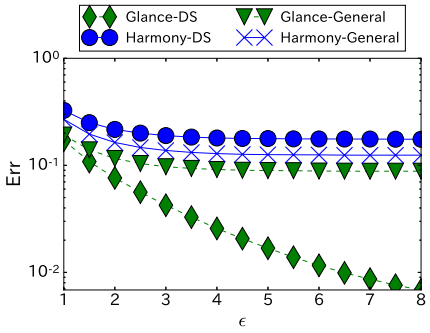

図 3: 縦軸は Err の値，横軸は ϵ を表す．提案法，既存法それぞれに対して $\mu = 0.995, 0.5$ で Err を測定した．密または疎ケースにおける提案法の誤差 (Glance-DS) が他の 3 つのパターンより早く小さくなっている．

小さくなる．

この実験では，アルゴリズムが出力した $\{\hat{\mu}_t\}_{t=1}^T$ に対して，指標 Err を次のように定義してアルゴリズムの正確性を測る：

$$\mathrm{Err} = \max_{t\in[T]} |\hat{\mu}_t - \mu_t|.$$

Err が小さいほど正確なアルゴリズムであるとする．この値は誤り率を無視した誤差だと言える．

全てのラウンドで μ_t は共通とし，$T = 50, N = 10,000$ は固定とする．$\beta = 0.1$ として計算すると $\mu_t = 0.9995$ の場合，任意の $1 \le \epsilon \le 8$ に対して条件 (2) を満たす．また，$\mu_t = 0.5$ の場合，$\epsilon \ge 2.5$ では条件 (2) を満たさない．各アルゴリズムを各パラメーターに対して 10,000 回ずつ実行して Err を計測して，その平均値を求める．

図 3 は結果をプロットしたものである．Glance-DS は $\mu_t = 0.9995$ での Glance の Err，Glance-General は $\mu_t = 0.5$ における Glance の Err，Harmony-DS は $\mu_t = 0.9995$ における Harmony の Err，Harmony-General は $\mu_t = 0.5$ における Harmony の Err を表す．Glance-DS 以外の Err は正の定数に向かって上から漸近しているが，Glance-DS の Err は ϵ に対して指数的小さくなっている．

7. Conclusion

我々は LPCC の新しいアルゴリズムとして Glance を提案した．我々は密または疎ケースにおいては提案法の誤差と既存法の誤差を解析し，十分に大きな ϵ に対しては既存法の誤差の下界よりも提案法の誤差の上界が小さくなることを示した．さらに，その場合において，提案法の誤差が既存法の誤差よりも小さいことを実験的に示した．

8. 謝辞

本研究の一部は JST CREST JPMJCR1302 および科学研究費 16H02864，18H04099 によりサポートされました．

参考文献

[1] Apple Inc.: Learning with Privacy at Scale (2017).
[2] Avent, B., Korolova, A., Zeber, D., Hovden, T. and Livshits, B.: Blender: Enabling local search with a hybrid differential privacy model, *Proceedings of the 26th USENIX Security Symposium*, pp. 747–764 (2017).
[3] Bassily, R. and Smith, A.: Local, private, efficient protocols for succinct histograms, *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, ACM, pp. 127–135 (2015).
[4] Chan, T.-H. H., Shi, E. and Song, D.: Private and continual release of statistics, *ACM Transactions on Information and System Security (TISSEC)*, Vol. 14, No. 3, p. 26 (2011).
[5] Ding, B., Kulkarni, J. and Yekhanin, S.: Collecting Telemetry Data Privately, *Advances in Neural Information Processing Systems*, pp. 3574–3583 (2017).
[6] Dwork, C., McSherry, F., Nissim, K. and Smith, A.: Calibrating noise to sensitivity in private data analysis, *TCC*, Vol. 3876, Springer, pp. 265–284 (2006).
[7] Dwork, C., Naor, M., Pitassi, T. and Rothblum, G. N.: Differential privacy under continual observation, *Proceedings of the forty-second ACM symposium on Theory of computing*, ACM, pp. 715–724 (2010).
[8] Erlingsson, Ú., Pihur, V. and Korolova, A.: Rappor: Randomized aggregatable privacy-preserving ordinal response, *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, ACM, pp. 1054–1067 (2014).
[9] Hsu, J., Khanna, S. and Roth, A.: Distributed private heavy hitters, *Proceedings of the 39th international colloquium conference on Automata, Languages, and Programming-Volume Part I*, Springer-Verlag, pp. 461–472 (2012).
[10] Kairouz, P., Oh, S. and Viswanath, P.: Extremal mechanisms for local differential privacy, *Advances in Neural Information Processing Systems*, pp. 2879–2887 (2014).
[11] Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S. and Smith, A.: What can we learn privately?, *SIAM Journal on Computing*, Vol. 40, No. 3, pp. 793–826 (2011).
[12] Nguyên, T. T., Xiao, X., Yang, Y., Hui, S. C., Shin, H. and Shin, J.: Collecting and analyzing data from smart device users with local differential privacy, *arXiv preprint arXiv:1606.05053* (2016).
[13] Qin, Z., Yang, Y., Yu, T., Khalil, I., Xiao, X. and Ren, K.: Heavy hitter estimation over set-valued data with local differential privacy, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ACM, pp. 192–203 (2016).
[14] Tang, J., Korolova, A., Bai, X., Wang, X. and Wang, X.: Privacy Loss in Apple's Implementation of Differential Privacy on macOS 10.12, *arXiv preprint arXiv:1709.02753* (2017).
[15] Wang, T., Blocki, J., Li, N. and Jha, S.: Locally differentially private protocols for frequency estimation, *Proceedings. of the 26th USENIX Security Symposium*, pp. 729–745 (2017).
[16] Warner, S. L.: Randomized response: A survey technique for eliminating evasive answer bias, *Journal of the American Statistical Association*, Vol. 60, No. 309, pp. 63–69 (1965).
[17] 小野元，福地一斗，佐久間淳：局所差分プライバシー制約下における逐次 heavy hitters 検知，第 10 回データ工学と情報マネジメントに関するフォーラム (2018).