ポイズニング攻撃に対してロバストなEMアルゴリズムを用いたkey-valueデータにおけるLDPプロトコル

堀込 光[1]　菊池 浩明[2]　Chia-Mu Yu[3]

概要：局所差分プライバシは，ユーザが信頼性のないサーバに自身の情報を送信する際に自身のデバイス内で情報をランダマイズすることで，サーバに対して自身の持つ情報を秘匿化する技術である．しかし一方で，局所差分プライバシはユーザが情報をランダマイズするため，悪意のあるユーザが特定の情報をサーバに送信することで分析結果を操作するポイズニング攻撃に対して脆弱である．2022 年に Wu らは key-value データにおける局所差分プライバシ PrivKV に対する 3 種類のポイズニング攻撃手法を提案した．そこで我々は，ポイズニング攻撃に対してより頑強性が期待できる EM(Expectation Maximization) アルゴリズムに着目し，新しい LDP プロトコル emPrivKV を提案する．本論文では，emPrivKV と PrivKV に対して 3 種類のポイズニング攻撃を行い，PrivKV と比較することで，emPrivKV のポイズニング攻撃に対する強度を調査する．

キーワード：局所差分プライバシー，EM アルゴリズム，PrivKV

EM estimation from LDP protocol for key-value data that is robust against poisoning attacks

Hikaru Horigome[1]　Hiroaki Kikuchi[2]　Chia-Mu Yu[3]

Abstract: Local differential privacy is a technology that randomizes locally private value before sending to an untrusted server. However, since the randomizations are performed at the user side by their selves, it has more chance to modify the randomized value so that malicious users try to manipulate the result. In 2022, Wu et al. proposed three kinds of poisoning attacks to PrivKV, one of state-of-art LDP protocol for key-value data. To make the LDP more robust against poisoning, we propose a new LDP protocol emPrivKV, using EM (Expectation Maximization) algorithm. In this paper, we conduct an experiment that applies the three poisoning attacks to our algorithm and report the strength of our work against poisoning, in compassion with the PrivKV.

Keywords: Local Differential Privacy，EM algorithm，PrivKV

1. はじめに

近年大幅に普及したスマートデバイスにより，サービス事業者は人々のあらゆる行動を分析できるようになった．例えば，Amazon などのオンライン商取引サービスでは，全利用者の購入履歴を取集し，購入頻度に基づいて利用者が購入した商品に関連する推薦商品を提供している．しかし，サービス事業者は全利用者の正確な行動履歴を保有しており，過失による情報漏洩や不正な内部犯行者によるプライバシ侵害の危険性がある．

個人情報の保護技術の一つに差分プライバシ [1] がある．これは、収集した情報の統計値を公開する際に確率的なノ

[1] 明治大学 先端数理科学研究科
Guraduate School of Advanced Mathematical Sciences, Meiji University
[2] 明治大学 総合数理学部
School of Interdisciplinary Mathematical Science, Meiji University
[3] Department of Information Management and Finance, National Yang Ming Chiao Tung Univercity

イズを付与するなどして出力プライバシーを保護する理論的な枠組みである．値を曖昧にする匿名加工情報よりも統計的な値の保護に適している．匿名加工情報は，再識別の禁止などの法的な規則と安全管理措置が適切である仮定の下で管理されており，プライバシ保護に関して理論的な保証はない．一方，収集の際の評価値などを保護する技術に局所差分プライバシ [2] がある．局所差分プライバシは，スマートデバイスからの情報を収集する際に確率的なノイズを付与するという技術である．これにより，サービス事業者でさえもユーザの真の値は分からない．

離散値と連続値の代表的な局所差分プライバシアルゴリズムとしてそれぞれ，Warner らによる Randomized Response(RR)[3] と Nguyên らによる Harmony[4] が知られている．さらに，Ye らは，Randomized Response(RR) と Harmony を組み合わせることで，離散値と連続値の組み合わせである key-value データセットに対して局所差分プライバシを満たす局所差分プライバシー方式 PrivKV[8] を提案した．これにより，各アイテムの頻度とその平均値の相関を維持した状態で統計値を推定することを可能にした．

しかし一方で，局所差分プライバシはユーザが情報をランダマイズするため，悪意のあるユーザが，意図的に加工した情報をサーバに送信することで分析結果を操作するポイズニング攻撃に対して脆弱であることが指摘されている [5]．また，多数の架空のユーザを作って分析結果を操作できることも課題である．2022 年に Wu らは key-value データにおける局所差分プライバシ PrivKV に対する 3 種類のポイズニング攻撃手法，Maximal Gain Attack(M2GA)，Random Message Attack(RMA)，Random Key-Value pair Attack(RKVA) を提案した [6]．既存の局所差分プライバシプロトコル PrivKV などでは，推定に最尤推定法が用いられており，ユーザの集計から最尤値を算出し推定値とするため，これらのポイズニング攻撃に対して推定値が操作されやすいことが考えられる．

そこで我々は，ポイズニング攻撃に対してより頑強性が期待できる EM(Expectation Maximization) アルゴリズム [7] に着目し，新しい LDP プロトコル emPrivKV を提案する．本論文では，emPrivKV と PrivKV に対して 3 種類のポイズニング攻撃を行い，PrivKV と比較することで，emPrivKV のポイズニング攻撃に対する強度を明らかにする．

2. 準備

2.1 基本定義

局所差分プライバシプロトコルでは，各ユーザが自身のデータに対してノイズを付与し，そのデータを収集者へ送信する．収集者は，各ユーザから得られたデータを集計し，度数や平均値を推定する．ユーザ数を n とし，ユーザの集合を $U = \{u_1, u_2, ..., u_n\}$ とする．各ユーザは離散値，連続値，または key-value データを保持している．取扱う d 種類の離散値の集合を $K = \{k_1, k_2, ..., k_d\}$，$[-1, 1]$ の連続値の集合を V とする．プライバシー費用を ϵ とし，ある入力を t に対してランダムアルゴリズム M を適用することを $M(t, \epsilon)$ と記述する．

2.2 局所差分プライバシ (LDP)

任意の異なる 2 つの入力に対して，M の出力が同一になる確率に差がないことを保証している．これにより，出力を参照してもユーザの正確な入力を特定することができず，ユーザのプライバシーを保護する．ランダムアルゴリズム M についての局所差分プライバシは以下のように定義される．

定義 1. 局所差分プライバシ

D を入力の集合，Z を出力の集合とする．M を入力 $t \in D$ に対して $z \in Z$ を出力するランダムアルゴリズムとする．任意の 2 つの入力 $t, t' \in D$ と任意の出力 $z \in Z$ に対して，

$$Pr[M(t, \epsilon) = z] \leq e^{\epsilon} Pr[M(t', \epsilon) = z]$$

が成立するとき，ランダムアルゴリズム M は ϵ-局所差分プライバシを満たすという．

2.3 PrivKV

Ye らは離散値と連続値の 2 次元データである key-value データについての局所差分プライバシアルゴリズム PrivKV[8] を提案した．key-value データの例は $\{\langle \text{YouTube}, 0.5\rangle, \langle \text{Twitter}, 0.1\rangle, \langle \text{Instagram}, 0.2\rangle\}$ のような離散値 (アプリケーション名など) と連続値 (使用時間など) の組み合わせデータである．key-value データの離散値に RR を，連続値に Harmony をそれぞれ独立に適用してしまうと離散値と連続値の相関が失われてしまう．そこで PrivKV では，入力が遷移するとき，離散値と連続値を同時に変化させることで，離散値と連続値の相関を維持した状態でデータ収集を行う．また，収集者は key に対する頻度と value に対する平均値を推定する．

本節では，PrivKV のランダマイズ方式と統計値 (度数，平均値) の推定方式を説明する．i 番目のユーザ u_i が持つ l_i 個の key-value 対の集合を $S_i = \{\langle k_j, v_j\rangle | 1 \leq j \leq l_i, k_j \in K, v_j \in V\}$ とする．S_i を key-vlaue 集合，S_i の h 番目の $\langle k_h, v_h\rangle$ を key-value データと呼ぶ．

入力 d 種類の key-value データの収集を考える．$S'_i = \{\langle k'_s, v'_s\rangle | 1 \leq s \leq d, k'_s \in K, v'_s \in V\}$ とする．ある $k' \in K$ について，key-value データが $\langle k_j, v_j\rangle \in S_i$ の場合，$\langle k'_s, v'_s\rangle = \langle 1, v_j\rangle$ とし，対応する key-value データが S_i にない場合，$\langle k'_s, v'_s\rangle = \langle 0, 0\rangle$ と符号化する．

例えば，$d = 5$ で

$$S_i = \{\langle k_1, v_1\rangle, \langle k_4, v_4\rangle, \langle k_5, v_5\rangle\}$$

であったとき，S_i' は，

$$S_i' = (\langle 1, v_1\rangle, \langle 0,0\rangle, \langle 0,0\rangle, \langle 1, v_4\rangle, \langle 1, v_5\rangle)$$

となり，このとき，$|S_i'| = d = 5$ である．こうして得られた S_i' を入力とする．key-value データの摂動には，value を摂動する工程と key を摂動する工程がある．

摂動　長さ d の key-value セット S_i' からランダムに 1 つの key-value データ $\langle k_a', v_a'\rangle \in S_i'$ を選択する．

・value の摂動　$k_a' = 0$ の場合，v_a' を $[-1,1]$ からランダムに選択する．まず，value の値を v_a' に依存する確率で 2 値化する．2 値化した値を v_a^* とする．

$$v_a^* = \begin{cases} 1 & w/p \quad \frac{1+v_a'}{2}, \\ -1 & w/p \quad \frac{1-v_a'}{2} \end{cases}$$

次に v_a^* を以下の確率でランダマイズし，v_a^+ とする．

$$v_a^+ = \begin{cases} v_a^* & w/p \quad p_2 = \frac{e^{\epsilon_2}}{1+e^{\epsilon_2}}, \\ -v_a^* & w/p \quad q_2 = \frac{1}{1+e^{\epsilon_2}} \end{cases}$$

・key の摂動　PrivKV では，key が遷移するとき value も同時に変化させる．key のランダマイズは以下のように行う．$k_a' = 1$ の場合，

$$\langle k_a^*, v_a^+\rangle = \begin{cases} \langle 1, v_a^+\rangle & w/p \quad p_1 = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}, \\ \langle 0,0\rangle & w/p \quad q_1 = \frac{1}{1+e^{\epsilon_1}} \end{cases}$$

となり，$k_a' = 0$ の場合，

$$\langle k_a^*, v_a^+\rangle = \begin{cases} \langle 0,0\rangle & w/p \quad p_1 = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}, \\ \langle 1, v_a^+\rangle & w/p \quad q_1 = \frac{1}{1+e^{\epsilon_1}} \end{cases}$$

となる．摂動化 $\langle k_a^*, v_a^+\rangle$ と選択した key-value データのインデックス a を送信する．このとき，PrivKV 全体の ϵ は $\epsilon = \epsilon_1 + \epsilon_2$ となり，本稿では，$\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$ と仮定する．

集計　n 人ユーザからインデックス a と key-value データ $\langle k_a^*, v_a^+\rangle$ を収集する．PrivKV では，$k_i \in K$ に対する頻度と $v_i \in V$ に対する平均値の推定を目的として最尤推定法を適用する．

・頻度推定　収集した key-value データ $\langle k_a^*, v_a^+\rangle$ の中で，$k_i = 1$ の度数を f_i' とし，k_i の真の度数を f_i とする．k_i の度数の最尤値 $\hat{f}_i$ は，

$$\hat{f}_i = \frac{p_1 - 1 + f_i'}{2p_1 - 1}, \text{where } p_1 = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}$$

と推定される．

・平均値推定　収集した key-value データ $\langle k_a^*, v_a^+\rangle$ の中で，$\langle k_i, v_i\rangle = \langle 1,1\rangle$ の度数を n_{1i}'，$\langle k_i, v_i\rangle = \langle 1,-1\rangle$ の度数を n_{2i}' とする．$\langle k_i, v_i\rangle = \langle 1,1\rangle$ の推定度数 $\hat{n}_{1i}$，$\langle k_i, v_i\rangle = \langle 1,-1\rangle$ の推定度数 $\hat{n}_{2i}$ は，

$$\begin{aligned} N &= n_{1i}' + n_{2i}' \\ \hat{n}_{1i} &= \frac{N(p_2-1) + n_{1i}'}{2p_2 - 1} \\ \hat{n}_{2i} &= \frac{N(p_2-1) + n_{2i}'}{2p_2 - 1}, \text{where } p_2 = \frac{e^{\epsilon_2}}{1+e^{\epsilon_2}} \end{aligned}$$

となり，平均値 $\hat{m}_i$ は，

$$\hat{m}_i = \frac{\hat{n}_{1i} - \hat{n}_{2i}}{N}$$

と推定される．

2.3.1 PrivKVM

Ye らは，PrivKV の対話型アルゴリズム PrivKVM[8] を提案している．key-value データのサンプリングで $\langle k_a', v_a'\rangle \in S_i' = \langle 0,0\rangle$ が選択された場合，v_a' は $[-1,1]$ からランダムに値が付与される．度数の少ない key では，v_a' がランダムに付与される割合が大きいため，平均値は 0 に近似する．PrivKVM では，算出した平均値をユーザに送り返し，2 回目以降の摂動で，$\langle k_a', v_a'\rangle \in S_i' = \langle 0,0\rangle$ のとき，$v_a' = \hat{m}_a$ とすることでこの問題を改善している．

ユーザとの対話回数を $c(\geq 2)$ とし，c 回目の推定度数，推定平均値をそれぞれ $\hat{f}_i^{(c)}$，$\hat{m}_i^{(c)}$ とする．また，key のランダマイズと value のランダマイズで対話ごとに割り振る ϵ をそれぞれ，$\epsilon_{11}, \epsilon_{12}, ..., \epsilon_{1c}$ と $\epsilon_{21}, \epsilon_{22}, ..., \epsilon_{2c}$ とする．1 回目の収集では，PrivKV を用いて推定値 $f_i^{(1)}, m_i^{(1)} = PrivKV(S_i', (\epsilon_{11} + \epsilon_{21}))$ を算出する．2 回目以降の収集では，$\langle k_a', v_a'\rangle \in S_i' = \langle 0,0\rangle$ の場合，$v_a' = m_i{}^{(c-1)}$ とする．c 回の対話のあと，$\hat{f}_i^{(1)}$，$\hat{m}_i^{(c)}$ を推定値とする．[8] では，

$$\begin{cases} \epsilon_{11} = \epsilon_1, \quad \epsilon_{12} = \epsilon_{13} = .. = \epsilon_{1c} = 0 \\ \epsilon_{21} = \epsilon_{22} = ... = \epsilon_{1c} = \frac{\epsilon_2}{c} \end{cases}$$

のように，ϵ を割り振っている．またこのとき，$\epsilon_1 = \sum_c^{n=1} \epsilon_{1n}$，$\epsilon_2 = \sum_c^{n=1} \epsilon_{2n}$ となる．本稿でも同様の大きさで ϵ を割り振る．

2.4 PrivKV に対するポイズニング攻撃

2.4.1 ポイズニング攻撃

ポイズニング攻撃とは，攻撃者がある key に対して特定の情報をサーバに送信することで，その key の推定値を操作する攻撃である．局所差分プライバシでは，ユーザが自身の情報をランダマイズし，送信情報を作成するためポイズニングを自在にできる．

攻撃者は，サーバが収集するアイテムの集合と局所差分プライバシプロトコルを参照できるものとし，システム上で複数人の偽ユーザを容易に作成できるものとする．攻撃者は，m 人の偽ユーザを作成し，偽ユーザから特定の情報を送信する．偽ユーザの集合を $U = \{u_{n+1}, u_{n+2}, ..., u_{n+m}\}$ とする．また，偽ユーザ i の出力を y_i とし，その集合を $Y = \{y_i\}_{i=n+1}^{n+m}$ とする．攻撃者が操作する r 個の key をターゲット key とし，その集合を $T = \{k_1, k_2, ..., k_r\}$ とする．サーバは，n 人の真ユーザと m 人の偽ユーザを合わせた $n+m$ 人の出力から統計値を推定する．

n 人の真ユーザの key k に対する推定度数を $\hat{f}_k$，偽ユーザを含めた $n+m$ 人の key k に対する推定度数を

$\tilde{f}_k$ とする．ポイズニング攻撃による推定度数の変化量を $\Delta\hat{f}_k = \tilde{f}_k - \hat{f}_k$ とし，ターゲット key に対するその総和を頻度利得 (frequency gain) $G_f(Y) = \sum_{k\in T}\mathbb{E}[\Delta\hat{f}_k]$ と呼ぶ．また，同様に n 人の真ユーザの key k に対する推定平均値を $\hat{m}_k$，偽ユーザを含めた $n+m$ 人の key k に対する推定平均値を $\tilde{m}_k$ とする．ポイズニング攻撃による推定平均値の変化量を $\Delta\hat{m}_k = \tilde{m}_k - \hat{m}_k$ とし，値利得 (mean gain) $G_m(Y) = \sum_{k\in T}\mathbb{E}[\Delta\hat{m}_k]$ とする．Wu らが提案するポイズニング攻撃は次の 3 つである．

・**Maximal Gain Attack(M2GA)**　サーバに送信する局所差分プライバシプロトコルの出力である key-value データとそのインデックスを意図的に加工する．

・**Random Message Attack(RMA)**　サーバに送信する key-value データとそのインデックスをランダムに選択する．

・**Random Key-Value pair Attack(RKVA)**　摂動する前の key-value データとそのインデックスを操作して定められた摂動化を行う．

2.4.2 M2GA

M2GA は，出力を操作する攻撃である．また，偽ユーザは頻度利得と値利得の両方が最大になるように出力を作成する．このとき，頻度利得と値利得は最尤推定法により算出されたものとする．n_1^k，n_{-1}^k をそれぞれ真のユーザの中で $\langle k,1\rangle$，$\langle k,-1\rangle$ を出力した人数とする．また，$\tilde{n}_1^k$，$\tilde{n}_{-1}^k$ をそれぞれ偽ユーザの中で $\langle k,1\rangle$，$\langle k,-1\rangle$ を出力した人数とする．PrivKV ではユーザは，1 つの key-value データとそのインデックスを出力する．

まず，頻度利得が最大になる条件を考える．頻度利得 $G_f(Y)$ は以下のように式変形できる．

$$\begin{aligned} G_f(Y) &= \sum_{k\in T}\mathbb{E}[\tilde{f}_k] - \mathbb{E}[\hat{f}_k] \\ &= \sum_{k\in T}\left\{\mathbb{E}\left[\frac{(n_1^k + n_{-1}^k + \tilde{n}_1^k + \tilde{n}_{-1}^k)/(n+m) - q_1}{p_1 - q_1}\right]\right. \\ &\quad \left. - \mathbb{E}\left[\frac{(n_1^k + n_{-1}^k)/n - q_1}{p_1 - q_1}\right]\right\} \end{aligned}$$

なので，偽ユーザの出力の集合 Y は，$\sum_{k\in T}\mathbb{E}[\frac{\tilde{n}_1+\tilde{n}_{-1}}{(n+m)(p_1-q_1)}]$ に影響する．また，$(n+m)(p_1-q_1)$ は定数であるため，頻度利得 $G_f(Y)$ を最大にするためには，$\sum_{k\in T}\mathbb{E}[\tilde{n}_1+\tilde{n}_{-1}]$ を最大にする必要がある．偽ユーザの全てがターゲット key T から意図的に特定のインデックス k を選択し，$\langle 1,1\rangle$ か $\langle 1,-1\rangle$ を送信したとき，$\sum_{k\in T}(\mathbb{E}[\tilde{n}_1^k]+\mathbb{E}[\tilde{n}_{-1}^k]) = m$ となり，頻度利得 $G_f(Y)$ が最大となる．

次に，同様に値利得が最大になる条件を考える．値利得 $G_m(Y)$ は以下のように式変形できる．

$$\begin{aligned} G_m(Y) &= \sum_{k\in T}\mathbb{E}[\tilde{m}_k] - \mathbb{E}[\hat{m}_k] \\ &= \sum_{k\in T}\left\{\mathbb{E}\left[\frac{n_1^k - n_{-1}^k + \tilde{n}_1^k - \tilde{n}_{-1}^k}{(p_2-q_2)(n_1^k + n_{-1}^k + \tilde{n}_1^k + \tilde{n}_{-1}^k)}\right]\right. \\ &\quad \left. - \mathbb{E}\left[\frac{n_1^k + n_{-1}^k}{(p_2-q_2)(n_1^k + n_{-1}^k)}\right]\right\} \end{aligned}$$

$c_1^k = nf_k(p_2-q_2)m_k$，$c_2^k = nf_k$ とすると，値利得 $G_m(Y)$ を最大にするためには，$\sum_{k\in T}\frac{\mathbb{E}[\tilde{n}_1-\tilde{n}_{-1}]+c_1^k}{\mathbb{E}[\tilde{n}_1+\tilde{n}_{-1}]+c_2^k}$ を最大にする必要がある．なので，偽ユーザの全てがターゲット key T から意図的に特定のインデックス k を選択し，value= 1 を送信したとき，$\mathbb{E}[\tilde{n}_1^k] = \frac{m}{r}$，$\mathbb{E}[\tilde{n}_{-1}^k] = 0$ となり，値利得 $G_m(Y)$ が最大となる．つまり，M2GA では，ターゲット key の頻度利得と値利得の両方を最大にするために，全偽ユーザはターゲット key を 1 つ任意に選択し，$\langle 1,1\rangle$ を送信する．

2.4.3 RMA

RMA は，出力をランダムに選択する攻撃である．RMA では偽ユーザは K から key を一つランダムに選択し，出力に関しても出力候補の中からランダムに選択する．つまり，$\frac{1}{2}$ の確率で，$\langle 0,0\rangle$ を送信し，$\frac{1}{4}$ の確率で，$\langle 1,1\rangle$ か $\langle 1,-1\rangle$ を送信する．このとき，偽ユーザは $\frac{1}{2d}$ の確率で，key k に対する推定値を操作する．つまり，$E[\tilde{n}_1^k] = E[\tilde{n}_{-1}^k] = \frac{m}{4d}$ となる．

2.4.4 RKVA

RKVA は，摂動対象を操作する攻撃である．RKVA では，偽ユーザはターゲット key$T$ から key を一つランダムに選択する．また，偽ユーザは，$\langle 1,1\rangle$ を摂動対象として PrivKV を適用し，得られた出力を送信する．このとき，$\mathbb{E}[\tilde{n}_1^k] = \frac{me^{\epsilon_1}e^{\epsilon_2}}{r(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)}$，$\mathbb{E}[\tilde{n}_{-1}^k] = \frac{me^{\epsilon_1}}{r(e^{\epsilon_1}+1)(e^{\epsilon_2}+1)}$ となる．

3. 提案手法

PrivKV の集計手法では，最尤推定法が用いられている．最尤推定法では，集計データの最尤値を推定値とするためデータの偏りがある場合，推定誤差が大きくなる問題がある．特にポイズニング攻撃のように偽ユーザが特定の key-value データを送信する場合，推定誤差が大きくなると考えられる．また，PrivKVM では，偽ユーザにより操作された平均値を用いて摂動を行うため，対話ごとに平均値の誤差が大きくなると考えられる．そこで本稿では，ポイズニング攻撃に対して強固であると考えられる EM アルゴリズムに着目し，key-value データにおける局所差分プライバシプロトコル PrivKV に対して EM アルゴリズムを適用した局所差分プライバシプロトコル emPrivKV を提案する．

3.1 EM アルゴリズム

EM アルゴリズムとは，ベイズの定理を利用した反復方

式 [7] である．d 種類の入力の集合を $X = \{x_1, x_2..., x_d\}$，d' 種類の出力の集合を $Z = \{z_1, z_2..., z_{d'}\}$ とする．n 人のユーザがそれぞれ自身の持つ値 $x_i \in X$ を入力とし，ランダムアルゴリズムを適用し，出力 $z_j \in Z$ を送信する．反復回数を t として，出力の集計から d 個の入力について度数推定を行う．t 回目の x_i 推定値を $\theta^{(t)} = ({\theta_1}^{(t)}, {\theta_2}^{(t)}, ..., {\theta_d}^{(t)})$ とし，初期値を $\theta^{(0)} = (\frac{1}{d}, \frac{1}{d}, ..., \frac{1}{d})$ とする．また，ユーザ u の t 回目の X に対する推定値を $\hat{\theta}_u^{(t)} = (\hat{\theta}_{u,1}^{(t)}, \hat{\theta}_{u,2}^{(t)}, ..., \hat{\theta}_{u,d}^{(t)})$ とする．

入力 x_i に対して出力 z_j となる条件付き確率は，

$$Pr[z_j|x_i] = \frac{Pr[z_j, x_i]}{Pr[x_i]}$$

となる．また，ベイズの定理より，出力 z_j で条件付けられたとき入力が x_i である確率は，

$$Pr[x_i|z_j] = \frac{Pr[z_j|x_i]Pr[x_i]}{\sum_{s=1}^{|X|} Pr[z_j|x_s]Pr[x_s]}$$

となり，$t-1$ 回目の推定出力 z_j に対する入力 x_i の t 回目の推定確率は，

$$\hat{\theta}_{u,i}^{(t)} = Pr[x_i|z_j] = \frac{Pr[z_j|x_i]{\theta_i}^{(t-1)}}{\sum_{s=1}^{|X|} Pr[z_j|x_s]{\theta_s}^{(t-1)}}$$

で更新される．$t-1$ 回目の推定値 $\theta^{(t-1)}$ を用いて，ユーザごとに $\hat{\theta}_u^{(t-1)}$ を計算し，$\hat{\theta}_u^{(t-1)}$ の平均値を $\theta^{(t)}$ として更新する．

$$\theta^{(t)} = \frac{1}{n}\sum_{u=1}^{n} \hat{\theta}_u^{(t-1)}$$

これをあらかじめ定めた閾値 $\eta > 0$ に対して，${\theta_i}^{(t)}$ が $|{\theta_i}^{(t)} - {\theta_i}^{(t-1)}| \leq \eta$ となって収束するまで繰り返す．これにより，入力 x_i の度数を推定する．

3.2 emPrivKV の提案

ユーザ i の入力 S_i' について，摂動化する key-value データ $\langle k'_a, v'_a \rangle \in S_i'$ の vlaue の値は $v'_a \in V(= [-1, 1])$ の連続値であるため，value が離散化された出力 $\langle {k^*}_a, {v^+}_a \rangle$ から $\langle k'_a, v'_a \rangle$ を推定することは困難である．そこで，出力 $\langle {k^*}_a, {v^+}_a \rangle$ から摂動工程の中の VPP を適用し value を 2 値化した key-value データ $\langle k'_a, {v^*}_a \rangle$ の度数を EM アルゴリズムを用いて推定する．このとき，推定する度数の集合 X は，$X = \{\langle 1,1\rangle, \langle 1,-1\rangle, \langle 0,1\rangle, \langle 0,-1\rangle\}$ となり，出力の集合 Z は，$Z = \{\langle 1,1\rangle, \langle 1,-1\rangle, \langle 0,0\rangle\}$ となる．n 人のユーザから出力 $\langle {k^*}_a, {v^+}_a \rangle \in Z$ を観測する．$k_a \in K$ について初期値は $\theta^{(0)} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ とする．t 回の反復を行なった結果得られた入力 X の推定度数を $\theta^{(t)} = (\theta^{(t)}_{\langle 1,1\rangle}, \theta^{(t)}_{\langle 1,-1\rangle}, \theta^{(t)}_{\langle 0,1\rangle}, \theta^{(t)}_{\langle 0,-1\rangle})$ とする．

$\langle {k'}_a, {v^*}_a \rangle$ が $\langle 1,1\rangle, \langle 1,-1\rangle$ であることは，k_a について v_a の値を保有していることを示し，$\langle {k'}_a, {v^*}_a \rangle$ が $\langle 0,1\rangle$，$\langle 0,-1\rangle$ であることは k_a について v_a の値を保有していないことを示す．k_a について，n 人のユーザの中で，$\langle {k'}_a, {v^*}_a \rangle = \langle 1,1\rangle$ の割合を $\delta_{\langle 1,1\rangle}$，$\langle {k'}_a, {v^*}_a \rangle = \langle 1,-1\rangle$ の割合を $\delta_{\langle 1,-1\rangle}$ とする．また，n 人のユーザの入力 S_i' の中で，k_a についての値 v_a を保有しているユーザの割合を δ^* とする．$d(= |S_i'|)$ 種類の key からランダムに 1 つの key-value データ $\langle {k'}_a, {v^*}_a \rangle$ を選択するとすると，

$$\mathbb{E}[\delta^*] = \mathbb{E}[\delta_{\langle 1,1\rangle} + \delta_{\langle 1,-1\rangle}]$$

となる．$\delta_{\langle 1,1\rangle}$ の推定値を $\theta_{\langle 1,1\rangle}$，$\delta_{\langle 1,-1\rangle}$ の推定値を $\theta_{\langle 1,-1\rangle}$ とすると，k_a について，ユーザが v_a の値を保有する確率は，$\theta_{\langle 1,1\rangle} + \theta_{\langle 1,-1\rangle}$ となるため，推定度数 $\hat{f}_a$ は，

$$\hat{f}_a = \theta_{\langle 1,1\rangle} + \theta_{\langle 1,-1\rangle}$$

となる．また，v_a の平均値 m_a の期待値は，

$$\mathbb{E}[m_a] = \mathbb{E}\left[\frac{\delta_{\langle 1,1\rangle} - \delta_{\langle 1,-1\rangle}}{\delta_{\langle 1,1\rangle} + \delta_{\langle 1,-1\rangle}}\right]$$

となるので，推定平均値 $\hat{m}_a$ は

$$\hat{m}_a = \frac{\theta^{(t)}_{\langle 1,1\rangle} - \theta^{(t)}_{\langle 1,-1\rangle}}{\theta^{(t)}_{\langle 1,1\rangle} + \theta^{(t)}_{\langle 1,-1\rangle}}$$

となる．提案手法を図 1 に示す．

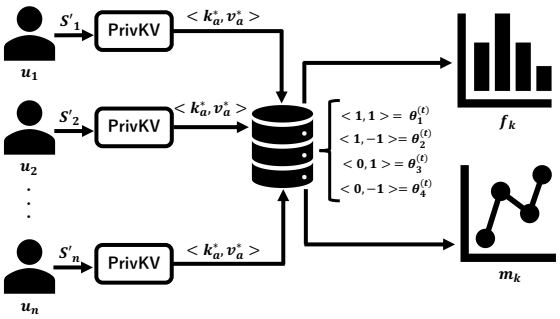


図 1: システム構成図

数値例 k_a の度数 $\hat{f}_a$ と平均値 $\hat{m}_a$ を推定することを考える．ユーザ u の出力 $\langle {k^*}_a, {v^+}_a \rangle \in Z$ が $z_1 = \langle 1,1\rangle$ であったとする．度数を推定する key-value データ $X = (\langle 1,1\rangle, \langle 1,-1\rangle, \langle 0,1\rangle, \langle 0,-1\rangle)$ の度数初期値を $\theta^{(0)} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ とする．出力 $\langle {k^*}_a, {v^+}_a \rangle$ が $z_1 = \langle 1,1\rangle$ であったとき，value を 2 値化した key-value データ $\langle {k'}_a, {v^*}_a \rangle$ が $x_1 = \langle 1,1\rangle$ である確率は，ベイズの定理を用いて，

$$\begin{aligned}
Pr[x_1|z_1] &= \frac{Pr[z_1|x_1]Pr[x_1]}{\sum_{s=1}^{4} Pr[z_1|x_s]Pr[x_s]} \\
&= \frac{Pr[z_1|x_1]{\theta_1}^{(0)}}{\sum_{s=1}^{4} Pr[z_1|x_s]{\theta_s}^{(0)}} \\
&= \frac{\frac{1}{4}p_1p_2}{\frac{1}{4}p_1p_2 + \frac{1}{4}p_1q_2 + \frac{1}{4}q_1p_2 + \frac{1}{4}q_1q_2} \\
&= \frac{p_1p_2}{p_1(p_2+q_2) + q_1(p_2+q_2)} \\
&= p_1p_2 = \frac{e^{\epsilon_1}e^{\epsilon_2}}{(1+e^{\epsilon_1})(1+e^{\epsilon_2})}
\end{aligned}$$

となる．$\epsilon = 1$，$\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$ とすると，$\hat{\theta}^{(1)}_{1,u}$ は，

$$\hat{\theta}^{(1)}_{1,u} \approx 0.387455$$

となる．出力 z_1 に対する入力 x_2, x_3, x_4 の確率についても同様に計算し，全てのユーザの平均を $\theta^{(1)}$ として更新する．

4. 実験

emPrivKV, PrivKV, PrivKVM に対して MG2A, RMA, RKVA の 3 手法でポイズニング攻撃を行い，ポイズニング攻撃に対する強度を調査する．PrivKVM では推定平均値を用いてユーザと 3 回の対話を行い，EM アルゴリズムを用いた推定では対話を行わない．

4.1 データセット

データセットには, key と value がガウス分布 ($\mu = 0, \sigma = 10$) に従う合成データを使用する．全てのユーザが 50 の key-value データを保持する．

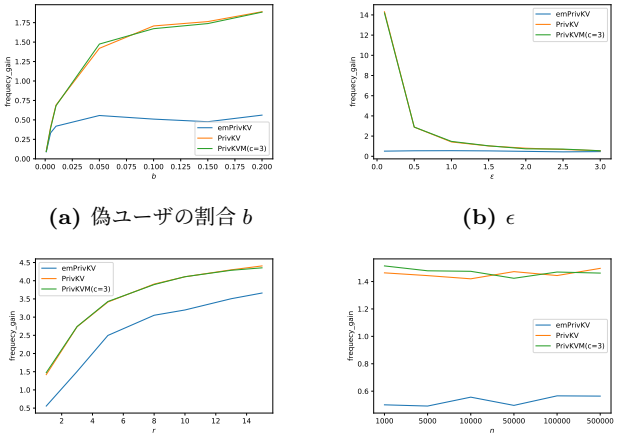
4.2 評価方法

安全性指標 ϵ，真ユーザに対する偽ユーザの割合 $b = \frac{m}{n}$，ターゲット key 数 r，ユーザ数 n を変化させ，頻度利得 (frequency gain) と値利得 (mean gain) を算出し，攻撃に対する強度を求める．頻度利得, 値利得が小さいほどポイズニング攻撃に対する変化量が小さく，ポイズニング攻撃に対して強固である．パラメータの初期設定を $\epsilon = 1$, $b = 0.05$, $r = 1$, $n = 10^4$ とする．ポイズニング攻撃によるターゲット key k の推定度数の変化量を $\Delta\hat{f}_k$，推定平均値値の変化量を $\Delta\hat{m}_k$ とし，頻度利得 (frequency gain) $G_f(Y) = \sum_{k \in T} E[\Delta\hat{f}_k]$，値利得 (mean gain) $G_m(Y) = \sum_{k \in T} E[\Delta\hat{m}_k]$ を算出する．パラメータごとに 50 回の試行を行い，結果の平均を評価値とする．

4.3 実験結果

4.3.1 M2GA

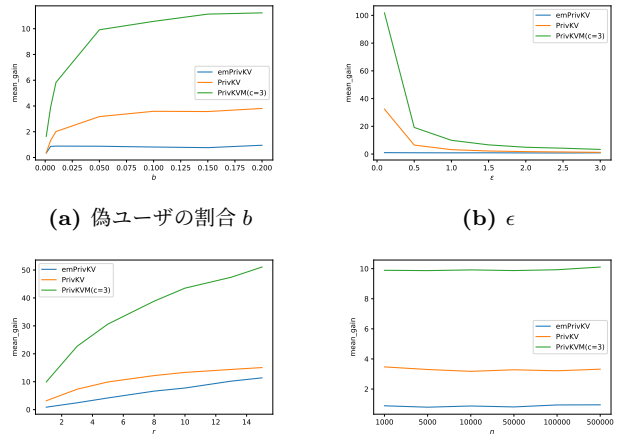
M2GA の各パラメータによる頻度利得の変化を図 2 に，値利得の変化を図 3 に示す．偽ユーザの割合 b を増加させると，PrivKV と PrivKVM の頻度利得は増加するが，emPrivKV では $b \geq 0.05$ で頻度利得は増加せず，最も攻撃による影響が小さい．偽ユーザが増えるほど，emPrivKV と PrivKV の頻度利得の差が大きくなる．安全性指標 ϵ を変化させた結果，emPrivKV では，ϵ が小さい時でも他プロトコルに比べ頻度利得が小さい．また，ターケット key 数 r の数にかかわらず，emPrivKV の頻度利得が 3 プロトコルの中で最も安定して小さい．3 プロトコルとも真ユーザ数 n を増加させても，頻度利得に大きな変化は見られない．emPrivKV が平均で 63.7%攻撃の影響を改善した．また，値利得はどの変数の組み合わせでも emPrivKV が M2GA に対して強度がある．



(a) 偽ユーザの割合 b **(b)** ϵ

(c) ターゲット key 数 r **(d)** 真ユーザ数 n

図 2: frequency gain (M2GA)



(a) 偽ユーザの割合 b **(b)** ϵ

(c) ターゲット key 数 r **(d)** 真ユーザ数 n

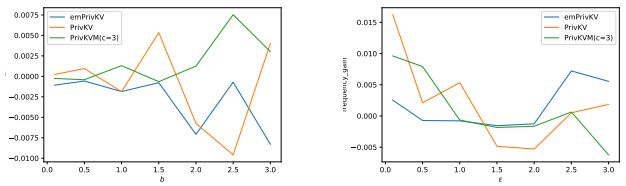
図 3: mean gain (M2GA)

4.3.2 RMA

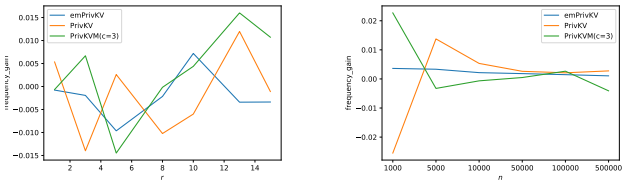
RMA の各パラメータによる頻度利得の変化を図 4 に，値利得の変化を図 5 に示す．頻度利得は，$n = 0.1$ のとき PrivKV が -0.028 であるが，真ユーザ数 n が小さいときでも，emPrivKV は攻撃の影響を受けていない．その他のパラメータを変化させたとき，3 つのプロトコルでの頻度利得は小さく，度数推定に大きな影響を受けていない．値利得では，PrivKVM は ϵ が小さいときやターゲット key 数 r が大きいときに値利得が大きくなっているが，どのパラメータの場合でも emPrivKV と PrivKV に差はない．

4.3.3 RKVA

RKVA の各パラメータによる頻度利得の変化を図 6 に，値利得の変化を図 7 に示す．RKVA の頻度利得では M2GA の頻度利得と同様に，どのパラメータの組み合わせでも emPrivKV が他のプロトコルに比べ小さい．特に ϵ が小さいとき，頻度利得の差が大きいくなる．また，値利得は，真ユーザに対する偽ユーザの割合が大きくなるほど emPrivKV と PrivKV の差が大きくなる．しかし，ター

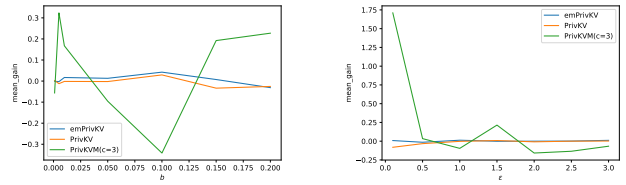


(a) 偽ユーザの割合 b **(b)** ϵ

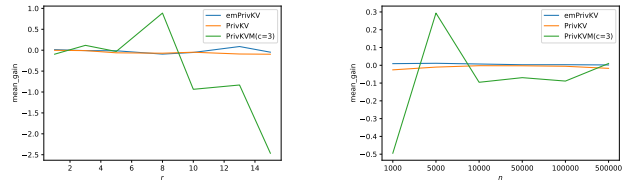


(c) ターゲット key 数 r **(d)** 真ユーザ数 n

図 4: frequency gain (RMA)

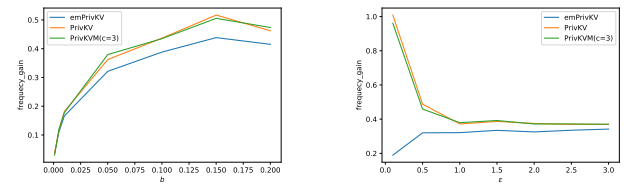


(a) 偽ユーザの割合 b **(b)** ϵ

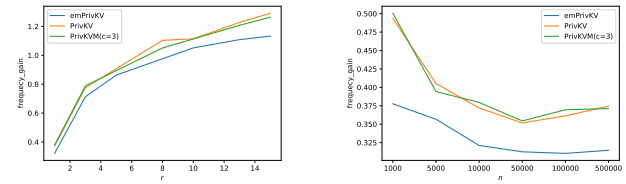


(c) ターゲット key 数 r **(d)** 真ユーザ数 n

図 5: mean gain (RMA)



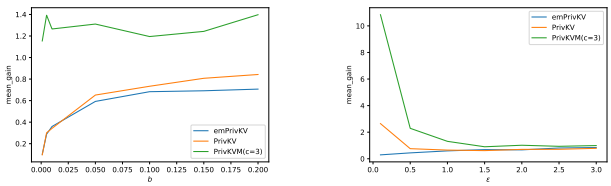
(a) 偽ユーザの割合 b **(b)** ϵ



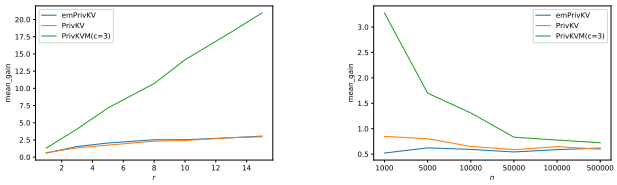
(c) ターゲット key 数 r **(d)** 真ユーザ数 n

図 6: frequency gain (RKVA)

ゲット key 数 r を変化させても emPrivKV と PrivKV では値利得に差は見られなかった．



(a) 偽ユーザの割合 b **(b)** ϵ



(c) ターゲット key 数 r **(d)** 真ユーザ数 n

図 7: mean gain (RKVA)

4.4 考察

度数と平均値の推定に最尤推定法を用いる PrivKV では，最尤値を推定値とする．そのため，図 2 に示すように，ポイズニング攻撃のようなデータの偏りを大きくする攻撃に対して影響を受けやすいと考えられる．また，平均値推定に関して PrivKV では，出力 v_k^+ から $v'_k = 1$ と $v'_k = -1$ の推定を行っており，$v'_k = 0$ の場合を考慮していない．そのため，図 3 に示すように，偽ユーザの全てが $\langle 1, 1\rangle$ を送信する M2GA では，攻撃の影響を強く受け値利得が大きくなると考えられる．一方で，emPrivKV では，ベイズの定理を用いた反復手法であるため，データの偏りに対する影響が小さい．また，出力 $\langle 1, 1\rangle$ から摂動対象が $\langle 1, 1\rangle$ や $\langle 1, -1\rangle$ の割合だけでなく，$\langle 0, 0\rangle$ の割合も考慮しているため，偽ユーザからの出力が全て $\langle 1, 1\rangle$ であっても，攻撃の影響を受けづらいと考える．

4.5 ポイズニング攻撃の対策

本稿では，PrivKV に EM アルゴリズムを適用し，ポイズニング攻撃に対してより強固である emPrivKV を提案した．しかし，M2GA のような攻撃に対しては，emPrivKV でさえ推定値が大きく変化した．Wu らは，PrivKVM のような対話型のプロトコルに関して，対話ごとの出力であるインデックスを参照することで偽ユーザを識別する手法 [6] を提案した．しかし，emPrivKV はユーザとの対話を行わないため，出力から偽ユーザを検出することは困難である．

そこで，ターゲット key の分析結果を操作するポイズニング攻撃，M2GA や RKVA への対策として，サーバ側が各ユーザがランダマイズする key-value データのインデックスをランダムに選択する手法が有効であると考える．紛失通信などの適切な暗号アルゴリズムを用いることで，ユーザはサーバが選択したインデックスの key-value データを偽ることなく送信することが可能である．このように変更

して，ユーザが特定のインデックスの key-value データをランダマイズしても局所差分プライバシを満たす．サーバー側がインデックスを選択することで，ユーザは特定の key に対する分析結果を操作することができず，度数利得や値利得を小さくすることができる．また，サーバが選択したインデックスと出力のインデックスを比較することで，偽ユーザを検出することができる．

5. おわりに

局所差分プライアバシは攻撃者が複数の偽ユーザを偽り特定の情報を送信し，推定結果を操作するポイズニング攻撃に対して脆弱である．key-value データにおける局所差分プライバシアルゴリズム PrivKV では，推定に最尤推定法を用いているためポイズニング攻撃の影響を大きく受けてしまう．そこで本稿では，ポイズニング攻撃に対して強固であると考えられる EM アルゴリズムに注目し，PrivKV に EM アルゴリズムを適用した局所差分プライバシプロトコル emPrivKV を提案した．emPrivKV，PrivKV，PrivKVM に対して 3 種のポイズニング攻撃を行い，攻撃による影響を調査した．

その結果，emPrivKV では PrivKV と比較してポイズニング攻撃による影響が減少した．特に推定値に最も影響を与える M2GA では，emPrivKV を適用することで，偽ユーザの割合 b が 0.2 のとき，frequency gain は PrivKV と比較し 70.3%改善した．また，mean gain は偽ユーザの割合 b が 0.2 のとき，PrivKV と比較して 75%改善し，PrivKVM と比較し 91.5%改善した．

しかし，emPrivKV でさえも M2GA のようなポイズニング攻撃により推定値が大きく変化してしまう．そこで，ユーザが出力するインデックスをサーバ側がランダムに選択する手法を提案した．実データを用いたポイズニング攻撃に対する emPrivKV の強度の調査と正確に偽ユーザを検出する手法の検討を今後の課題とする．

謝辞 本研究は，JSPS 科研費 JP18H04099 と JST CREST JPMJCR21M1 の助成を受けたものである．

参考文献

[1] C. Dwark, F. McSherry, K. Nissim, A. Smith, “Calibrating noise to sensitivity in private data analysis”, TCC, Vol. 3876, pp. 265-284, 2006.

[2] J. C. Duchi, M. I. Jordan, M. J. Wainwright, “Local privacy and statistical minimax rates”, FOCS, pp. 429-438, 2013.

[3] S. L. Warner, “Randomized response: A survey technique for eliminating evasive answer bias”, Journal of the American Statistical Association, pp. 63-69, 1965.

[4] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, J. Shin, “Collectiong and analyzing data from smart device users with local differential privacy”, arXiv:1606.05053, 2016.

[5] X. Cao, J. Jia, N. Z. Gong, “Data poisoning attacks to local differential privacy protocols”, USENIX Security Symposium, pp. 947-964, 2021.

[6] Y. Wu, X. Cao, J. Jia, N. Z. Gong, “Poisoning Attacks to Local Differential Privacy Protocols for Key-Value Data”, USENIX Security Symposium, pp. 519-536, 2022.

[7] 宮川雅巳, “EM アルゴリズムとその周辺”, 応用統計学, Vol. 16, No. 1, pp. 1-19, 1987.

[8] Q. Ye, H. Hu, X. Meng, H. Zheng, “PrivKV : Key-Value Data Collection with Local Differential Privacy”, IEEE S&P, pp. 294-308, 2019.