Key-ValueデータのLDPプロトコルPCKVの推定値操作攻撃の提案と評価

谷口 輝海[1,a)]　菊池 浩明[2,b)]

概要：スマートデバイスの普及にともに，企業はユーザの行動データを収集し，活用することで，サービス向上を図っている．しかし，ユーザの端末から収集されるデータは個人識別性が高く，データのプライバシ保護が重要な課題となっている．そこで，データをユーザの端末上でランダム化してサーバに送信する局所差分プライバシー (LDP) 技術が注目されている．しかし，LDP は悪意のあるユーザが細工したデータを送信することで，推定統計値を歪めるポイズニング攻撃に対して脆弱である．2024 年に Li らは，平均値推定を行う LDP 方式に対し，平均値を攻撃者の意図した値に操作する推定値操作 (Fine-Grained Poisoning attack) が可能であることを示した．そこで，我々は，単純な一次元の平均値推定ではなく，Key-Value データに対して LDP の安全性を保証したプロトコル PCKV に対する新たな推定値操作攻撃を提案する．推定平均と推定頻度を同時に歪める攻撃を理論的，実験的に評価する．

キーワード：局所差分プライバシー，PCKV

Proposal and Evaluation of Estimated Statistics Manipulation Attack on LDP Protocol PCKV for Key-Value Data

Terumi Yaguchi[1,a)]　Hiroaki Kikuchi[2,b)]

Abstract: With the spread of smart devices, companies improve their services by collecting and utilizing users' behavioral data. However, the collected data from the user's device is subject to be identify individuals, and hence privacy protection is required. Local Differential Privacy (LDP) is a technique that perturbs the user's data before sending to the server so that the server is not able to have access to private data. Unfortunately, LDP is vulnerable to a poisoning attack in which a set of malicious users disrupt the estimated statistics by sending crafted data. In 2024, Li et al. showed that fine-grained manipulation of the estimated means is feasible. In this work, we study a new fine-grained attack to a multidimensional data with LDP known as PCKV for Key-Value data. We evaluate the proposed fine-grained attack to PCKV from both theoretical and empirical viewpoints.

Keywords: Local Differential Privacy, PCKV

1. はじめに

近年，あらゆるデバイスがインターネットに接続されるようになり，サービス事業者はユーザの行動データを収集し，活用することでサービス向上を図っている．しかし，行動データには機微な情報が含まれており，ユーザのデータをそのまま収集することはプライバシ侵害に繋がる可能性がある．そこで，Duchi[1] らによって局所差分プライバシが提案された．局所差分プライバシでは，データを収集する際，データに確率的なノイズを加えることで，収集者に対して真のデータを秘匿することができる．これにより，真のデータはユーザのデバイスにのみ保存されるため，例

[1] 明治大学 大学院
Meiji University Graduate School
[2] 明治大学
Meiji University
[a)] cs242036@meiji.ac.jp
[b)] kikn@meiji.ac.jp

えサーバのデータが漏洩したとしても第三者が真のデータを知ることはない.

しかし，局所差分プライバシのモデルでは，ユーザのデバイス上でローカルにランダマイズを行うため，悪意のあるユーザが意図的に細工したデータを送信するポイズニング攻撃に対して脆弱であることが知られている [4][5]. 2023年に Li らは，連続値の平均値と分散を推定する局所差分プライバシプロトコルに対して，2 つの推定値を任意の値に操作する Fine-grained Poisoning Attack(推定値操作攻撃)[3] を提案した.

本研究では，Li らの研究に着目する．Li らの攻撃の対象は，連続値の 1 次元データであることが仮定されていた．従って，離散値の key に対する連続値の value が管理されている key-value データに対して，適用するのは自明ではない．標的とする key や value に依っては，多量の不正者を動員する必要があり，攻撃が困難な場合が考えられるためである．加えて，key-value データは，情報推薦や大規模データベースなどに広く活用されており，応用範囲も広く，その安全性を高めることは有用である．そこで，本研究では，最新の，Key-Value データのための局所差分プライバシプロトコルである PrivKV[12] と PCKV[11] を対象とした提案攻撃が成功するための十分条件を明らかにし，オープンデータを用いてその有効性を評価する.

2. 準備

2.1 局所差分プライバシ

局所差分プライバシでは，任意の異なる入力に対し，ランダム化アルゴリズム M の出力が等しくなる確率が近いことを保証する．これにより，データ収集者を含む第三者は M の出力から真の入力値を特定することができず，ユーザのプライバシが保証される．ランダム化アルゴリズム M について，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

$\mathcal{X}$ を入力値の集合，$\mathcal{Y}$ を出力値の集合とする．M を入力 $x \in \mathcal{X}$ に対して $y \in \mathcal{Y}$ を出力するランダム化アルゴリズムとする．$\epsilon \in \mathbb{R}^+$ が与えられたとき，任意の 2 つの入力 $x, x' \in \mathcal{X}$ と任意の出力 $y \in \mathcal{Y}$ に対して，

$$Pr[M(x,\epsilon) = y] \leq e^{\epsilon} Pr[M(x',\epsilon) = y]$$

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

2.2 PrivKV

Ye らは，離散値と連続値の組で構成される Key-Value データに対する局所差分プライバシプロトコルとして，PrivKV[12] を提案した. PrivKV では，Key の値の集合から一様ランダムに key を 1 つサンプリングした後，key に対して Randomize Response(RR)[2] を，Value に対して確率的丸め込みメカニズムのひとつである Harmony を適用することでランダム化を行う．しかし，RR と Harmony を key と value に独立して適用してしまうと，key と value の間の相関が失われてしまうという課題がある．そこで，PrivKV では，key と value を同時にランダム化することで相関を保ったままデータの収集を行う．収集者は key k の頻度と key k に関する value の平均値を推定する.

key の集合を $\mathcal{K} = \{1, \ldots, d\}$ とし，要素数を $d = |\mathcal{K}|$ とする. また，value の定義域の集合を $[-1, 1]$ で表す. i 番目のユーザ u_i が持つ ℓ_i 個の Key-Value ペア (以下 KV ペア) の集合を $S_i = \{\langle k_j, v_j \rangle | j \in \{1, \ldots, \ell_i\}, k_j \in \mathcal{K}, v_j \in \mathcal{V}\}$ とする.

入力　ユーザ i の KV ペアの集合 S_i を入力とする.

サンプリング　$\mathcal{K}$ からランダムに一つ key j を選択し，ユーザは key j について報告を行う．このとき，key j を持つ KV ペアが S_i の中に存在するとき，$(j, \langle 1, v_j \rangle)$，存在しないとき，$(j, \langle 0, 0 \rangle)$ をランダム化アルゴリズムの入力とする.

摂動　$(j, \langle k'_j, v_j \rangle)$ を入力とする．ただし，$k'_j \in \{0, 1\}, j \in \mathcal{K}$ である.

value の摂動　$k'_j = 0$ の場合，v_j を $[-1, 1]$ からランダムに選択する．次に，v_j を v_j に依存する確率で v'_j に 2 値化する.

$$v'_j = \begin{cases} 1 & \text{w.p. } \frac{1+v_j}{2}, \\ -1 & \text{w.p. } \frac{1-v_j}{2}. \end{cases}$$

次に，確率 p_2, q_2 に基づき，ランダム化した結果を v^*_j とする.

$$v^*_j = \begin{cases} v'_j & \text{w.p. } p_2, \\ -v'_j & \text{w.p. } q_2, \end{cases}$$

ただし，

$$p_2 = \frac{e^{\epsilon_2}}{1 + e^{\epsilon_2}}, q_2 = \frac{1}{1 + e^{\epsilon_2}}$$

である.

key の摂動　key と value の相関を保つために, PrivKV では key の摂動にあわせて value も変化させる.

$k'_j = 1$ の場合,

$$\langle k^*_j, v^+_j \rangle = \begin{cases} \langle 1, v^+_j \rangle & \text{w.p. } p_1, \\ \langle 0, 0 \rangle & \text{w.p. } q_1, \end{cases}$$

$k'_j = 0$ の場合,

$$\langle k^*_j, v^+_j \rangle = \begin{cases} \langle 0, 0 \rangle & \text{w.p. } p_1, \\ \langle 1, v^+_j \rangle & \text{w.p. } q_1, \end{cases}$$

となる. ただし,

$$p_1 = \frac{e^{\epsilon_1}}{1 + e^{\epsilon_1}}, q_1 = \frac{1}{1 + e^{\epsilon_1}}$$

である. 摂動化結果 $\langle k_j^*, v_j^+ \rangle$ と key j をサーバに送信する．このとき PrivKV における全体の ϵ は, $\epsilon = \epsilon_1 + \epsilon_2$ となり，本稿では $\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$ を仮定する．

集計 n 人のユーザから収集した key j と KV ペア $\langle k_j^*, v_j^+ \rangle$ を用いて各 key の真の頻度と平均を推定する．

頻度推定 収集したデータの中で，インデックスが j であるもののうち，$k_j = 1$ であるペアの頻度を f_j' とすると，真の頻度は,

$$\hat{f}_j = \frac{p_1 - 1 + f_j'}{2p_1 - 1}$$

と推定される．

平均値推定 収集したデータの中で，インデックスが j であるもののうち，$v_j^+ = 1$ の度数を n_1^j，$v_j^+ = -1$ の度数を n_{-1}^j とする．$v_j = 1$ である KV ペアの推定度数 $\hat{n}_1^j$ と，$v_j = -1$ である KV ペアの推定度数 $\hat{n}_{-1}^j$ は,

$$N = n_1^j + n_{-1}^j$$
$$\hat{n}_1^j = \frac{N(p_2 - 1) + n_1^j}{2p_2 - 1}$$
$$\hat{n}_{-1}^j = \frac{N(p_2 - 1) + n_{-1}^j}{2p_2 - 1}$$

となり，key k の平均値 $\hat{\mu}_k$ は,

$$\hat{\mu}_k = \frac{\hat{n}_1^j - \hat{n}_{-1}^j}{N}$$

と推定される．

2.3 PCKV

Gu らは，PrivKV を改良したモデルとして，PCKV[11] を提案した．PrivKV では，Key の数が多い場合や，ユーザ数が少ない場合に推定精度が下がるという課題があった．PCKV ではサンプリング手法として，Padding-and-Sampling(PaD) を採用することでこれらの課題に対処している．

入力 ユーザ i の KV ペアの集合 S_i とパディング長 ℓ を入力とする．

サンプリング PCKV では PaD と呼ばれるサンプリング手法を用いて各ユーザが送信するデータをサンプリングする．PaD ではまず，B を $\eta = \frac{|S_i|}{\max(|S_i|, \ell)}$ についての $Bernoulli(\eta)$ 試行とする．$B = 1$ のとき，S_i からランダムに KV ペア $\langle k, v^* \rangle$ をサンプリングし，$B = 0$ のとき，$v^* = 0$ とし，k を $\{d+1, \ldots, d+\ell\}$ からランダムに選択する．v を確率 $\frac{1+v^*}{2}$ で 1，$\frac{1-v^*}{2}$ で -1 とし，$\langle k, v \rangle$ をランダム化アルゴリズムの入力とする．

摂動 $\langle k, v \rangle$ を入力とする．ただし，$k \in \mathcal{K} \cup \{d+1, \ldots, d+\ell\}, v \in \{-1, 1\}$ である．

PCKV-UE PCKV-UE は Unary Encoding[8] に基づいた摂動方法である．出力は $d + \ell$ 次元ベクトル $\boldsymbol{y} \in \{1, -1, 0\}^{d+\ell}$ であり，$\boldsymbol{y}$ の各要素は，k について

$$\boldsymbol{y}_k = \begin{cases} v & \text{w.p. } ap, \\ -v & \text{w.p. } a(1-p), \\ 0 & \text{w.p. } 1-a \end{cases}$$

$i \neq k$ について

$$\boldsymbol{y}_i = \begin{cases} 1 & \text{w.p. } \frac{b}{2}, \\ -1 & \text{w.p. } \frac{b}{2}, \\ 0 & \text{w.p. } 1-b \end{cases}$$

で定める. ただし，$i \in [d+\ell] \setminus \{k\}$ で,

$$a = \frac{1}{2}, b = \frac{2}{e^\epsilon + 3}, p = \frac{e^\epsilon}{e^\epsilon + 1}$$

である．

PCKV-GRR PCKV-GRR は Generalized Randomized Response[8] に基づいた摂動方法である．PCKV-GRR の出力 $\langle k', v' \rangle$ は,

$$\langle k', v' \rangle = \begin{cases} \langle k, v \rangle & \text{w.p. } ap, \\ \langle k, -v \rangle & \text{w.p. } a(1-p), \\ \langle i, 1 \rangle & \text{w.p. } \frac{b}{2}, \\ \langle i, -1 \rangle & \text{w.p. } \frac{b}{2} \end{cases}$$

で定める. ただし，$i \in [d+\ell] \setminus \{k\}$ で一様に選び,

$$a = \frac{\ell(e^\epsilon - 1) + 2}{\ell(e^\epsilon - 1) + 2(d+\ell)}, b = \frac{1-a}{d+\ell-1}, p = \frac{\ell(e^\epsilon - 1) + 1}{\ell(e^\epsilon - 1) + 2}$$

である．

集計 key k が与えられたとき，n_1^k, n_{-1}^k を PCKV-UE と PCKV-GRR のそれぞれについて，次のように定める．PCKV-UE では，n_1^k を $\boldsymbol{y}_k = 1$ であるユーザの数，n_{-1}^k を $\boldsymbol{y}_k = -1$ であるユーザの数と定義する．PCKV-GRR では，n_1^k を $\langle k, 1 \rangle$ であるユーザの数，n_{-1}^k を $\langle k, -1 \rangle$ であるユーザの数と定義する．このとき，PCKV-UE と PCKV-GRR の頻度と平均値を次のように推定する．

頻度推定 key k の推定頻度 $\hat{f}_k$ は,

$$\hat{f}_k = \frac{\frac{n_1^k + n_{-1}^k}{n} - b}{a - b} \ell$$

となる．

平均値推定 key k の推定平均値 $\hat{\mu}_k$ は,

$$\hat{\mu}_k = \frac{\hat{n}_1^k - \hat{n}_{-1}^k}{n \hat{f}_k} \ell$$

となる. ただし,

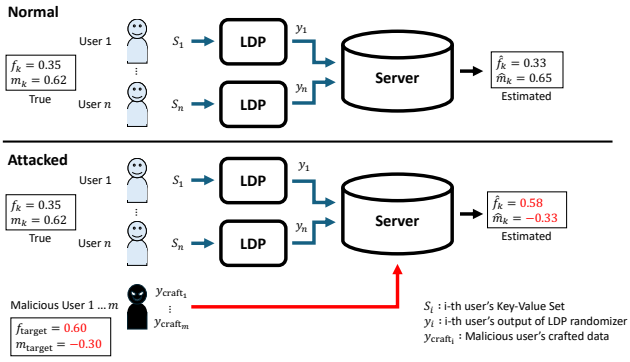

図 1　推定値操作攻撃の概要図

$$\begin{bmatrix} \hat{n}_1^k \\ \hat{n}_{-1}^k \end{bmatrix} = A^{-1} \begin{bmatrix} n_1^k - \frac{nb}{2} \\ n_{-1}^k - \frac{nb}{2} \end{bmatrix},$$

$$A = \begin{bmatrix} ap - \frac{b}{2} & a(1-p) - \frac{b}{2} \\ a(1-p) - \frac{b}{2} & ap - \frac{b}{2} \end{bmatrix}$$

である．なお，推定値を算出する際に，$\hat{f}_k$ は平均値推定に用いるために $[\frac{1}{n}, 1]$ にクリッピングし，$\hat{n}_1^k, \hat{n}_{-1}^k$ は $[0, \frac{n\hat{f}_k}{\ell}]$ にクリッピングを行う．

3. 提案方式

Li ら [3] は，一次元データの平均値推定を行う LDP プロトコルである Stocatstic Rounding[13] と Piecewise Mechanism[14] に対して推定平均と推定分散を攻撃者の意図した値に操作する推定値操作攻撃 (Fine-Graind Poisoning Attack) を提案した．我々はこれを応用し，より高次元の Key-Value データ収集のための LDP モデルである PrivKV, PCKV-GRR, PCKV-UE に対して，推定頻度と推定平均を標的値に操作する推定値操作攻撃を提案する．

3.1 概要

3.1.1 脅威モデル

図 1 に提案手法の概要を示す．攻撃者は LDP のシステムに対して m 人の偽ユーザを紛れ込ませることができるとする．また，攻撃者は LDP プロトコルのパラメータや推定方式をアクセスでき，摂動ステップを回避して意図したデータをサーバへ直接送信する．

攻撃者は攻撃対象となる key k について，推定頻度 $\hat{f}_k$ と推定平均値 $\hat{\mu_k}$ を任意の標的値 $f_{k,t}$ と $\mu_{k,t}$ に近づけることを目的とする．具体的には，偽ユーザを m 人追加することで，$\mathbb{E}[\hat{f}_k] = f_{k,t}, \mathbb{E}[\hat{\mu_k}] = \mu_{k,t}$ となるように偽ユーザの送信するデータを決定する．

3.2 定義

本節で使用する記号を表 1 に示す．

3.2.1 PrivKV に対する攻撃

PrivKV において，ランダム化後に送信するデータは，どの key について報告するかのインデックス j に加え，Key と Value の組，$\{(1,1),(1,-1),(0,0)\}$ のいずれかである．したがって，ターゲットとなる key k について，1 を送信する偽ユーザ数 m_1^k，-1 を送信する偽ユーザ数 m_{-1}^k，0 を送信する偽ユーザ数 m_0^k を定めればよい．推定頻度について，

$$\begin{aligned} \mathbb{E}[\hat{f}_k] &= \mathbb{E}\left[\frac{(n_1^k + n_{-1}^k + m_1^k + m_{-1}^k)/(n^k + m^k) - q_1}{p_1 - q_1}\right] \\ &\simeq \frac{(\mathbb{E}[n_1^k + n_{-1}^k] + m_1^k + m_{-1}^k)/(\mathbb{E}[n^k] + m^k) - q_1}{p_1 - q_1} \\ &= f_{k,t} \end{aligned}$$

となる．m_1^k, m_{-1}^k について整理し，

$$\mathbb{E}[n_1^k + n_{-1}^k] = \frac{nf_k}{d}, \mathbb{E}[n^k] = \frac{n}{d}$$

を代入すると，

$$m_1^k + m_{-1}^k = ((p_1 - q_1)f_{k,t} + q_1)(m^k + \frac{n}{d}) - \frac{nf_k}{d} \quad (1)$$

を得る．ただし，確率変数の商の期待値の近似値として，平均値周りでテイラー展開 [15] をしたときの第一項，すなわち，確率変数 X と Y についての，$\mathbb{E}[\frac{X}{Y}]$ を

$$\mathbb{E}[\frac{X}{Y}] \simeq \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}$$

により与えている．次に，推定平均値についても同様に，

$$\begin{aligned} \mathbb{E}[\hat{\mu}_k] &= \mathbb{E}\left[\frac{n_1^k - n_{-1}^k + m_1^k - m_{-1}^k}{(p_2 - q_2)(n_1^k + n_{-1}^k + m_1^k + m_{-1}^k)}\right] \\ &\simeq \frac{1}{p_2 - q_2} \cdot \frac{\mathbb{E}[n_1^k - n_{-1}^k] + m_1^k - m_{-1}^k}{\mathbb{E}[n_1^k + n_{-1}^k] + m_1^k + m_{-1}^k} \\ &= \mu_{k,t} \end{aligned}$$

であるので，m_1^k, m_{-1}^k について整理し，

$$\mathbb{E}[n_1^k - n_{-1}^k] = \frac{nf_k(p_2 - q_2)\mu_k}{d}$$

を代入すると，

$$q_2 m_1^k - p_2 m_{-1}^k = \frac{nf_k(p_2 - q_2)(\mu_{k,t} - \mu_k)}{d} \quad (2)$$

を得る．最終的に，(1), (2) 式と以下の連立方程式を解くこ

表 1　記号一覧

Notation	Description
β	偽ユーザの割合 ($\beta = \frac{m}{m+n}$)
ϵ	プライバシ予算
n_v^k	key k に関して value v を報告した真ユーザの数
m_v^k	key k に関して value v を報告した偽ユーザの数
n^k, m^k	key k に関して報告した真/偽ユーザの数
n, m	真/偽ユーザの総数
$f_{k,t}, \mu_{k,t}$	key k の目標頻度/平均
$\tilde{f}_k, \tilde{\mu}_k$	推定頻度/平均
f_k, μ_k	真の頻度/平均
$f_{k,e}, \mu_{k,e}$	攻撃者が推定する頻度/平均

とで，偽ユーザの送信するデータを決定する．

$$\begin{cases} m_1^k + m_{-1}^k = ((p_1 - q_1)f_{k,t} + q_1)(m^k + \frac{n}{d}) - \frac{nf_k}{d} & \text{(3a)} \\ q_2 m_1^k - p_2 m_{-1}^k = \frac{nf_k(p_2 - q_2)(\mu_{k,t} - \mu_k)}{d} & \text{(3b)} \\ m_1^k + m_{-1}^k + m_0^k = m^k & \text{(3c)} \\ 0 \le m_1^k, m_{-1}^k, m_0^k \le m^k & \text{(3d)} \end{cases}$$

3.2.2 PCKV に対する攻撃

PCKV において，ユーザがランダム化後に送信するデータは，PCKV-UE の場合は d 次元のベクトルであり，各要素の値は $\{1, -1, 0\}$ のいずれかであるので，ベクトルの k 番目の要素が $1, -1, 0$ である偽データの個数をそれぞれ，m_1^k, m_{-1}^k, m_0^k とする．PCKV-GRR の場合は，ユーザがランダム化後に送信するデータは，value が $\{1, -1\}$ のいずれかである KV ペアである．したがって，key が k であるデータのうち，value が $1, -1$ である偽データの数をそれぞれ，m_1^k, m_{-1}^k とし，key が k 以外である偽データの数を m_0^k とする．

推定頻度について，

$$\begin{aligned} \mathbb{E}[\hat{f}_k] &= \mathbb{E}\left[\frac{(n_1^k + n_{-1}^k + m_1^k + m_{-1}^k)/(n+m) - b}{a-b} \cdot \ell\right] \\ &= \frac{(\mathbb{E}[n_1^k + n_{-1}^k] + m_1^k + m_{-1}^k)/(n+m) - b}{a-b} \cdot \ell \\ &= f_{k,t} \end{aligned}$$

となる．m_1^k, m_{-1}^k について整理し，

$$\mathbb{E}[n_1^k + n_{-1}^k] = n\left(\frac{f_k}{\ell}a + \left(1 - \frac{f_k}{\ell}\right)b\right)$$

を代入すれば，

$$m_1^k + m_{-1}^k = \frac{a-b}{\ell}\left((m+n)f_{k,t} - f_k\right) + mb \qquad (4)$$

を得る．次に，推定平均値についても同様に，

$$\begin{aligned} \mathbb{E}[\hat{\mu}_k] &= \mathbb{E}\left[\frac{l}{(n+m)\hat{f}_k} \cdot \frac{n_1^k - n_{-1}^k + m_1^k - m_{-1}^k}{a(2p-1)}\right] \\ &\simeq \frac{l}{a(2p-1)(n+m)} \cdot \frac{\mathbb{E}[n_1^k - n_{-1}^k] + m_1^k - m_{-1}^k}{\mathbb{E}[\hat{f}_k]} \\ &= \mu_{k,t} \end{aligned}$$

であるので，m_1^k, m_{-1}^k について整理し，

$$\mathbb{E}[n_1^k - n_{-1}^k] = \frac{nf_k}{\ell} \cdot a(2p-1)\mu_k$$

を代入すると，

$$m_1^k - m_{-1}^k = \frac{a(2p-1)}{\ell} \cdot ((n+m)f_{k,t}\mu_{k,t} - nf_k\mu_k) \quad (5)$$

を得る．最終的に，式 (4), (5) と以下の連立方程式を解くことで，偽ユーザの送信するデータを決定する．

$$\begin{cases} m_1^k + m_{-1}^k = \frac{a-b}{\ell}\left((m+n)f_{k,t} - f_k\right) + mb & \text{(6a)} \\ m_1^k - m_{-1}^k = \frac{a(2p-1)}{\ell} \cdot ((n+m)f_{k,t}\mu_{k,t} - nf_k\mu_k) & \\ m_1^k + m_{-1}^k + m_0^k = m^k & \\ 0 \le m_1^k, m_{-1}^k, m_0^k \le m^k & \end{cases}$$

4. 理論評価

本節では，key k に対して攻撃を実行する際に必要な偽ユーザの数 m^k の範囲について議論する．

4.1 PrivKV に対する攻撃における偽ユーザ数

各パラメータが与えられたとき，攻撃のための偽ユーザのデータを決定するためには連立方程式 (3a), (3b), (3c), (3d) が解けなければならない．式 3a の右辺を $C_{PrivKV,+}(m^k)$，式 3b の右辺を $C_{PrivKV,-}(m^k)$ とする．このとき，m_1^k, m_{-1}^k, m_0^k は

$$\begin{aligned} m_1^k &= \frac{p_2 \cdot C_{PrivKV,+}(m^k) + C_{PrivKV,-}(m^k)}{p_2 + q_2} \\ m_{-1}^k &= \frac{q_2 \cdot C_{PrivKV,+}(m^k) - C_{PrivKV,-}(m^k)}{p_2 + q_2} \\ m_0^k &= m^k - m_1^k - m_{-1}^k \end{aligned}$$

と表せる．したがって，解があるための m^k の条件は，

$$\begin{aligned} 0 &\le \frac{p_2 \cdot C_{PrivKV,+}(m^k) + C_{PrivKV,-}(m^k)}{p_2 + q_2} \le m^k \\ 0 &\le \frac{q_2 \cdot C_{PrivKV,+}(m^k) - C_{PrivKV,-}(m^k)}{p_2 + q_2} \le m^k \\ 0 &\le m^k - m_1^k - m_{-1}^k \le m^k \end{aligned}$$

である．攻撃対象となる key が複数あるとき，攻撃対象の key の集合を T とすると，$m = \sum_{k \in T} m^k$ である．

4.2 PCKV に対する攻撃における偽ユーザ数

各パラメータが与えられたとき，攻撃のための偽ユーザのデータを決定するためには連立方程式 (6a), (6b), (6b), (6b) が解けなければならない．式 6a の右辺を $C_{PCKV,+}(m)$，式 6b の右辺を $C_{PCKV,-}(m)$ とする．このとき，m_1^k, m_{-1}^k, m_0^k は

$$\begin{aligned} m_1^k &= \frac{C_{PCKV,+}(m) + C_{PCKV,-}(m)}{2} \\ m_{-1}^k &= \frac{C_{PCKV,+}(m) - C_{PCKV,-}(m)}{2} \\ m_0^k &= m^k - m_1^k - m_{-1}^k \end{aligned}$$

と表せる．したがって，解があるための m の条件は，

$$\begin{aligned} 0 &\le \frac{C_{PCKV,+}(m) + C_{PCKV,-}(m)}{2} \le m^k \\ 0 &\le \frac{C_{PCKV,+}(m) - C_{PCKV,-}(m)}{2} \le m^k \\ 0 &\le m^k - m_1^k - m_{-1}^k \le m^k \end{aligned}$$

である．攻撃対象となる key が複数あるとき，PCKV-UE の場合，$m = \max_{k \in T} m^k$，PCKV-GRR の場合，$m = \sum_{k \in T} m^k$ である．

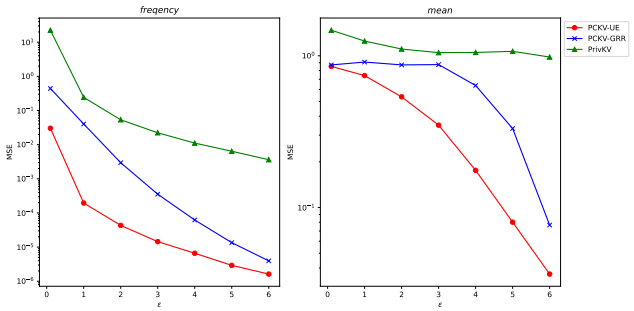

図 2 ϵ を変化させたときの推定頻度と推定平均の MSE(左: 頻度, 右: 平均値)

5. 実験評価

5.1 データセット

提案方式の評価には Clothing Fit Dataset[16] を用いる. 表 2 に使用したデータセットの概要を示す.

5.2 実験方法

各パラメータについて試行回数 $N = 50$ 回攻撃を実行し, 攻撃後の推定値と目標値の間の MSE を評価指標として用いる. すなわち, i 回目の攻撃後の推定頻度を $\hat{f}_{k,t_i}$, 推定平均を $\hat{\mu}_{k,t_i}$ とすると,

$$MSE_f = \frac{1}{N}\sum_{i=1}^{N}(f_k - \hat{f}_{k,t_i})^2$$

$$MSE_\mu = \frac{1}{N}\sum_{i=1}^{N}(\mu_k - \hat{\mu}_{k,t_i})^2$$

である.

実験で使用するパラメータを表 3 に示す. 攻撃者が用いる事前推定値 $f^*_{k,e}, \mu^*_{k,e}$ はランダムに選んだ 1000 ユーザの推定値である.

5.3 実験結果

5.3.1 推定精度

PrivKV, PCKV-UE, PCKV-GRR を用いて, 頻度の高い key の上位 50 個について, プライバシー予算 ϵ を変化させたときの推定頻度と推定平均値の MSE を図 2 に示す.

PrivKV では, サンプリング時に, 全ての key から一様ランダムに送信する key を決定する. そのため, ユーザが所持する KV ペアの数に対して, key の数が大きいとき, 各ユーザは高い確率で所持していないデータを送信することとなり, 推定精度が低下する. 一方, PCKV では, key の数の影響を受けずに, ユーザのデータをサンプリングすることが可能である. したがって, PrivKV は PCKV の 2 プロトコルに比べ, 頻度, 平均ともに MSE が大きく, 推定精度が低い.

また, PCKV の 2 プロトコルで推定精度を比較すると, key の値域の大きさの影響を受けない PCKV-UE が PCKV-GRR よりも高い精度で推定可能である.

5.3.2 攻撃精度

PCKV-UE, PCKV-GRR, PrivKV に対し, プライバシ予算 ϵ を変化させて推定値操作攻撃を行ったときの攻撃精度を図 3 に示す. 頻度の操作では, ϵ を増加させたとき, 3 つのすべてのプロトコルにおいて MSE が減少する. PCKV-UE と PCKV-GRR では, ϵ が小さいときは PCKV-UE の MSE が小さく, 高い精度で攻撃が可能であるが, ϵ が上昇するにつれて MSE の差は小さくなる. また, PCKV と PrivKV で比較すると, PrivKV の MSE は約 10^{-1} から 10^{-3} 倍ほど小さく, すべての ϵ で PrivKV が高精度で操作可能である. 平均値の操作では, PCKV の 2 つのプロトコルにおいて, ϵ の増加に伴い, MSE が単調に減少した. 一方, PrivKV は ϵ の値によらず MSE の値がほぼ一定である. また, 頻度推定と異なり, PrivKV の攻撃精度は最も低い.

PCKV-UE, PCKV-GRR, PrivKV に対し, 偽ユーザの割合 β を変化させて推定値操作攻撃を行ったときの攻撃精度を図 4 に示す. ϵ を変化させたときと同様に, 偽ユーザの割合が増加するほど MSE が減少する.

図 5, 6 に攻撃を実行するために, 攻撃者が初期推定する頻度 $f_{k,e}$ と平均値 $m_{k,e}$ についての攻撃精度を示す. 横軸は, 真の頻度 (平均値) と攻撃者の初期推定した頻度 (平均値) の差を示している. 頻度推定や, 平均値推定における PrivKV と PCKV-GRR による攻撃精度は, $f_{k,e}, m_{k,e}$ の初期推定精度に依存していない. 一方, 平均値推定における PCKV-UE では $f_{k,e}, m_{k,e}$ の推定誤差が 0 に近いとき, 攻撃誤差は小さくなり, 0 からの距離に比例して攻撃誤差が大きくなる.

5.3.3 十分条件

図 7, 8, 9 に目標値 $f_{k,t}, \mu_{k,t}$ を変更したときに連立方程式に解があるような m の範囲のうち, 最小の m を用いて β を計算したときの β の値を示す. ただし, $\epsilon = 1, f_{k,e} = f^*_{f,e}, \mu_{k,e} = \mu^*_{k,e}, m = 0.1n$ である. 図 8 より PCKV-UE では, 頻度を上昇させる方向に目標値を変化させるとき, 必要な偽ユーザ数が 0.1 程度まで上昇する. 図 7, 9 より PrivKV 及び PCKV-GRR についてはごく少数の偽ユーザで攻撃可能である. また, PCKV-GRR では $\mu_{k,t} = -1, f_{k,t} = 0$ に近づくとき必要な偽ユーザ数が上昇する.

5.4 考察

5.4.1 推定精度と攻撃精度

図 2, 3 によると, ϵ を大きくする (弱いプライバシ保証) ほど攻撃精度が上昇する. これは, LDP のノイズの大きさが, 推定値の操作に影響を及ぼすからであると考えられる.

また, 推定精度では頻度推定において, PCKV-UE, PCKV-GRR, PrivKV の順に MSE が小さかったが, 攻撃精度においては PCKV に比べ, PrivKV の MSE がおよそ

表 2 データセット

dataset	User 数	key 数	record 数
clothing	105,508	5850	192,198

表 3 実験で用いるパラメータ

dataset	target key k	$f_{k,t}$	$\mu_{k,t}$	$f^*_{k,e}$	$\mu^*_{k,e}$	f_k	μ_k
clothing	562	0.300	-0.300	0.022	0.905	0.021	0.740

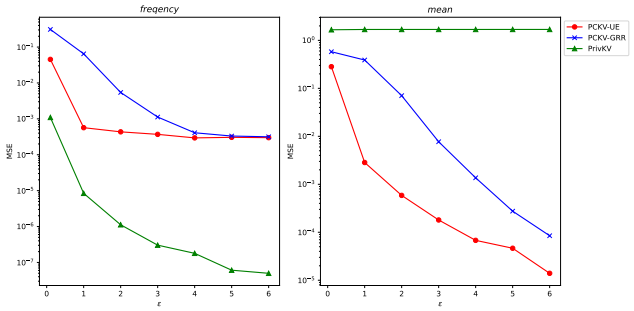

図 3 ϵ を変化させたときの攻撃精度 (左: 頻度，右: 平均値), $(m = 0.1n)$

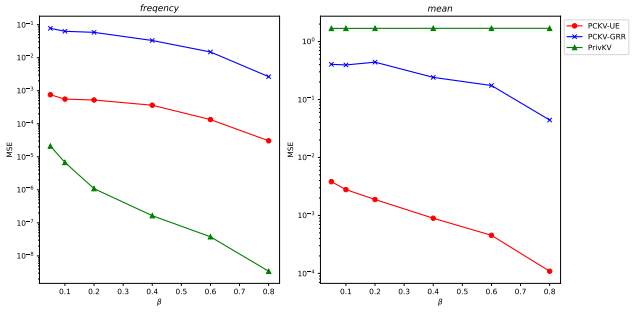

図 4 β を変化させたときの攻撃精度 (左: 頻度, 右: 平均値), $(\epsilon = 1.0)$

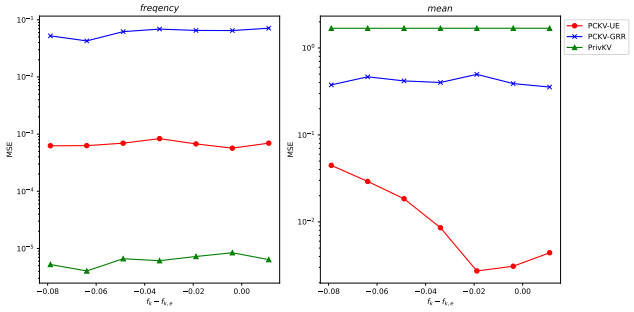

図 5 $f_{k,e}$ を変化させたときの攻撃精度 (左: 頻度，右: 平均値) $(\epsilon = 1.0, m_{k,e} = m^*_{k,e})$

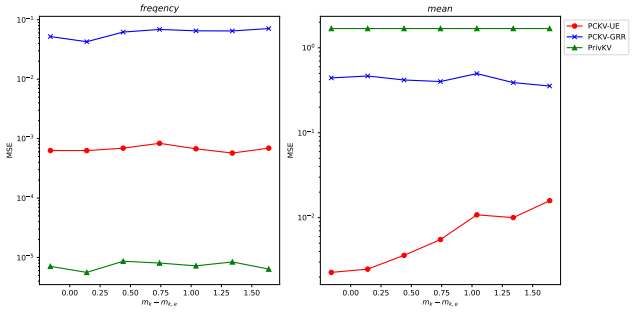

図 6 $m_{k,e}$ を変化させたときの攻撃精度 (左: 頻度，右: 平均値) $(\epsilon = 1.0, f_{k,e} = f^*_{k,e})$

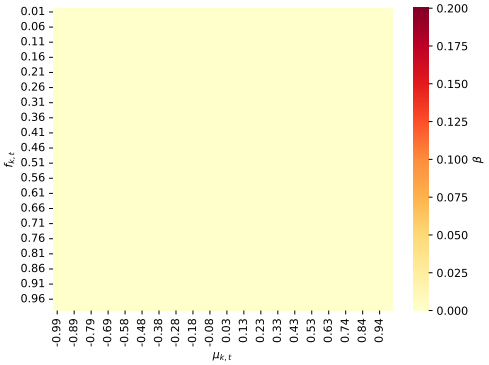

図 7 $f_{k,t}, \mu_{k,t}$ を変化させたときに方程式に解がある β の最小値:PrivKV

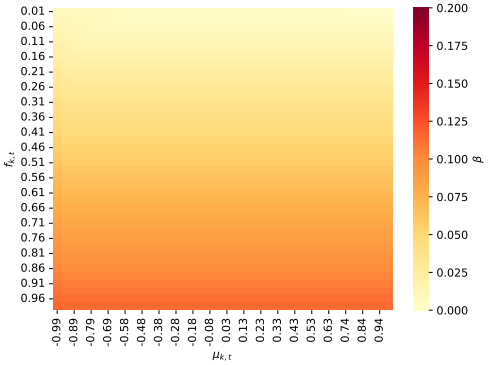

図 8 $f_{k,t}, \mu_{k,t}$ を変化させたときに方程式に解がある β の最小値:PCKV-UE

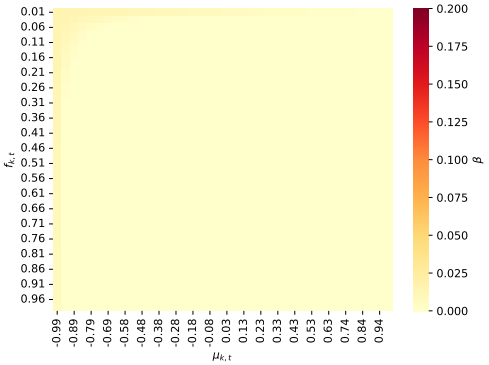

図 9 $f_{k,t}, \mu_{k,t}$ を変化させたときに方程式に解がある β の最小値:PCKV-GRR

2 桁以上小さいという結果になった．これはサンプリング後の key の収集方式の違いに原因があると考えられる．key の頻度推定において，PrivKV では 2 値の RR，PCKV-UE では UE，PCKV-GRR では d-RR を用いている．d-RR では値域の大きさ d が推定精度を決定する．Wang ら [8] によ

れば，頻度推定における推定誤差の理論値は $d = 5852$ の場合，2-RR，UE，d-RR の順に小さい．また，$\epsilon \approx 3.5$ で d-RR の推定精度が UE の推定精度とほぼ等しくなる．これは図 3 の結果と一致している．PrivKV ではサンプリング方式に起因して推定精度が下がるという問題があった．しかし，攻撃者はサンプリングと摂動のステップを回避してデータを収集者に送信することができる．したがって，頻度に対する攻撃の精度を左右するのは，key の収集に用いられる LDP 方式の精度差と考えられる．

平均値推定において，PrivKV は，外れ値を考慮しない合計値を用いて推定を行うため，ϵ の値を大きくしても精度があまり向上しない．PCKV では合計値や頻度の推定値をクリッピングすることで平均値が発散することを防ぎ、推定精度を向上させている．したがって，PCKV における推定ステップが PrivKV の推定ステップの精度を上回るため，攻撃精度においても PrivKV が低い MSE となったと考えられる．

5.4.2 LDP 方式の違いによる安全性の差

PrivKV と PCKV を比較すると，PrivKV では頻度推定，平均値推定においてともに推定精度が低いにも関わらず，頻度推定における推定値操作攻撃では最も脆弱である．一方，PCKV-UE と PCKV-GRR で比較すると，推定精度が高いほど，攻撃精度も高くなる整合性がある．そのため，Key-Value データの収集には推定精度とセキュリティの優先度合いに応じて PCKV-GRR もしくは PCKV-UE を用いるとよい．

6. 対策技術

堀込らは，PrivKV における推定を EM アルゴリズムを用いて行うことで，ポイズニング攻撃に対して強固な emPrivKV を提案した [17]．また，クラスタリングを用いて推定値を回復させる手法が提案されている [3]．クラスタリングベースの手法ではまず，収集したデータをいくつかのサブセットに分割し，サブセットの中で推定値を計算する．そして，算出された推定値に対して k-means などのクラスタリングを適用し，要素数が最も多いクラスタの平均値を最終的な推定値とする．

7. 結論

本稿では，Key-Value データのための局所差分プライバシプロトコルである PrivKV，PCKV-UE，PCKV-GRR に対して，推定頻度と推定平均値を攻撃者の意図した値に変化させる推定値操作攻撃を提案した．実験に基づく評価の結果，PCKV は PrivKV に比べ推定精度が良く，提案攻撃に対しても安全であり，トレードオフはない．

参考文献

[1] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. IEEE FOCS, pp.429-438, 2013.

[2] Stanley L. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, pp.63–69, 1965.

[3] Xiaoguang Li, Ninghui Li, Wenhai Sun, Neil Zhenqiang Gong, Hui Li. Fine-grained Poisoning Attack to Local Differential Privacy Protocols for Mean and Variance Estimation. USENIX Security, pp.1739-1756, 2023.

[4] Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Data poisoning attacks to local differential privacy protocols. USENIX Security, pp. 947-964, 2021.

[5] Albert Cheu, Adam Smith, and Jonathan Ullman. Manipulation attacks in local differential privacy. IEEE S&P, pp.883-900, 2021.

[6] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. ACM CCS, pp.1054-1067, 2014.

[7] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multi-dimensional data with local differential privacy. IEEE ICDE, pp.638-649, 2019.

[8] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. USENIX Security, pp.729-745, 2017.

[9] Zitao Li, Tianhao Wang, Milan Lopuhaä-Zwakenberg, Ninghui Li, and Boris Škoric. Estimating numerical distributions under local differential privacy. ACM SIGMOD, pp.621-635, 2020.

[10] Yongji Wu, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Poisoning Attacks to Local Differential Privacy Protocols for Key-Value Data. USENIX Security, pp.519–536, 2022.

[11] Xiaolan Gu, Ming Li, Yueqiang Cheng, Li Xiong, and Yang Cao. PCKV: Locally differentially private cor- related key-value data collection with optimized utility. USENIX Security Symposium, pp.967-984, 2020.

[12] Qingqing Ye, Haibo Hu, Xiaofeng Meng, and Huadi Zheng. Privkv: Key-value data collection with local differential privacy. IEEE S&P, pp. 294-308, 2019.

[13] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association, 182–201, 2018.

[14] NingWang, XiaokuiXiao, YinYang, JunZhao, SiuCheungHui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. IEEE ICDE, 2019.

[15] Haym Benaroya, Seon Mi Han, and Mark Nagurka. Probability Models in Engineering and Science. CRC Press, 2005, p166.

[16] Clothing fit dataset for size recommendation. https://www.kaggle.com/rmisra/ clothing-fit-dataset-for-size-recommendation.

[17] 堀込 光，菊池 浩明，Yu Chia-Mu. 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. コンピュータセキュリティシンポジウム 2022，pp.129-136，2022