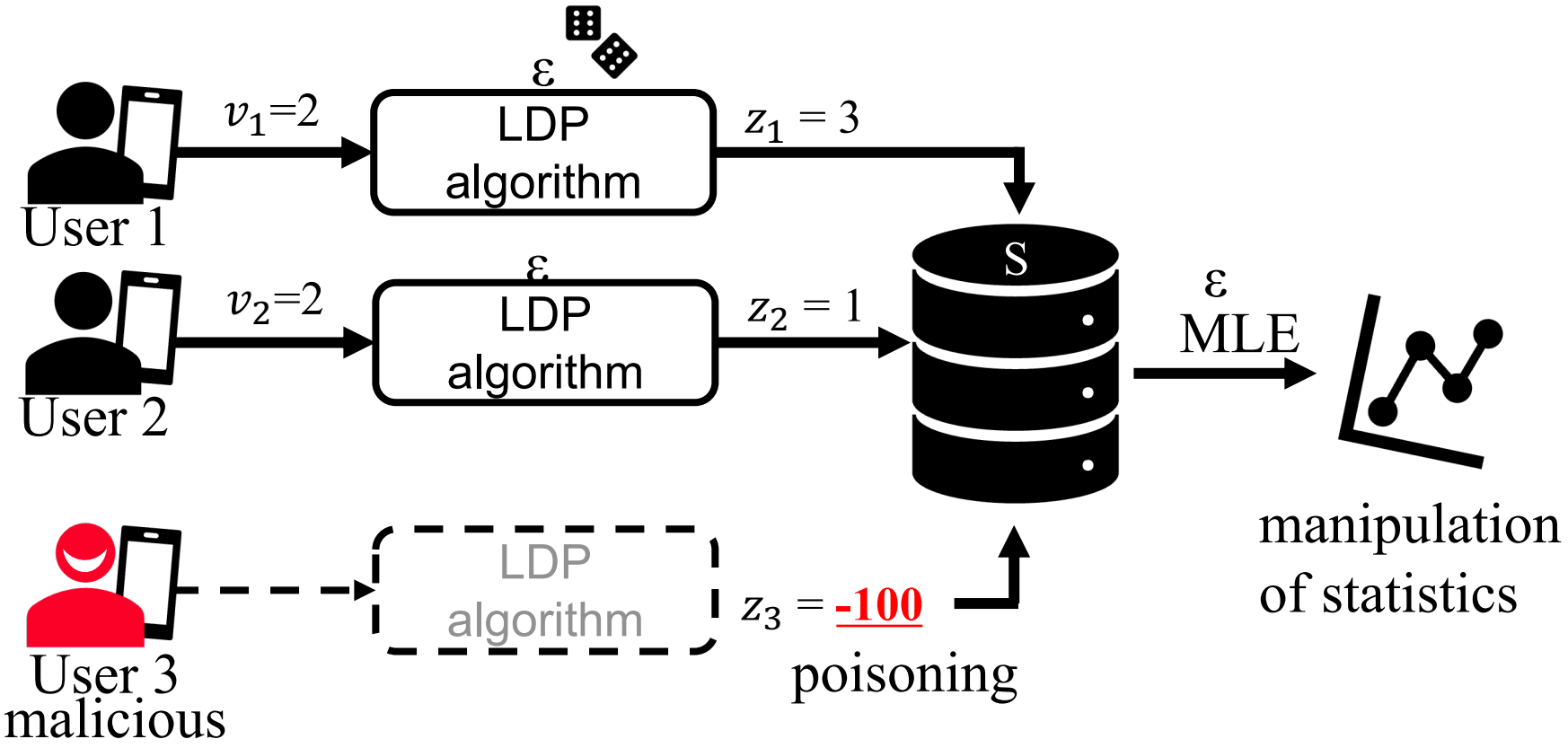
Harmless Opening for Continuous-Valued Local Differential Privacy

Hiroaki Kikuchi

Meiji University, Japan

DPM 2026, Roma

Issue of LDP: *Poisoning*

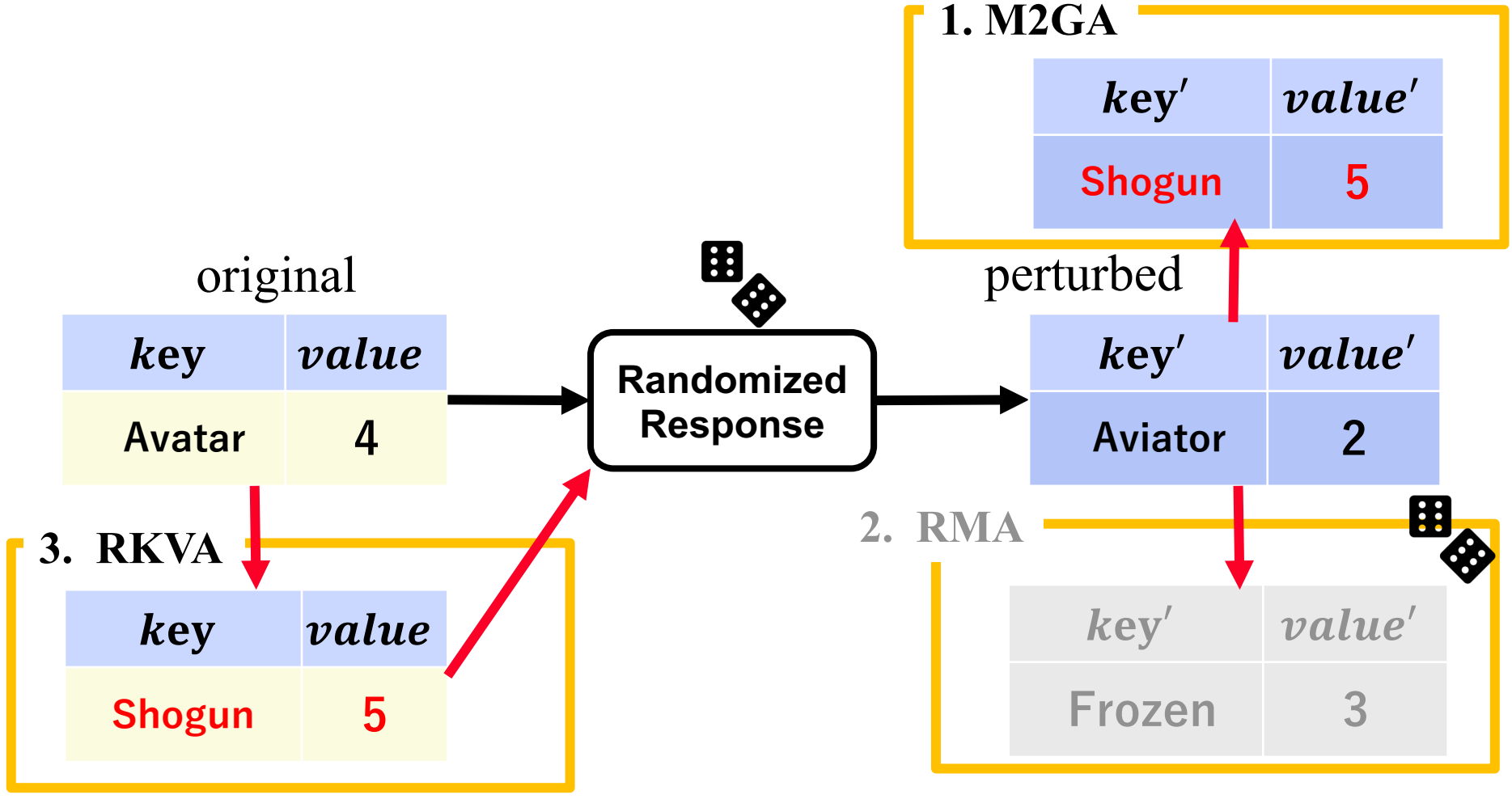


[Y.Wu, et al., 2022]

Poisoning attacks

Input Poisoning (**IPA**)

Output Poisoning (**OPA**)



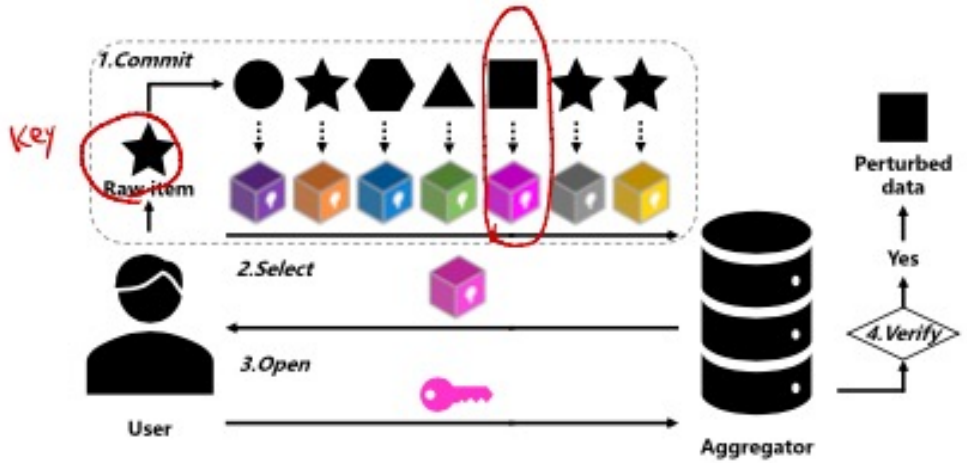
Related Works

- Poisonings
  - [7] Cheu S&P21
  - [8] Cao Usenix 21
  - [9] Wu Usenix 22
- Robust LDP protocols
  - [10] Zhao, S&P 25, RobustLDP
  - Ye, S&P 25, Subset Counting Query
- Cryptographical approach
  - [6] Kato, DBSec 2021
  - [1] Song, IEEE Trans. IFS 2023.

[1] Song+23

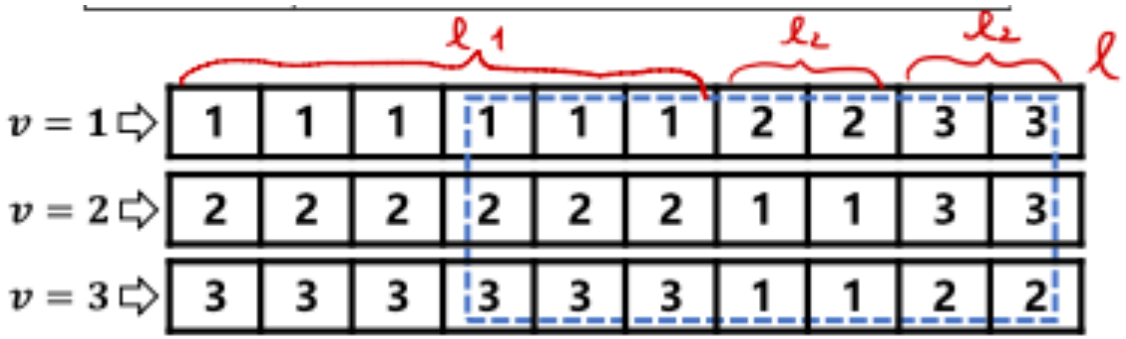
S. Song, L. Xu and L. Zhu, "Efficient Defenses Against Output Poisoning Attacks on Local Differential Privacy," in *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 5506-5521, 2023

- Interactive Perturbation
  - Cryptographic approach
    - » Commit: User sends commitment of his input value and fake input values
    - » Select: Aggregator uniformly choose one
    - » Open: Users opens it with some of fake values
    - » Verify : Aggregator verifies user follows protocol.



"*Harmless Opening*"

- VGRR (Verifiable GRR)
  - Commit ℓ_1 true value and ℓ_2 fake values
  - $\ell = \ell_1 + (d-1)\ell_2$
  - Example ℓ=10, d = 3
    - » Pr[x = 1] = 6/10
    - » Pr[x = 2] = Pr[x = 3] = 2/10
    - » Aggregator can not distinguish x = 1 and x = 2,3



VGRR (Verified GRR)

- Pros:
  - VGRR is ($\epsilon' = \ln \frac{\ell_1}{\ell_2}$)-LDP
  - VGRR is robust against OPA
    - » robust iff $Pr[S_{max} = S_{in}] > 1 - negl(k)$
    - » S_{in} = IPA, S_{max} = OPA
  - VGRR is lightweight (no ZK proof)
- Cons:
  - deals with only discrete values (frequency estimation)

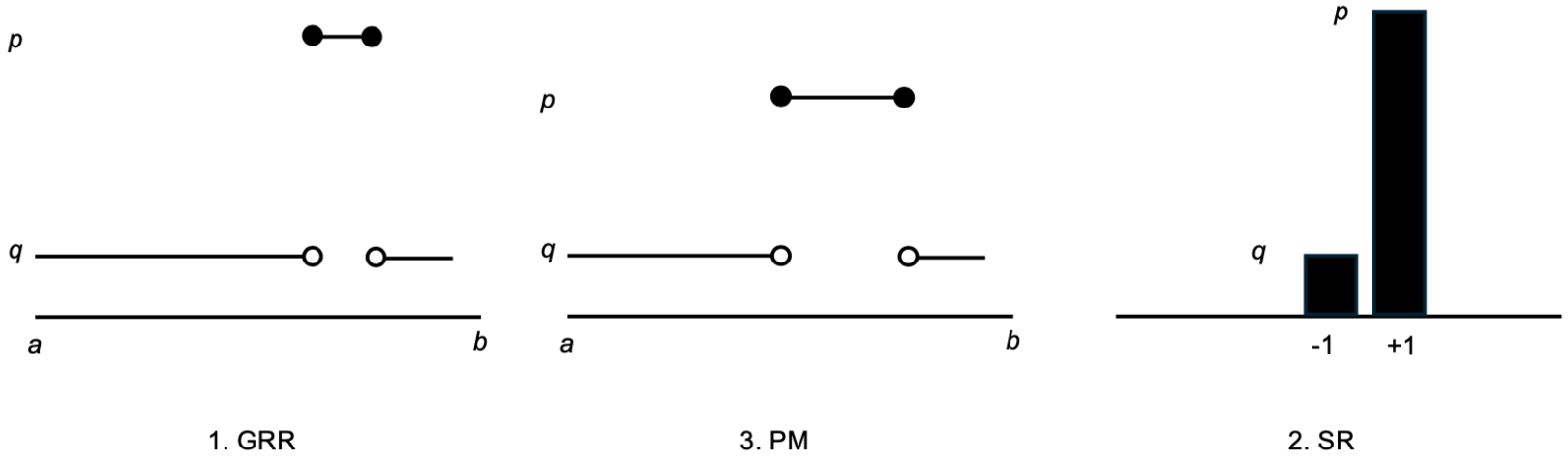
Our work

- Goal
  - Robust LDP for continuous value

Task	LDP	Poisoning	Mitigation	
			probabilistic	cryptographical
Frequency estimation (categorical value)	GRR, OUE, OLH, Rappor	[7] Cheu [8] Cao [9] Wu	[8], [12]	[1] Song, [6] Kato
Mean estimation (continuous value)	[3] SR [4] PM	[10] Zhao	[10] Zhao	Our work

Continuous-valued LDP

- LDP
  - [2] GRR (Generalized Randomized Response)
  - [3] SR (Stochastic Rounding)
  - [4] PM (Pricewise Mechanism)



SR (Stochastic Rounding)

- Input: x
  - x in [a,b] → x~ in [-1,1]
    - » x~ = -1 + (x-a)*2/(b-a)
- Perturbation: x'

$$\Pr[\Psi_{SR(\varepsilon)}(\tilde{x}) = x'] = \begin{cases} q + \frac{(p-q)(1-\tilde{x})}{2}, & \text{if } x' = -1 \\ q + \frac{(p-q)(1+\tilde{x})}{2}, & \text{if } x' = 1 \end{cases}$$

$$p = \frac{e^{\varepsilon}}{1+e^{\varepsilon}} \text{ and } q = 1 - p$$

- Unbiased estimator

$$\mathbb{E}\left(\frac{x'}{p-q}\right) = \tilde{x}.$$

PM (Piecewise Mechanism)

- Input
  - x in [a,b] $\rightarrow \tilde{x}$ in [-1,1]
- Perturbation
  - x' in [-s, +s]

$$\Pr[\Psi_{PM(\varepsilon)}(\tilde{x}) = x'] = \begin{cases} \frac{e^{\varepsilon/2}(e^{\varepsilon/2}-1)}{2(e^{\varepsilon/2}+1)}, & \text{if } x' \in [l(\tilde{x}), r(\tilde{x})] \\ \frac{e^{\varepsilon/2}-1}{2(e^{\varepsilon/2}+e^{\varepsilon})}, & \text{otherwise} \end{cases},$$

$$\text{where } -s \le l(\tilde{x}) < r(\tilde{x}) \le s, l(\tilde{x}) = \frac{e^{\varepsilon/2}\tilde{x}-1}{e^{\varepsilon/2}-1} \text{ and } r(\tilde{x}) = \frac{e^{\varepsilon/2}\tilde{x}+1}{e^{\varepsilon/2}-1}.$$

- Unbiased Estimator
  - $E[x'] = \tilde{x}$

Threat model

- *n* honest users
- *m* malicious users
  - rate β = m/(n+m)
  - gain = target mean (max) – true mean
  - IPA (input = targeted value, with LDP)
  - OPA (output = targeted value, without LDP, leaves if $x < \theta$)

Proposed Method

- Idea
  - binning (making into discrete bins) and applies Harmless Opening (VGRR)

base		Harmless Opening	
LDP	encoding	Binning	
GRR	D→D	*D* →*d* values {0,…,d-1}	VGRR-HO
SR	[a,b]→{-1,+1}	[a,b] → *d* values	SR-HO
PM	[a,b] →[-s, +s]	[-s,+s] → *d* values	PM-HO

Example 1) GRR

- 1. Commit
  - d = 5, x in {1,2,3,4,5}, x~ in {-1, -1/2, 0, ½, 1}, ε = log 2
  - x = 4, x~ = 1/2, p = $e^{\varepsilon}/(e^{\varepsilon}+d-1)$ =2/6, q = 1/6

1	2	2	3	4	5
C(1)	C(2)	C(2)	C(3)	C(4)	C(5)

- 2. A picks j = 2 in {1,..,6}
- 3. U opens c(1) c(2) c(3) c(4) c(5)
- 4. A checks C and c(1) c(2) c(3) c(4) c(5)

Example 2) SR-HO

- D = {0,1,2,3,4}, d = 5, ε =ln 2, ℓ = 12

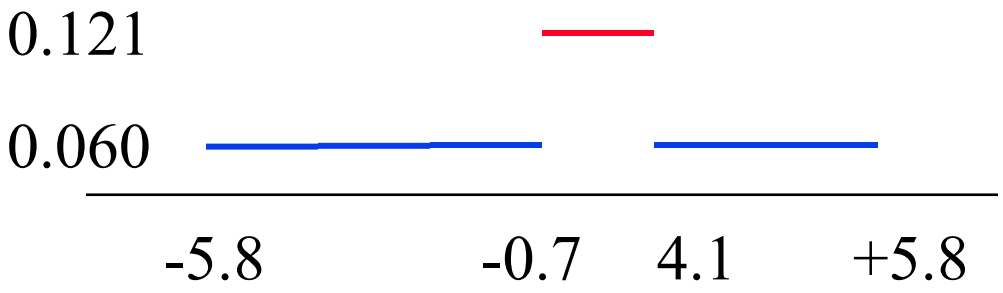
x	~x	P*	ℓ_{-1}	ℓ_1
0	-1	1/3	**4**	8
1	-1/2	5/12	5	7
2	0	1/2	6	6
3	1/2	7/12	7	5
4	1	2/3	**8**	4

x	1	2	3	4	5	6	7	8	9	10	11	12
0	C(-1)	C(-1)	C(-1)	C(-1)	C(1)	C(1)	C(1)	C(1)	C(1)	C(1)	C(1)	C(1)
1	C(-1)	C(-1)	C(-1)	C(-1)	C(-1)	C(1)	C(1)	C(1)	C(1)	C(1)	C(1)	C(1)

- HO proves that 4/12 ≦ p* ≦ 8/12 for all cases

Example 3) PM-HO

- Commit
  - d = 5, x in {0,1,2,3,4}, ~x in {-1, -1/2, 0, ½, 1}, ε = log 2,
  - x = 3, x~ = 1/2, s = $(e^{\varepsilon/2}+1)/(e^{\varepsilon/2}-1)$ = 5.8
  - ℓ(x) = 2x -1 /(2-1) = -0.7, r(x) = 4.1



Questions

- RQ1. Does Harmless Opening mitigates poisoning attacks?
- RQ2. How much accuracy reduction does come with the binning?
- RQ3. Which is the best LDP scheme in terms of accuracy and computational costs?

Security Analysis

- Privacy
  - GRR-HO, SR-HO and PM-HO satisfies ε-LDP.
- Robustness
  - Theorem 1 GRR-HO is robust
    - » $Pr[S_{max} = S_{in}] > 1 - negl(k)$
  - Proposition SR-HO is semi-robust?
    - » $Pr[S_{max} > S_{in}] > 1 - negl(k)$

bogus

Security of SR-HO

- Lemma 1. Any report accepted by the Harmless opening satisfies $E\left[\frac{x'}{p-q}\right] \leq 1$

  for mean-maximizing attack.
  - Mean-maximizing adversary would like to commit +1 value, which is limited at most ℓ_2 (ℓ_{-1}) times in HO verification process.
- Theorem 2. SR-HO is robust for mean estimation, i.e., gain $G_{max} = G_{IPA}$.

Evaluation

- Dataset
  - UCI Adult dataset , Age [a = 17, b = 90]
- Taks
  - Mean estimation
- Metrics
  - Accuracy: MSE
  - Robustness: gain

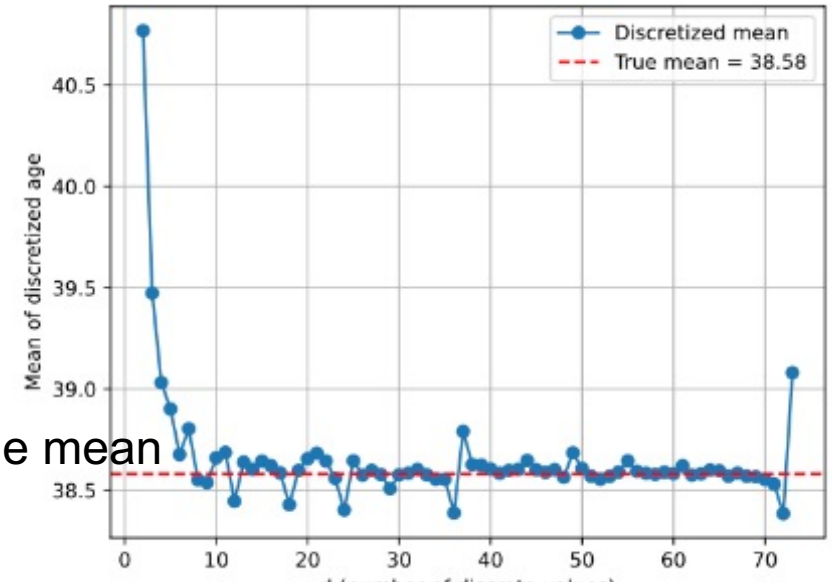
    gain = estimate with poisoning – estimate

Result 1. binning

Without perturbation

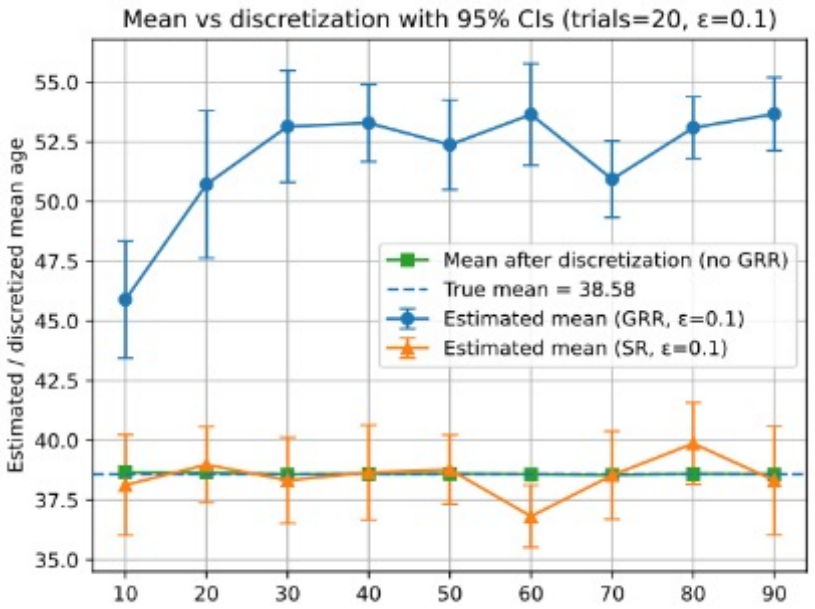
LDP (GRR, SR)

Estimated mean



True mean

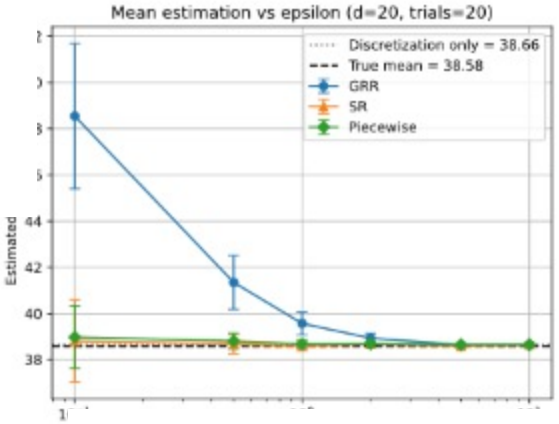
d (# values)



d (# values)

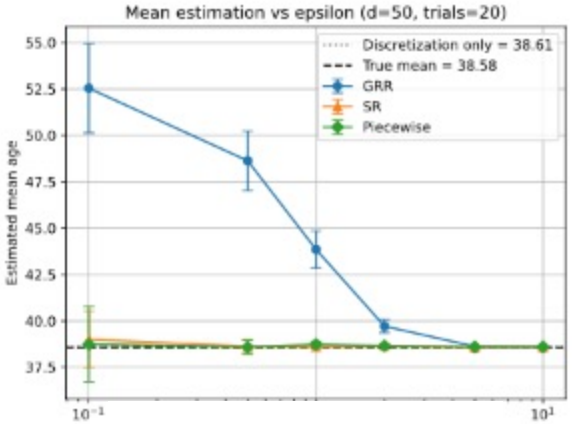
Result 2. Estimated Mean

Estimated mean (age)



Privacy budget ε

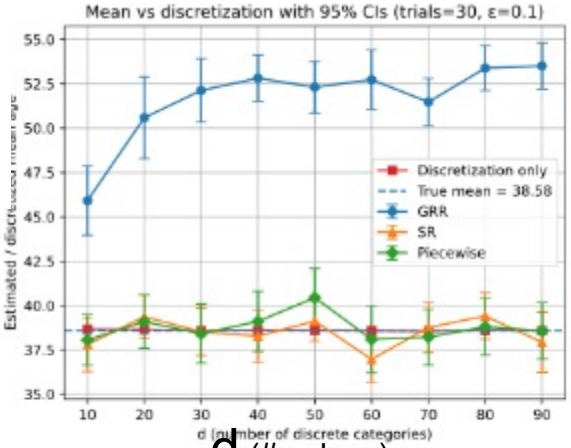
(a) $d = 20$



ε

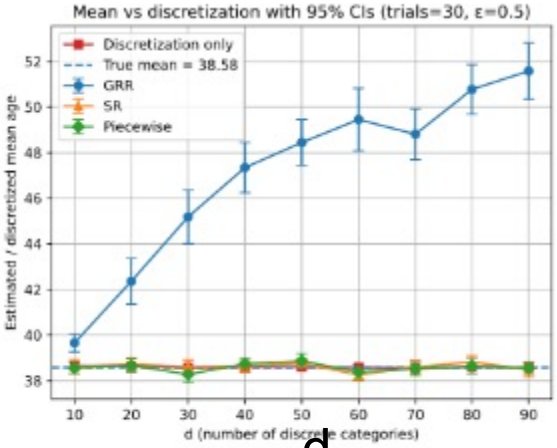
(b) $d = 50$

Estimated mean (age)



d (# values)

(a) $\epsilon = 0.1$



d

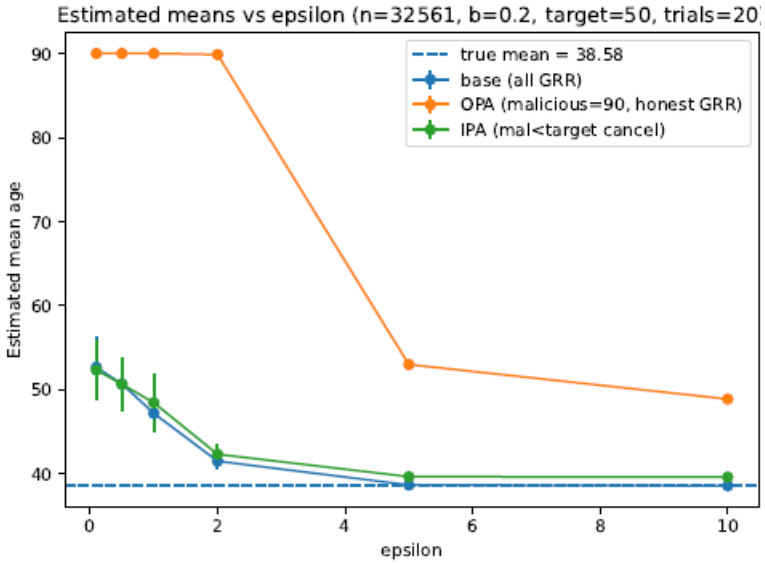
(b) $\epsilon = 0.5$

Result 3. Gain (=vulnerability)

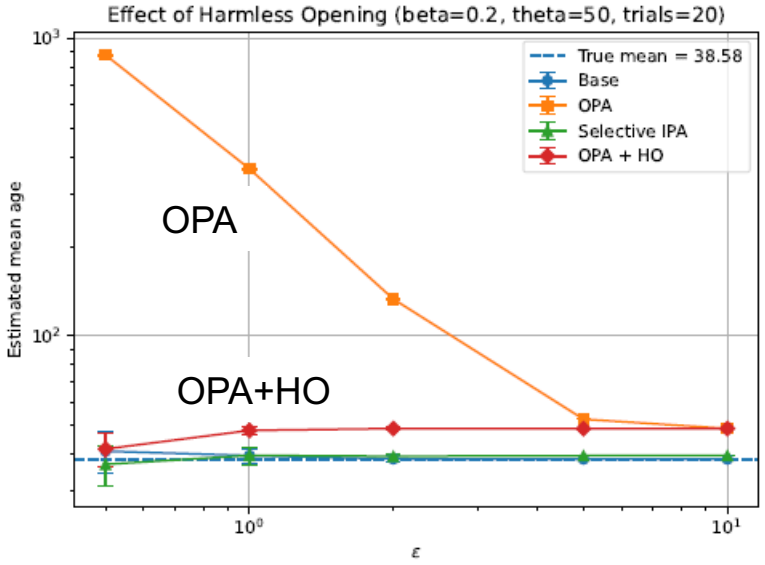
Gain

vulnerable

Robust

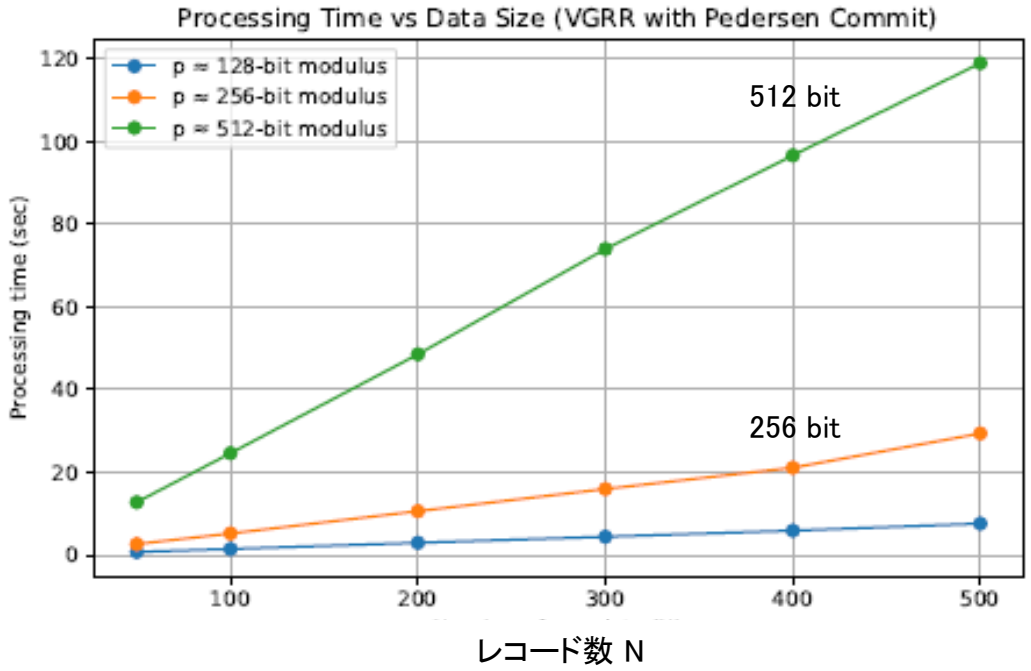


Privacy budget ε



ε

Result 4 processing time



Total time used for Pedersen Commitment
d = 73, ℓ = 200

Per user time is very small.
E.g. 75 s / 300 (N)
= 0.25 sec/user

Conclusions

- RQ1. does HO mitigate poisoning?
  - Yes. GRR and SM are robust.
- RQ2. Accuracy reduction? small
- RQ3. Best scheme? SR is robust and lightweighted
- Future work: robustness for PM and work with Robust LDP [Zhao 25]

	GRR	SR	PM
Robustness	Frequency	Mean	Mean
security	Robust	Robust	(open)
ℓ	ℓ_1+(d-1) ℓ_2	ℓ_1+ℓ_2	ℓ_1s + ℓ_2s
Cost	High	Low	High
Accuracy	Low	High	High
Gain	High	Low	Low

Pedersen Commitment

- 情報理論的安全性を保証したコミット
  - Z1, Z2: 有限群 (DLP仮定)
  - 公開鍵: (g, h) Z1, Z2の生成元
  - コミットする値: ρ
  - コミット: $C(\rho) = g^{\rho}h^{\tau}$
  - 準同型性

$$C_K(\rho_1; \tau_1)C_K(\rho_2; \tau_2) = C_K(\rho_1 + \rho_2; \tau_1 + \tau_2).$$

準同型性を用いたランダムサンプリン

- 初期値
  - A: x, B: y
  - A→B: C(x,r1)
  - B→A: C(y,r2)
  - C’ = C(x,r1)*C(y,r2) = C(x+y, r1+r2)を求めて，x, y をopen

VGRR

Protocol 1 VGRR

The user $\mathcal{U}$ possesses a raw item v. The aggregator $\mathcal{A}$ obtains a value y as the perturbed data from the user $\mathcal{U}$.

Initialize: The user $\mathcal{U}$ and the aggregator $\mathcal{A}$ share the public key K and the length parameters ℓ, ℓ_1 and ℓ_2.

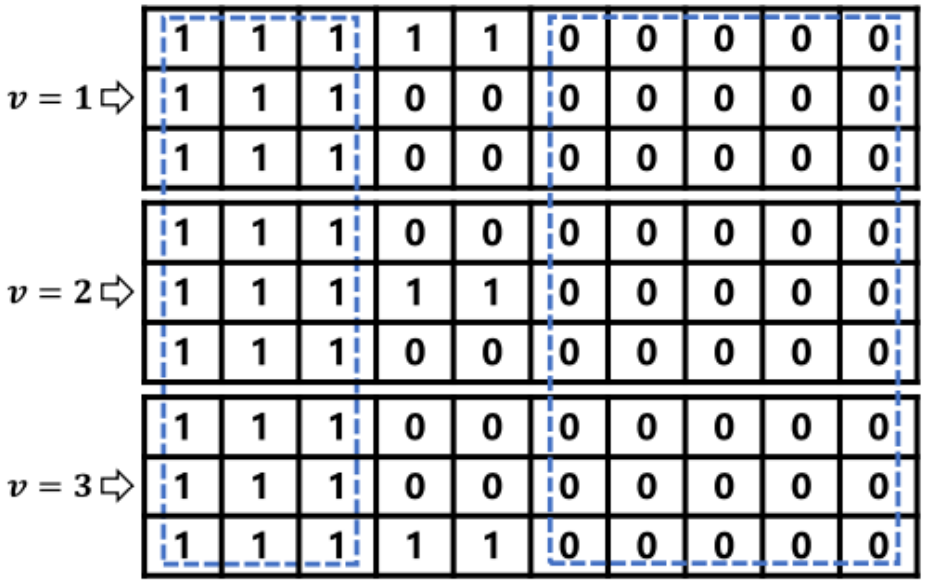
1) **Commit**
   a) $\mathcal{U}$ constructs an ℓ-dimensional vector $\boldsymbol{\mu}$ which satisfies $\theta_v^{\boldsymbol{\mu}} = \ell_1$ and $\theta_i^{\boldsymbol{\mu}} = \ell_2$ for $i \in [d] \backslash \{v\}$.
   b) A new ℓ-dimensional vector $\boldsymbol{\eta}$ is constructed by performing $\boldsymbol{\eta}[i] \triangleq C_K(\boldsymbol{\mu}[i]; \tau_i)$ for $1 \leq i \leq \ell$, where τ_i is drawn randomly from $\mathbb{Z}_{prime_2}$. The vector $\boldsymbol{\eta}$ is sent to $\mathcal{A}$.
2) **Select**
   $\mathcal{A}$ randomly selects j from $\{1, 2, \cdots, \ell\}$ and sends j to $\mathcal{U}$.
3) **Open**
   $\mathcal{U}$ picks $d\ell_2$ numbers $j_1, j_2, \cdots, j_{d\ell_2}$, where $j_1 = j$. For $1 \leq i \leq d\ell_2$, $\mathcal{U}$ opens the commitment $\boldsymbol{\eta}[j_i]$ by sending $\boldsymbol{\mu}[j_i]$ and τ_{j_i} to $\mathcal{A}$. After the $d\ell_2$ commitments $\boldsymbol{\eta}[j_1], \boldsymbol{\eta}[j_2], \cdots, \boldsymbol{\eta}[j_{dl_2}]$ are opened, the vector $\boldsymbol{\eta}$ need to satisfy $\varphi_i^{\boldsymbol{\eta}} = \ell_2$ for $i \in [d]$.
4) **Verify**
   a) $\mathcal{A}$ checks if $j_1 = j$ and $\boldsymbol{\eta}[j_i] = C_K(\boldsymbol{\mu}[j_i]; \tau_{j_i})$ for $1 \leq i \leq d\ell_2$.
   b) For $i \in [d]$, $\mathcal{A}$ checks if the vector $\boldsymbol{\eta}$ satisfies $\varphi_i^{\boldsymbol{\eta}} = \ell_2$ for $i \in [d]$
   c) If all the checks are affirmative, $\mathcal{A}$ takes $\boldsymbol{\mu}[j]$ as the output y from $\mathcal{U}$.

robustの定義

- 定義2 (識別不能性)
  - A* (真のアイテムを推測するアルゴリズム)
  - view (Aが観測する情報)
  - Pr(A*(view;y) = v) – Pr[A*(y) = v] < nelg(k) である時，Mは識別不能.
- 定義3 (ロバスト)
  - Smax = Overall scoreの最大値.
  - Pr[Smax = Sin] > 1 – nelg(k) である時，Mはrobust.
- 定義4 (semi-robust)
  - Pr[Sout > Smax > Sin] > 1 – nelg(k)

VOUE

- VGRRのOUEバージョン
  - $\ell = 2\ell_1$, $\ell_1 > \ell_2$
  - d個のそれぞれについて，jを選ぶ.



VOUEの性質

- 定理4 (LDP)
  - $\varepsilon' = \ln \frac{2\ell_1 - \ell_2}{\ell_2}$ LDPを満たす.
- 定理5 (識別不能性)
- 定理6 (安全性，少し弱い)
  - VOUEは semi-robust である.
    - » Sout = r ℓ_1 > Smax = r/2

VOUE+

- 集合への所属のZKIPの導入
  - [36, Kato], [40, Benarroh 2021]
  - 未開示のエレメントが正しく所定の集合に属していることを証明する.
- 定理9 (頑強性)
  - VOUE+はrobustである.