Longitudinal data collection based on local hashing for local differential privacy[1]

Carlos Pinzón

8th Franco-Japanese Cybersecurity Workshop, WG1, formal methods

ENS de Cognitique, Bordeaux. November 29, 2023.

Arcolezi et al. 2023[1]

Outline

1. Problem: collect longitudinal data preserving privacy
2. State of the art
3. Longitudinal local hashing

Problem: collect longitudinal data preserving privacy

	Single individual	Population
Single time	(one value)	Cross-sectional
Many times	Time series	**Longitudinal**

Problem

- n users, $\tau \gg 1$ time steps, k possible values: $x_t^{(u)} \in [1..k]$
- We want to collect data to estimate population frequencies:

$$\mathbf{P}(X_t = x) \text{ for each } x \in [1..k] \text{ and each } t \in [1..\tau]$$

- ...but individual values are private

Local Differential Privacy (LDP)

"Collected values will be partial, approximate or mere garbage"

$$\epsilon\text{-LDP:} \quad \forall (y, x^+, x^-), \quad \mathbf{P}(Y{=}y|X{=}x^+) \leq e^{\epsilon}\, \mathbf{P}(Y{=}y|X{=}x^-)$$

Trust barrier:

$$\text{Users} \xrightarrow[\text{mistrust}]{\text{LDP}} \text{Server} \longrightarrow \text{Public or third parties}$$

Local Differential Privacy (LDP)

"Collected values will be partial, approximate or mere garbage"

$$\epsilon\text{-LDP:} \quad \forall (y, x^+, x^-), \quad \mathbf{P}(Y{=}y|X{=}x^+) \leq e^\epsilon \, \mathbf{P}(Y{=}y|X{=}x^-)$$

Trust barrier:

$$\text{Users} \longrightarrow \text{Server} \xrightarrow[\text{mistrust}]{\text{DP}} \text{Public or third parties}$$

Local Differential Privacy (LDP)

"Collected values will be partial, approximate or mere garbage"

ϵ-LDP: $\forall (y, x^+, x^-), \quad \mathbf{P}(Y{=}y|X{=}x^+) \leq e^\epsilon \, \mathbf{P}(Y{=}y|X{=}x^-)$

Trust barrier:

$$\text{Users} \xrightarrow[\text{mistrust}]{\text{LDP}} \text{Server} \longrightarrow \text{Public or third parties}$$

Mechanisms for LDP

Categorical $\texttt{secret} \in [1..k]$

1. Randomized Response (RR)[1]:

$$\texttt{reported value} = \begin{cases} \texttt{secret} & \text{with some prob.} \\ \text{any other value (uniformly)} & \text{otherwise} \end{cases}$$

2. Unary Encoding[2]:

apply one-hot encoding, then RR to each bit

3. Local Hashing[3]:

partition $[1..k]$ into groups, then apply RR to the group labels

Warner 1965[1] Erlingsson et al 2014[2] Bassily and Smith 2015[3]

Naïve solution

$$x \overset{\epsilon}{\rightsquigarrow} x'_1$$
$$x \overset{\epsilon}{\rightsquigarrow} x'_2$$
$$\vdots$$

- The privacy guarantee is $t\epsilon$-LDP after t reports $[x'_1, .., x'_t]$.
- As $t \to \infty$, LDP breaks and we guess x.

Improvement: double randomization

$$
\begin{aligned}
x \overset{\epsilon_0}{\rightsquigarrow} x' &\overset{\epsilon}{\rightsquigarrow} x''_1 \\
x' &\overset{\epsilon}{\rightsquigarrow} x''_2 \\
x' &\overset{\epsilon}{\rightsquigarrow} x''_3 \\
&\vdots
\end{aligned}
$$

- As $t \to \infty$, we guess x' but not x (the user's value).
- ϵ_0-LDP is guaranteed.

Improvement: double randomization

$$
\begin{aligned}
x &\overset{\epsilon_0}{\rightsquigarrow} x' \overset{\epsilon}{\rightsquigarrow} x''_1 \\
x &\xrightarrow{\text{memo}} x' \overset{\epsilon}{\rightsquigarrow} x''_2 \\
x &\xrightarrow{\text{memo}} x' \overset{\epsilon}{\rightsquigarrow} x''_3 \\
&\vdots
\end{aligned}
$$

- As $t \to \infty$, we guess x' but not x (the user's value).
- ϵ_0-LDP is guaranteed.

Improvement: double randomization

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix} \overset{\text{memo}}{\underset{\epsilon_0}{\rightsquigarrow}} \begin{bmatrix} x_1' \\ x_2' \\ \vdots \end{bmatrix} \overset{\epsilon}{\rightsquigarrow} \overset{\epsilon}{\rightsquigarrow} \vdots \begin{bmatrix} x_1'' \\ x_2'' \\ \vdots \end{bmatrix}$$

- Is ϵ_0-LDP guaranteed?

Example of double randomization

$$
\begin{array}{lcccc}
x_1 = \mathtt{A} & \overset{\epsilon_0}{\rightsquigarrow} & \mathtt{A}' & \overset{\epsilon}{\rightsquigarrow} & x_1'' \\
x_2 = \mathtt{C} & \overset{\epsilon_0}{\rightsquigarrow} & \mathtt{C}' & \overset{\epsilon}{\rightsquigarrow} & x_2'' \\
x_3 = \mathtt{C} & \xrightarrow{\text{memo}} & \mathtt{C}' & \overset{\epsilon}{\rightsquigarrow} & x_3'' \\
x_5 = \mathtt{A} & \xrightarrow{\text{memo}} & \mathtt{A}' & \overset{\epsilon}{\rightsquigarrow} & x_5'' \\
x_9 = \mathtt{A} & \xrightarrow{\text{memo}} & \mathtt{A}' & \overset{\epsilon}{\rightsquigarrow} & x_9'' \\
x_4 = \mathtt{G} & \overset{\epsilon_0}{\rightsquigarrow} & \mathtt{G}' & \overset{\epsilon}{\rightsquigarrow} & x_4'' \\
x_6 = \mathtt{T} & \overset{\epsilon_0}{\rightsquigarrow} & \mathtt{T}' & \overset{\epsilon}{\rightsquigarrow} & x_6'' \\
x_7 = \mathtt{T} & \xrightarrow{\text{memo}} & \mathtt{T}' & \overset{\epsilon}{\rightsquigarrow} & x_7'' \\
 & \xrightarrow{\text{memo}} & \vdots & & \\
 & \xrightarrow{\text{memo}} & \vdots & &
\end{array}
$$

$\forall \mathtt{X} \in [1..k] = \{\mathtt{A}, \mathtt{G}, \mathtt{T}, \mathtt{C}\}$

- the fixed value $\mathtt{X}'$ is exposed to multiple queries.
- $\mathtt{X}$ is exposed only once.

But there are leakages about data changes and time patterns

Outline

1. Problem: collect longitudinal data preserving privacy
2. **State of the art**
3. Longitudinal local hashing

Real world deployments

00:51

RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

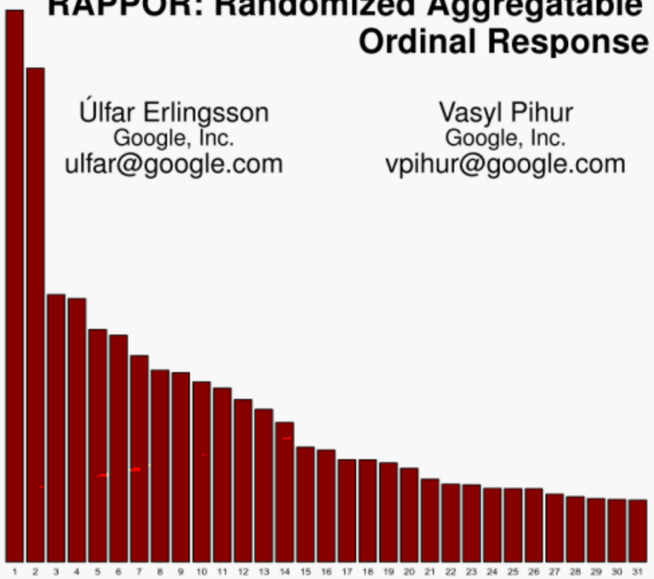
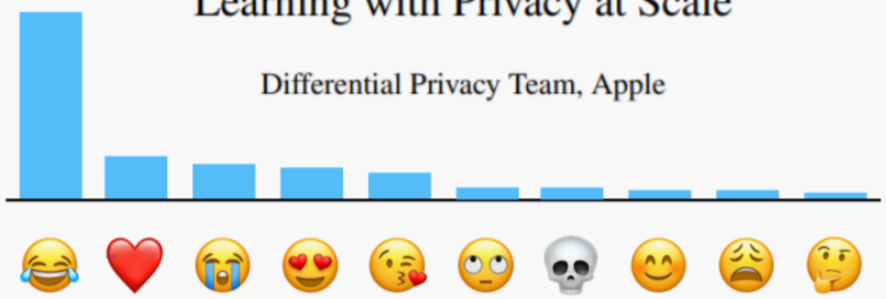


Figure 6: Relative frequencies of the top 31 unexpected Chrome homepage domains found by analyzing ∼14 million RAPPOR reports, excluding expected domains (the homepage "google.com", etc.).

Learning with Privacy at Scale

Differential Privacy Team, Apple



The Count Mean Sketch technique allows Apple to determine the most popular emoji to help design better ways to find and use our favorite emoji. The top emoji for US English speakers contained some surprising favorites.

Collecting Telemetry Data Privately

Bolin Ding, Janardhan Kulkarni, Sergey Yekhanin
Microsoft Research
{bolind, jakul, yekhanin}@microsoft.com

Windows Insiders in Windows 10 Fall Creators Update to protect users' privacy while collecting application usage statistics.

State of the art protocols

	Dom.	Preprocessing		Perm. rand.	Inst. rand.	Output dom.
L-RR	$[1..k]$		$[1..k]$	RR	RR	$[1..k]$
RAPPOR[1]	$[1..k]$	One hot	$\{0,1\}^k$	SUE	SUE	$\{0,1\}^k$
L-OSUE	$[1..k]$	One hot	$\{0,1\}^k$	OUE	SUE	$\{0,1\}^k$
dBitFlipPM[2], $d \leq k$	$[1..k]$	Ad-hoc	$\{0,1\}^d$	SUE		$\{0,1\}^d$
LOLOHA	$[1..k]$	Local hash	$[1..g]$	RR	RR	$[1..g]$

Erlingsson et al. 2014[1] Ding et al. 2017[2]

Outline

1. Problem: collect longitudinal data preserving privacy
2. State of the art
3. **Longitudinal local hashing**

LOLOHA: Longitudinal Local Hashing

The best of both worlds:

Protocol	Privacy strength	Parameters
RAPPOR[1]	double sanitization	e^{ϵ_∞} and $e^{\alpha\epsilon_\infty}$
d-BitFlipPM[2]	reduced domain	e^{ϵ_∞} and $d \ll k$
LOLOHA	(both)	e^{ϵ_∞} and $e^{\alpha\epsilon_\infty}$ and $g \ll k$

Parameter g:

1. Tuned for privacy $g = 2$
2. Tuned for utility $g = 1 + \max\left(1, \frac{1-e^{2\epsilon_\infty}+\sqrt{e^{4\epsilon_\infty}-14e^{2\epsilon_\infty}+12e^{(1+\alpha)\epsilon_\infty}(1-e^{(1+\alpha)\epsilon_\infty})+12e^{(3+\alpha)\epsilon_\infty}+1}}{6e^{\epsilon_\infty}-e^{\alpha\epsilon_\infty}}\right)$

Erlingsson et al. 2014[1] Ding et al. 2017[2]

LOLOHA: LOngitudinal LOcal HAshing

Setup:

1. Fix integer g
2. Each user u reports a random hash function $H_u : [1..k] \to [1..g]$

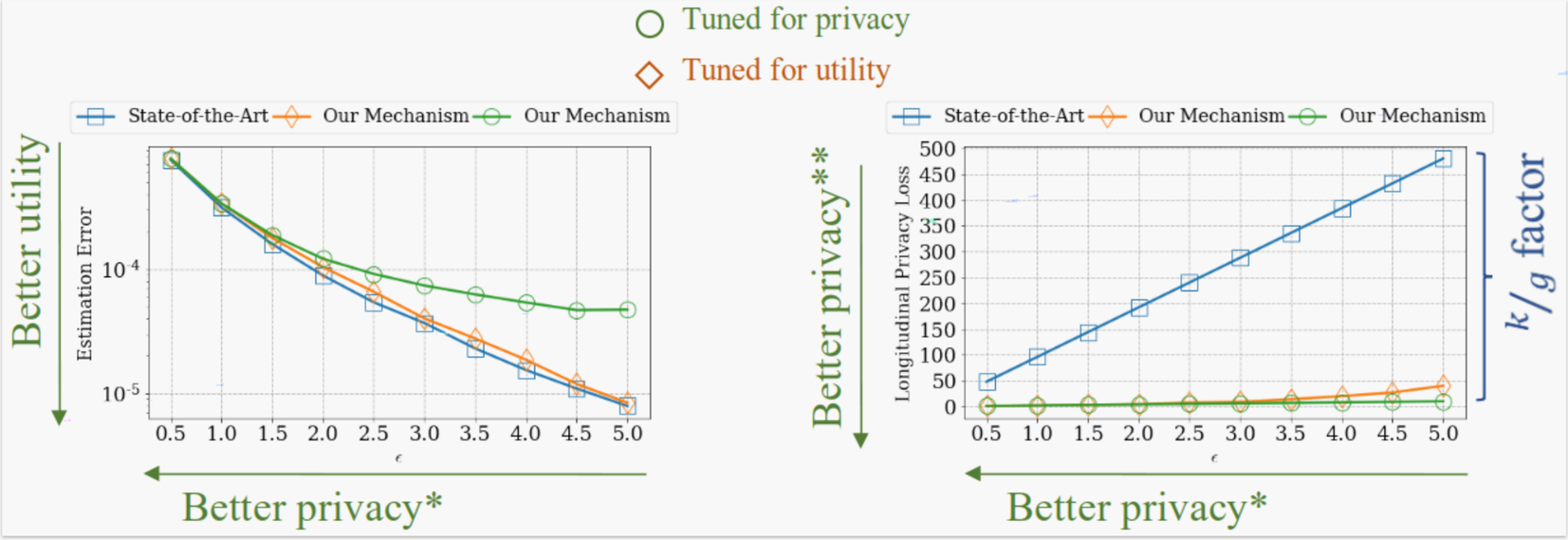
Execution (each user u at each time step t):

1. Hash the secret $x_t^{(u)}$ into $[1..g]$
2. Apply permanent g-RR and instantaneous g-RR

$$x \quad \mapsto \quad \mathrm{RR}\left(\mathrm{RR}_{\mathrm{memo}} \left(H_u(x_t^{(u)}) \bmod g \right) \right)$$

Results

Adult dataset (`hours_per_week` attribute, $n = 45422, k = 96, \tau = 260$)



ϵ_∞	$\in (0, \infty)$	*LDP privacy loss assuming constant input.
MSE	$\in [0, 1]$	$\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{1}{k}\sum_{x=1}^{k}(f(x) - \hat{f}(x))^2$
$\check{\epsilon}_{\text{avg}}$	$\in (0, k\epsilon_\infty)$	***"longitudinal loss"* permanent randomizations.

Conclusion

LOLOHA: a protocol for collecting evolving categorical data

Remarks

- LOLOHA combines double randomization with reduced domain
- Double randomization is decent but not LDP

Contributions

- Similar performance with less *“LDP on the users' values”*
- First performant method not based on Unary Encoding