PCKV: Locally Differentially Private Correlated Key-Value Data Collection with Optimized Utility

Xiaolan Gu*, Ming Li*, Yueqiang Cheng**, Li Xiong# and Yang Cao†

* University of Arizona       ** Baidu Security       # Emory University       † Kyoto University

USENIX Security Symposium, August 2020
Overview

• Background of LDP
• Problem Statement and Existing Mechanism
• Our Framework: PCKV
• Experiments
• Conclusion
Background

• Companies are collecting our private data to provide better services (Google, Facebook, Apple, Yahoo, Uber, …)

• However, privacy concerns arise

• Possible solution: locally private data collection model

  • Yahoo: massive data breaches impacted 3 billion user account, 2013
  • Facebook: 267 million users’ data has reportedly been leaked, 2019
  • …
Local Differential Privacy (LDP) [Duchi et al, FOCS' 13]

A mechanism $M$ satisfies $\epsilon$-LDP if and only if for any pair of inputs $x, x'$ and any output $y$

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

- $x, x'$: the possible input (raw) data (generated by the user)
- $y$: the output (perturbed) data (public and known by adversary)
- $\epsilon$: privacy budget (a smaller $\epsilon$ indicates stronger privacy)

An adversary cannot infer whether the input is $x$ or $x'$ with high confidence (controlled by $\epsilon$)
Applications of LDP

Enabling developers and organizations to use differential privacy
Thursday, September 5, 2019

Posted by Miguel Guevara, Product Manager, Privacy and Data Protection Office

Source:
https://developers.googleblog.com/2019/09/enabling-developers-and-organizations.html

Source:
https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html
LDP Protocol: Randomized Response

• Randomized Response (RR) [Warner, 1965]: reports the truth with some probability (for binary answer: yes-or-no)

  Advanced versions: Unary Encoding, Generalized RR, …

• Example: Is your annual income more than 100k?

| Truth \(x\) | Response \(y\) |
|--------------|---------------|
| 1            | w.p. \(p\)   |
| 0            | w.p. \(1-p\) |
| 0            | w.p. \(1-p\) |
| 0            | w.p. \(p\)   |

To satisfy \(\varepsilon\)-LDP: \(p = \frac{e^\varepsilon}{e^\varepsilon + 1}\) (since \(\frac{p}{1-p} = e^\varepsilon\))

\(\mathbb{E}[f] = f^*p + (1 - f^*)(1 - p) = (2p - 1)f^* + (1 - p)\)

Frequency estimation: \(\hat{f} = \frac{f - (1 - p)}{2p - 1}\)

Unbiasedness: \(\mathbb{E}[\hat{f}] = f^*\)
Extend RR for General Cases

• Assume the domain size is $d$ (taking $d = 5$ for example)

---

**Optimized Unary Encoding (OUE)**  
[Wang et al, USENIX Security’ 17]

**Input**

| Bit vector | 1 | 0 | 1 | 0 | 0 |
---|---|---|---|---|---|

**Encode**

| Output | 1 | 0 | 1 | 1 | 0 |
---|---|---|---|---|---|

**Perturb**

- Flip 1 w.p. 0.5
- Flip 0 w.p. $1 - p$

To satisfy $\epsilon$-LDP: $p = \frac{e^\epsilon}{e^\epsilon + 1}$

---

**Staircase or Generalized RR (GRR)**  
[Kairouz et al, NeurIPS’ 16]

**Input**

| 1 | 2 | 3 | 4 | 5 |
---|---|---|---|---|

**Encode**

| Output | 1 | 2 | 3 | 3 | 5 |
---|---|---|---|---|---|

**Perturb**

- Directly perturb
- Flip 1 w.p. 0.5
- Flip 0 w.p. $1 - p$

To satisfy $\epsilon$-LDP: $p = \frac{e^\epsilon}{e^\epsilon + d - 1}$

---

RR, OUE and GRR are building block mechanisms for frequency aggregation
Key-Value Data Collection

A motivating example (movie rating system)

| Movies            | # Ratings | Avg. Rating |
|-------------------|-----------|-------------|
| Man in Black      | 1200      | 4.1         |
| Spider-Man        | 1000      | 3.3         |
| The Godfather     | 200       | 4.7         |

Ratings are in the range [1, 5]

- Data Type: each user has multiple key-value pairs
- Data Domain: key in \{1, 2, \ldots, d\}, value in \([-1,1]\)
- Task: frequency and mean estimation
- Threat Model: honest-but-curious server
- Objectives: good privacy-utility tradeoff

Challenges

1. Each user has different number of key-value pairs.
2. If a fake key is reported, how to report the corresponding value?
3. How to design an optimal mechanism with the best privacy-utility tradeoff?
Existing Mechanism: PrivKVM [Ye et al, S&P’ 19]

Step 1. Convert key-value pairs into a vector

In each round, each user 1) randomly samples an index $j$ from $\{1, \ldots, d\}$; 2) privately reports the $j$-th pair (if a fake key is reported, then the value will be perturbed from the estimated mean by the server)

Limitations of PrivKVM

- **Multiple rounds** requires all users to be always online and the privacy budget in each round is very small (thus large error).
- The naive sampling protocol may not work well for a large domain.
- No improved privacy budget composition (although key and value are perturbed with some correlation).

Our Mechanism

- Only one round
- Advanced sampling protocol
- Tight privacy budget composition (and optimized budget allocation)
Outline

• Background of LDP

• Problem Statement and Existing Mechanism

• Our Framework: PCKV

• Experiments

• Conclusion
Overview of PCKV

- **Set Up**
  - Privacy Budget Allocation and Perturbation Probability Computation
    - $\epsilon$: the total privacy budget
    - PCKV-UE: $\epsilon \rightarrow \{\epsilon_1, \epsilon_2\} \rightarrow \{a, b, p\}$
    - PCKV-GRR: $\epsilon \rightarrow \{\epsilon_1, \epsilon_2\} \rightarrow \{a\}$
  - $\epsilon_1$: budget for key perturbation
  - $\epsilon_2$: budget for value perturbation
  - $a, b, p$: perturbation probabilities

- **User-Side**
  - Sampling
    - $S \rightarrow x = \langle k, v \rangle$
  - Perturbation
    - PCKV-UE: $x \rightarrow y$ (vector)
    - PCKV-GRR: $x \rightarrow y' = \langle k', v' \rangle$
  - Sampling
    - $S$: the set of key-value pairs
    - $x$: the sampled key-value pair
  - Perturbation
    - $y$ or $y'$: the output of each user

- **Server-Side**
  - Aggregation
    - PCKV-UE: $y[k] \rightarrow \{n_1, n_2\}$
    - PCKV-GRR: $y' \rightarrow \{n_1, n_2\}$
  - Estimation
    - $(n_1, n_2) \rightarrow \{f_k, m_k\}$

- **Joint perturbation and privacy analysis can improve privacy-utility tradeoff** (due to tight privacy budget composition)

- **Optimized budget allocation further improves the utility**

- **Advanced sampling protocol**: each user pads her keys into a uniform length $\ell$ by some dummy keys

- **Joint privacy analysis**: in an end-to-end way (instead of directly using sequential composition)

- **Optimized allocation** of $\epsilon_1$ and $\epsilon_2$: by minimizing MSE of estimation under tight budget composition

We use Padding-and-Sampling [S&P’18] to improve sampling efficiency

We theoretically evaluate the utility by MSE of estimation
Perturbation and Privacy Analysis

Joint/Correlated Perturbation

Unbiased map to 1 and -1

With privacy budget $\epsilon_1$

Value Discretization

Key Perturbation

If a fake key is reported?
  Yes
  Report value as 1 and -1 w.p. 0.5
  It also results in less information/privacy leakage
  No

With privacy budget $\epsilon_2$

Joint Privacy Analysis

The final privacy budget is less than $\epsilon_1 + \epsilon_2$

- PCKV-UE has tighter privacy budget composition than directly using sequential composition

$\epsilon = \max\{\epsilon_2, \epsilon_1 + \ln[2/(1 + e^{-\epsilon_2})]\} \leq \epsilon_1 + \epsilon_2$

(because $\epsilon_1 \geq 0$ and $\frac{2}{1+e^{-\epsilon_2}} \leq e^{\epsilon_2}$)

- PCKV-GRR has similar tight budget composition and additional privacy benefit from sampling.

- PrivKVM does not have tight budget composition (because the fake value is reported with two different probabilities).
Aggregation and Estimation

- The server aggregates the supporting numbers of value 1 and −1 for the $k$-th key.

- Estimated frequency $\hat{f}_k$: multiplied by $\ell$ due to sampling, where $\mathbb{E}[\hat{f}_k] = f_k^*$

- Estimated mean $\hat{m}_k = \frac{\text{calibrated sum}}{\text{calibrated counts}}$, where $\mathbb{E}[\hat{m}_k] \rightarrow m_k^*$ when $n \rightarrow \infty$

- The MSEs of $\hat{f}_k$ and $\hat{m}_k$ depend on how to balance $\epsilon_1$ and $\epsilon_2$ under a fixed total privacy budget $\epsilon$

Unbiased

Asymptotically Unbiased

Tractability of theoretical analysis
Optimized Privacy Budget Allocation

\[ \min \text{ MSE} + \text{Tight Composition} \overset{\Rightarrow}{\longrightarrow} \text{Optimized Allocation} \]

- A function of \( \epsilon_1, \epsilon_2 \)
- Relationship among \( \epsilon_1, \epsilon_2 \) and \( \epsilon \)
- How to optimally determine \( \epsilon_1, \epsilon_2 \) when given \( \epsilon \)

\[ \epsilon_1 = \ln((e^{\epsilon} + 1)/2), \quad \epsilon_2 = \epsilon \]
\[ \epsilon_1 = \ln(\ell \cdot (e^{\epsilon} - 1)/2 + 1), \quad \epsilon_2 = \ln(\ell \cdot (e^{\epsilon} - 1) + 1) \]

Summary of PCKV

- Step 1. Choose the advanced sampling protocol
- Step 2. Jointly perturb key-value and jointly analyze the privacy (which provides tight privacy budget composition)
- Step 3. Optimally put things together (i.e., optimized privacy budget allocation under a fixed total budget)

Final Perturbation (after sampling)
Experiments

- The theoretical results close (dashed lines) to the empirical results (solid lines)
- Our mechanisms outperforms existing ones on both frequency and mean estimation

Improvements of PCKV
- Advanced sampling protocol
- Tight budget composition
- Optimized budget allocation
Experiments

Benefit from each improvement
- Tight Budget Composition v.s. Sequential Composition
- Optimized Budget Allocation v.s. Non-optimized

Success of top frequent keys identification (varying domain size)
- PCKV mechanisms outperforms other ones
- PCKV-UE has smaller impact from large domain size
Real-world Data

Amazon Dataset

# ratings: 2M
# users: 1M
# keys: 249K

Data source: [https://www.kaggle.com/skillmuggler/amazon-ratings](https://www.kaggle.com/skillmuggler/amazon-ratings)

Movie Dataset

# ratings: 20M
# users: 138K
# keys: 26K

Data source: [https://www.kaggle.com/ashukr/movie-rating-data](https://www.kaggle.com/ashukr/movie-rating-data)
Conclusion

• The advanced sampling protocol can improve the sampling efficiency and the utility.

• Joint/correlated perturbations of key and value (rather than independent ones) can provide more options for mechanism design and the chance to choose the optimized one.

• Joint privacy analysis can lead to better privacy-utility tradeoff (because it results in tighter privacy budget composition than sequential composition)

Future work

• Study the optimized strategy of choosing $\ell$ in Padding-and-Sampling protocol.

• Extend the correlated perturbation and tight composition analysis to other general types of multi-dimensional data.
Thanks for your attention!

Q&A

Contact Information: xiaolang@email.arizona.edu