Slowly Scaling Per-Record Differential Privacy

Dr. Brian Finley[1], Anthony Caruso[1], Dr. Justin Doty[1], Dr. Ashwin Machanavajjhala[2], **Dr. Mikaela Meyer[3]**, Dr. David Pujol[2], Dr. Zachary Terner[3]

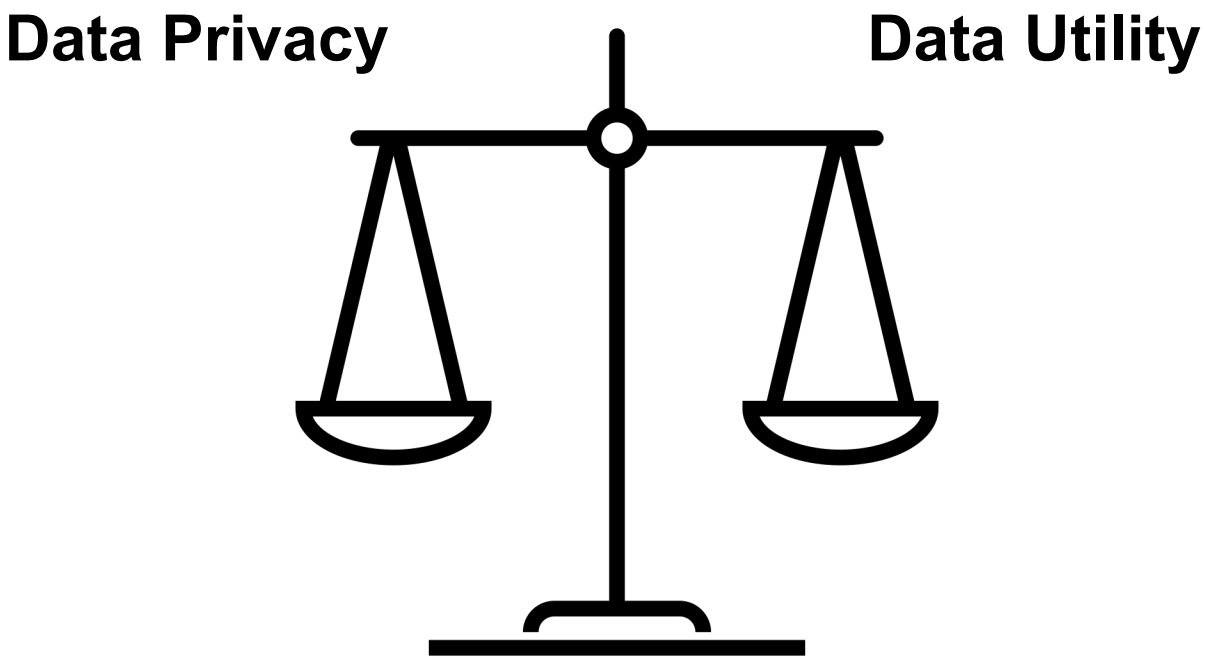
[1]United States Census Bureau, [2]Tumult Labs, [3]The MITRE Corporation

Government and Public Sector R Conference, October 2024

Any opinions and conclusions expressed herein are those of the authors and do not reflect the views of the U.S. Census Bureau

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD®

Data Privacy and Utility: A Delicate Balance



APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Introduction (1/2)

ID	Q1 Revenue	Avg. Salary
1	$7,000,000	$60,000
2	$8,000,000	$75,000
3	$7,500,000	$62,500

Dataset D

ID	Q1 Revenue	Avg. Salary
1	$7,000,000	$60,000
3	$7,500,000	$62,500

Dataset D'

- Some attacker wants to know whether they are querying D or D' (i.e., whether ID #2 is present).
- If they run a query on a dataset, they may be able to infer whether a record is in or out of the dataset.
- The broad goal of **differential privacy** is to add noise to queries to bound the probability that an attacker can tell whether they are querying D or D'; ideally, this probability is low

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Introduction (2/2)

ID	Q1 Revenue	Avg. Salary
1	$7,000,000	$60,000
2	$8,000,000	$75,000
3	$7,500,000	$62,500

Dataset D

ID	Q1 Revenue	Avg. Salary
1	$7,000,000	$60,000
3	$7,500,000	$62,500

Dataset D'

- If an attacker cannot tell whether they are querying D or D', we can say that

$$d_\alpha\left(M(D)\|M(D')\right) \leq \rho\alpha \text{ for all } \alpha \in (1,\infty).$$

- d_α represents the distance between the two outputs from applying mechanism M to D and D'
- ρ represents the privacy loss.
- We inhibit the attacker's inferences by ensuring that the difference between $\mathrm{M}(D)$ and $\mathrm{M}(D')$ is acceptably small.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

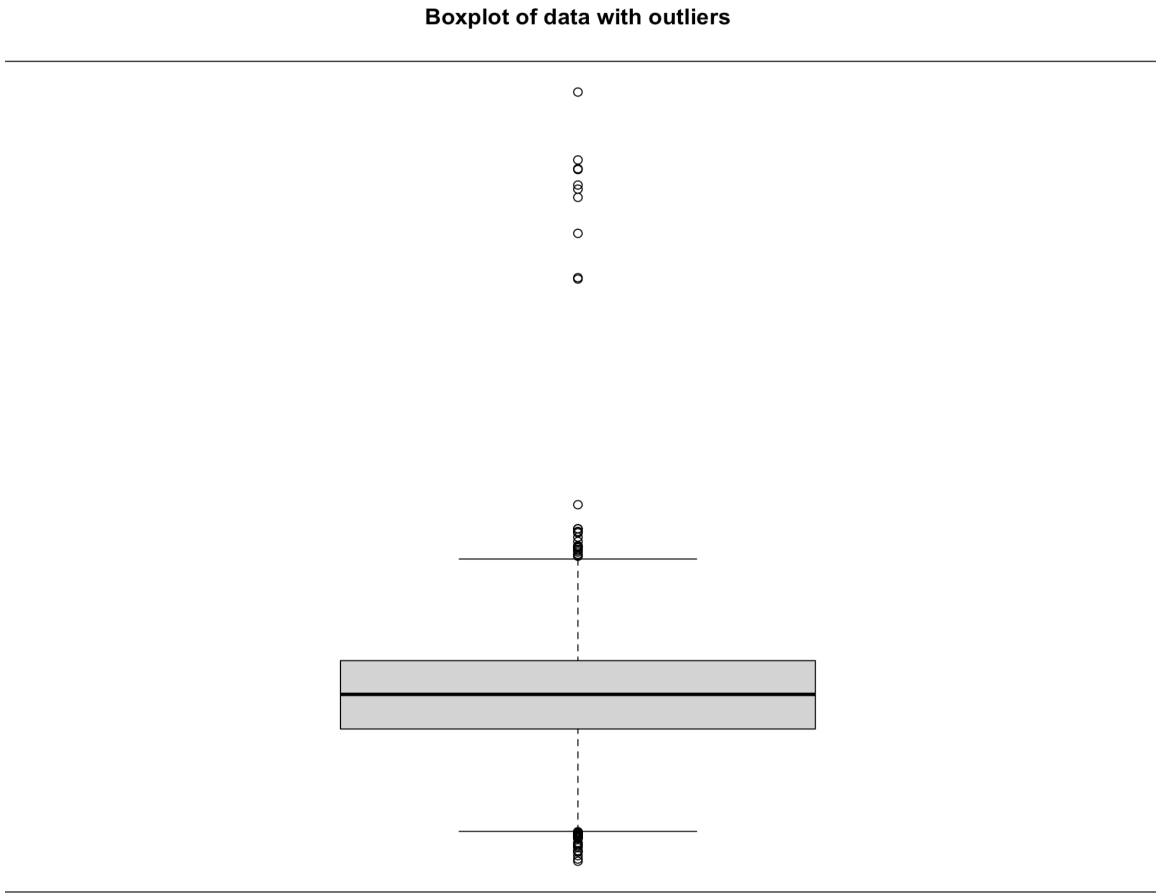
Motivation

- With traditional differential privacy, if you have a dataset with a few large outliers, you need to add noise on the scale of the largest outliers to mask their presence.

- This can harm the utility of the dataset.

- Data about business sizes can follow this pattern.

- Census wants to protect privacy in these cases.



APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

We propose *slowly scaling per-record differential privacy*

- Rather than give all observations in the dataset the same amount of protection, we can protect some records (i.e., smaller records) more strongly than others by allowing the upper bound to vary with the record's value.
- Slowly scaling per-record differential privacy:

$$D \ominus D' = \{r\} \implies d_\alpha(M(D)||M(D')) \leq \alpha P(r) \quad \forall \alpha \in (1, \infty)$$

- $P(r)$ is a *record-specific* policy function
- We can add noise that scales with the size of the record or add noise from fat tails.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Our mechanisms fall into two categories: *additive* and *transformation*

- Additive mechanisms: add noise from distributions with thicker tails than the Gaussian distribution
  - We derive properties of two distributions: Generalized Gaussian and Exponential Poly Log
- Transformation mechanisms:
  - Pass the query result *q(D)* through a transformation function *f*
  - Add Gaussian noise
  - Invert the transformation function with inverse function *g*

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Mechanisms Considered in this Paper

Transformation Mechanism with Transformation f

Name	$f(x)$	Variance	PRzCDP Policy Function
Identity	x	σ^2	$\frac{\Delta(r)^2}{2\sigma^2}$
Square Root	$\sqrt{x}$	$2\sigma^4 + 4\sigma^2(q(D) + a)$	$\frac{(\sqrt{\Delta(r)+a}-\sqrt{a})^2}{2\sigma^2}$
Fourth Root	$\sqrt[4]{x}$	$24\sigma^8 + 96\sigma^6(q(D)+a)^{\frac{1}{2}} + 72\sigma^4(q(D)+a) + 16\sigma^2(q(D)+a)^{\frac{3}{2}}$	$\frac{(\sqrt[4]{\Delta(r)+a}-\sqrt[4]{a})^2}{2\sigma^2}$
kth Root	$\sqrt[k]{x}$	$\sum_{i=0}^{k-1}\binom{k}{i}^2(k-i)!\sigma^{2(k-i)}(q(D)+a)^{\frac{2i}{k}}$	$\frac{(\sqrt[k]{\Delta(r)+a}-\sqrt[k]{a})^2}{2\sigma^2}$
Log	$\ln(x)$	$(e^{\sigma^2}-1)(q(D)+a)^2$	$\frac{(\ln(\Delta(r)+a)-\ln(a))^2}{2\sigma^2}$

The variance of transformation mechanisms scales with the query value.

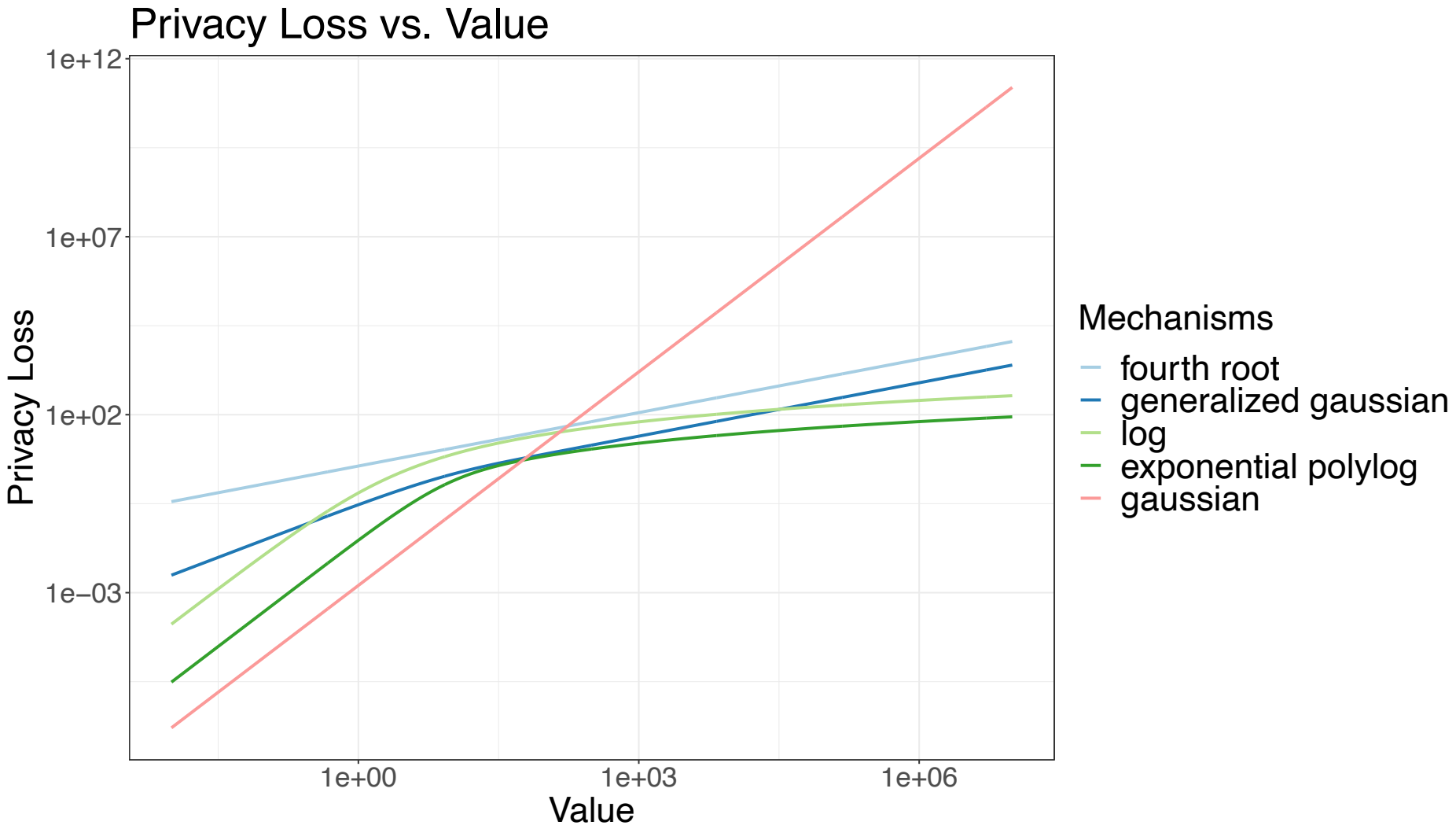
Additive Mechanism with Noise Density $\propto e^{f(|z|)}$

Name	$f(\lvert z\rvert)$	Variance	PRDP Policy Function
Generalized Gaussian	$-\left(\frac{\lvert z\rvert}{\sigma}\right)^p$	$\sigma^2\frac{\Gamma(\frac{3}{p})}{\Gamma(\frac{1}{p})}$	$(\frac{\Delta(r)}{\sigma})^p$
Exponential Polylog ($p=2$)	$-d\ln(\frac{\lvert z\rvert}{\sigma}+a)^2$	$\sigma^2\frac{e^{\frac{2}{d}}[1-\kappa(a,\frac{3}{2})]-2ae^{\frac{3}{4d}}[1-\kappa(a,1)]}{[1-\kappa(a,\frac{1}{2})]}+a^2\sigma^2$	$d(\ln(\frac{\Delta(r)}{\sigma}+a)^2-\ln(a)^2)$

This column represents the policy functions, $P(r)$.

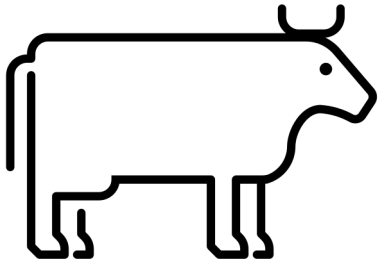
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Our mechanisms guarantee less privacy loss when the dataset has large outliers

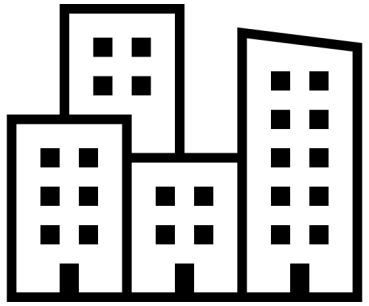


APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

We empirically evaluate our mechanisms using two datasets



USDA's 2019 Cattle Inventory Survey

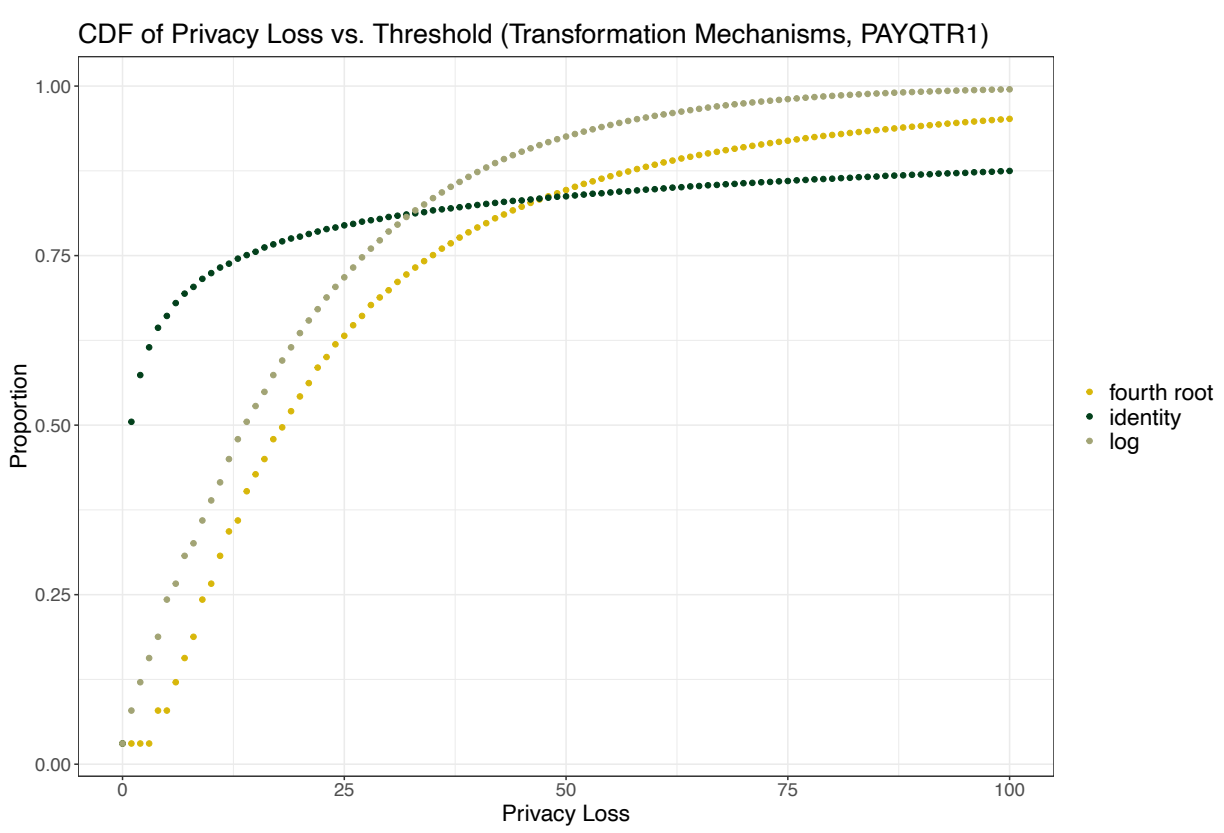


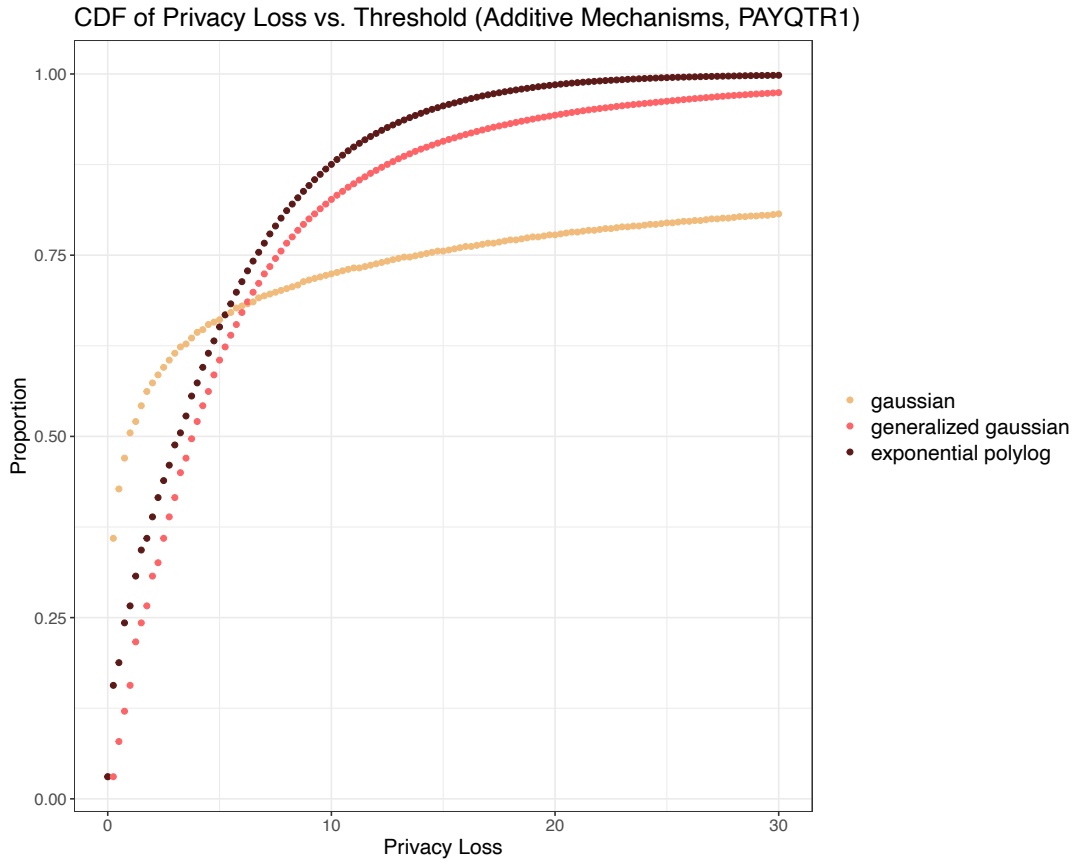
Simulated County Business Patterns Data

- Both datasets have large outliers (e.g., large farms and large businesses)
- Queries:
  - Number of cattle in each state
  - Number of employees, total annual pay, and total Q1 pay for each county and industry
- We use a Python package developed by Tumult Labs to implement the mechanisms and make all of our plots in ggplot using Wes Anderson palettes

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Cumulative distribution functions (CDFs) show that our mechanisms incur less privacy loss



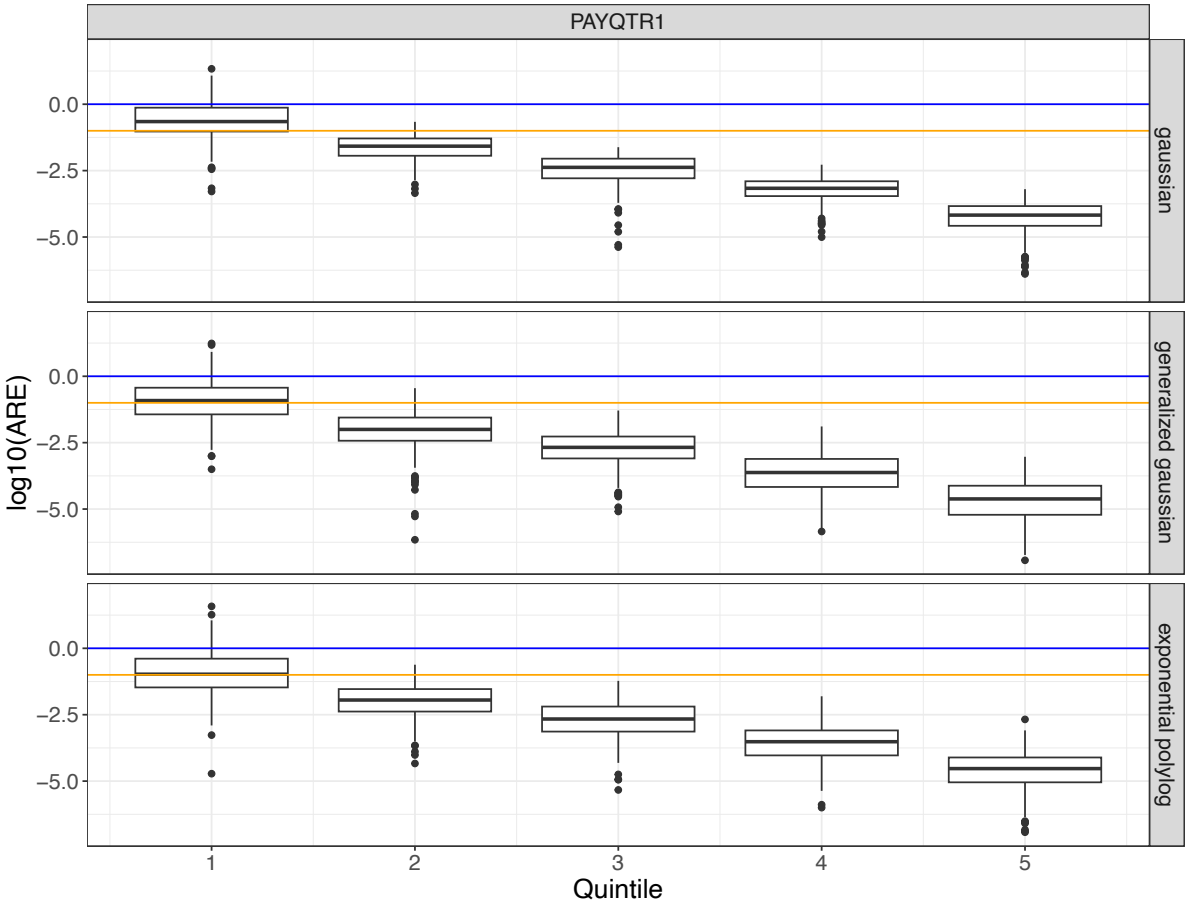


APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

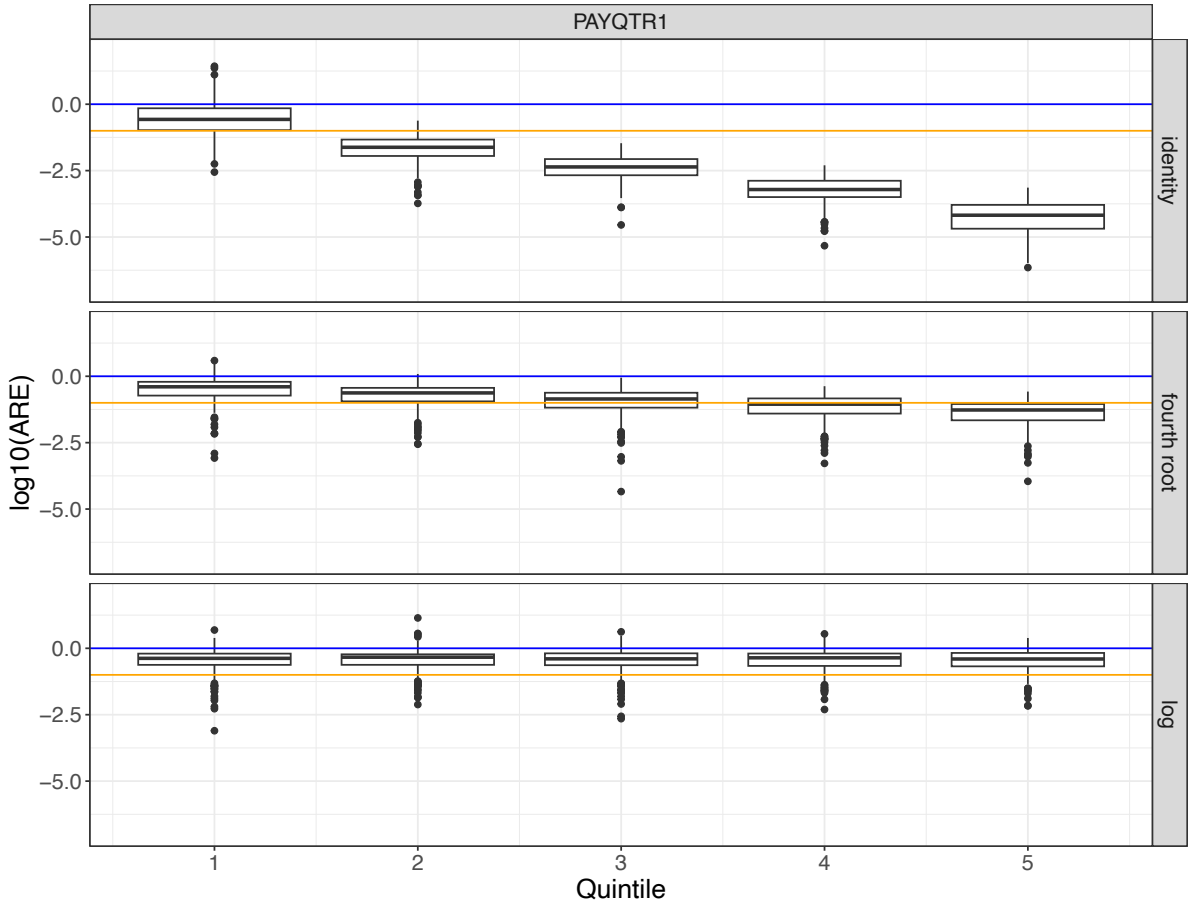
Some of our mechanisms incur too much relative error

$$ARE = \frac{|Noisy - True|}{True}$$

Log10(ARE) vs. True Sum Quintiles: Additive Mechanisms (Simulated)



Log10(ARE) vs. True Sum Quintiles: Transformation Mechanisms (Simulated)



APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

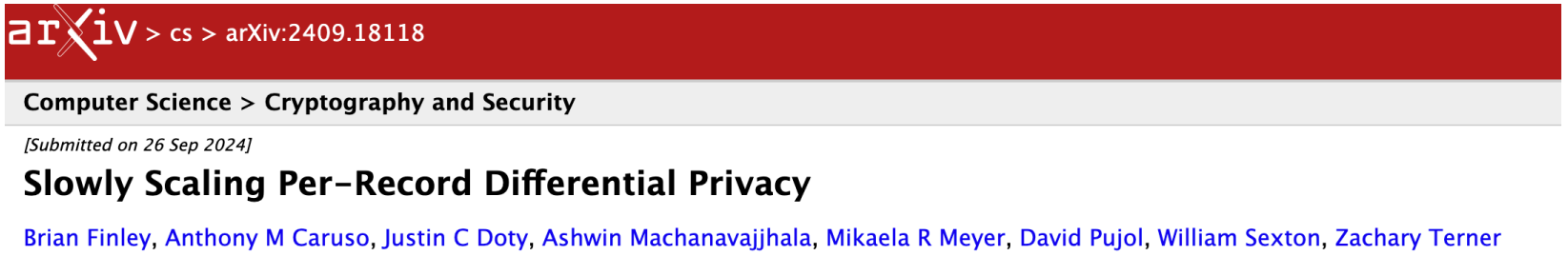
Discussion

- We developed formally private mechanisms with *slowly scaling* privacy loss to better handle positively skewed distributions.
- Additive Mechanisms: add noise from fat-tailed distributions, which lets privacy loss scale slowly.
- Transformation Mechanisms: Privacy loss scales slowly by transforming the query, adding noise, and inverting the transformation.
- Neither type of mechanism uniformly dominates.
- Further work is needed to determine the conditions for choosing a specific mechanism and policy function.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182

Thank you!

- Our paper is posted on arXiv! arxiv.org/abs/2409.18118
- Corresponding author: Brian Finley, brian.finley@census.gov
- Presenter: Mikaela Meyer, mrmeyer@mitre.org
  - Personal website: mikaelameyer.netlify.app


arXiv > cs > arXiv:2409.18118

Computer Science > Cryptography and Security

[Submitted on 26 Sep 2024]

Slowly Scaling Per-Record Differential Privacy

Brian Finley, Anthony M Caruso, Justin C Doty, Ashwin Machanavajjhala, Mikaela R Meyer, David Pujol, William Sexton, Zachary Terner

QR Code to arXiv

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PUBLIC RELEASE CASE NUMBER 24-3182