

# Differential Privacy for Economic Statistics

Margaret Beckom, *U.S. Census Bureau*

Anthony Caruso, *U.S. Census Bureau*

William Sexton, *Tumult Labs*



Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection [Project No: P-7520953, Disclosure Review Board (DRB) approval number: CBDRB-FY23-ESMD002-035].

# What will be covered...

- County Business Patterns (CBP) Background
- High-level Overview of Per-Record Differential Privacy (PRDP)
- Accuracy & Privacy
- Future Work
- Outstanding Research

# County Business Patterns (CBP)

A Brief Overview on the Program and Current Disclosure Avoidance Methodology

# CBP Background

- Includes the following estimates
  - Counts
    - Establishments
  - Magnitude
    - Employment during the week of March 12
    - First quarter payroll
    - Annual payroll
- Data is useful for studying the economic activity in small areas
- Current methodology: multiplicative noise
  - Applied only to magnitude data

# A Look into Per-Record DP

How to handle heavy-tailed distributions

**Objective: Accurately release key economic indicators from heavy-tailed distributions with modernized privacy protection.**

## **Challenges:**

- Differential privacy (DP) provides strong privacy protection but does not handle heavy-tailed distributions very well.
- “Per-Record” DP (PRDP) provides high data utility and formal privacy protection, but the privacy protection is not as strong as differential privacy.

# The challenge of applying DP to the CBP: How many employees work in ZIP 89506?

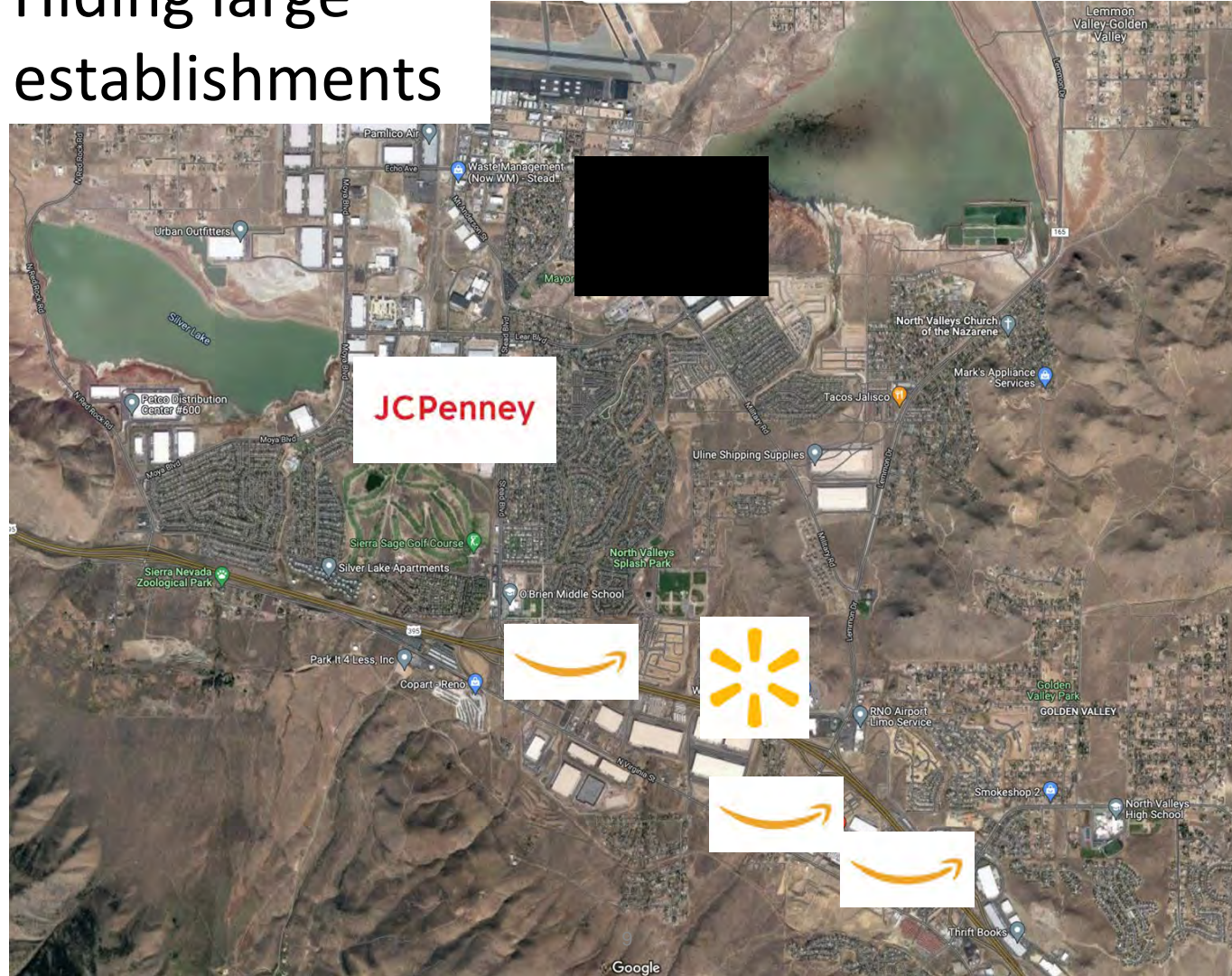


# Hiding large establishments





# Hiding large establishments



# Overkill for small establishments



Solution: Design a formal privacy framework that provides “sliding” protection to establishments.

**Protection against “fact-of-filing”:** adversaries should not be able to easily infer whether an establishment is represented in the CBP dataset.

**Protection against exact inference:** adversaries should not be able to deterministically infer exact attributes about an establishment, such as number of employees or annual payroll.

**Protection for firms:** the privacy properties of firms (i.e., collections of related establishments) should be inherited from the privacy properties of their individual establishments.

**“Sliding” establishment protection:** Allow the privacy guarantees to vary by establishment. In particular, allow privacy guarantees to degrade as the influence of an establishment grows.

# Hide small establishments?



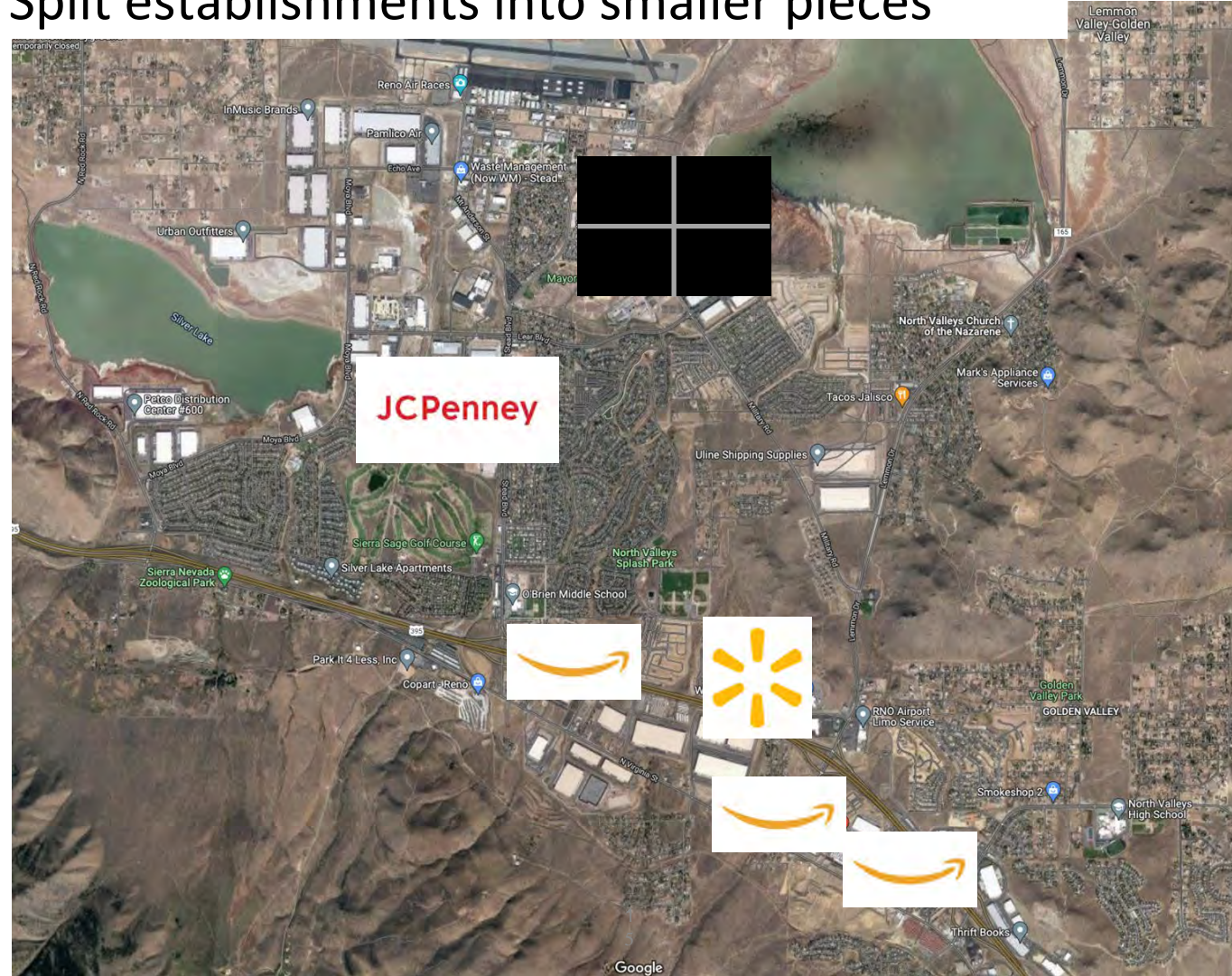
# Fully hide small establishments



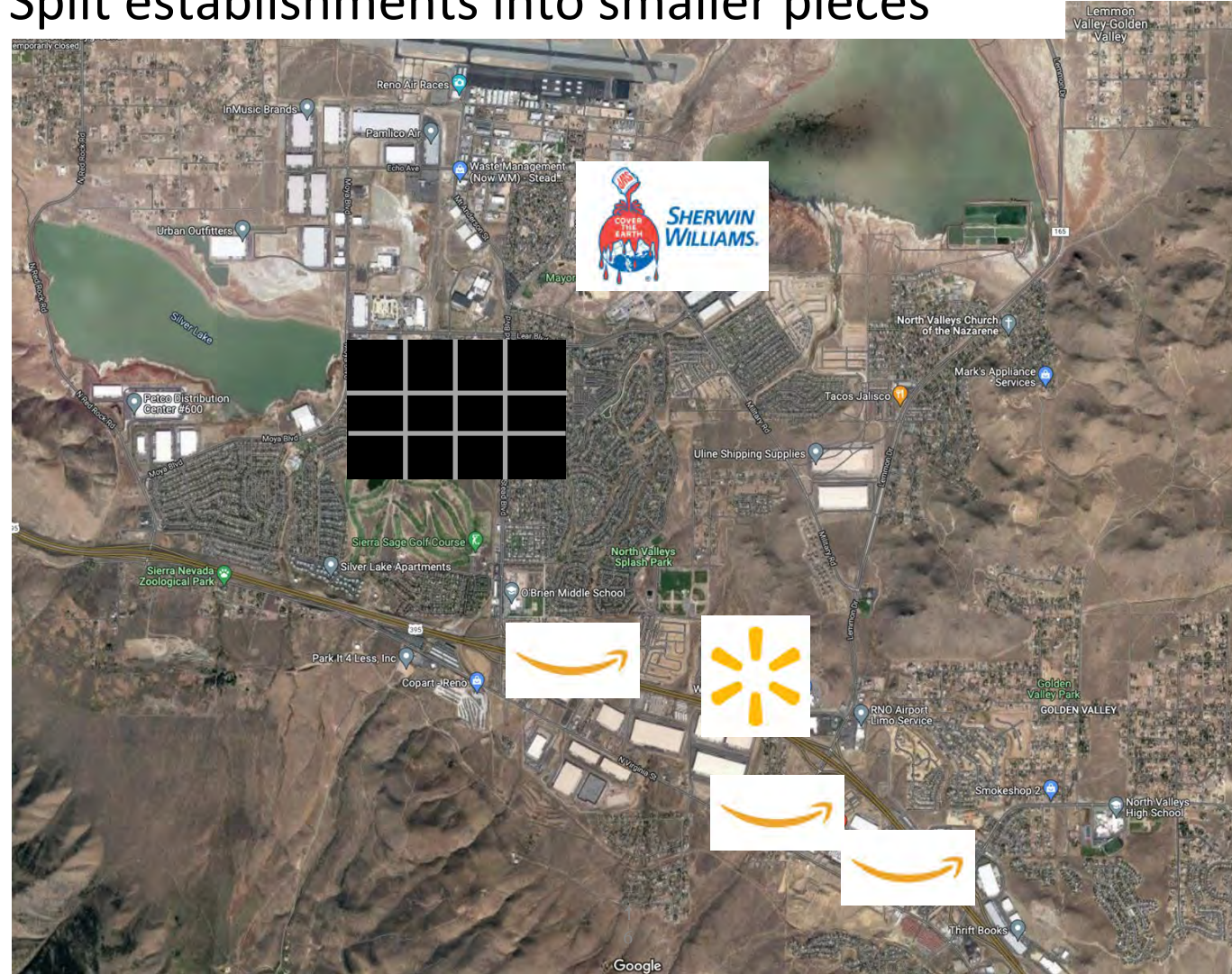
# Hide large establishments?



# Split establishments into smaller pieces



# Split establishments into smaller pieces





# Group Privacy for $\epsilon$ -DP Algorithms

- $\epsilon$ -DP algorithm privacy guarantee extends to groups of size  $k$
- When input databases differ by adding or removing up to  $k$  records:
  - Output distributions are bounded by  $k \times \epsilon$

# PRDP for Establishments

- When input databases differ by adding or removing a particular record,  $r$ 
  - Output distributions are bounded by a function  $P(r)$
- With establishment splitting  $A$  followed by DP mechanism  $M$ 
  - $P(r) = |A(r)| \times \varepsilon$  where  $|A(r)|$  is the number of pieces record  $r$  gets split into

# Data Utility Comparison

## **No establishment splitting**

- Bias-variance trade off
- Outliers drive noise requirements
- Outliers dominate a large share of the aggregations

## **Establishment splitting**

- No bias
- Noise requirements can be calibrated based on smaller establishments

# Accuracy and Privacy

# Accuracy and Privacy

- **The tradeoff between accuracy and privacy remains.**

Less privacy for establishments → More accurate estimates

More privacy for establishments → Less accurate estimates

- Challenge: Find a balance where quality is preserved, and all establishments are adequately protected.

# Median CV of County x 3-Digit NAICS Annual Payroll Estimates by Standard Deviation of Noise and Cell Size

Standard deviation of noise	All	Cell size (number of establishments)				
		a) 1-2	b) 3-9	c) 10-24	d) 25-99	e) 100+
10	0.00	0.03	0.01	0.00	0.00	0.00
20	0.01	0.07	0.02	0.00	0.00	0.00
50	0.02	0.17	0.04	0.01	0.00	0.00
100	0.04	0.35	0.08	0.02	0.01	0.00
200	0.09	0.69	0.16	0.04	0.01	0.00
500	0.22	>1.00	0.40	0.11	0.03	0.00
1,000	0.44	>1.00	0.80	0.22	0.06	0.01
2,000	0.88	>1.00	>1.00	0.44	0.13	0.02
5,000	>1.00	>1.00	>1.00	>1.00	0.31	0.04
10,000	>1.00	>1.00	>1.00	>1.00	0.63	0.09
20,000	>1.00	>1.00	>1.00	>1.00	>1.00	0.18
# of cells	187,446	53,264	61,799	32,569	26,901	12,913

Median CV
0.00 – <0.05
0.05 – <0.10
0.10 – <0.15
0.15 – <0.50
0.50+

# Median CV of County x 3-Digit NAICS Annual Payroll Estimates by Standard Deviation of Noise and Cell Size

Standard deviation of noise	All	Cell size (number of establishments)				
		a) 1-2	b) 3-9	c) 10-24	d) 25-99	e) 100+
10	0.00	0.03	0.01	0.00	0.00	0.00
20	0.01	0.07	0.02	0.00	0.00	0.00
50	0.02	0.17	0.04	0.01	0.00	0.00
100	0.04	0.35	0.08	0.02	0.01	0.00
200	0.09	0.69	0.16	0.04	0.01	0.00
500	0.22	>1.00	0.40	0.11	0.03	0.00
1,000	0.44	>1.00	0.80	0.22	0.06	0.01
2,000	0.88	>1.00	>1.00	0.44	0.13	0.02
5,000	>1.00	>1.00	>1.00	>1.00	0.31	0.04
10,000	>1.00	>1.00	>1.00	>1.00	0.63	0.09
20,000	>1.00	>1.00	>1.00	>1.00	>1.00	0.18
# of cells	187,446	53,264	61,799	32,569	26,901	12,913

Median CV
0.00 – <0.05
0.05 – <0.10
0.10 – <0.15
0.15 – <0.50
0.50+

# Ensuring Adequate Protection

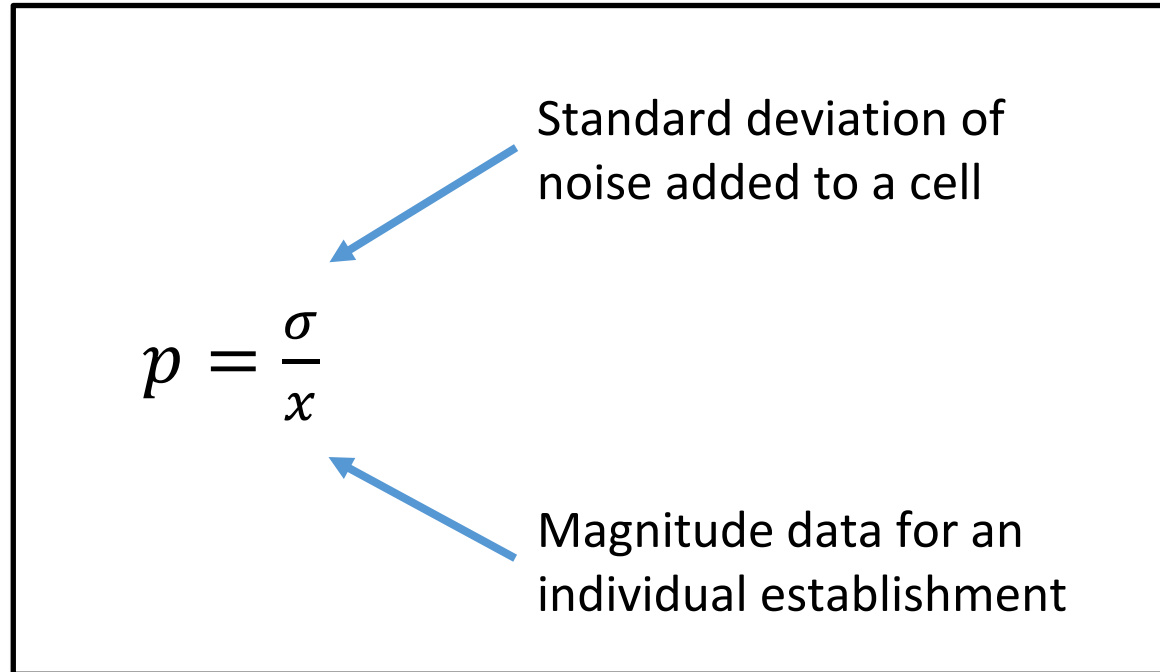
- PRDP with quality-preserving parameters will provide adequate protection for most establishments' magnitude data.
- Very large establishments will not have adequate protection.

Establishment	Employment
#1	75
#2	150
#3	30,000
<b>Total</b>	<b>30,225</b>

PRDP with  $\theta = 10$   
and  $\rho = 0.1$  → **30,200 ( $\sigma = 22$ )**



# Relative Protection



# Relative Protection

$$p = \frac{22}{30,000} \approx 0.0007$$

Standard deviation of noise added to a cell

Magnitude data for an individual establishment

# Proportion of Establishments (N=7,960,386) with Relative Protection\* Meeting or Exceeding Selected Levels

Standard deviation of noise	p ≥ 0.001	p ≥ 0.01	p ≥ 0.02	p ≥ 0.05	p ≥ 0.10	p ≥ 0.15	p ≥ 0.20	p ≥ 1.00
10	0.987	0.877	0.792	0.633	0.489	0.402	0.346	0.111
20	0.994	0.932	0.877	0.758	0.633	0.550	0.489	0.191
50	0.998	0.972	0.945	0.877	0.792	0.728	0.676	0.346
100	0.999	0.987	0.972	0.932	0.877	0.831	0.792	0.489
200	1.000	0.994	0.987	0.965	0.932	0.903	0.877	0.633
500	1.000	0.998	0.996	0.987	0.972	0.958	0.945	0.792
1,000	1.000	0.999	0.998	0.994	0.987	0.980	0.972	0.877
2,000	1.000	1.000	0.999	0.998	0.994	0.991	0.987	0.932
5,000	1.000	1.000	1.000	0.999	0.998	0.997	0.996	0.972
10,000	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.987
20,000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.994

Proportion
0.95+
0.90 – <0.95
0.85 – <0.90
0.75 – <0.85
0.00 – <0.75

\* Relative protection for annual payroll

## Number of Establishments (N=7,960,386) with Relative Protection\* Less than Selected Levels

Standard deviation of noise	p < 0.001	p < 0.01	p < 0.02	p < 0.05	p < 0.10	p < 0.15	p < 0.20	p < 1.00
10	103,250	982,779	1,655,356	2,920,852	4,067,289	4,761,364	5,205,203	7,074,254
20	45,312	542,070	982,779	1,926,842	2,920,852	3,585,832	4,067,289	6,439,113
50	14,304	221,237	441,006	982,779	1,655,356	2,168,765	2,578,852	5,205,203
100	5,651	103,250	221,237	542,070	982,779	1,346,866	1,655,356	4,067,289
200	2,169	45,312	103,250	278,455	542,070	774,739	982,779	2,920,852
500	502	14,304	34,458	103,250	221,237	334,079	441,006	1,655,356
1,000	147	5,651	14,304	45,312	103,250	162,836	221,237	982,779
2,000	42	2,169	5,651	18,995	45,312	73,773	103,250	542,070
5,000	6	502	1,549	5,651	14,304	23,987	34,458	221,237
10,000	2	147	502	2,169	5,651	9,766	14,304	103,250
20,000	1	42	147	737	2,169	3,785	5,651	45,312

\* Relative protection for annual payroll

## Number of Establishments (N=7,960,386) with Relative Protection\* Less than Selected Levels

Standard deviation of noise	p < 0.001	p < 0.01	p < 0.02	p < 0.05	p < 0.10	p < 0.15	p < 0.20	p < 1.00
10	103,250	982,779	1,655,356	2,920,852	4,067,289	4,761,364	5,205,203	7,074,254
20	45,312	542,070	982,779	1,926,842	2,920,852	3,585,832	4,067,289	6,439,113
50	14,304	221,237	441,006	982,779	1,655,356	2,168,765	2,578,852	5,205,203
100	5,651	103,250	221,237	542,070	982,779	1,346,866	1,655,356	4,067,289
200	2,169	45,312	103,250	278,455	542,070	774,739	982,779	2,920,852
500	502	14,304	34,458	103,250	221,237	334,079	441,006	1,655,356
1,000	147	5,651	14,304	45,312	103,250	162,836	221,237	982,779
2,000	42	2,169	5,651	18,995	45,312	73,773	103,250	542,070
5,000	6	502	1,549	5,651	14,304	23,987	34,458	221,237
10,000	2	147	502	2,169	5,651	9,766	14,304	103,250
20,000	1	42	147	737	2,169	3,785	5,651	45,312

\* Relative protection for annual payroll

## Number of Establishments (N=7,960,386) with Relative Protection\* Less than Selected Levels

Standard deviation of noise	p < 0.001	p < 0.01	p < 0.02	p < 0.05	p < 0.10	p < 0.15	p < 0.20	p < 1.00
10	103,250	982,779	1,655,356	2,920,852	4,067,289	4,761,364	5,205,203	7,074,254
20	45,312	542,070	982,779	1,926,842	2,920,852	3,585,832	4,067,289	6,439,113
50	14,304	221,237	441,006	982,779	1,655,356	2,168,765	2,578,852	5,205,203
100	5,651	103,250	221,237	542,070	982,779	1,346,866	1,655,356	4,067,289
200	2,169	45,312	103,250	278,455	542,070	774,739	982,779	2,920,852
500	502	14,304	34,458	103,250	221,237	334,079	441,006	1,655,356
1,000	147	5,651	14,304	45,312	103,250	162,836	221,237	982,779
2,000	42	2,169	5,651	18,995	45,312	73,773	103,250	542,070
5,000	6	502	1,549	5,651	14,304	23,987	34,458	221,237
10,000	2	147	502	2,169	5,651	9,766	14,304	103,250
20,000	1	42	147	737	2,169	3,785	5,651	45,312

\* Relative protection for annual payroll

## Number of Establishments (N=7,960,386) with Relative Protection\* Less than Selected Levels

Standard deviation of noise	p < 0.001	p < 0.01	p < 0.02	p < 0.05	p < 0.10	p < 0.15	p < 0.20	p < 1.00
10	103,250	982,779	1,655,356	2,920,852	4,067,289	4,761,364	5,205,203	7,074,254
20	45,312	542,070	982,779	1,926,842	2,920,852	3,585,832	4,067,289	6,439,113
50	14,304	221,237	441,006	982,779	1,655,356	2,168,765	2,578,852	5,205,203
100	5,651	103,250	221,237	542,070	982,779	1,346,866	1,655,356	4,067,289
200	2,169	45,312	103,250	278,455	542,070	774,739	982,779	2,920,852
500	502	14,304	34,458	103,250	221,237	334,079	441,006	1,655,356
1,000	147	5,651	14,304	45,312	103,250	162,836	221,237	982,779
2,000	42	2,169	5,651	18,995	45,312	73,773	103,250	542,070
5,000	6	502	1,549	5,651	14,304	23,987	34,458	221,237
10,000	2	147	502	2,169	5,651	9,766	14,304	103,250
20,000	1	42	147	737	2,169	3,785	5,651	45,312

\* Relative protection for annual payroll

# Decreasing Privacy Loss / Increasing Relative Protection

- Second-stage noise
  - Added post-PRDP
  - Scaled to noisy sums
  - **Not** scaled to the largest establishment in each cell
  - **Not** formally private
- Slow-scaling PRDP
  - Noise is added to *transformed* data
  - Greater control over relative noise levels



# Slow-Scaling PRDP

- Privacy loss grows more slowly as establishment employment size increases

Establishment	Employment
#1	75
#2	150
#3	30,000
<b>Total</b>	<b>30,225</b>

# Slow-Scaling PRDP

- Privacy loss grows more slowly as establishment employment size increases

Establishment	Employment
#1	75
#2	150
#3	30,000
<b>Total</b>	<b>30,225</b>

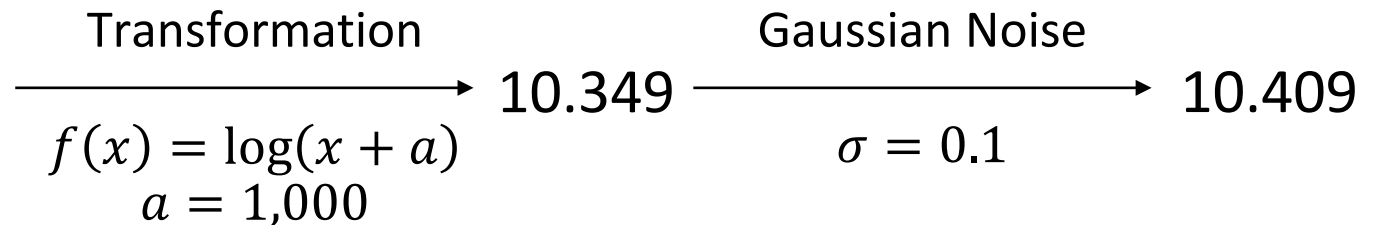
Transformation

$$\xrightarrow{\quad} 10.349$$
$$f(x) = \log(x + a)$$
$$a = 1,000$$

# Slow-Scaling PRDP

- Privacy loss grows more slowly as establishment employment size increases

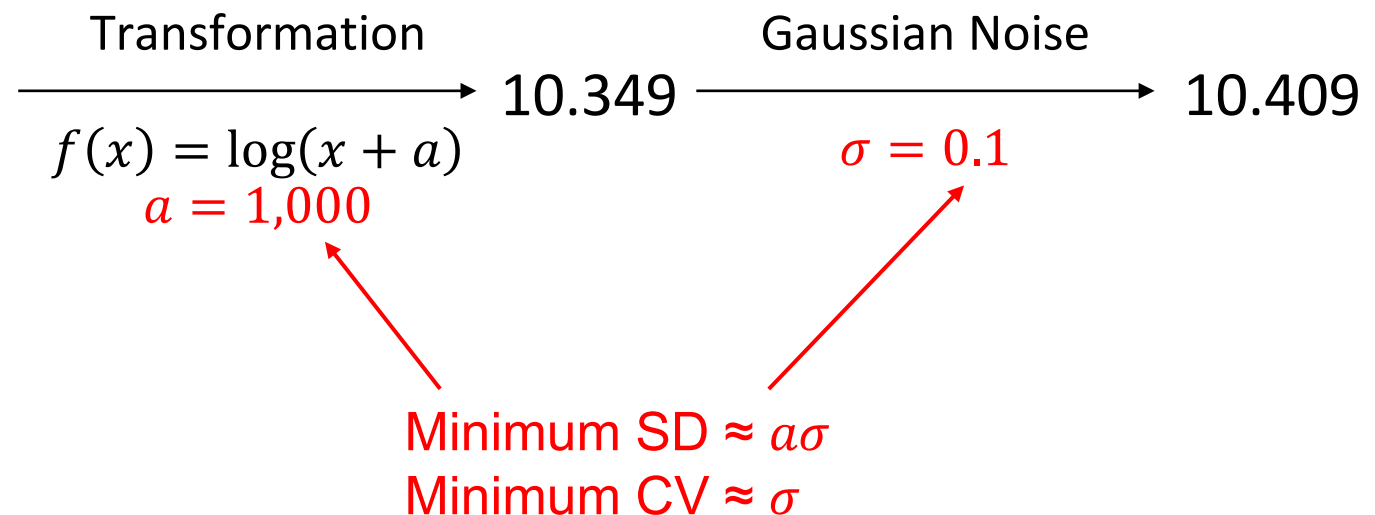
Establishment	Employment
#1	75
#2	150
#3	30,000
<b>Total</b>	<b>30,225</b>



# Slow-Scaling PRDP

- Privacy loss grows more slowly as establishment employment size increases

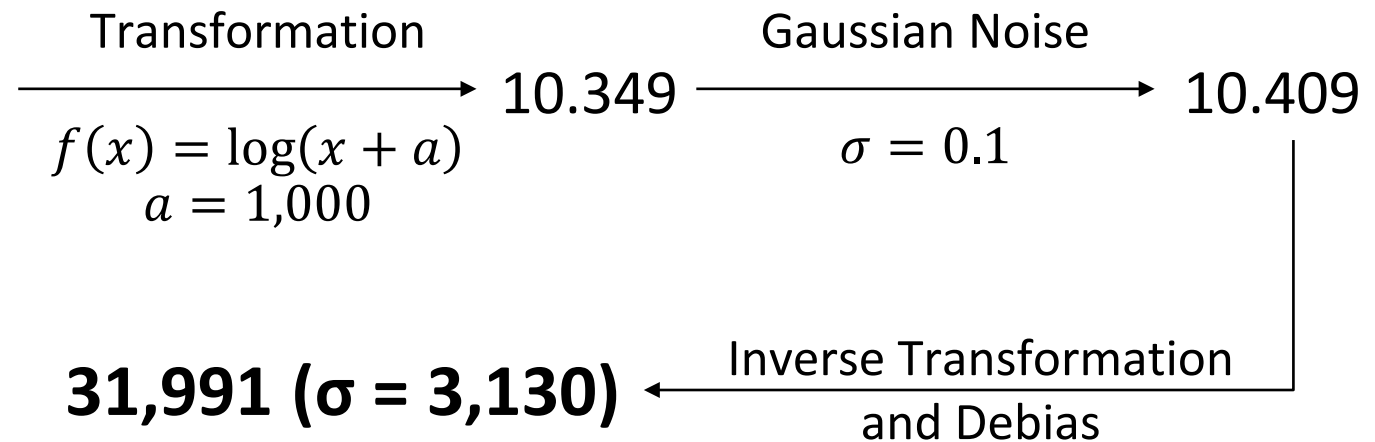
Establishment	Employment
#1	75
#2	150
#3	30,000
<b>Total</b>	<b>30,225</b>



# Slow-Scaling PRDP

- Privacy loss grows more slowly as establishment employment size increases

Establishment	Employment
#1	75
#2	150
#3	30,000
<b>Total</b>	<b>30,225</b>



# Future Work & Outstanding Research

A Look into the Future

# Future Work

- Slow-scaling PRDP
- Parameter tuning
  - *95% of cells with 25 or more establishments will have a CV of 10% or less*
  - Ratios
  - Trends
- Extending the work to other data products
- Communication about the methods

# Outstanding Research

- Protecting sample-based estimates
- Privacy-conserving approaches for firm counts
- Implications for benchmarking
- Privacy-protection algorithms for functions of the data other than sums
- Required privacy protections for product-level statistics



# Contact Us

- Margaret Beckom  
Dissemination Standards Branch  
Economic Management Division

Email: [margaret.m.beckom@census.gov](mailto:margaret.m.beckom@census.gov)  
with the subject “Econ Disclosure Feedback”

Phone: 301-763-7522