Formal Privacy in Census Data

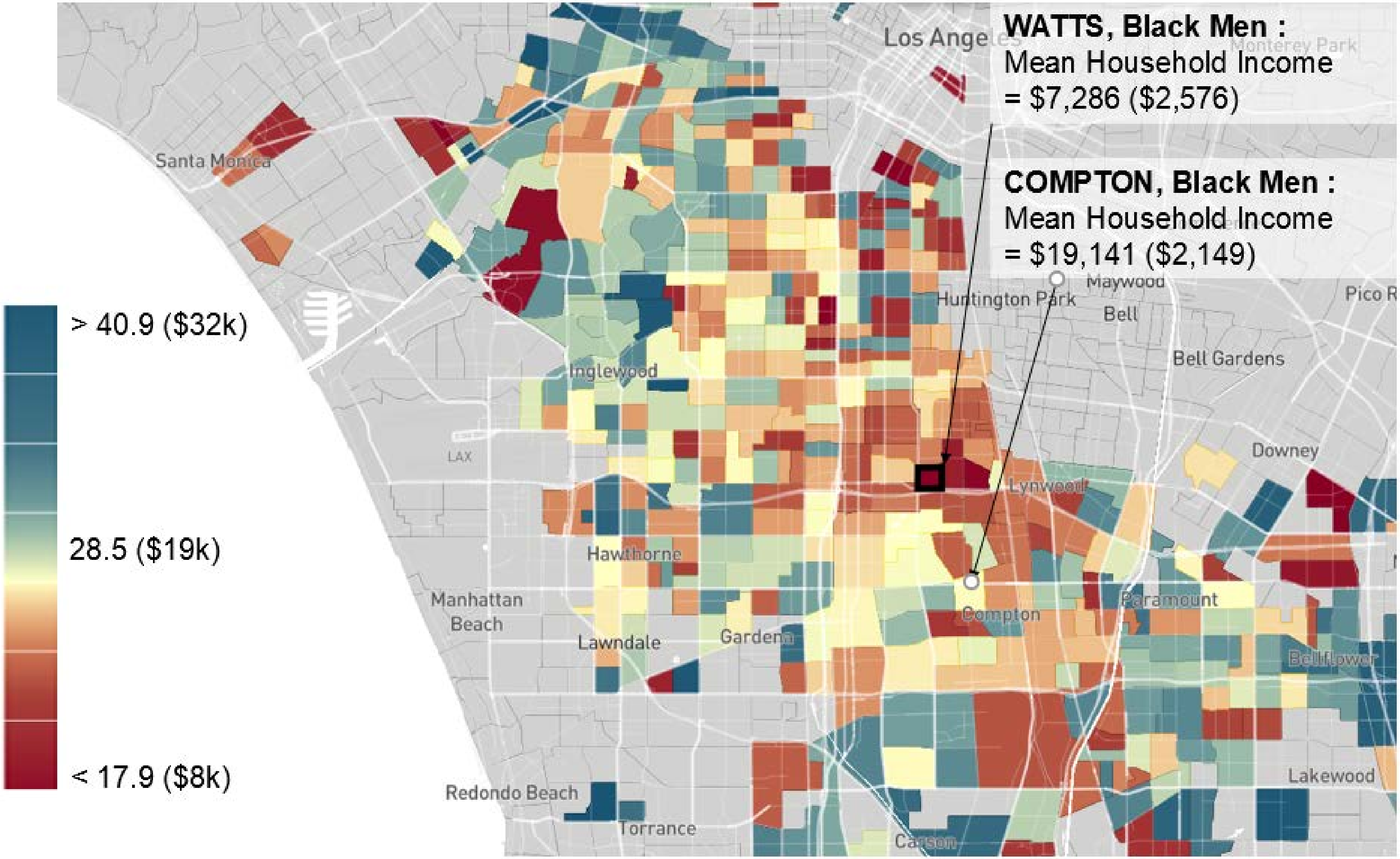
Ian M. Schmutte
Department of Economics
Terry College of Business
University of Georgia

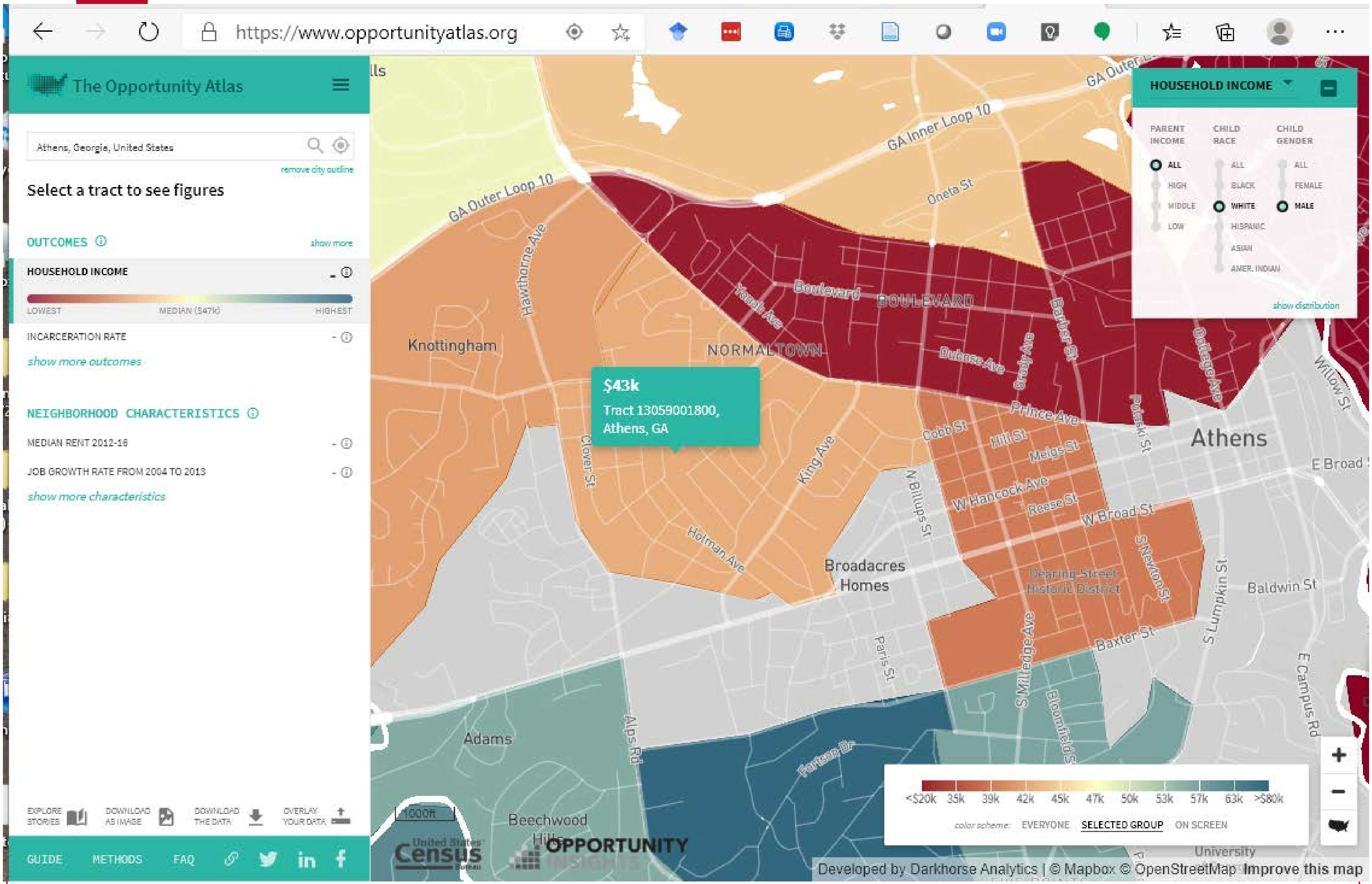
NBER Methods Lecture
July 17, 2020

Building Cool Stuff

1. Opportunity Atlas and the MOSE
2. LODES and EE-ER Privacy
3. IMI Hot Reports
4. Post-Secondary Employment Outcomes
5. Veterans Employment Outcomes

Raj Chetty
John N. Friedman




https://www.opportunityatlas.org
The Opportunity Atlas
Athens, Georgia, United States
remove city outline
Select a tract to see figures
OUTCOMES
show more
HOUSEHOLD INCOME
LOWEST
MEDIAN ($47k)
HIGHEST
INCARCERATION RATE
show more outcomes
NEIGHBORHOOD CHARACTERISTICS
MEDIAN RENT 2012-16
JOB GROWTH RATE FROM 2004 TO 2013
show more characteristics
EXPLORE STORIES
DOWNLOAD AS IMAGE
DOWNLOAD THE DATA
OVERLAY YOUR DATA
GUIDE
METHODS
FAQ
HOUSEHOLD INCOME
PARENT INCOME
ALL
HIGH
MIDDLE
LOW
CHILD RACE
ALL
BLACK
WHITE
HISPANIC
ASIAN
AMER. INDIAN
CHILD GENDER
ALL
FEMALE
MALE
show distribution
$43k
Tract 13059001800, Athens, GA
GA Outer Loop 10
GA Inner Loop 10
Oneta St
Hawthorne Ave
Boulevard
BOULEVARD
Knottingham
NORMALTOWN
Prince Ave
Cobb St
Hill St
Meigs St
W Hancock Ave
Reese St
W Broad St
King Ave
Clover St
Holman Ave
N Billups St
Broadacres Homes
Dearing Street Historic District
Baxter St
Athens
E Broad St
Baldwin St
S Lumpkin St
S Milledge Ave
Adams
Alps Rd
Forrest Dr
Beechwood Hills
Barber St
Pulaski St
Willow St
E Campus Rd
University
1000ft
<$20k
35k
39k
42k
45k
47k
50k
53k
57k
63k
>$80k
color scheme:
EVERYONE
SELECTED GROUP
ON SCREEN
United States Census Bureau
OPPORTUNITY INSIGHTS
Developed by Darkhorse Analytics | © Mapbox © OpenStreetMap Improve this map

Source of Opportunity Atlas Data

Data for people, i, in (race, gender, tract) group g

$$D_g = (\boldsymbol{y_i}, rank_i)_{i \in g}$$

Fit least-squares regression models per g

$$y_i = \alpha_g + \beta_g rank_i + \nu_i$$

Queries of interest

$$\theta_g(rank) \equiv q\left(D_g, rank\right) = \widehat{\alpha_g} + \widehat{\beta_g} rank$$

- Very small cells
- High sensitivity

Calibrating Noise to Sensitivity

Using Laplace mechanism, publish

$$\widetilde{\theta_g}(rank) = q\left(D_g, rank\right) + \omega_g$$

Where

$$\omega_g \sim Lap(0, \frac{\Delta q}{\epsilon})$$

Properties

- Satisfies ϵ-differential privacy
- Parallel composition across groups means that total privacy loss is

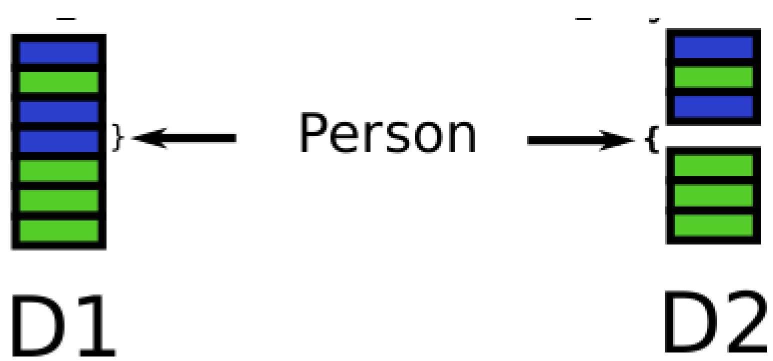
$$\epsilon = \max_g(\epsilon_g)$$

This Won't Work

Recall:

DP depends on how much output can change when evaluated on

ANY two different datasets



The (global) sensitivity is too dang high

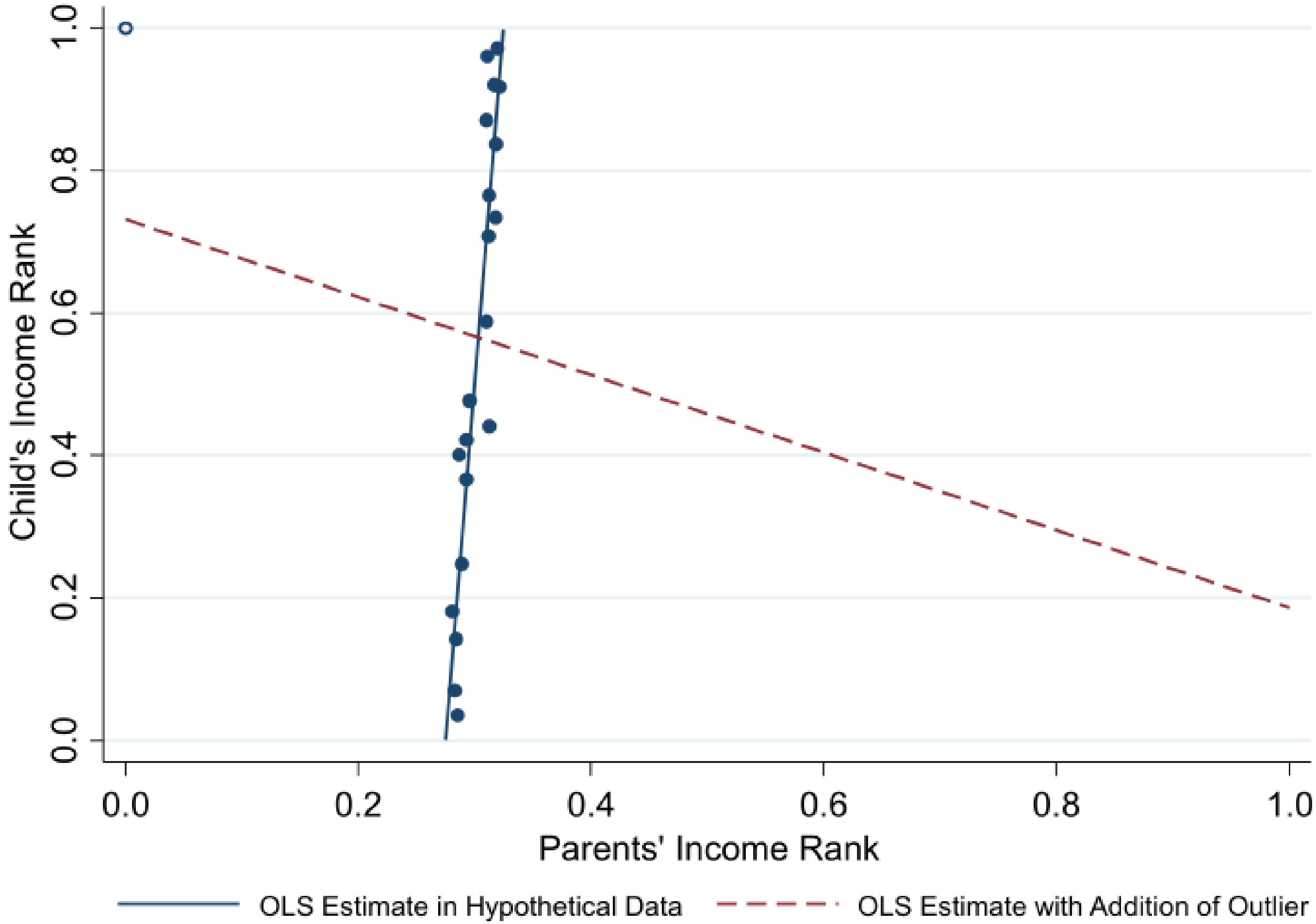
For Opportunity Atlas, sensitivity is:

how much could conditional mean of child earnings rank change

if I added or removed any legal value

from any conceivable dataset?

Answer: a lot



1.0
0.8
0.6
0.4
0.2
0.0
Child's Income Rank
0.0
0.2
0.4
0.6
0.8
1.0
Parents' Income Rank
OLS Estimate in Hypothetical Data
OLS Estimate with Addition of Outlier

How about local sensitivity?

Global requirement is overkill

Local sensitivity:

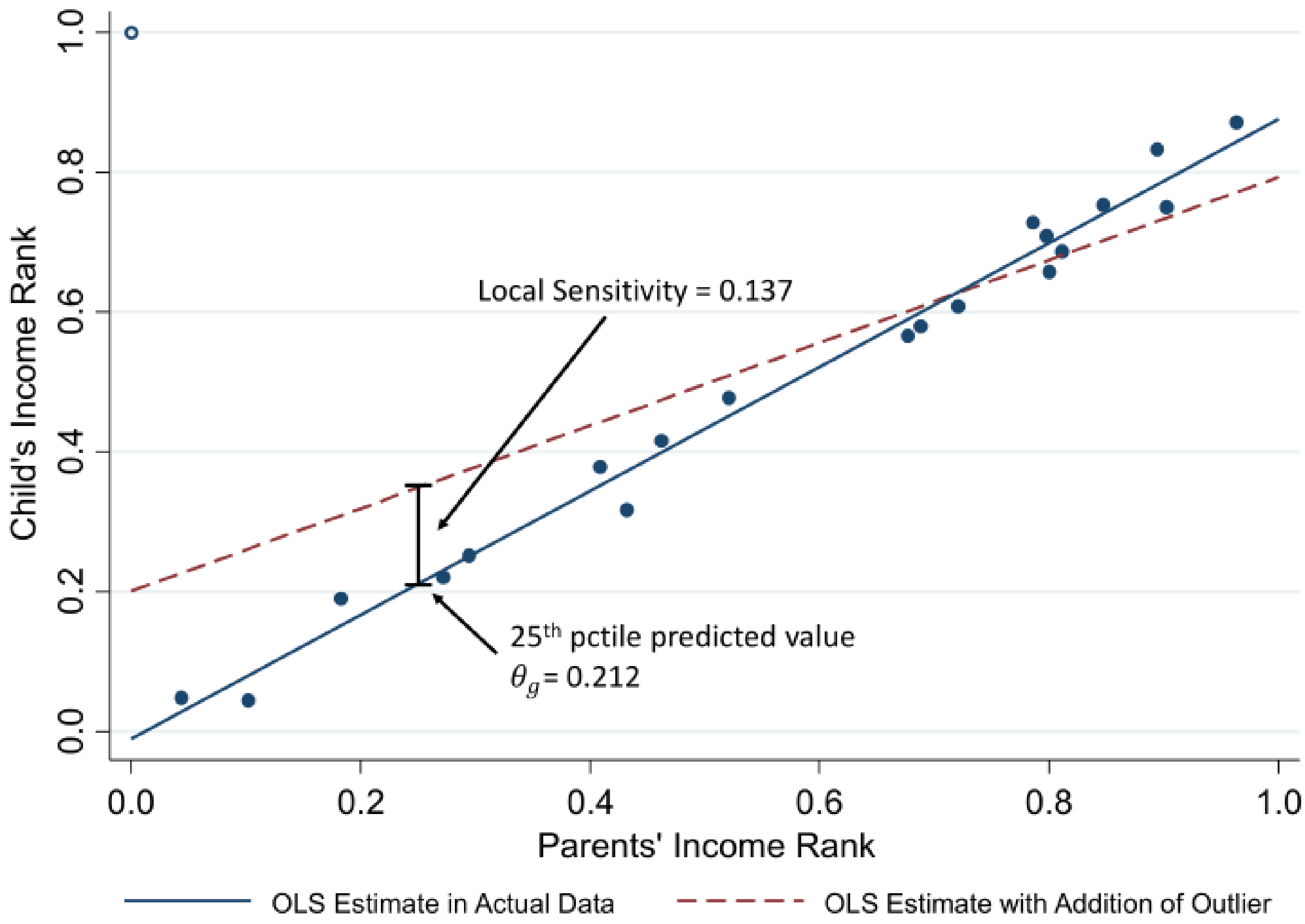
How much could conditional mean of child earnings rank change

if I added or removed any legal value

~~from any conceivable dataset~~

from the observed dataset, D_g

Answer: not as much



New Method

Publish

$$\widetilde{\theta_g}(rank) = q\left(D_g, rank\right) + \omega_g$$

Where

$$\omega_g \sim Lap(0, \frac{\Delta_{LS}^{g} q}{\epsilon})$$

$$\Delta_{LS}^{g} q = \max_{D' \in N\left(D_g\right)} \left| q\left(D_g, rank\right) - q(D', rank) \right|$$

Properties

- ~~Satisfies ϵ-differential privacy~~
- Parallel composition across groups means that total privacy loss is

$$\epsilon = \max_g(\epsilon_g)$$

Terry College of Business
UNIVERSITY OF GEORGIA

This won't work, either

Privacy-aware analysis requires knowledge of

$$var(\omega_g) = \frac{\Delta_{LS}^{g} q}{\epsilon}$$

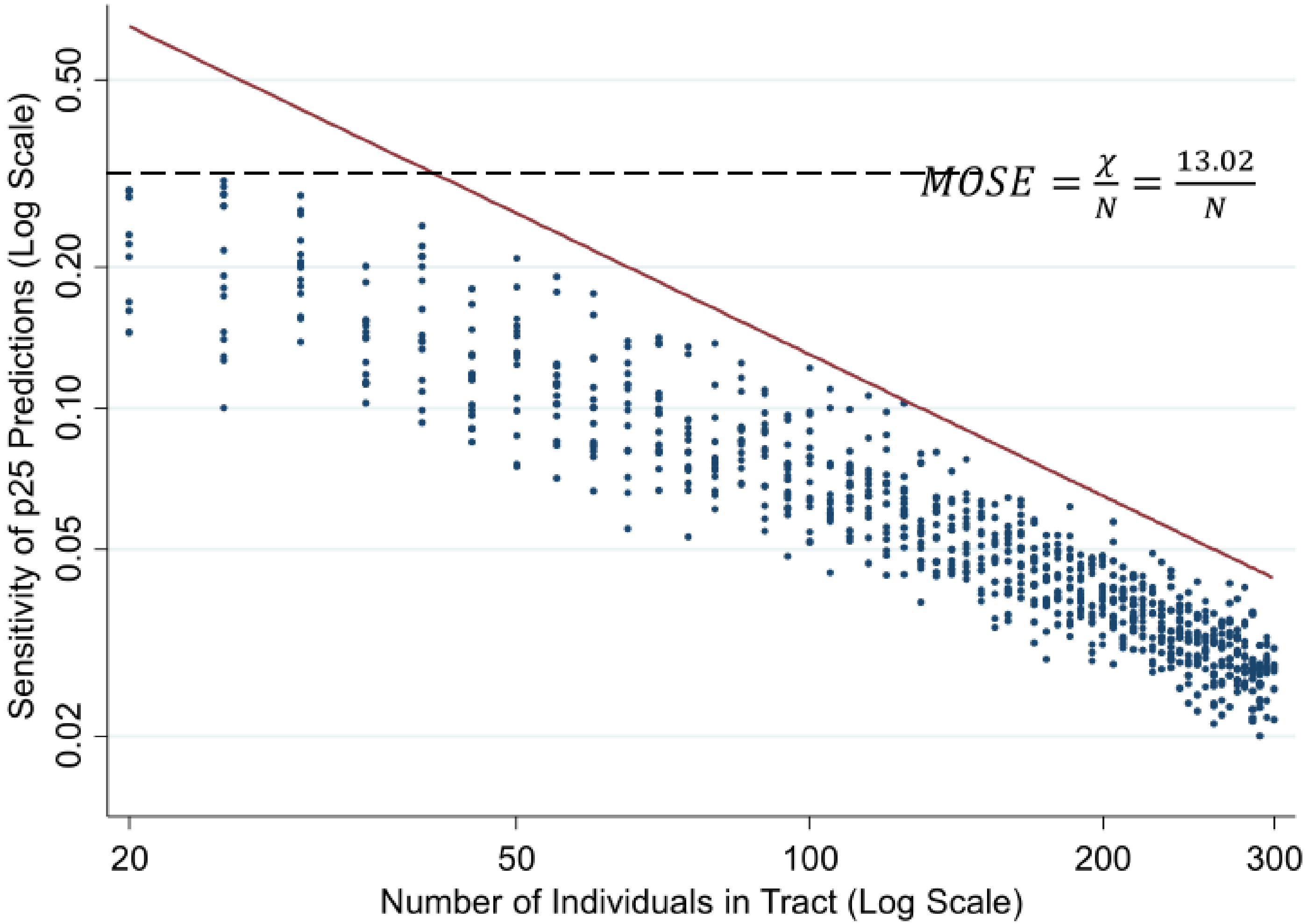
But

$$\Delta_{LS}^{g} q = \max_{D' \in N(D_g)} |\, q(D_g, rank) - q(D', rank)|$$

is a function of D_g

…it is also a population statistic
…which also has a privacy cost

MOSE profile



Goldilocks Solution: Maximum Observed Sensitivity

Using Laplace mechanism, publish

$$\widetilde{\theta_g}(rank) = q\left(D_g, rank\right) + \omega_g$$

Where

$$\omega_g \sim Lap\left(0, \frac{\Delta_{MOSE}(N_g)}{\epsilon}\right)$$

$$\Delta_{MOSE}(N_g) = \frac{\chi}{N_g}$$

for

$$\chi = \max_g [N_g \times \Delta_{LS}^g]$$

Properties

- NOT ϵ-differential privacy
- HOWEVER, conditional on χ
    - Satisfies DP guarantee
    - Parallel composition across groups

Implementation details

Local sensitivity further controlled through Winsorization

Scaling parameter χ estimated separated for state-gender-race groups

Set privacy loss parameter at

$$\epsilon = 8$$

Based on accuracy measure:

- Probability of correctly classifying tracts into top or bottom tail

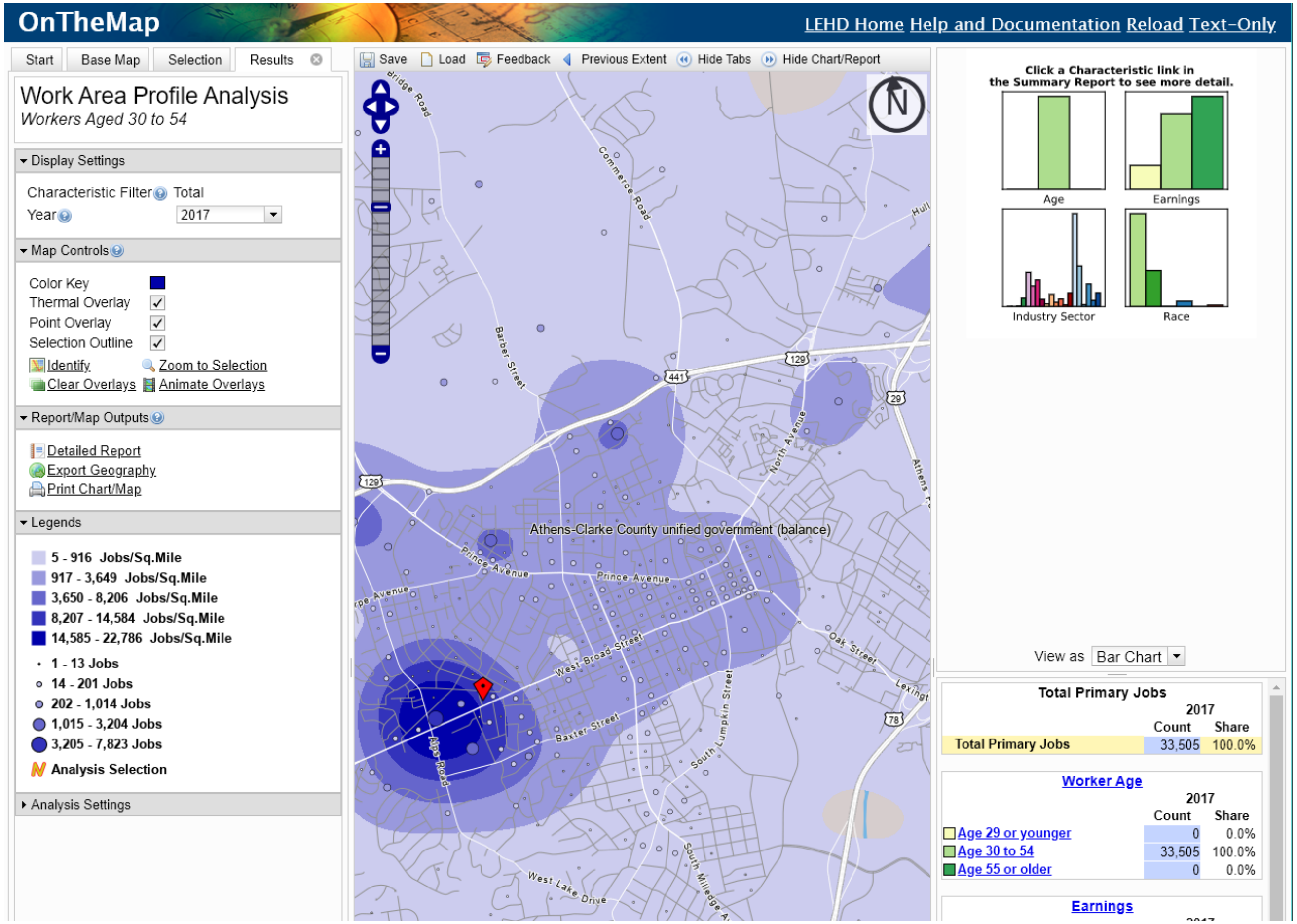
Takeaways

- MOSE “hack” solves issue of high global sensitivity
- Hard to imagine these data being published under conventional SDL
- Chetty-Friedman show cell suppression is far worse (see last talk)
- Latest research (Alabi et al. https://arxiv.org/abs/2007.05157) gives full differential privacy results for this class of problems

Issues

- Noise scales in data size.
  Cell counts are not always publishable
- Not formally private (unless Alabi et al. methods are used)

Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics



Samuel Haney
Duke University
shaney@cs.duke.edu

Ashwin Machanavajjhala
Duke University
ashwin@cs.duke.edu

John M. Abowd
U.S. Census Bureau, U.S.A.
john.maron.abowd@census.gov

Matthew Graham
U.S. Census Bureau, U.S.A.
matthew.graham@census.gov

Mark Kutzbach
U.S. Census Bureau, U.S.A.
mark.j.kutzbach@census.gov

Lars Vilhuber
Cornell University
lars.vilhuber@cornell.edu

Terry College of Business
UNIVERSITY OF GEORGIA

How to protect LODES?

LODES = LEHD Origin-Destination Employment Statistics

Tabulation of jobs

- Workplace characteristics
    - Location (block)
    - Industry
    - Ownership Type
- Worker Characteristics
    - Age
    - Race
    - Ethnicity
    - gender

Problem features

- Data are sparse
- Employment data are right-skewed
- Need to protect both WORKERS and EMPLOYERS
- What is the data, D?
- How to think about neighbors?

New Approach (Pufferfish; Kifer Machanavajjhala 2014)
- Decide what needs to be protected
- Define neighboring databases in terms of protected characteristics
- Devise provably private algorithms

What must be protected

1. No re-identification of individuals. Should not learn too much about whether an employee
   - is in the database or not
   - works for a specific type of employer
   - has particular demographic characteristics

2. No precise inference of establishment size
   - existence is not private (for employer businesses)
   - industry and location are not private
   - coarse size is not private, but exact size is

3. No precise inference of workforce composition
   - e.g., can't infer the share of female employees

Formalization: Protected from Whom?

The adversary knows

- Set of all employer establishments, E, and their public attributes
- Set of all workers, U
- Each worker, $w \in U$ has private attributes, $A_1, A_2, \ldots, A_k$ (including where they work and whether they are not in the data)

- Adversary's beliefs
  - π_w, a distribution over attributes
  - $\theta = \Pi_{w \in U} \pi_w$: beliefs over all workers
  - $\Theta = \{\theta\}$

DEFINITION 4.1 (EMPLOYEE PRIVACY REQUIREMENT). *For randomized algorithm A, if for some $\epsilon \in (0, \infty)$, and for every employee $w \in U$, for every adversary $\theta \in \Theta$, for every $a, b \in \mathcal{T}$ such that $Pr_\theta[w = a] > 0$ and $Pr_\theta[w = b] > 0$, and for every output $\omega \in range(A)$:*

$$\log\left(\frac{Pr_{\theta,A}[w = a|A(D) = \omega]}{Pr_{\theta,A}[w = b|A(D) = \omega]} \bigg/ \frac{Pr_\theta[w = a]}{Pr_\theta[w = b]}\right) \leq \epsilon \qquad (3)$$

Then the algorithm A protects employees against informed attackers at privacy-loss level ϵ.

DEFINITION 4.2 (EMPLOYER SIZE REQUIREMENT). *Let e be any establishment in $\mathcal{E}$. A randomized algorithm A protects establishment size against an informed attacker at privacy level (ϵ, α) if, for every informed attacker $\theta \in \Theta$, for every pair of numbers x, y, and for every output of the algorithm $\omega \in range(A)$,*

$$\left| \log \left(\frac{Pr_{\theta,A}[|e| = x | A(D) = \omega]}{Pr_{\theta,A}[|e| = y | A(D) = \omega]} \Big/ \frac{Pr_{\theta}[|e| = x]}{Pr_{\theta}[|e| = y]} \right) \right| \leq \epsilon \quad (4)$$

whenever $x \leq y \leq \lceil (1+\alpha)x \rceil$ and $Pr_{\theta}[w = x], Pr_{\theta}[w = y] > 0$.

Differential Privacy

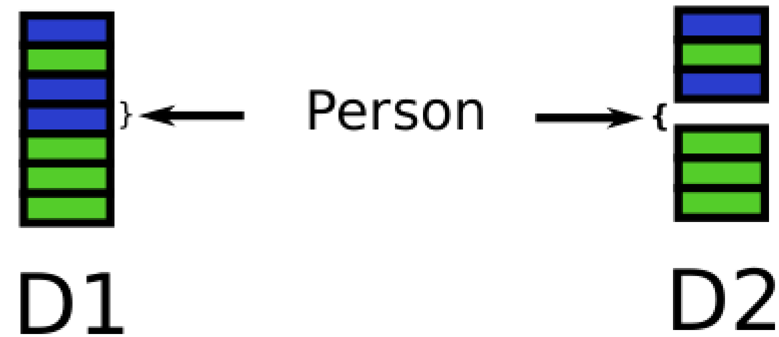
Need a concept of neighboring databases

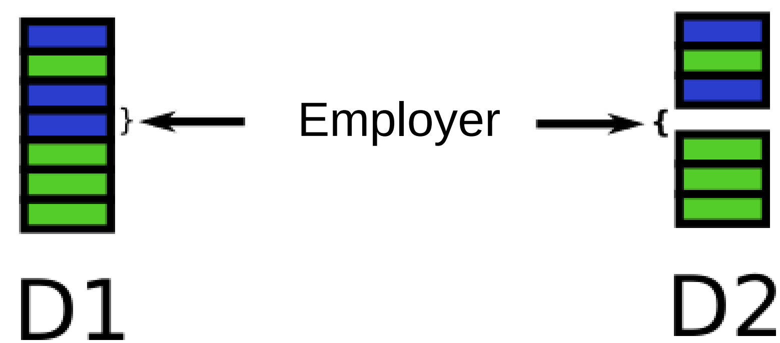
Option 1: Neighbors add or remove a single worker

- Queries are counts
- Laplace mechanism with sensitivity 1
- FAILS employer size requirement

Option 2: Neighbors add or remove a single employer

- Queries include sums of workers
- Can satisfy all requirements
- Quality is atrocious





Goldilocks Solution

Neighbor Definition: Strong α-Neighbors

- Two databases, D and D' are *Strong α-Neighbors* if they
    - Differ in the employment attribute of exactly one record, e
    - Let x be the number of workers at e in D
    - Let x' be the number of workers at e in D'
    - $x \leq x' \leq \max((1+\alpha)x, x+1)$

- Similar to original LEHD specification for Quarterly Workforce Indicators

New privacy concept

DEFINITION 7.2 (((α, ϵ)-ER-EE PRIVACY). *A randomized algorithm $\mathcal{M}$ is said to satisfy (α, ϵ)-ER-EE Privacy, if for every set of outputs $S \subseteq range(M)$, and every pair of strong α-Neighbors D and D', we have*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S]$$

- *Sufficient for* worker and establishment size requirements
- Satisfies sequential and parallel composition in ϵ

Application

Global sensitivity can still be high

Key query: Total employment

Let $q(D)$ be such a counting query.

Sensitivity,

$$\Delta_q = \max|q(D) - q(D')| = \max_{e \in E}(\alpha x_e)$$

(with D and D' strong α-neighbors)

Application

Sensitivity,

$$\Delta_q = \max|q(D) - q(D')| = \max_{e \in E}(\alpha x_e)$$

Essentially unbounded.

However,

$$\Delta_{log q} = \max|\log q(D) - \log q(D')| = 1 + \alpha$$

Algorithm 1 Log-Laplace Mechanism

Input: : n : the sum of employment counts for a set of cells, α, ϵ: privacy parameters

Output: : $\tilde{n}$: the noisy employment count

Set $\gamma \leftarrow 1/\alpha$

$\ell \leftarrow \ln(n + \gamma)$

Sample $\eta \sim Laplace(2\ln(1 + \alpha)/\epsilon)$

$\tilde{n} \leftarrow e^{\ell+\eta} - \gamma$

Result:

- Log-Laplace Mechanism satisfies strong (α, ϵ)-privacy for employer attributes
- Biased

Other Mechanisms

Smooth Sensitivity: Complementary approach to the "Goldilocks" problem

Idea: Derive function, $S(x)$, such that

$$S(D) \geq LS_q(D)$$

While

$$S(D) \leq e^{\alpha} S(D')$$

For all D' neighbors of D

tl;dr, can add noise proportional to $\max_e(\alpha x_e)$ over all employers, e in D

Algorithm 2: Smooth Gamma

- Satisfies strong (α, ϵ)-EE-ER privacy
- Unbiased

Algorithm 3: Smooth Laplace

- Satisfies strong $(\alpha, \epsilon, \delta)$-EE-ER privacy [approximate]
- unbiased

Data

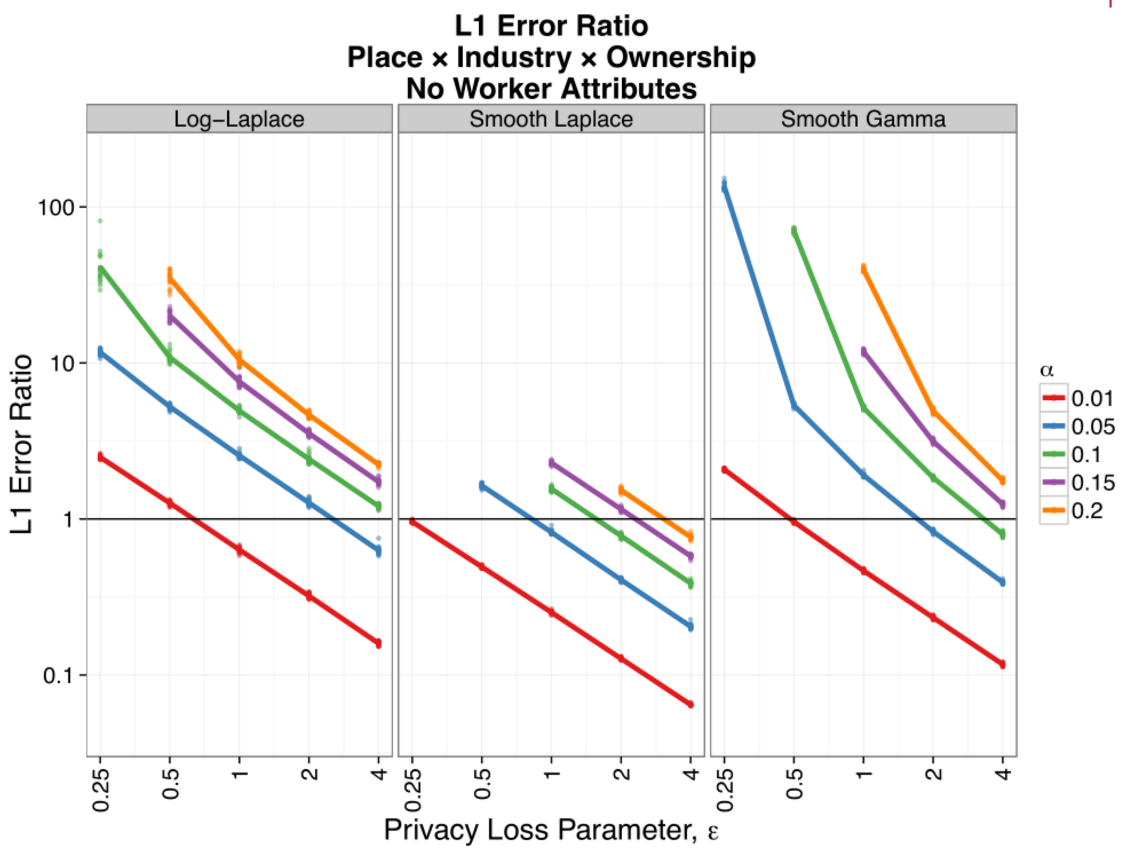
- 11 mill. jobs;
- 527K employers

Queries: all margins of

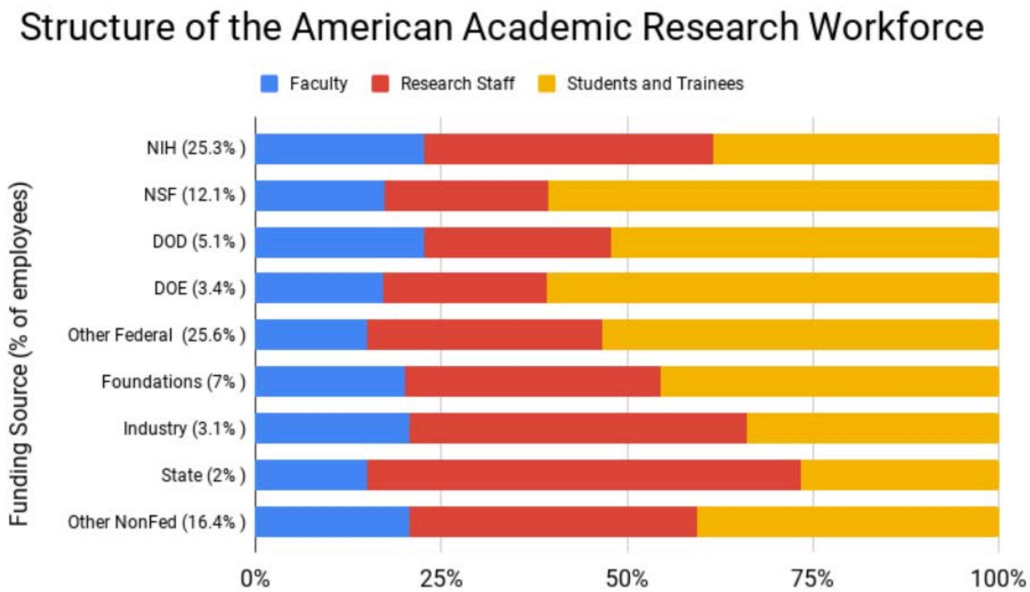
- Place = city/town
- NAICS Sector
- Ownership

Compare L1 Error using

- Orig. system
- Proposed systems



UMETRICS Employee Profile Reports



Goal:
Track employment and earnings outcomes of grant-funded employees

Method:
Link UMETRICS data to W2, LEHD, BR

IRIS INSTITUTE FOR RESEARCH ON INNOVATION & SCIENCE

United States Census Bureau

Desired outputs

Cells

- Title (e.g. faculty, grad student)
- Sector of employment [3 categories]
- Years since leaving [up to 10]

Statistics

- Employment
- Average Wage

…one table per University!

Privacy requirements

- Protect university employees against re-identification on the basis of
  - Inclusion in the data
  - All attributes of employment history
- Neighboring databases add or remove a single employee and their entire employment history
- Simpler if we were just protecting single jobs…

Methods

- Laplace mechanism for employment counts (sensitivity 1)
- Modified *MOSE* for average earnings (Chetty-Friedman 2019)
- MOS at job title-by-sector level (9 values)
- Upper bound MOSE

Accuracy Requirement

- Target a threshold for

$$APD_c = \frac{|true_c - noisy_c|}{true_c}$$

Privacy Analysis

Composition possibilities

- Each worker only appears in one (job title)-by-(sector) pair [parallel]
- Each worker can appear in multiple years [serial]
- Each record is used to compute both employment and earnings [serial]

Define

- $\epsilon_{emp,t}^{s}$ (for employment queries t years out at university s)
- $\epsilon_{earn,t}^{s}$ (for mean earnings queries t years out at university s)

The total privacy loss associated with Employee Report for University s:

$$\epsilon^{s} = \sum_{t=1..T^{s}} (\epsilon_{emp,t}^{s} + \epsilon_{earn,t}^{s})$$

Other Examples

Post-Secondary Employment Outcomes:
https://lehd.ces.census.gov/data/pseo_experimental.html

Veterans Employment Outcomes:
https://lehd.ces.census.gov/applications/veo/service

Technical documentation on privacy protection: Foote et al. Releasing Earnings Distributions using Differential Privacy, Journal of Privacy and Confidentiality, 2019,
DOI: https://doi.org/10.29012/jpc.722.

Thank You!

Ian M. Schmutte
https://ianschmutte.org
Schmutte@uga.edu