CS208: Applied Privacy for Data Science
Implementing (Centralized) Differential Privacy: One-Shot Releases

James Honaker & Salil Vadhan

School of Engineering & Applied Sciences
Harvard University

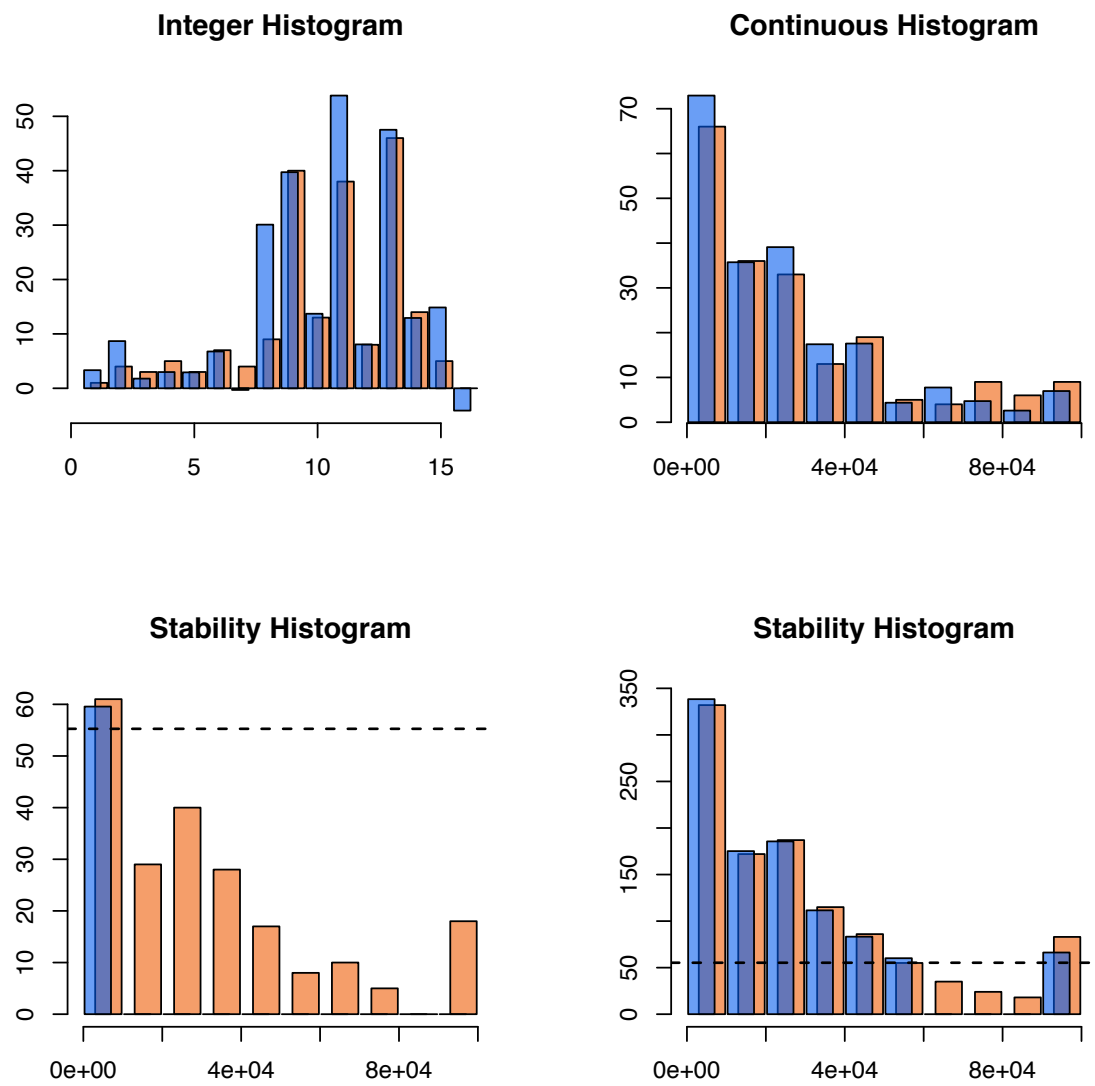
March 8, 2019

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

Review: Doing Better than Composition

- Not all sequences of k queries require error growing as $\sqrt{k}$.

- Example: histograms
  - Let $B_1, \dots, B_k \subseteq \mathcal{X}$ be disjoint bins.
  - Define $q_j : \mathcal{X}^n \to \{0,1\}$ by $q_j(x) = \#\{i : x_i \in B_j\}$.
  - Define $M(x) = (q_1(x) + Z_1, q_2(x) + Z_2, \dots, q_k(x) + Z_k)$ where the Z_j's are independent $\mathrm{Lap}(2/\varepsilon)$ or $\mathrm{Geo}(2/\varepsilon)$.
  - Then M is ε-DP.


Integer Histogram
0
10
20
30
40
50
0
5
10
15
Continuous Histogram
0
10
30
50
70
0e+00
4e+04
8e+04
Stability Histogram
0
10
20
30
40
50
60
0e+00
4e+04
8e+04
Stability Histogram
0
50
150
250
350
0e+00
4e+04
8e+04

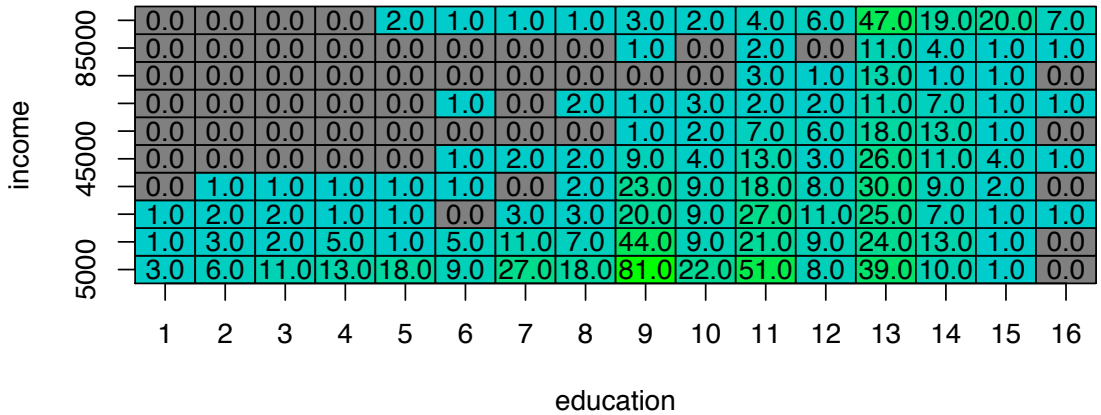
Synthetic Data via DP Histograms

- Use singleton bins $B_y = \{y\}$ for each $y \in \mathcal{Y}$.
- Construct a DP histogram $\left(a_1, \ldots, a_{|\mathcal{X}|}\right) \leftarrow M_{\text{hist}}(x)$, where $a_y \approx \#\{i : x_i = y\}$.
- Output synthetic dataset $\hat{x}$ with a_y copies of each element y.

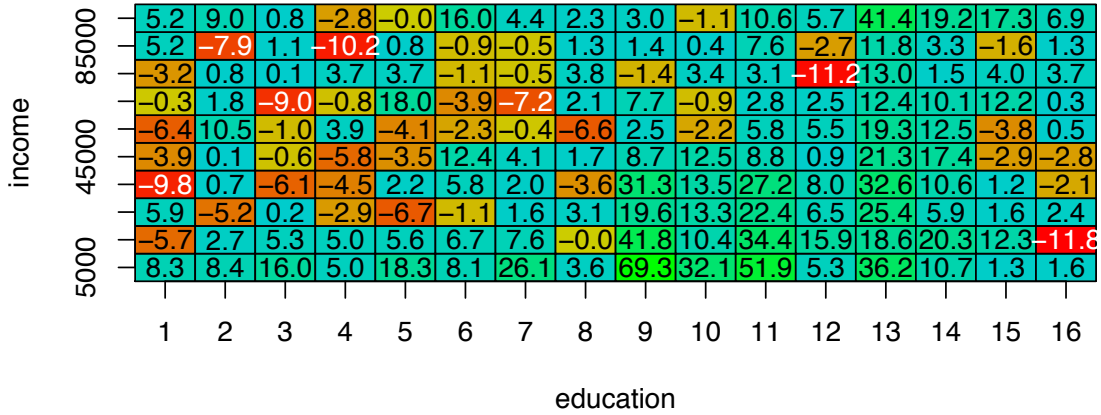
Difficulties?

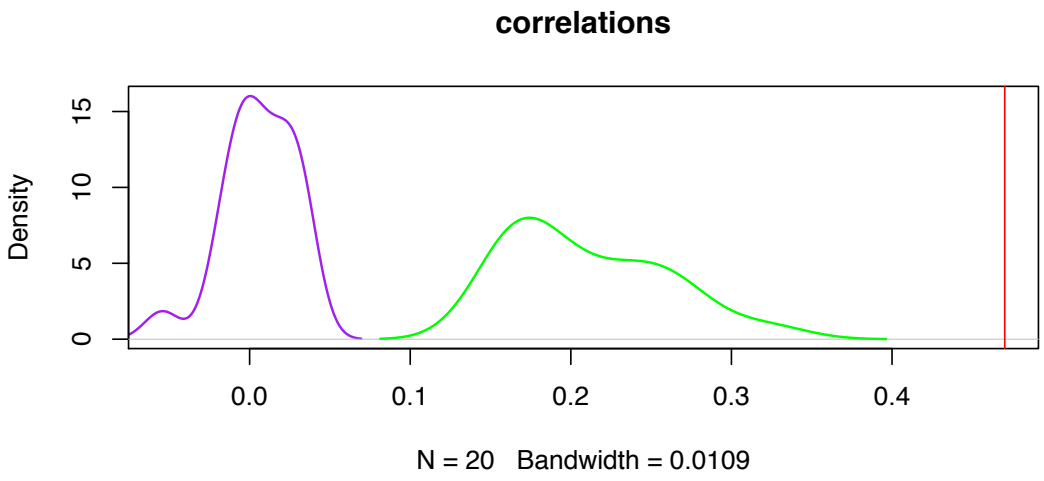
- a_y's may not be nonnegative integers.
  - Soln 1: use Geometric Mechanism and clamp at 0.
  - Soln 2: use Exponential Mechanism with range $\{0, \ldots, n\}$.
- Poor utility & efficiency when $\mathcal{X}$ is large.

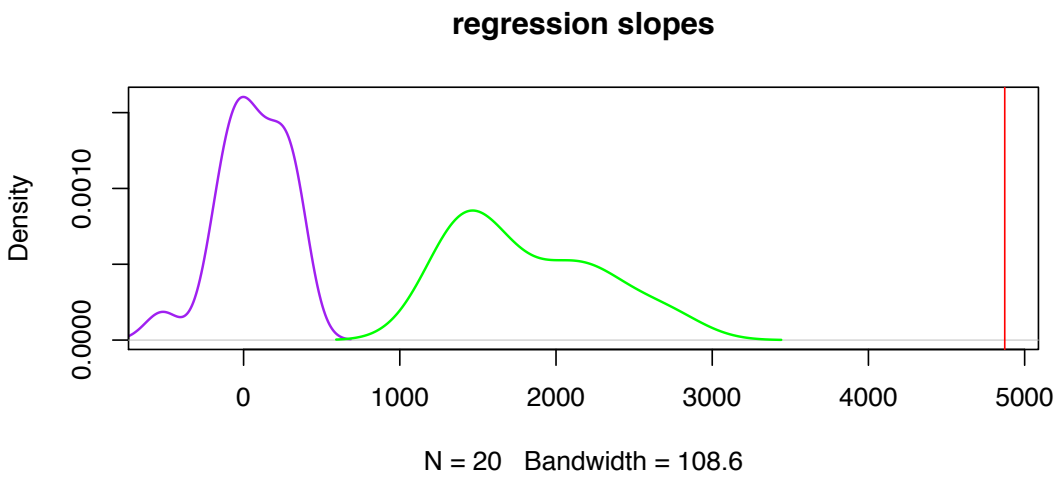
Sensitive Data



DP Release




correlations
Density
0
5
10
15
0.0
0.1
0.2
0.3
0.4
N = 20 Bandwidth = 0.0109


regression slopes
Density
0.0000
0.0010
0
1000
2000
3000
4000
5000
N = 20 Bandwidth = 108.6

A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Samples

Raj Chetty, Harvard University and NBER
John N. Friedman, Brown University and NBER

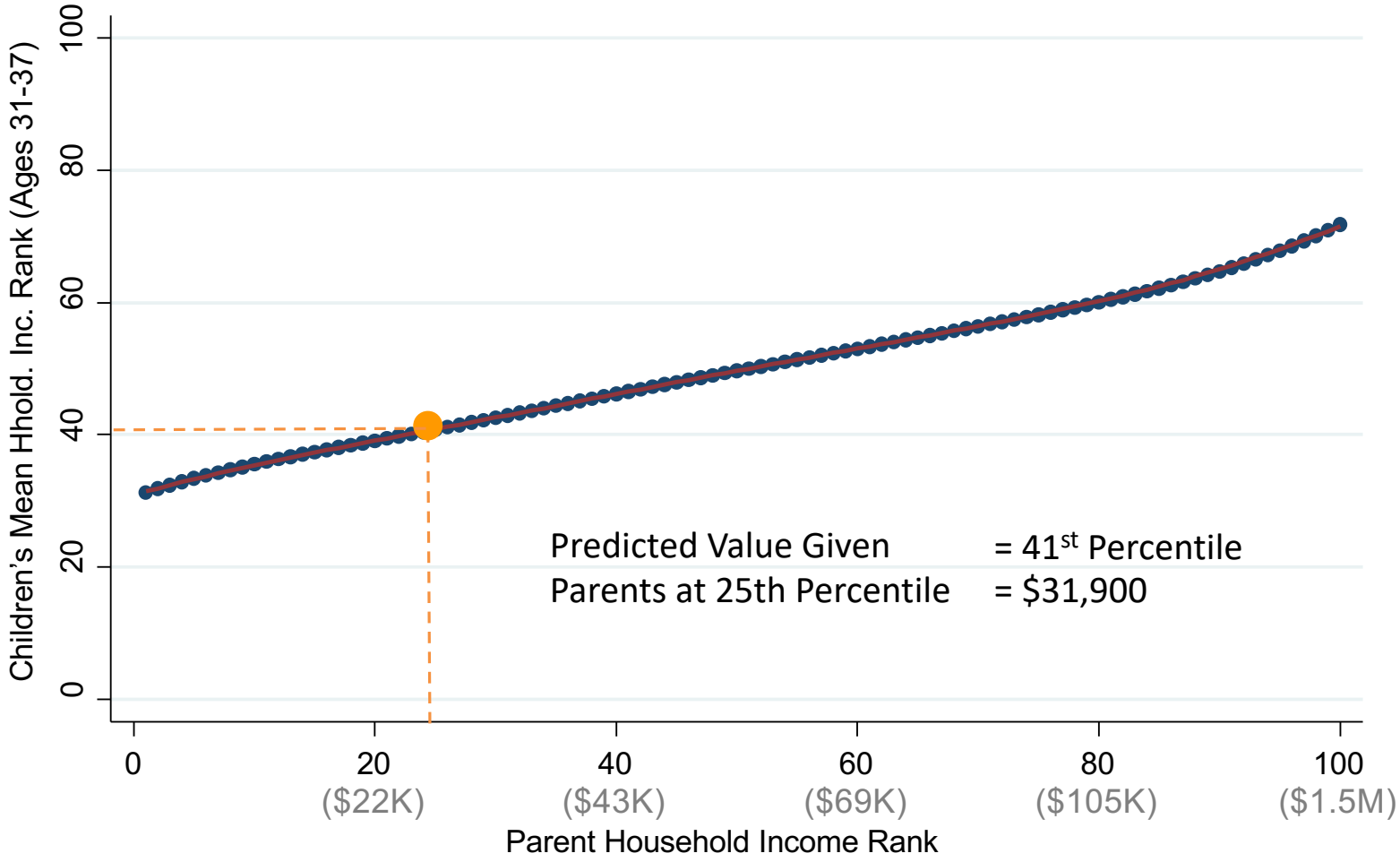
March 2019

Publishing Statistics Based on Small Cells

- Social scientists increasingly use confidential data to publish statistics based on cells with a small number of observations
  - Causal effects of schools or hospitals [e.g., Angrist et al. 2013, Hull 2018]
  - Local area statistics on health outcomes or income mobility [e.g., Cooper et al. 2015, Chetty et al. 2018]

Intergenerational Mobility in the United States

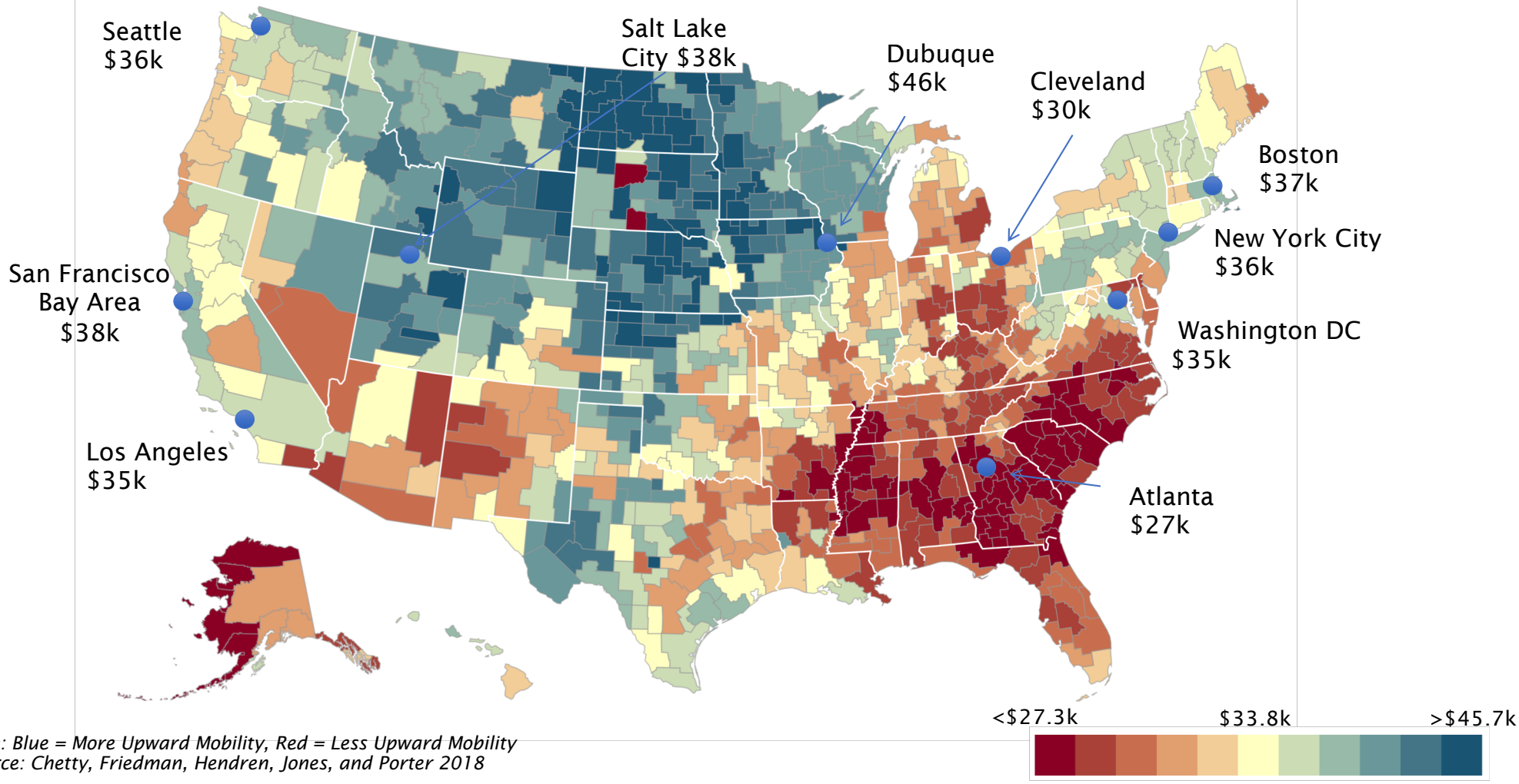
Mean Child Household Income Rank vs. Parent Household Income Rank



Source: Chetty, Friedman, Hendren, Jones, Porter (2018)

Geography of Upward Mobility in the United States

Average Income at Age 35 for Children whose Parents Earned $25,000 (25th percentile)



Note: Blue = More Upward Mobility, Red = Less Upward Mobility
Source: Chetty, Friedman, Hendren, Jones, and Porter 2018

Controlling Privacy Loss

- Problem with releasing such estimates at smaller geographies (e.g., Census tract): risk of disclosing an individual's data

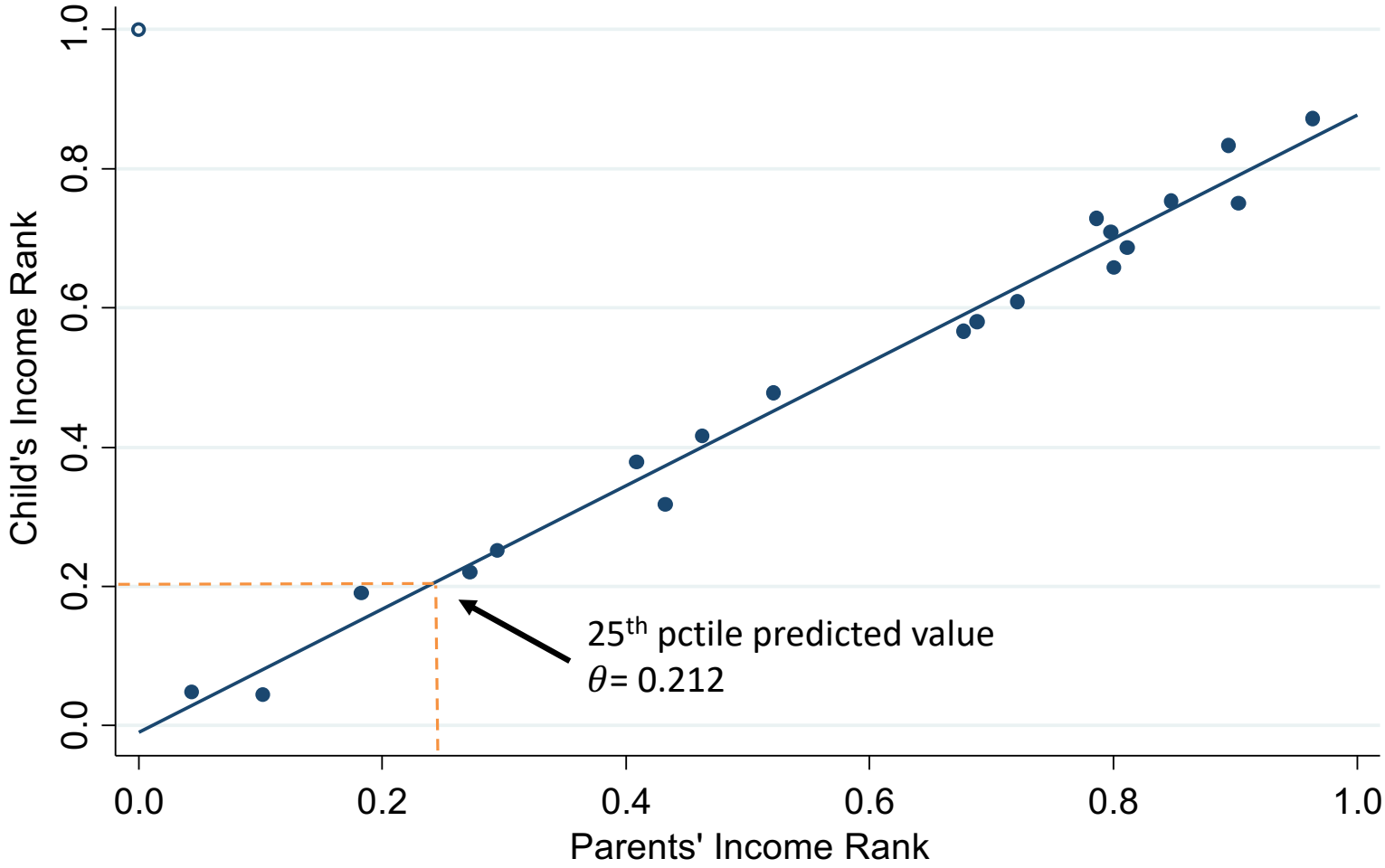
- Literature on differential privacy has developed practical methods to protect privacy for simple statistics such as means and counts [Dwork 2006, Dwork et al. 2006]

- But methods for disclosing more complex estimates, e.g. regression or quasi-experimental estimates, are not feasible for many social science applications [Dwork and Lei 2009, Smith 2011, Kifer et al. 2012]

This Paper: A Practical Method to Reduce Privacy Loss

- We develop and implement a simple method of controlling privacy loss when disclosing arbitrarily complex statistics in small samples
  - The “Maximum Observed Sensitivity” (MOS) algorithm
- Method outperforms widely used methods such as cell suppression both in terms of privacy loss and statistical accuracy
  - Does not offer a formal guarantee of privacy, but potential risks occur only at more aggregated levels (e.g., the state level)

Example Regression from One Small Cell

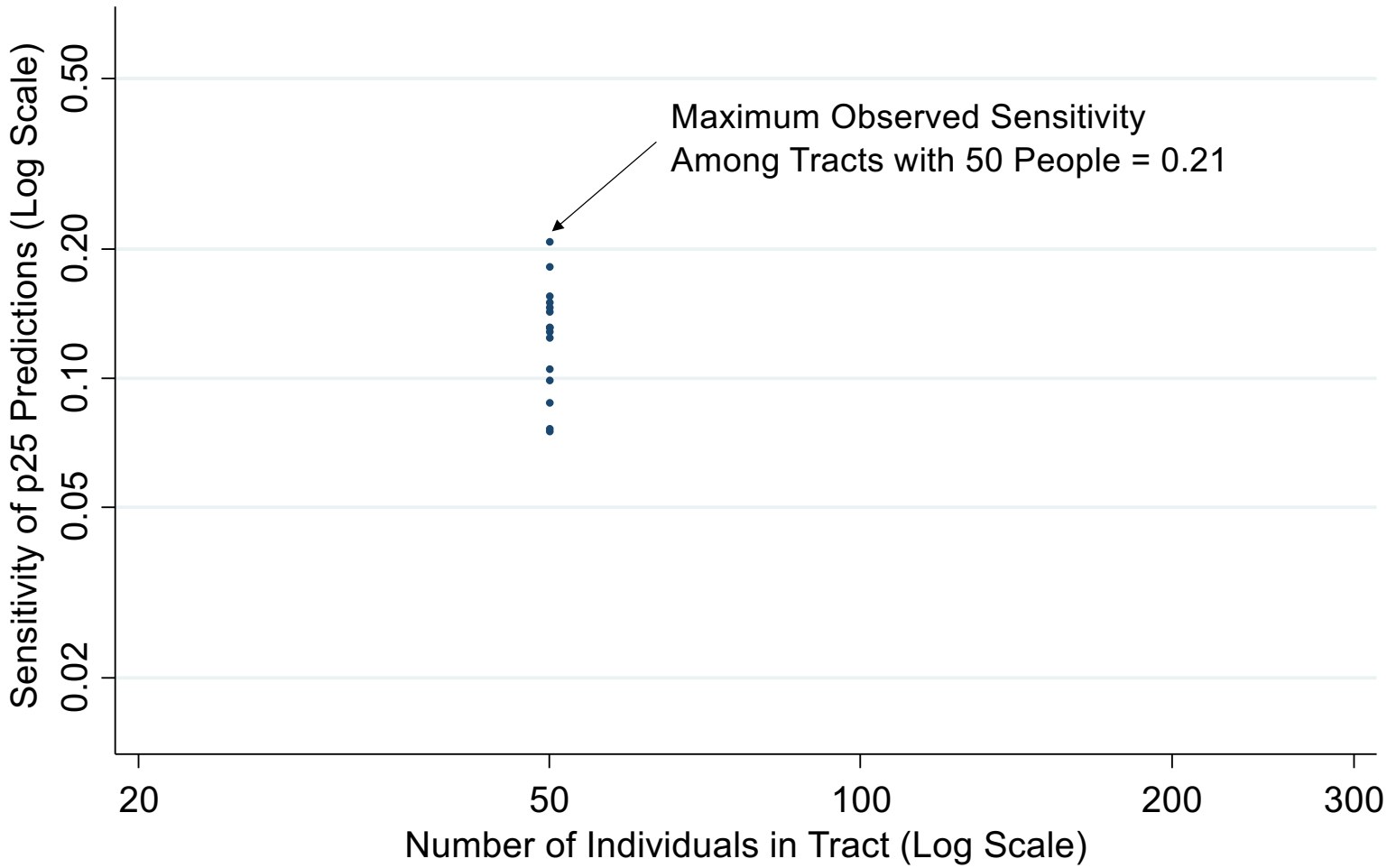


Source: Authors' simulations.

Maximum Observed Sensitivity

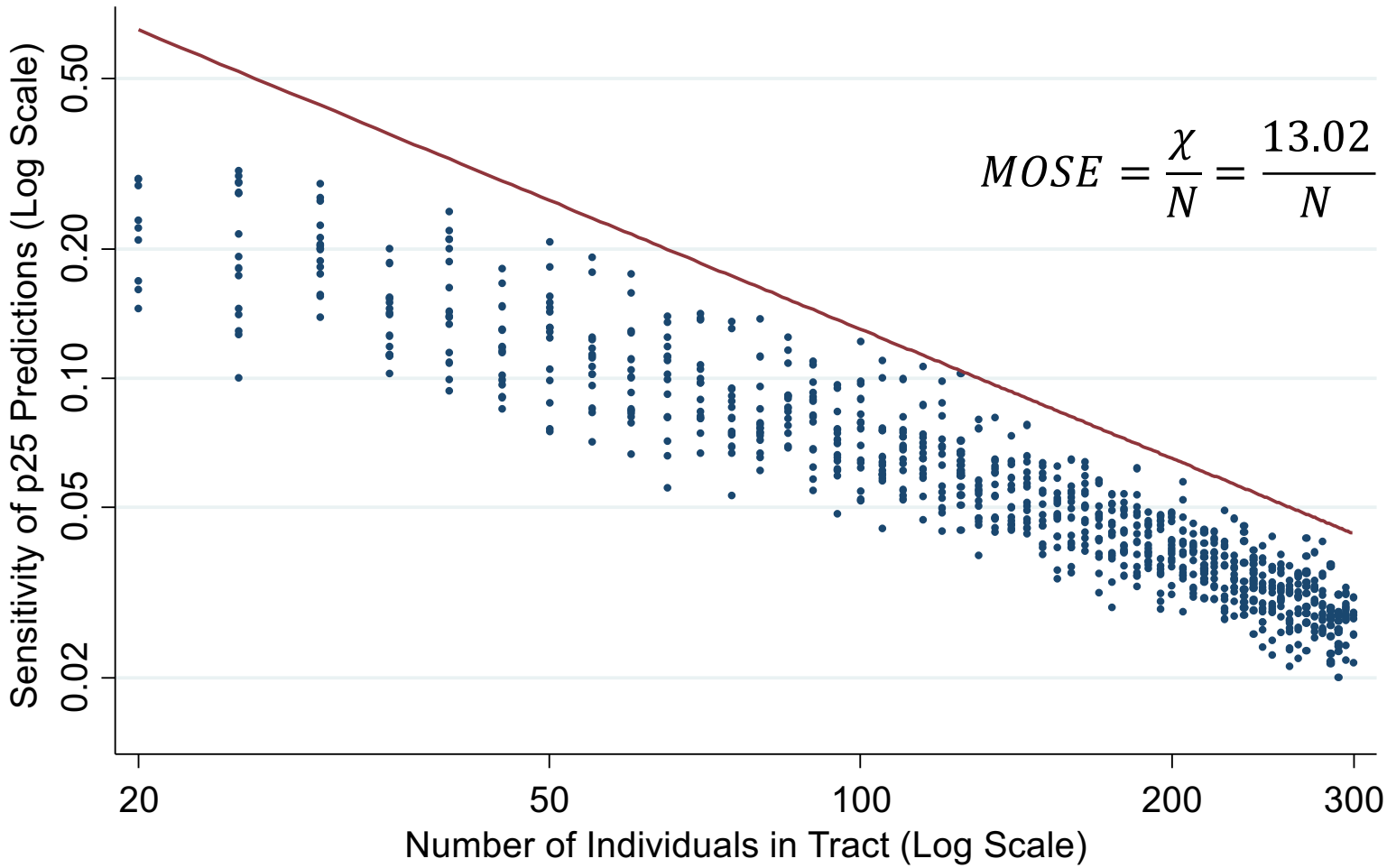
- Our method: use the maximum observed local sensitivity across all cells in the data
  - In geography of opportunity application, calculate local sensitivity in every tract
  - Then use the maximum observed sensitivity (MOS) across all tracts within a given state as the sensitivity parameter for every tract in that state
- Analogous to Empirical Bayes approach of using actual data to construct prior on possible realizations rather than considering all possible priors

Maximum Observed Sensitivity Envelope



Source: Authors' simulations.

Computing Maximum Observed Sensitivity



Source: Authors' simulations.

Producing Noise-Infused Estimates for Public Release

- Use max observed sensitivity χ, tract counts, and exogenously specified privacy parameter ε to add noise and construct public estimates:

$$\tilde{\theta}_g = \theta_g + L\left(0, \frac{\chi}{\varepsilon N_g}\right) \qquad \tilde{N}_g = N_g + L\left(0, \frac{1}{\varepsilon}\right)$$

- This method not "provably private," but it reduces privacy risk to release of the single max observed sensitivity parameter (χ)
  - Privacy loss from release of regression statistics themselves is controlled below risk tolerance threshold ε

- Critically, χ can be computed at a sufficiently aggregated level that disclosure risks are considered minimal ex-ante
  - Ex: Census Bureau currently does not consider most statistics released at state or

Comparison to Alternative Methods: Statistical Bias

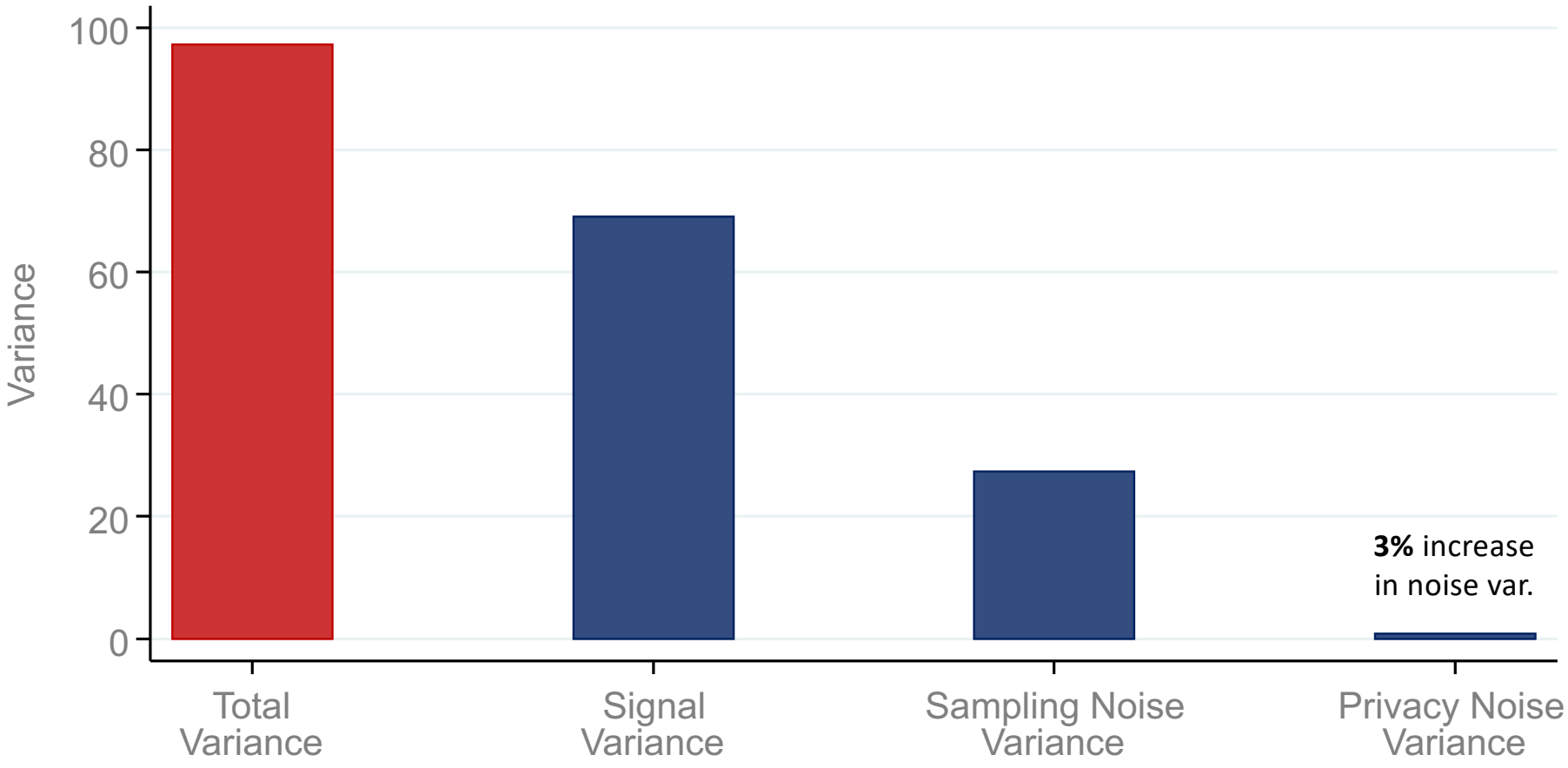
- In noise-infused data, regression provides an unbiased estimate of the (strong positive) relationship between teenage-birth rates for black women and single-parent share.
  - More generally, can adjust for noise using the “signal correlation”
- In contrast, count-based suppression generates bias that eliminates the result, since induces correlated measurement error from two sources:
  - Suppressing cells with few teenage births mechanically omits tracts with low teenage birth rates, which are concentrated in areas with few single parents.
  - Areas with a smaller black population (i.e., less diversity) have fewer teenage births and fewer areas with few single parents
- Identifying and correcting for these biases would be very difficult if one only had access to the post-suppression data

Comparison to Alternative Methods: Statistical Precision

- Primary concern of end users: will estimates be too noisy to be useful?
  - In Atlas, noise added to protect privacy was similar to inherent noise due to sampling error → estimates remain highly accurate
  - E.g., added privacy noise reduces reliability (i.e., fraction of total variance that is signal) only from 71.8% to 71.0%

Variance Decomposition for Tract-Level Estimates

Teenage Birth Rate For Black Women With Parents at 25th Percentile



Source: Chetty, Friedman, Hendren, Jones, Porter (2018)

Conclusion

- Main lesson: tools from differential privacy literature can be adapted to control privacy loss while improving statistical inference
  - Opportunity Atlas has been used by half a million people, by housing authorities to help families move to better neighborhoods, and in downstream research
    [Creating Moves to Opportunity Project; Morris et al. 2018]
- The MOS algorithm can be practically applied to any empirical estimate
  - Example: difference-in-differences or regression discontinuity
  - Even when there is only one quasi-experiment, pretend that a similar change occurred in other cells of the data and compute MOS across all cells

Future Work

- Two areas for further work that could increase use of differential privacy methods in social science:

  1. Developing formal metrics for risk of privacy loss for algorithms in which a single statistic (e.g., sensitivity) is released at a broader level of aggregation

  2. Developing techniques that can be applied to many estimators without requiring users to develop new algorithms for each application