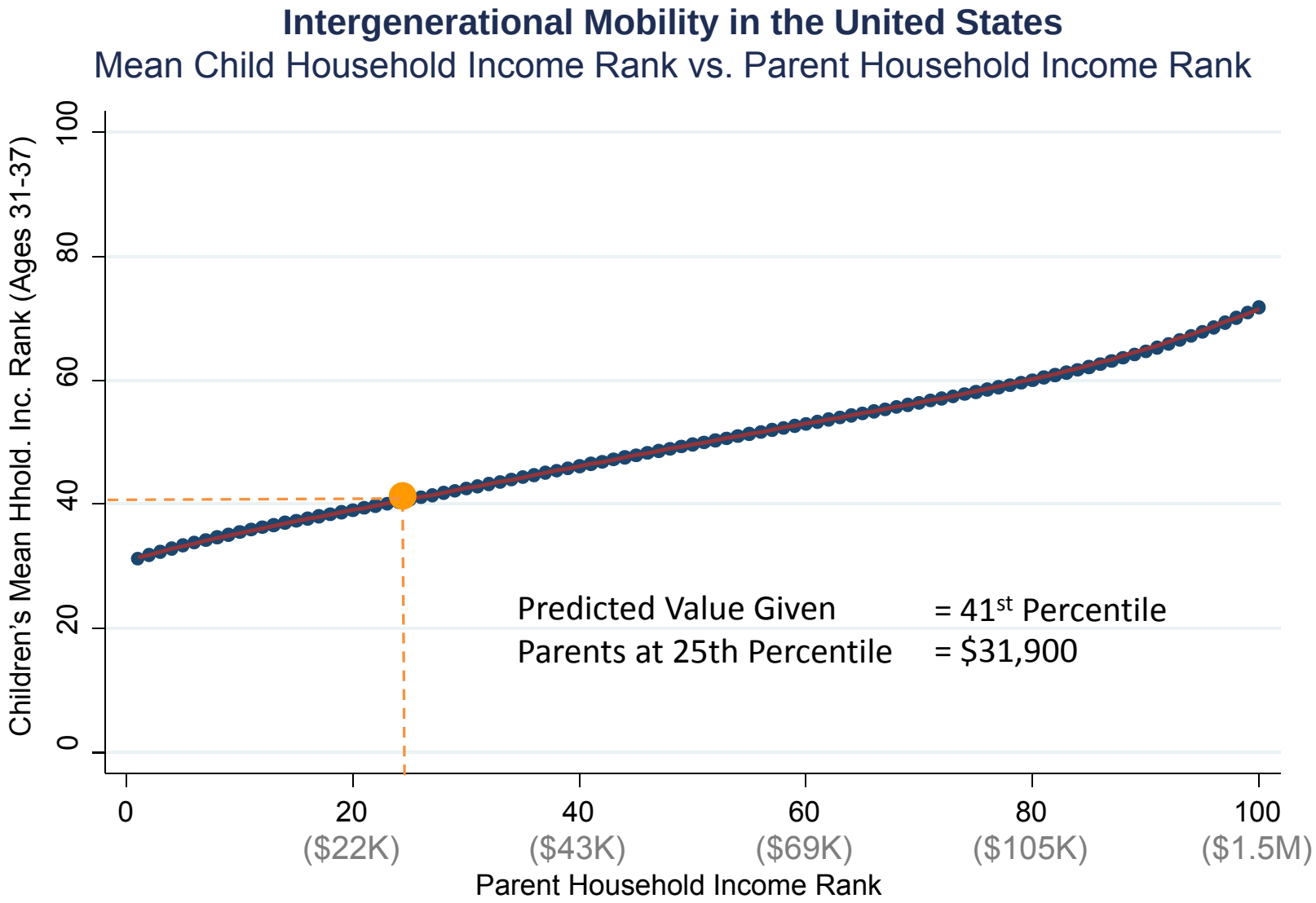
# A Practical Method to Reduce Privacy Loss when Disclosing Statistics Based on Small Samples

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

CNSTAT Workshop, June 2019

## Publishing Statistics Based on Small Cells

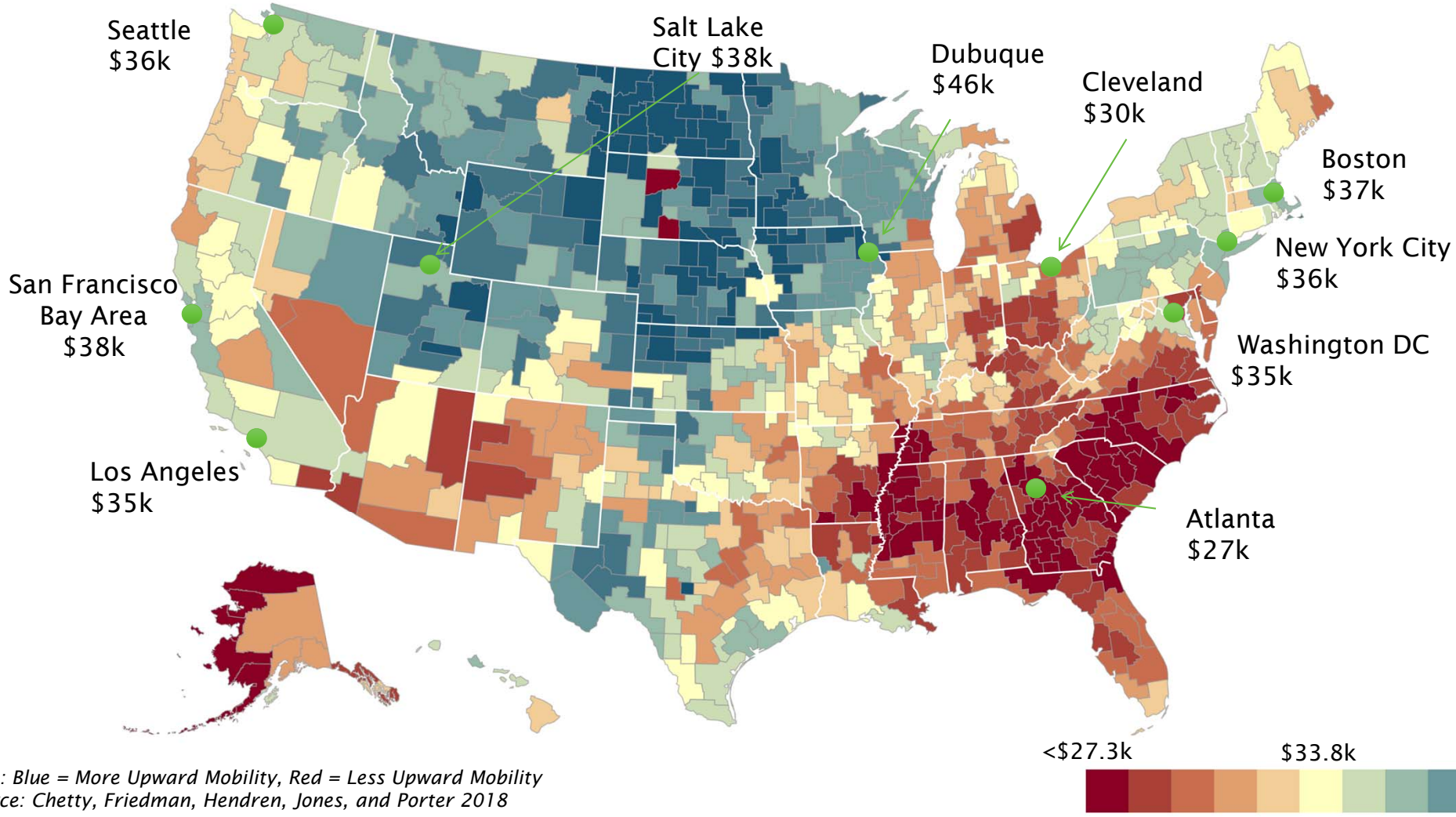
- Social scientists increasingly use confidential data to publish statistics based on cells with a small number of observations
  - Causal effects of schools or hospitals [e.g., Angrist et al. 2013, Hull 2018]
  - Local area statistics on health outcomes or income mobility [e.g., Cooper et al. 2015,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




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)

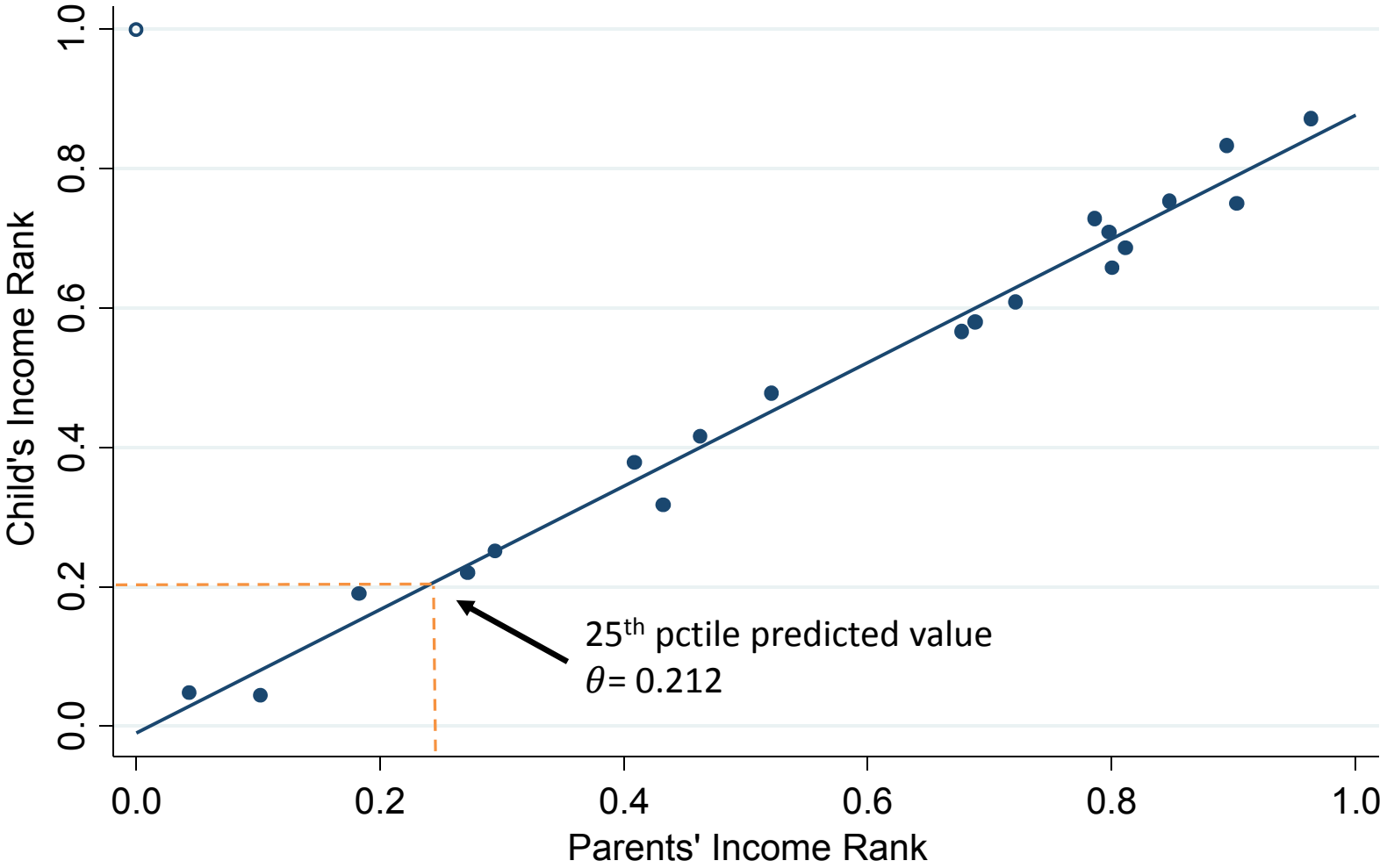
1. Method: Maximum Observed Sensitivity
2. Application: Opportunity Atlas
3. Comparison with Traditional Methods

1. **Method: Maximum Observed Sensitivity**
2. **Application: Opportunity Atlas**
3. **Comparison with Traditional Methods**

# Privacy Protection via Noise Infusion

- Goal: release predicted values from univariate regressions ($\theta$) in small cells

## Example Regression from One Small Cell




Source: Authors' simulations.

# Privacy Protection via Noise Infusion

- Goal: release predicted values from univariate regressions ($\theta$) in small cells
- Follow LaPlace mechanism: add i.i.d. random noise $\omega$ to these statistics:

$$\tilde{\theta} = \theta + \omega$$

- When $\omega \sim L\left(0, \frac{\Delta\theta}{\varepsilon}\right)$, can bound the privacy loss (measured as the log-likelihood ratio):

$$\log \frac{f\left(\tilde{\theta} - \theta(D_1)\right)}{f\left(\tilde{\theta} - \theta(D_2)\right)} \leq \varepsilon$$

- Intuitively, this ratio measures whether a published statistic is more likely given dataset $D_1$ vs $D_2$, for two adjacent datasets (i.e., that differ by just one element)
  - The more noise that is added, the closer to 0 this log-likelihood ratio becomes, decreasing the ability to distinguish between the underlying datasets from the statistic that is released

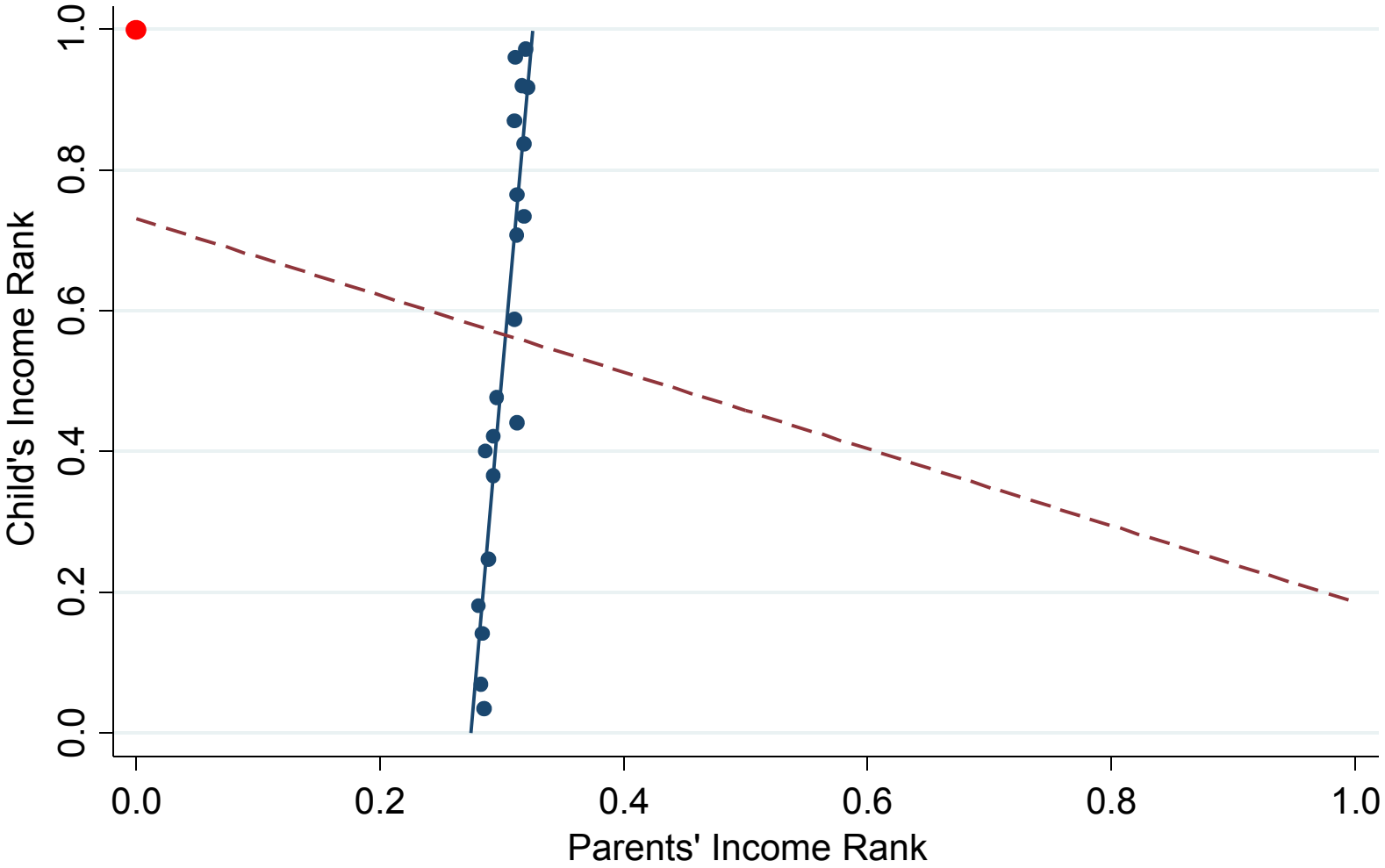
## Calculating Sensitivity

- Key remaining question: how do we compute sensitivity $\Delta\theta$?

- Standard approaches in differential privacy literature do not function well in practice in our setting:

  - Measure global (or smooth) sensitivity: Typically infinite in a regression setting

## Global Sensitivity: How Much Can A Single Observation Change the Estimate in Any Dataset?




Source: Authors' simulations.

## Calculating Sensitivity

- Key remaining question: how do we compute sensitivity $\Delta\theta$?

- Standard approaches in differential privacy literature do not function well in practice in our setting:

  - Measure global (or smooth) sensitivity: Typically infinite in a regression setting

  - Robust regression techniques: Poor downstream properties (e.g., no iterated expectations with medians, cannot re-aggregate the data)

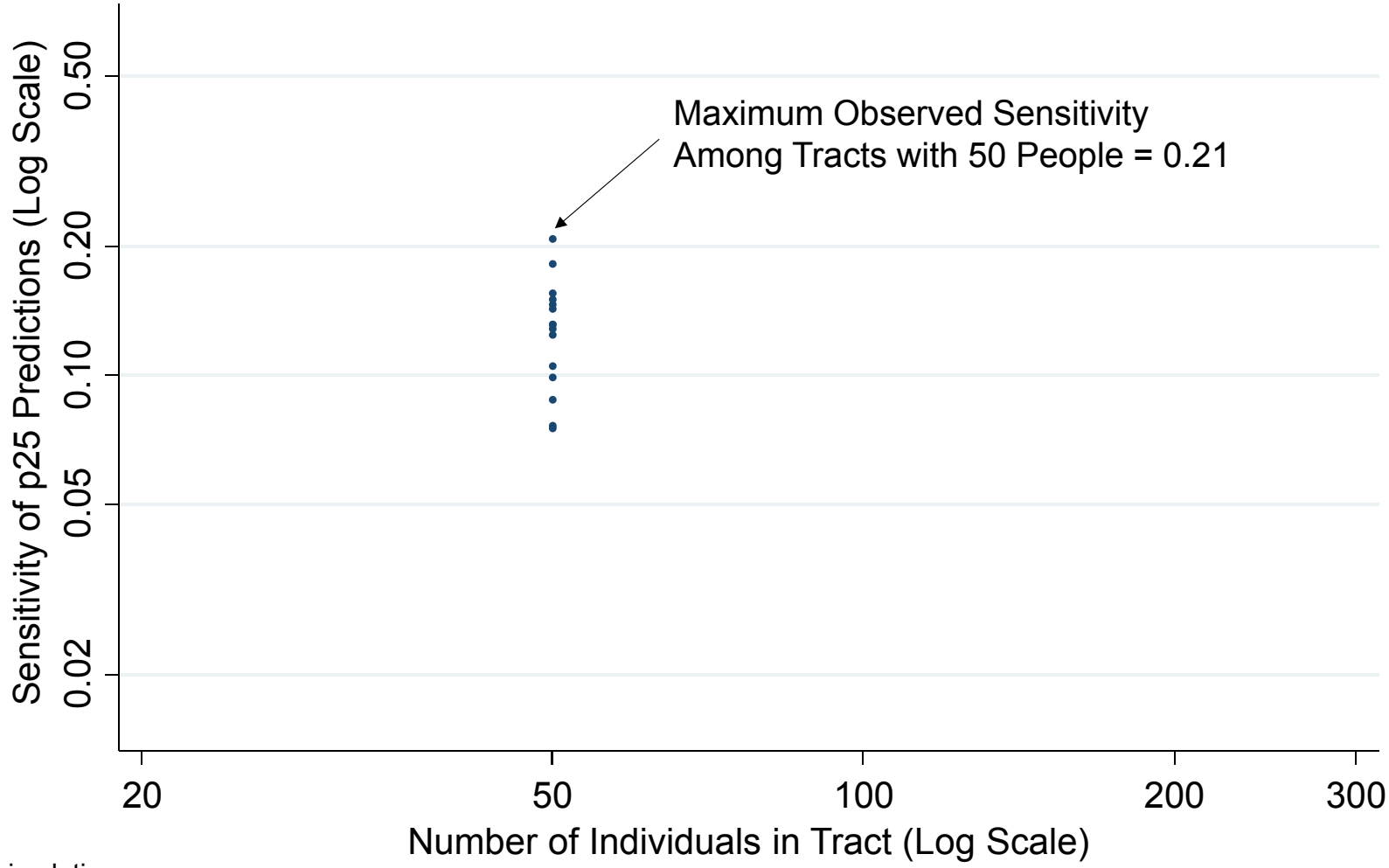
  - Compose regression estimates from noise-infused variance and covariance: Generates bias, unstable estimates due to noise in the denominator

→ How can we proceed?

## Maximum Observed Sensitivity

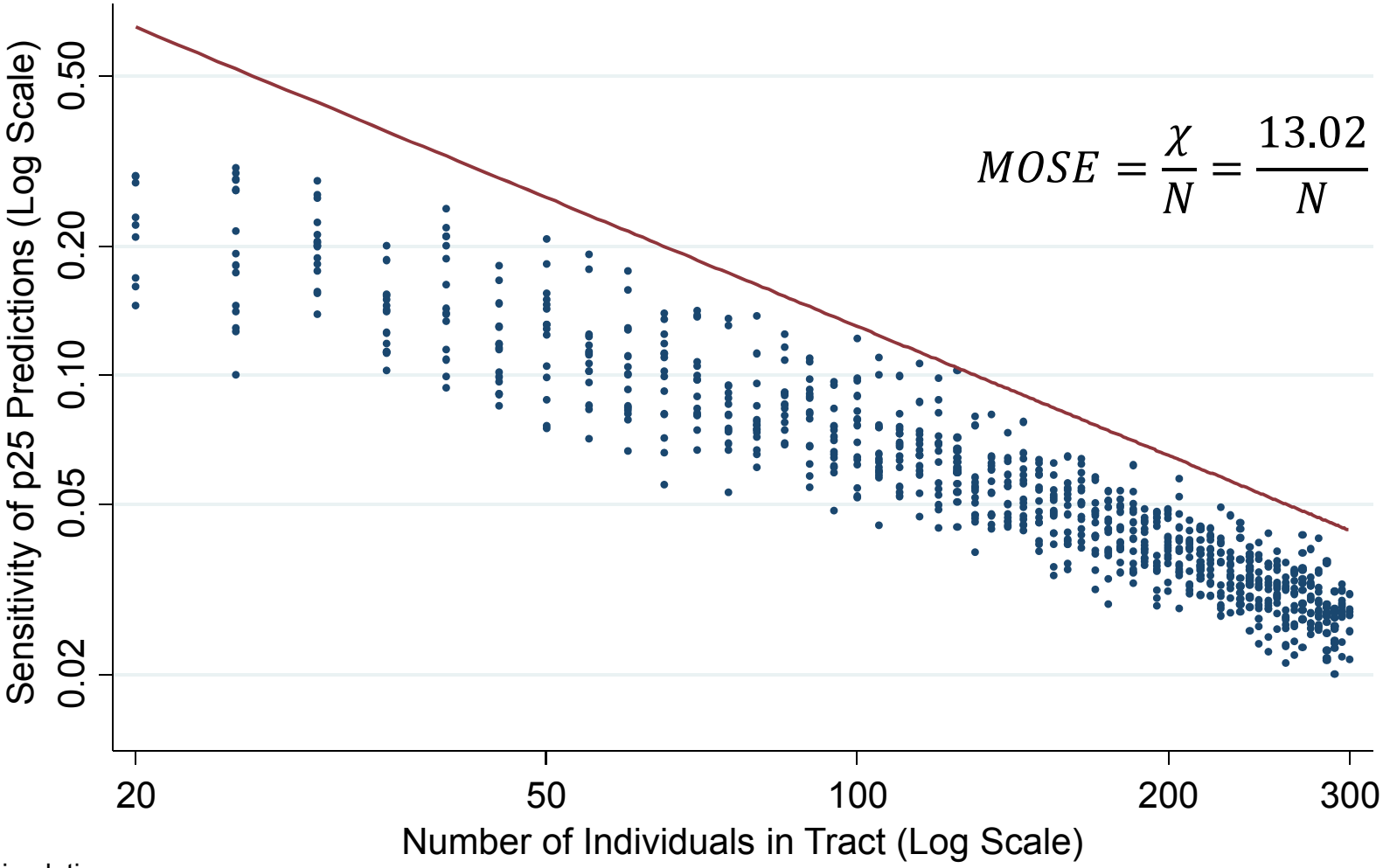
- Our method: use the maximum observed local sensitivity across all cells in the data
  - In geography of opportunity application, calculate local sensitivity in every tract
  - Then use the maximum observed sensitivity (MOS) across all tracts within a given state as the sensitivity parameter for every 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 $\chi$, tract counts, and exogenously specified privacy parameter $\varepsilon$ to add noise and construct public estimates:

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

- This method not "provably private," but it reduces privacy risk to release of the single max observed sensitivity parameter ($\chi$)
  - Privacy loss from release of regression statistics themselves is controlled below risk tolerance threshold $\varepsilon$
- Critically, $\chi$ 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 higher level to pose a privacy risk

1 Method: Maximum Observed Sensitivity

2 Application: Opportunity Atlas

3 Comparison with Traditional Methods

## Application: Opportunity Atlas

- Set risk tolerance $\varepsilon$ following Abowd and Schmutte (2019) approach of weighing privacy losses against social benefits

- Two definitions of social benefit:

  1. Mean-squared error loss when predicting tract-level outcomes

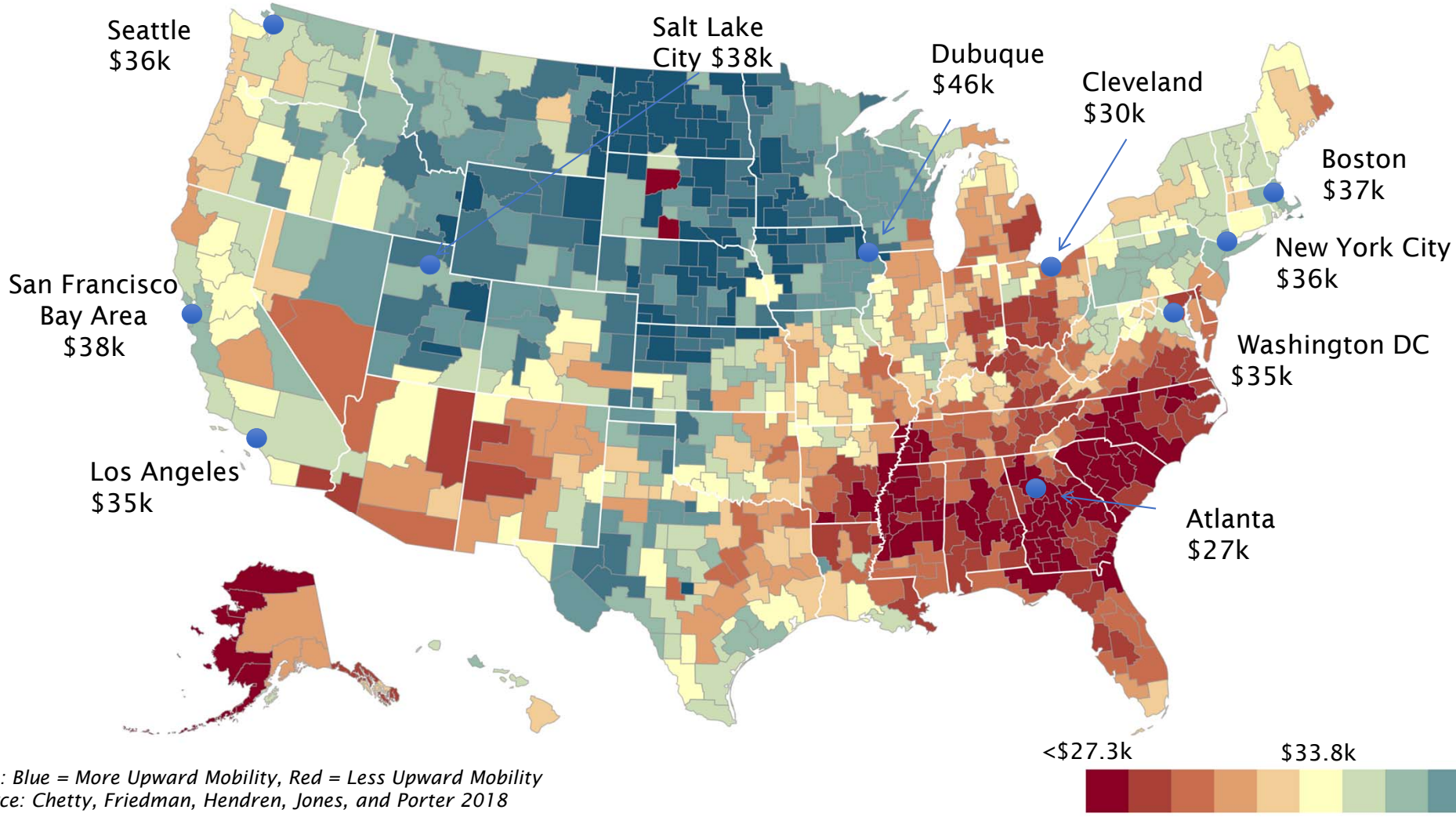
  2. Accuracy of information when predicting the best and worst tracts

     - E.g., consider a family seeking to move to a neighborhood with high upward mobility

     - Operationalize e.g. as $f(Actual\ Quantile \mid Public\ Quantile > 0.95)$

# Geography of Upward Mobility in the United States

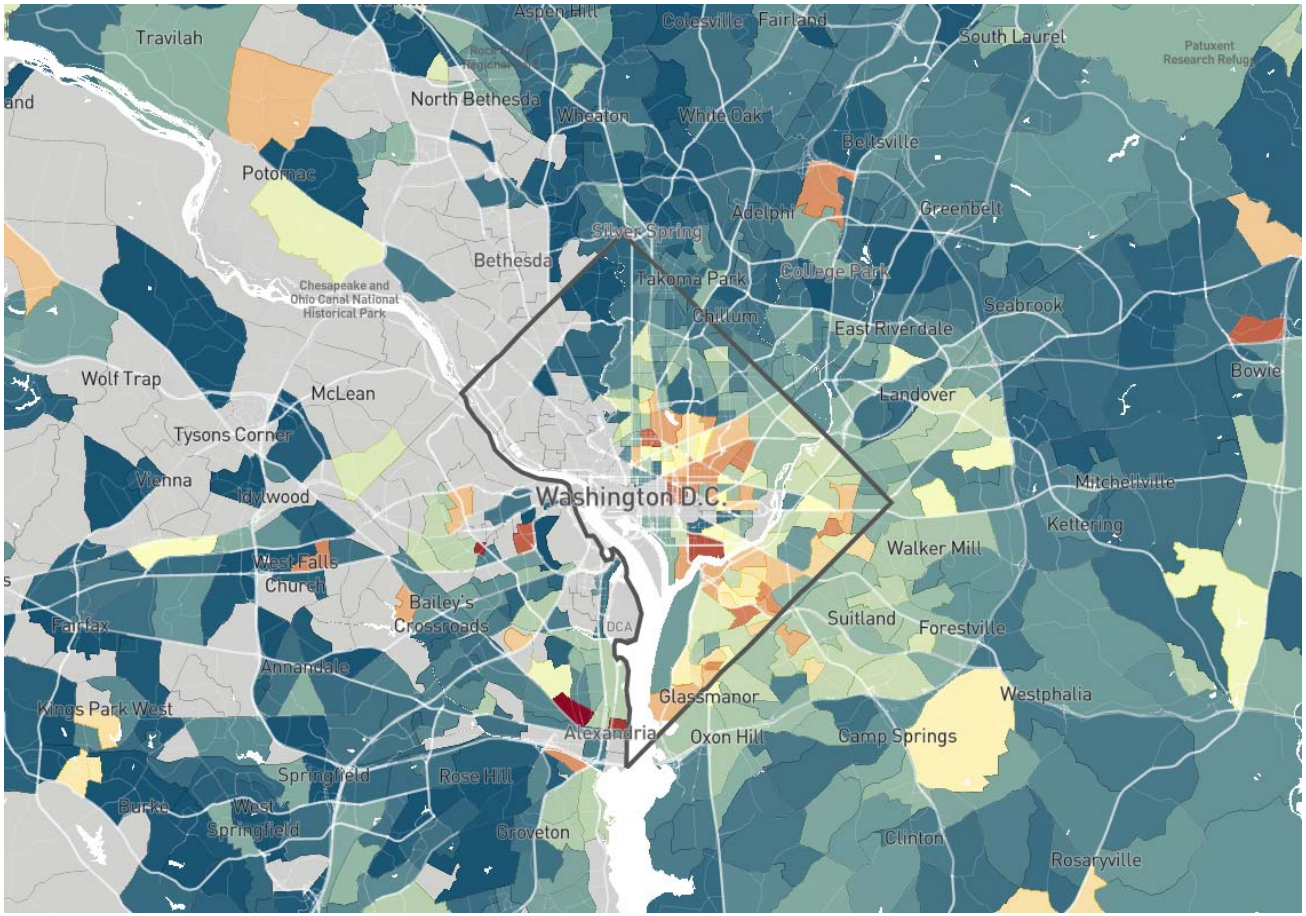
## Average Income at Age 35 for Children whose Parents Earned $25,000 (25th percentile)




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

## Geography of Upward Mobility for Black Children in Washington, D.C.

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*

1 Statement of the Problem

2 Method: Maximum Observed Sensitivity

3 Comparison with Traditional Methods

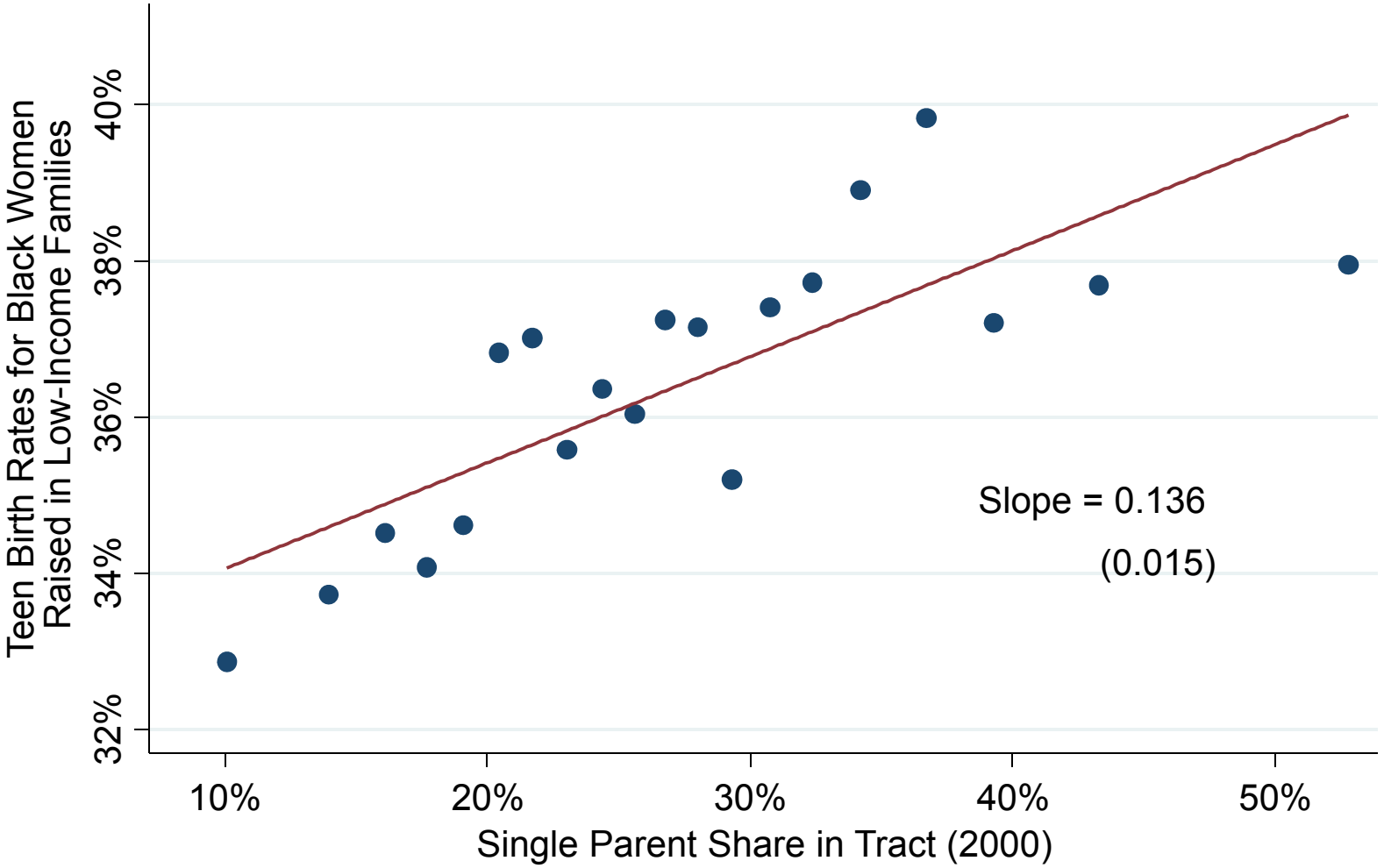
## Comparison to Alternative Methods

- We now compare the properties of our noise-infusion approach to existing methods (such as count-based cell suppression).

- Evaluate three key metrics:

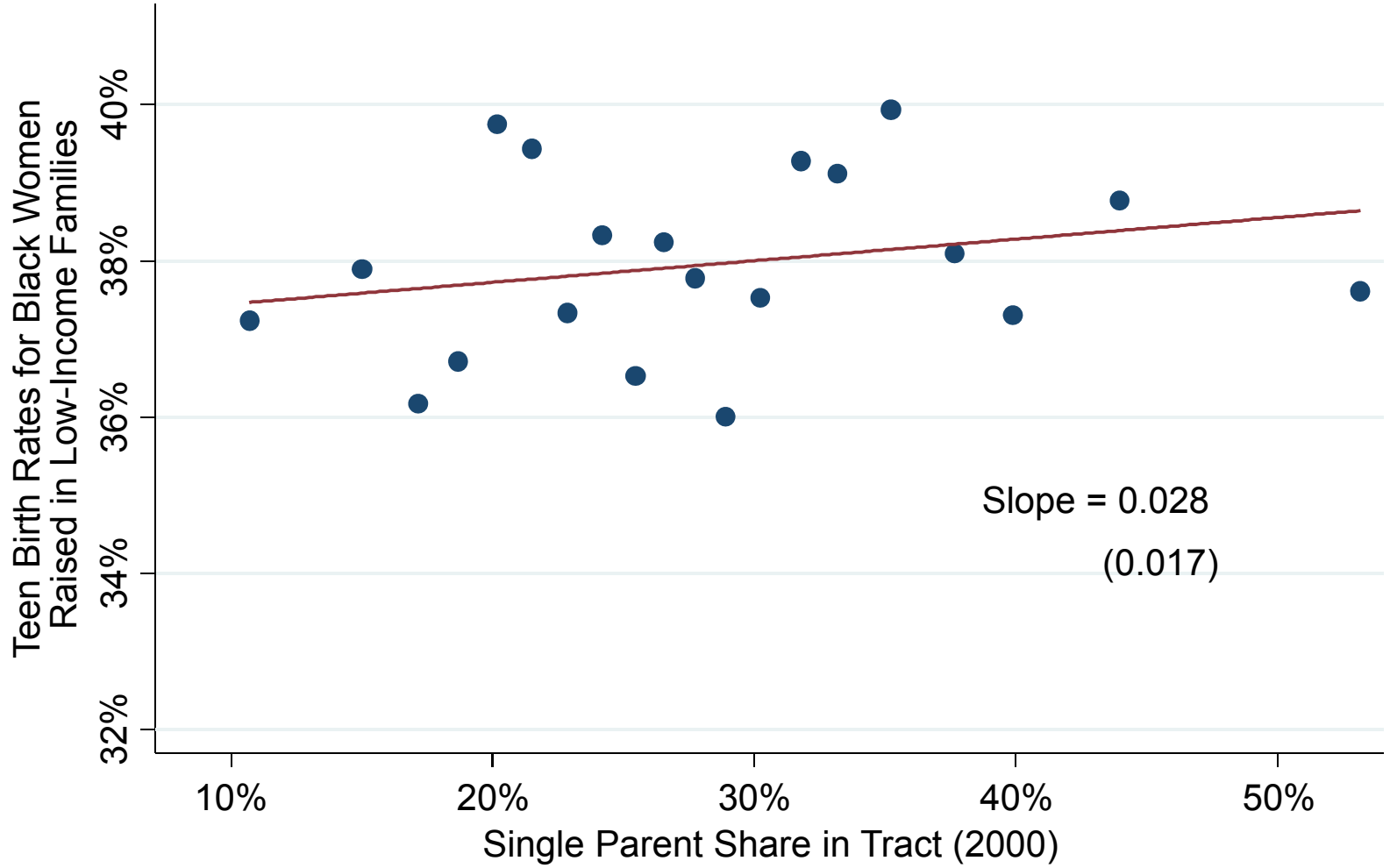
    1. Privacy loss
    2. Statistical bias
    3. Statistical precision

## Association between Teenage Birth and Two-Parent Share: Noise-Infused Data




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

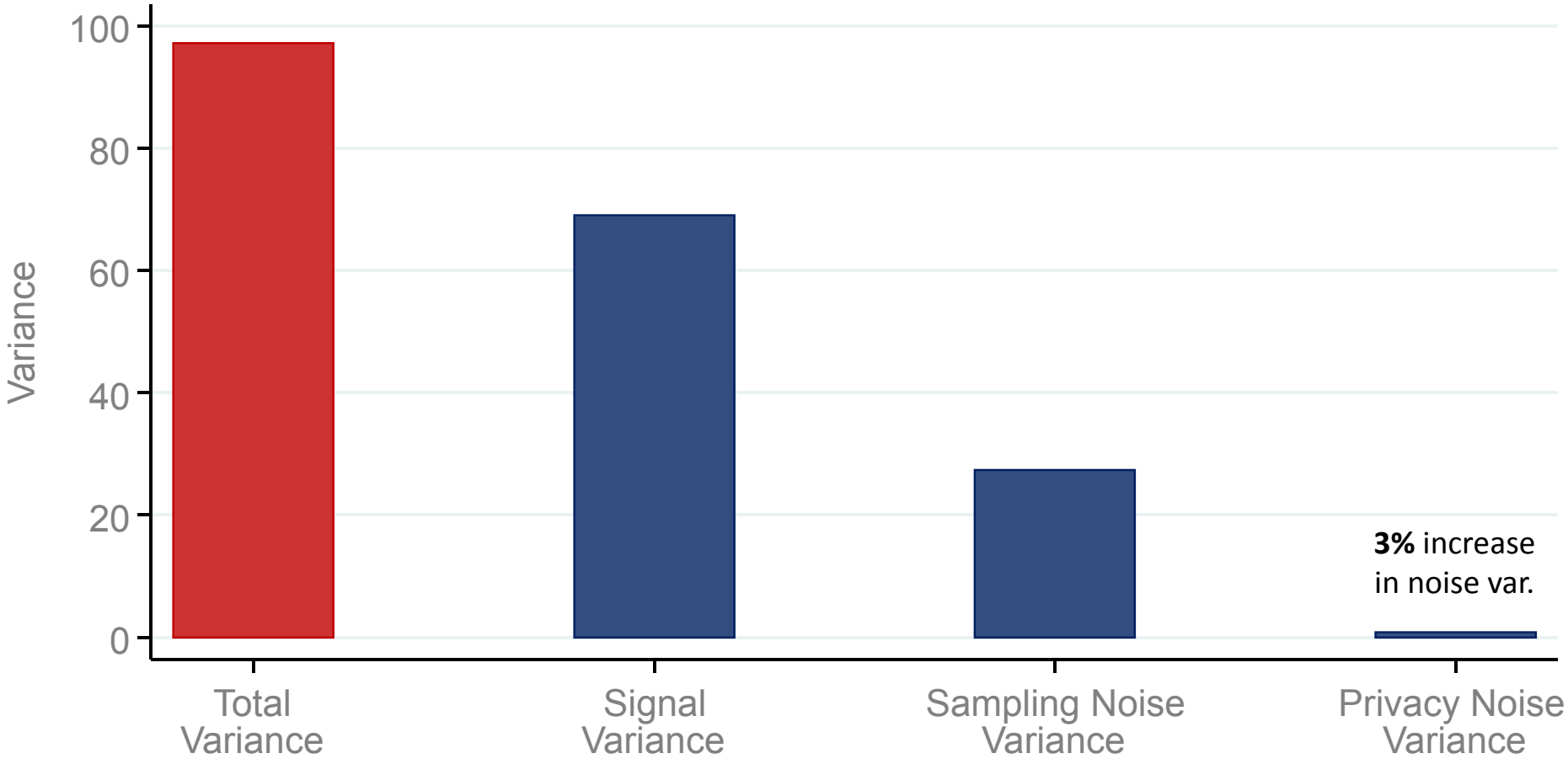
## Association between Teenage Birth and Two-Parent Share: Count-Suppressed Data




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

**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

# Appendix Slides

## Formal Definition of Differential Privacy

- More formally, consider two datasets $D_1$ and $D_2$ that differ by just one element, and an algorithm $\mathcal{A}(D)$ that produces a statistic $\theta$.
  - Let $X$ be the dataset that produced $\theta$.
- The algorithm is "$\varepsilon$ -differentially private" if:

$$\Pr[X = D_1] \leq e^{\varepsilon} \times \Pr[X = D_2]$$

- Intuitively, it is not much more likely that the true underlying dataset is $D_1$ or $D_2$, where the probability is calculated over the randomness from the algorithm.

## Summary: Maximum Observed Sensitivity Disclosure Algorithm

1. Calculate the local sensitivity $LS_{\theta,g}$ for the statistic in each cell $g$ of your data

2. Compute the maximum observed sensitivity envelope scaling parameter $\chi$ :

$$\chi = \max_{g}\{N_g \times LS_{\theta,g}\}$$

3. Determine the privacy parameter $\varepsilon$ .
4. Add random noise proportional to and pre-specified privacy parameter to each statistic:

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

5. Release the noise-infused statistics $\{\tilde{\theta}_g\}$, $\{\widetilde{N}_g\}$, $\varepsilon$ and $\chi$ publicly.

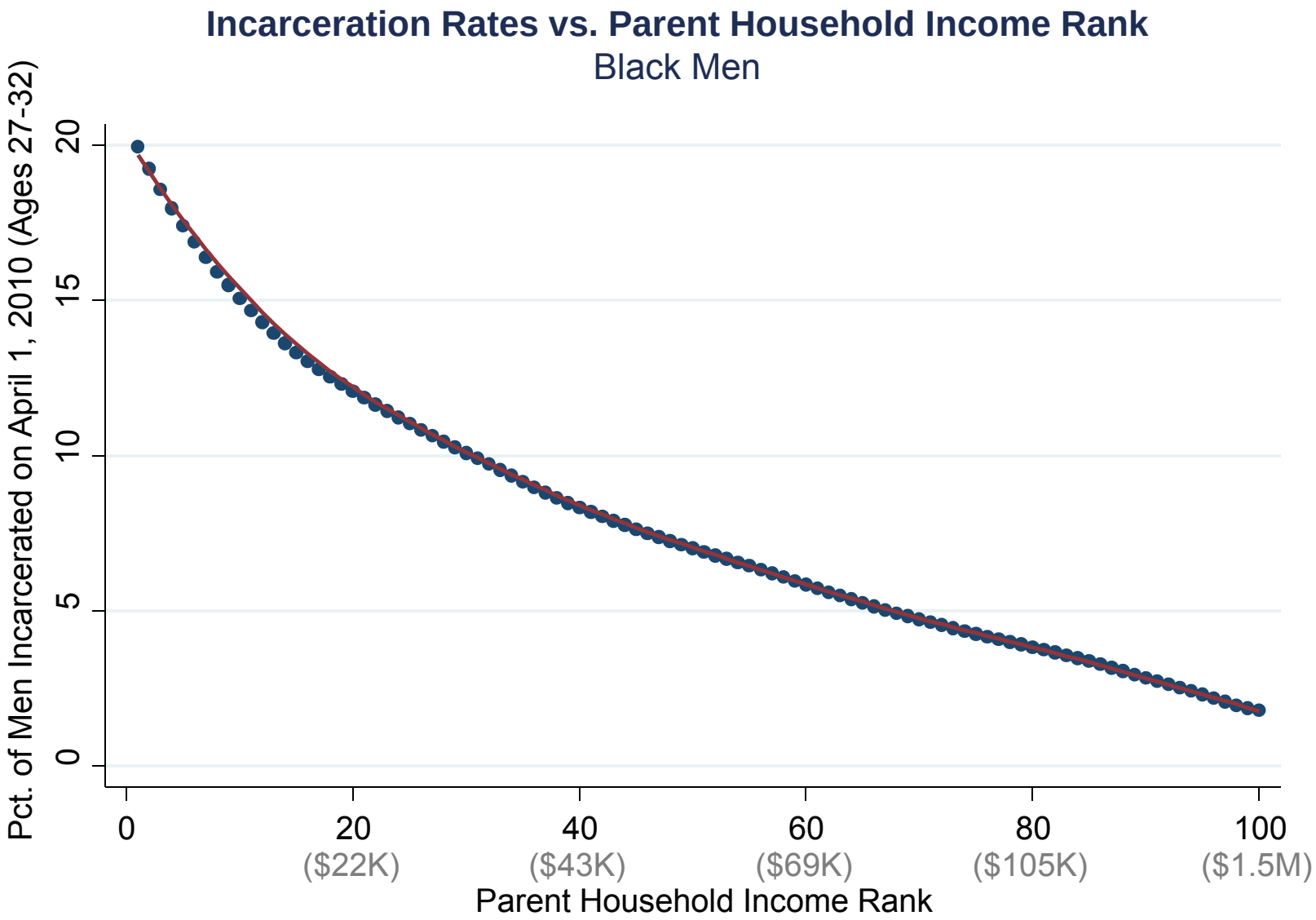
   – Can release standard errors through similar procedure.

## Specification Details

- Measuring Incomes:
  - Parents' pre-tax household incomes: mean Adjusted Gross Income from 1994-2000, assigning non-filers zeros.
  - Children's pre-tax incomes measured in 2014-15 (ages 31-37)

- To mitigate lifecycle bias, focus on percentile ranks in **national** distribution:
  - Rank children relative to their birth cohort and parents relative to other parents
  - Address non-linearities in a linear regression framework:

$$y_{ic} = \alpha_c + \beta_c \times f(p_{ic}) + \varepsilon_{ic}$$




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

## Data Sources and Sample Definitions

- Data sources: Census data (2000, 2010, ACS) covering U.S. population linked to federal income tax returns from 1989-2015
- Link children to parents based on dependent claiming on tax returns
- Target sample: Children in 1978-83 birth cohorts who were born in the U.S. or are authorized immigrants who came to the U.S. in childhood
- Analysis sample: 20.5 million children, 96% coverage rate of target sample

## Specification Details

- Measuring Incomes:
  - Parents' pre-tax household incomes: mean Adjusted Gross Income from 1994-2000, assigning non-filers zeros.
  - Children's pre-tax incomes measured in 2014-15 (ages 31-37)

- To mitigate lifecycle bias, focus on percentile ranks in **national** distribution:
  - Rank children relative to their birth cohort and parents relative to other parents
  - Address non-linearities in a linear regression framework:

$$y_{ic} = \alpha_c + \beta_c \times f(p_{ic}) + \varepsilon_{ic}$$

- Define cells for the MOS parameter at the race-by-state-by-gender level; e.g., white women in Utah.

## Other Practicalities for Privacy Method Implementation

- **Predicted Values at the 25$^{th}$ and 75$^{th}$ percentiles**
- **Winsorize**
- **Exclude Small Cells**
- **Gaussian Noise**
- **Weighted Average over Time Spent in Each Neighborhood**

## Other Practicalities for Privacy Method Implementation

- **Predicted Values** We produce predicted values at the 25$^{th}$ and 75$^{th}$ percentiles of parent income, so that we can estimate the full line
  - Instead predict the 50$^{th}$ and 1$^{st}$ (100$^{th}$) for tracts with less than 10% of obs above (below) median parent income

## Other Practicalities for Privacy Method Implementation

- **Predicted Values**

- **Winsorize** the Data to reduce the influence of outliers, sensitivity

  – Must calculate sensitivity, MOS on the composed function including Winsorization

## Other Practicalities for Privacy Method Implementation

- **Predicted Values**
- **Winsorize**
- **Define Cells** for the MOS scaling parameter at the state X race X gender level. E.g., white women in Utah.

## Other Practicalities for Privacy Method Implementation

- **Predicted Values**
- **Winsorize**
- **Define Cells**
- **Exclude Small Cells** to comply with current IRS regulations.
  - Censor cells with fewer than 20 obs; better would be to censor on public counts to avoid further privacy “leaks”

## Other Practicalities for Privacy Method Implementation

- **Predicted Values**
- **Winsorize**
- **Define Cells**
- **Exclude Small Cells**
- **Gaussian Noise** In practice, Normally distributed noise is more convenient for downstream statistical inference, e.g., the construction of confidence intervals or Bayesian shrinkage estimators.
  - Instead add $N\left(0, \sqrt{2}\frac{\chi}{\varepsilon N_g}\right)$, though will not conform exactly to privacy loss bounds in the tails.

## Comparison to Alternative Methods: Privacy

- Our method is likely to reduce the risk of privacy loss substantially relative to count-based cell suppression (like most noise-infusion algorithms)
  - Even if one suppresses cells with counts below some threshold, can recover information about a single individual from similar datasets.
  - Hence, statistics released after cell suppression still effectively have infinite (uncontrolled) privacy risk.
- In contrast, our maximum observed sensitivity approach reduces uncontrolled privacy risks to one number ($\chi$)
  - Can typically estimate in a sufficiently large sample that poses negligible privacy risk.

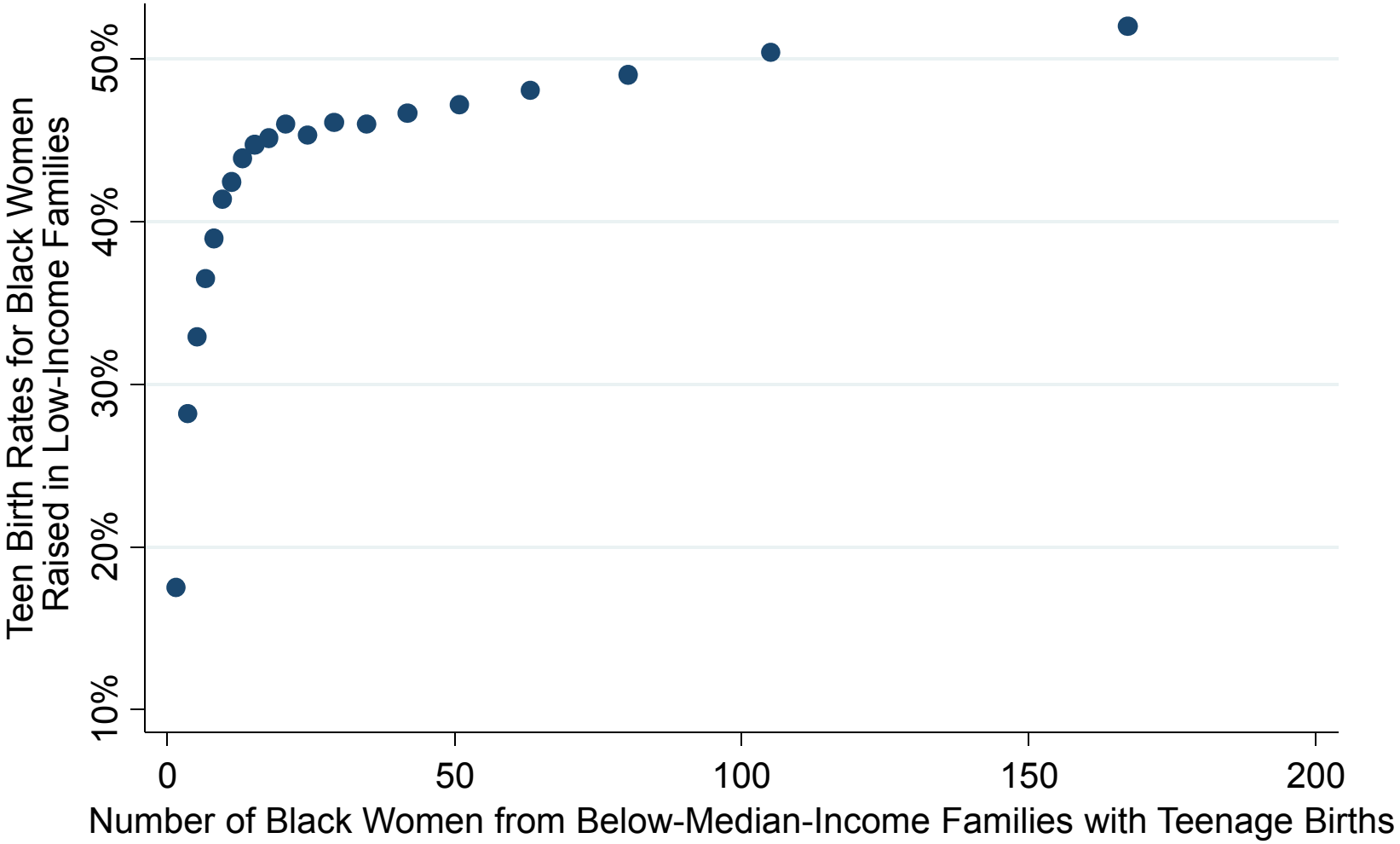
## Comparison to Alternative Methods: Statistical Bias

- Noise infusion via known parameters offers significant advantages in downstream statistical inference.
  - Easy to extract unbiased estimates of any downstream parameter using standard measurement error correction techniques
  - In contrast, count-based suppression can create bias in ways that cannot be easily identified or corrected ex-post.
- Illustrate by comparing how actual results reported in Chetty et al. (2018) would have changed had count-based suppression been used instead of noise infusion
  - Are teenage birth rates higher for those who grow up in neighborhood with a higher share of single parents?

## Comparison to Alternative Methods: Statistical Bias

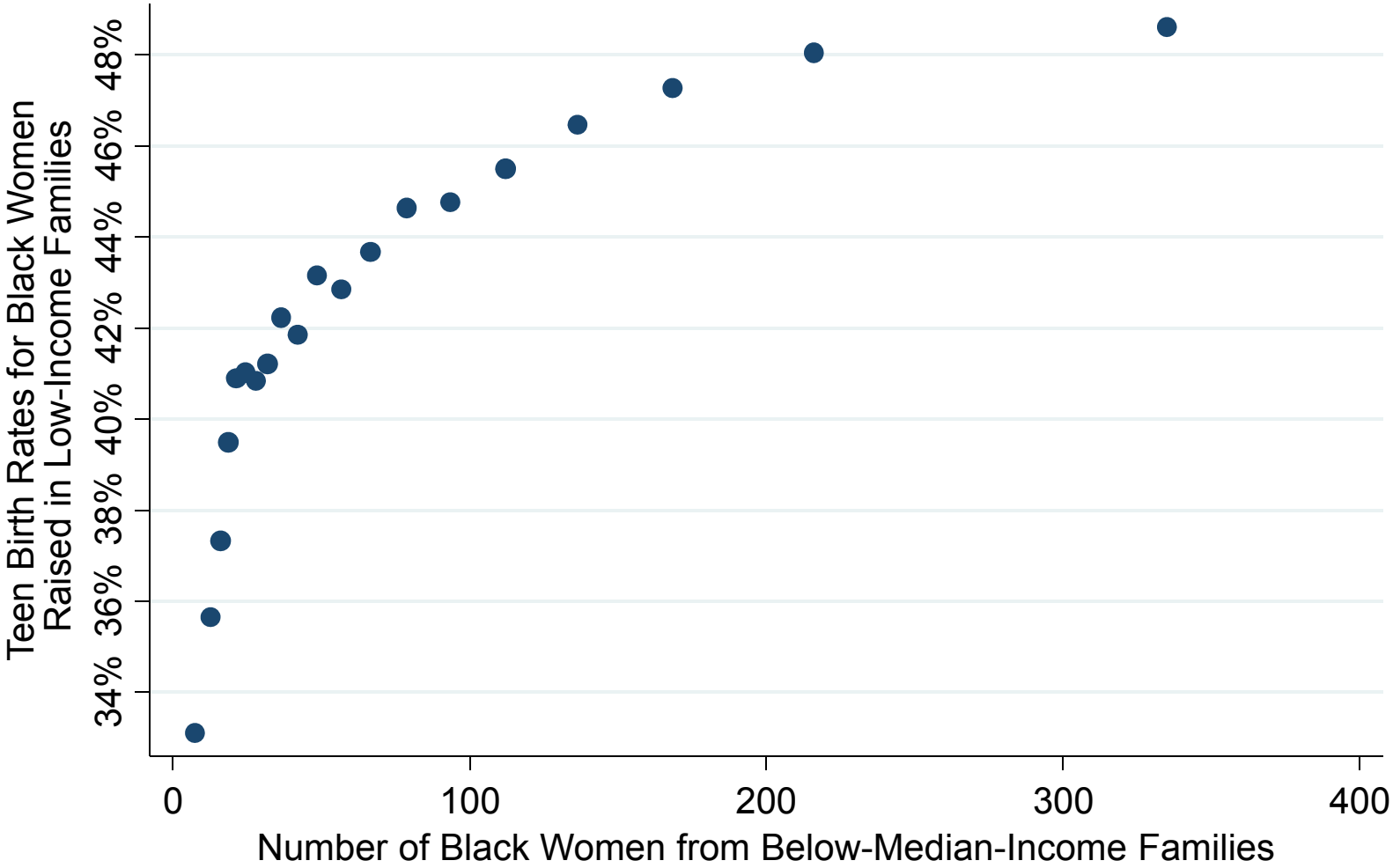
- In noise-infused data, regression provides an unbiased estimate of the (strong positive) relationship between teenage-birth rates for black women and single-parent share.
  - More generally, can adjust for noise using the “signal correlation”
- In contrast, count-based suppression generates bias that 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

## Teenage Birth Rates for Black Women vs. Number of Black Women with Teenage Births in Tract




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

## Teenage Birth Rates for Black Women vs. Number of Black Women in Tract




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

## 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%