CBAMS: A Case Study in Differential Privacy at the U.S. Census Bureau

Caleb Floyd and Rolando Rodríguez

Center for Enterprise Dissemination – Disclosure Avoidance

U.S. Census Bureau

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

The Census Bureau is committed to data quality

- The Census Bureau's *mission* is to serve as the nation’s leading provider of **quality** data about its people and economy
- The Census Bureau’s *goal* is to provide the best mix of timeliness, relevancy, **quality** and cost for the data we collect and services we provide

The Census Bureau is also committed to data privacy protection

- The Census Bureau operates, collects data, and publishes statistics under the authority of several titles of the U.S. Code
- Title 13, Sec.9: Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof […] may […] make any publication whereby the **data furnished by any particular establishment or individual** under this title **can be identified**

The Census Bureau is modernizing the way we protect data

Traditional

- Include techniques such as
  - Geographic Aggregation
  - Cell Suppression
  - Variable Top-Coding
  - Category Collapsing
- Methods are ad hoc and based on known risks
- May require secrecy of methods and parameters

Modern

- Guarantee privacy against broad classes of attacks
- Do not depend on which datasets are now or will be available
- Have a calculable, global privacy loss for a given set of releases at a given accuracy
- Allow for transparency about the method, data accuracy, and privacy loss

The Census Bureau is pioneering noise injection techniques

- Disciplined and careful noise injection can help provide estimates with favorable properties
- Differential privacy requires that statistical disclosure avoidance techniques, such as noise injection, meet mathematically defined bounds on privacy loss
- We must weigh noise injection against alternative disclosure avoidance methods
  - Even when the properties of the noise injection are sub-optimal, the method can still outperform alternatives
  - Especially true when the alternative is providing less output

The use of formal privacy is expanding at the Census Bureau

In Use or Planned

- OnTheMap
- 2020 Census
- Post-Secondary Education Outcomes (PSEO)
- Census Barriers Attitudes and Motivators Study (CBAMS)

Ongoing Research

- American Community Survey
- Many more to come!

The Census Barriers Attitudes and Motivators Study (CBAMS) is critical for the 2020 Census

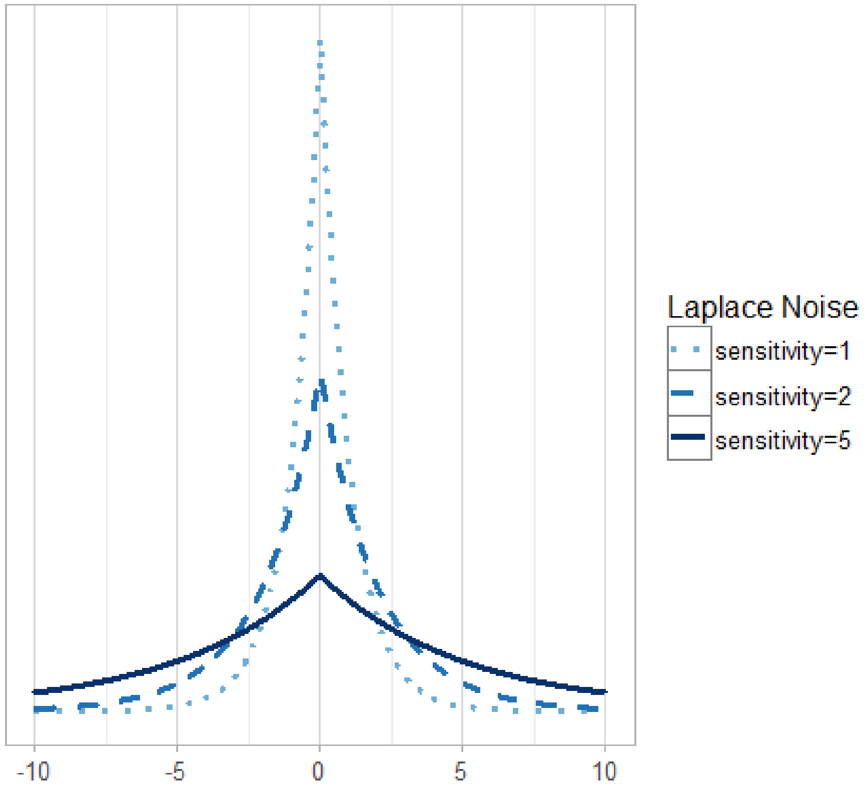
- CBAMS is a nationwide survey designed to identify barriers, attitudes, and motivators toward Census response
- Critical for an accurate and cost efficient Census
  - Emphasizes hard-to-count populations
  - Used to help allocate the media budget and pursue maximum impact of the media campaign
- The 2020 communication campaign requires sharing CBAMS data with external partners

We choose modern methods for CBAMS

- The communications team needs tens of thousands of estimates from crosstabs
- Cell suppression was considered but it would be complex and time consuming
  - Anytime a cell needs to be suppressed, adjacent cells within a table and across linked tables would need to be suppressed to avoid reconstruction of the suppressed cell
  - Combining categories to reduce the number of offending cells would result in some proportion of lost information
- We provide a protected microdata file using differentially private noise injection
  - Provides partners with higher quality data than they would have otherwise received with traditional methods
  - The communication strategy would need to be drastically changed without this solution
- We employ the local model with two differentially private noise injection mechanisms to produce the protected CBAMS microdata file

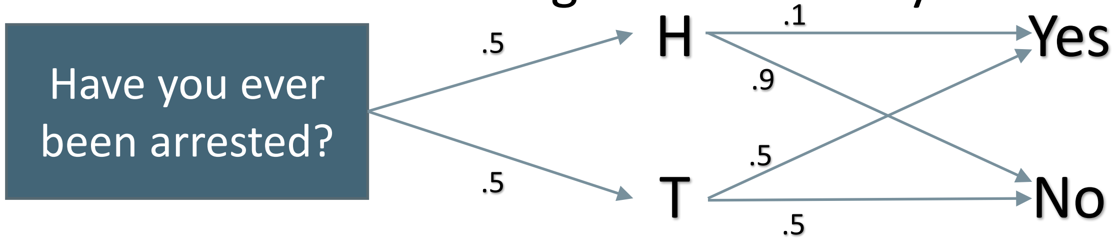
We add Laplace noise to continuous variables

- For a value x we return $x + \omega$, where $\omega \sim \text{Laplace}(0, \frac{S_x}{\epsilon})$
- S_x is called the sensitivity
  - For this model, it is calculated as the difference between the maximum and minimum possible values
  - All continuous values are percentages, so we have $S_x = 100 – 0 = 100$
- ϵ is the privacy parameter



Randomized response is a valuable differentially private mechanism for categorical variables

- Originally developed as a survey method that allows respondent to respond to sensitive questions while maintaining confidentiality
- Example:



  - Assume that the true percentage of respondents who have been arrested is 10%. Then we'll have the expected proportion of answers:

YES	.5(.1) + .5(.5) = .3
NO	.5(.9) + .5(.5) = .7

  - Given that the true proportion of 'yes' answers is unknown, it can be reconstructed as

$$YES = \frac{YES_{answered} - P(T)P(H)}{P(H)}$$

We apply randomized response to discrete variables

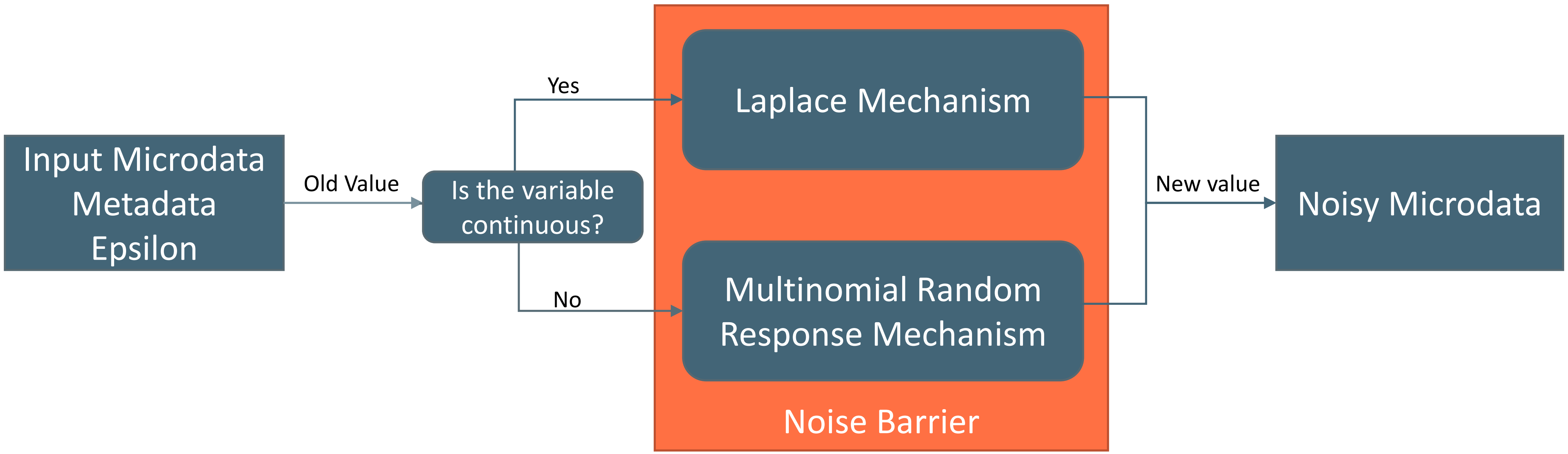
- Many survey questions have categorical responses
  - Example: Is the Census used to decide how much money communities will get from the government?
    1. (1) Yes, used for this
    2. (2) No, not used
    3. (3) Don't know
- For a two-category variable the noise injection is displayed in the following design matrix, where ϵ is the differential privacy parameter

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} \frac{e^{\epsilon}}{1+e^{\epsilon}} & \frac{1}{1+e^{\epsilon}} \\ \frac{1}{1+e^{\epsilon}} & \frac{e^{\epsilon}}{1+e^{\epsilon}} \end{pmatrix}$$

- We use a multinomial version of randomized response to perturb each categorical variable

We proceed record by record to create the microdata

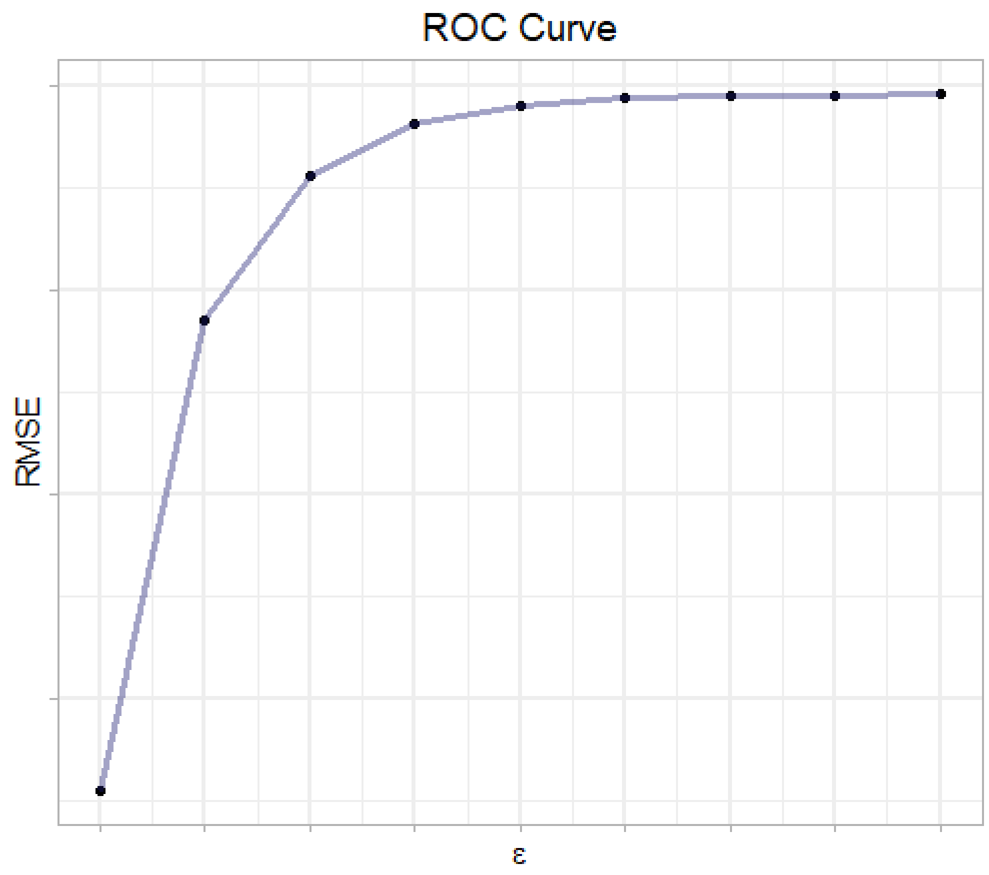
- Implement noise injection on each record variable-by-variable



- All data items must pass through the noise barrier

The ROC curve displays the necessary tradeoff

- Shows the tradeoff between the amount of privacy lost and the accuracy of the data
  - ϵ is our privacy parameter
  - The root mean squared error (RMSE) is our measure of accuracy
- Quantifying the tradeoff is a key feature of differentially private methods
- Allows policy makers to balance important social goods – data accuracy and data privacy
- We use $\epsilon = 7$ for each variable
- Publicly available via the Census Bureau's FOIA page



New methods provide an opportunity for greater data privacy and data accuracy

- These techniques allow us to deliver a wealth of data to support the 2020 Census communication team, and to help ensure the success of the Census itself
- The Census Bureau is actively expanding the use of new methods
- These methods allow us to fulfill our obligations to data users and to respondents
- These changes will let the Census Bureau better serve data users
  - Transparency
  - Data-driven decisions about balancing data privacy with data accuracy
  - Continue to have a quality, trusted, reputable product for years to come

Thank You!

Caleb Floyd

caleb.r.floyd@census.gov

Rolando Rodríguez

rolando.a.rodriguez@census.gov

FOIA link

https://www2.census.gov/programs-surveys/decennial/2020/program-management/census-research/cbams/2020-CBAMS-Survey.zip