PURDUE UNIVERSITY | Department of Computer Science

Brandeis UNIVERSITY

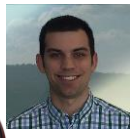
Issues, and Solutions, to Providing Formal Privacy for ACS

Chris Clifton, Shawn Merrill *Purdue University*

Eric Hanson *Norwegian University of Science & Technology*

Keith Merrill *Brandeis University*

CERIAS Center for Education and Research in Information Assurance and Security



This work supported by the U.S. Census Bureau under CRADA CB16ADR0160002

The views and opinions expressed in this talk are those of the authors and not the U.S. Census Bureau.

Indiana Center for Database Systems™

PURDUE UNIVERSITY
Department of Computer Science

Outline

Brandeis UNIVERSITY

- Differential Privacy for ACS harder than Decennial
  - Higher Dimensionality
    - *Several known approaches to mitigate impact of Differential Privacy for high dimensionality, e.g., [Qardaji14,McKenna18]*
  - Stratified survey rather than enumeration
- Solutions - Smooth sensitivity for:
  - Missing data imputation
  - Post-stratification
- Other Challenges

2

PURDUE UNIVERSITY
Department of Computer Science

ACS Challenges

- Considerable variance in response quality and rates across different groups
  - Geographic
  - Demographic
- Statistical methods used to reduce bias and variance
  - Stratified sampling
  - Missing data imputation
  - Weighting approaches

  These make differential privacy more challenging

3

PURDUE UNIVERSITY
Department of Computer Science

ACS Challenges

- Differential privacy based on hiding the impact of any single individual on a published value
- Statistical methods used to reduce bias and variance
  - *These can impact how much one response influences an estimate!*
- Simple example: Weighting
  - Samples with high weights have greater impact on outcomes
  - For a count, one person changes result by at most 1
  - *But for an estimate from a weighted sample, one person can change the value by their weight, so must add noise to cover highest possible weight*

4

PURDUE UNIVERSITY
Department of Computer Science

These problems are solvable!

- Formal privacy techniques to reduce bias given biased samples
  - Missing data imputation
  - Post-stratification

  We've developed approaches that address each of these

5

PURDUE UNIVERSITY
Department of Computer Science

Real-world DP: What is the Sensitivity?

- Sensitivity: Maximum change in query/statistic from adding or deleting one individual
- Examples of sensitivity
  - Count (e.g., how many high income individuals in this room): 1
  - Average (e.g., average income in this room):

$$\frac{\sum_{all} income}{n} - \frac{\sum_{all+1} income}{n} \leq \frac{Maximum\ income}{Number\ of\ individuals}$$

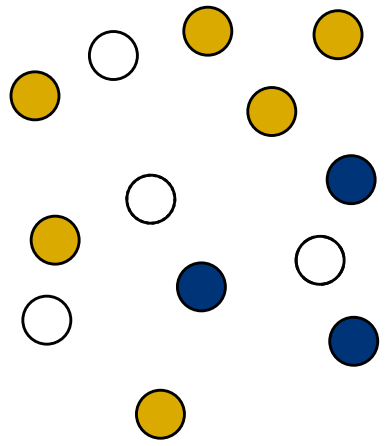
maximum ***possible***

Minimum possible

6

PURDUE UNIVERSITY
Department of Computer Science

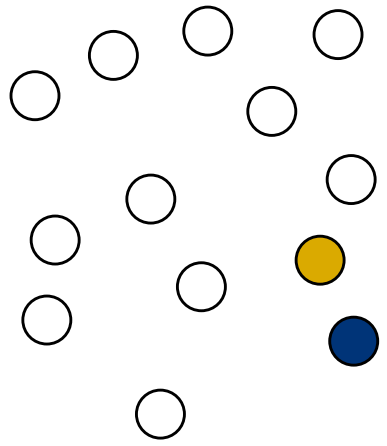
Problem: Missing Data
And does this impact privacy?



- Assignment
  - Missing value determined from other characteristics of individual
  - Assuming rules not derived from other records, no impact on sensitivity
- Allocation
  - Copy values from similar "donor" individual
  - *Increases Sensitivity!*

PURDUE UNIVERSITY
Department of Computer Science

Global Sensitivity High

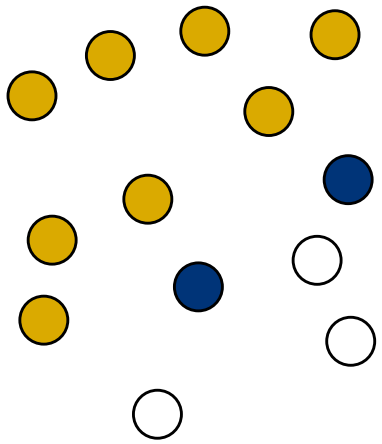


- Can construct situations where changing one individual dramatically changes result
- Global Sensitivity of count query ≈ size of dataset
  - Any mechanism based on global sensitivity will have untenably high variance.

PURDUE UNIVERSITY
Department of Computer Science

Solutions

- Ignore missing data
  - No impact on sensitivity ➔ low variance solution
  - But biased result
    - Data frequently not missing completely at random
- Low sensitivity allocation method?
  - Will this effectively reduce bias?
- *Smooth Sensitivity*



PURDUE UNIVERSITY
Department of Computer Science

Smooth Sensitivity

(Nissim, Raskhodnikova & Smith STOC'07)

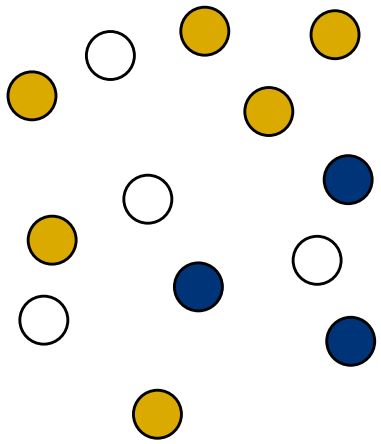
- Idea: Neighbors of current database don't induce big changes
  - Local sensitivity low
- Pathological datasets far away from the real dataset have substantially less impact on privacy.
- Sensitivity based on distance from actual data
  - For $\beta > 0$, the $\beta-$smooth sensitivity of f is

$$S^*_{f,\beta}(x) = \max_{y \in D^n} (LS_f(y) \cdot e^{-\beta d(x,y)})$$

PURDUE UNIVERSITY
Department of Computer Science

Problem statement

- (Deterministic) Nearest neighbor
- Determine upper bound on local sensitivity of function on data and allocated values
- Solutions for count, mean, variance
  - *Technical details in a paper under review*



11

PURDUE UNIVERSITY
Department of Computer Science

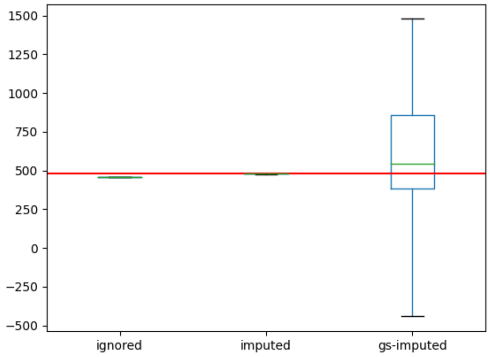
Evaluation Testbed

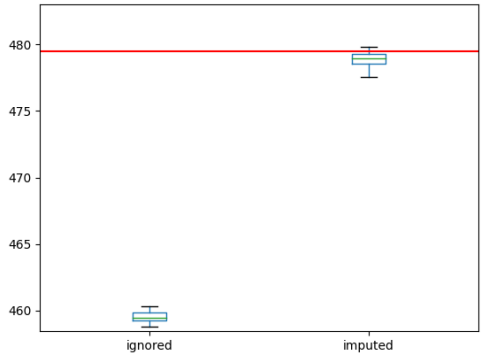
- Idea: *Model* missing data based on ACS Public Use Microdata
  - We know what was missing
  - Use to build model of what makes data missing
- Apply model to 1940 Census Data
  - Complete dataset (no weighted sample issues)
  - Known ground truth

12

PURDUE UNIVERSITY
Department of Computer Science

1940 Mean Individual Income Ages 20-59



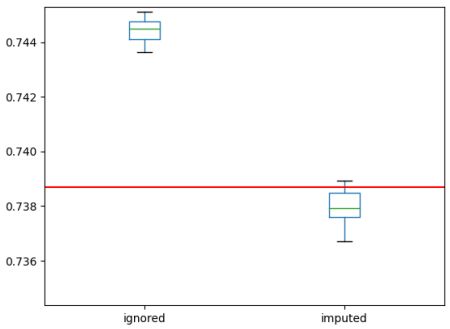


13

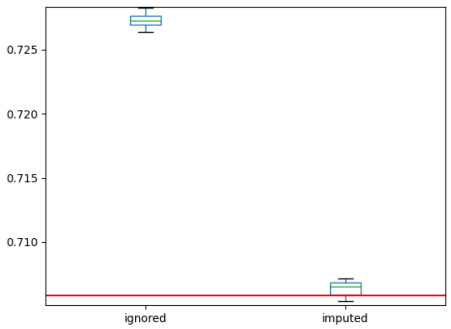
PURDUE UNIVERSITY
Department of Computer Science

Proportion of Individuals unable to support Family of 4 Above Poverty Line

Ages 20-29



Ages 40-49

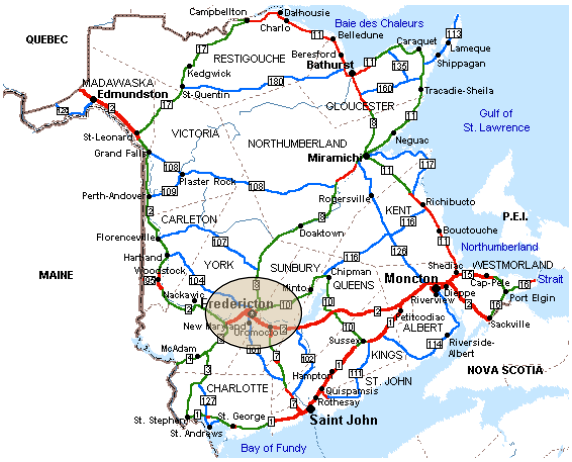


14

PURDUE UNIVERSITY
Department of Computer Science

Post-Stratification

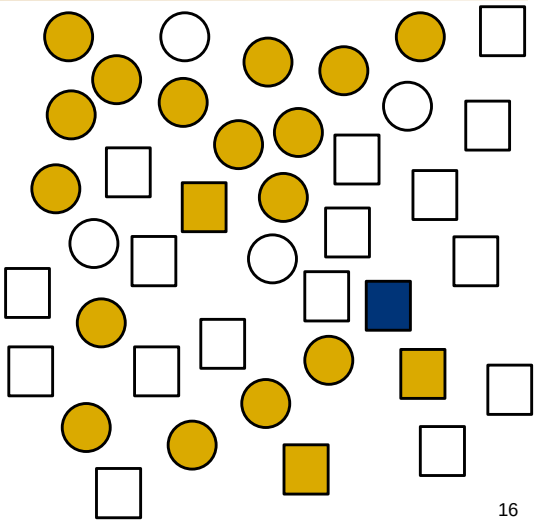
- Idea: Determine stratification based on known statistics
  - Number of francophones in Fredericton
- Weight francophones in sample so results match known values
  - Improves correlated survey results for which values unknown (e.g., if occupation varies by language)
- *Global Sensitivity off the charts!*
  - Hypothetical survey that only captures one francophone – represents ~4000 individuals



PURDUE UNIVERSITY
Department of Computer Science

Challenge with Limiting Post-Stratification Weight

- Known: Count of Circles, Squares
  - Limit maximum weight to 4
  - 20 known squares, 4 in sample – weight exceeds limit
  - Weight based on total count only: 40 / 20 in sample = 2
- Query: Number of blue
  - One in sample
  - One individual represents 2, so query results = 2
- Neighboring Database: Additional blue square in sample
  - 5 squares in sample, so weight = 4
- Number of blue = 2 * 4 = 8
  - Sensitivity 8-2 = 6!



PURDUE UNIVERSITY
Department of Computer Science

Solution: Stochastic Post-Stratification

(with Aref Dajani, Stephen Clark, Rolando Rodriguez-Rivera, U.S. Census Bureau)

- Select post-stratification binning P for dataset D
- Balance maximum weight and fine granularity
  - $f(D,P) = \frac{|P|}{W_0^P(D)}$
- Choose among possible P with probability $e^{\epsilon f(D,P)}$
  - Differentially private selection avoids one individual having too great an impact on how post-stratification done
- *Currently running experiments to evaluate bias reduction / variance tradeoff*

17

PURDUE UNIVERSITY
Department of Computer Science

Conclusions

- Formal Privacy for ACS is challenging
  - But solvable
- Requires new techniques, which could be
  - New approaches to differential privacy
  - New techniques for bias/variance reduction that are more amenable to formal privacy
  - New formal privacy definitions and methods
- *The research community is making advances on all of these*

18