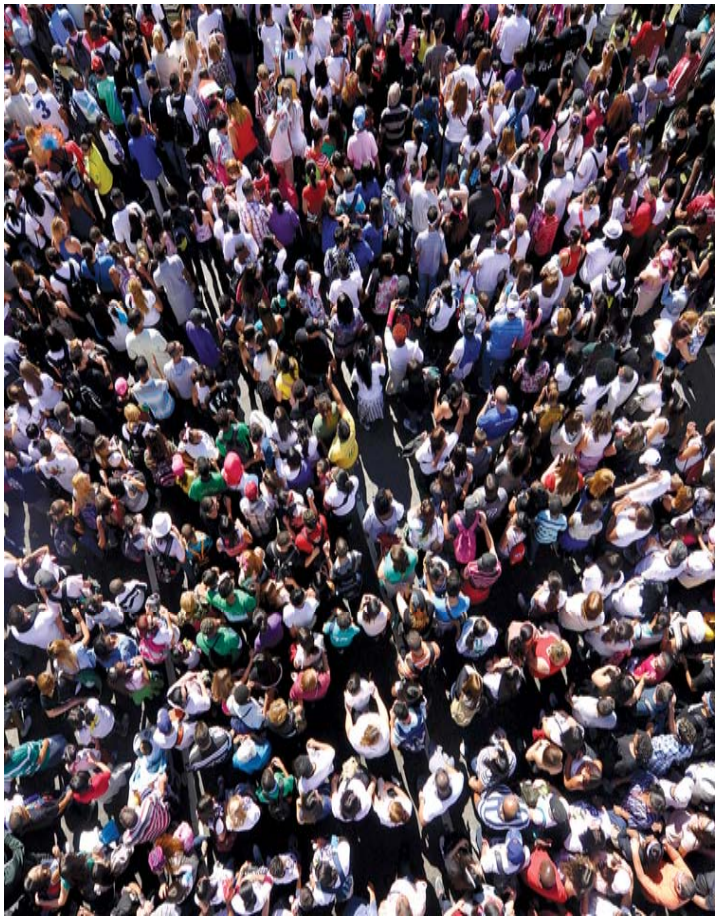

Commentary on Privacy, Utility, and Potential Application of Differential Privacy to Census Data

Kirk Wolter, Federal Economic Statistics Advisory Committee

December 14, 2018

NORC at the UNIVERSITY of CHICAGO

I'll discuss…

- A couple of preliminaries
- Four concerns about potential application of DP to census data
- Two questions
- Summary

Preliminaries

- Tension between privacy and utility
  - Privacy is very important
  - Utility is very important
  - Calls for balance, within the applicable legal framework of the census

Preliminiaries

- Masking/differential privacy (DP) applied to census data
  - y is a raw, unadjusted statistic of interest
  - The Census Bureau would release

    $$Y = y + e$$

  - e is the DP error
    - $e \sim Laplace(0, b)$ or similar
    - $E\{e\} = 0$
    - $Var\{e\} = \sigma^2 = 2b^2$
    - $b = \Delta y / \epsilon$ is specified by census experts

Concerns

1. Effect of DP on various uses of census data
2. Reconstruction does not equate to identification
3. Application to skewed populations
4. Census needs a communications strategy

Concern 1

- Effect of DP on survey design and estimation
  - On the between PSU component of variance
  - On the oversampling of rare populations
  - On the estimation procedure
  - Bottom line
    - Given fixed budget, variances increase and policy and business decisions degrade
    - Given fixed variance, costs of data collection and analysis increase
- Effect of DP on denominators in death and other rates

Concern 1

- Effect of DP on multivariate analysis
  - Errors-in-variables problem
    - $y = x\beta$
    - $Y = y + e$ is observed
    - $X = x + u$ is observed
    - Standard analysis results in a biased estimator of β
    - If the Census Bureau actually implements DP, it must publish the covariance matrix of (e, u) and provide instruction to users on how to conduct correct analysis
  - General multivariate analysis
    - y is now a vector of statistics
    - $Y = y + e$ is released to the public
    - $\Sigma_{YY} = \Sigma_{yy} + \Omega_{ee}$
    - Correlations are depressed

Concern 1

- Propagation of the error injected under DP
  - Consider the estimated difference between two domains 1 and 2, e.g., compare housing density in Chicago and New York
    - $D = \frac{Y_1}{X_1} - \frac{Y_2}{X_2}$ with $Var\{D\} = O(4\sigma^2)$
    - $\Delta_t = D_t - D_{t-1}$ with $Var\{\Delta_t\} = O(8\sigma^2)$

Concern 2

- DP is concerned with the question of database reconstruction
  - With enough computing power, time, money, expertise, and motive, can a data intruder reconstruct person-level census records?
- Disclosure of new information about a census individual requires the data intruder have access to an external database (or equivalent)
- Here is the process of disclosure
  - The reconstructed census record: (X, Y)
  - The external database known to the data intruder: $(Name, X, Z)$
  - Following a match on X, the data intruder's merged result: $(Name, X, Y, Z)$
  - The data intruder now knows $Name$'s value of Y

Concern 2

- Consideration of DP requires consideration of various questions
  - What are potential external databases?
  - Are they available to the data intruder?
  - If an external database exists but is not available to the data intruder, has a disclosure occurred or is privacy at risk?
  - How do the resulting risks of disclosure balance against the loss of utility brought by DP?
- **Reconstruction does not necessarily imply identification!**

Concern 3

- Application of pure DP to skewed populations may result in unusable, worthless data
- Examples: manufacturers' shipments, household income
- Pure DP requires the standard error of noise e be large enough to protect the large respondents in the tail of the distribution
- Obliterates most of the information
- Leaves us working with the distribution of Y, which now contains virtually no information about the distribution of y

Concern 3

- With or without DP, privacy demands standard census practices must continue
  - Aggregation
  - Categorization or coarsening
  - Top-coding
- Future considerations -- $e \sim Laplace\left(0, ay^b\right)$ with $b \in \left[\frac{1}{2}, 2\right]$

Concern 4

- Census Bureau needs a DP communications strategy
- Test of DP on 2010 data and transparent release of the result for public review and comment

Questions

1. To what extent are census data already protected by the various errors they embody?
2. How does the Census Bureau think about application of DP to ACS data?

Question 1

- Response errors
- Nonresponse/imputation errors
- Coverage errors (gross undercounts and overcounts)
- Geocoding errors
- Given DP, the public now observes $Y = y + e$, where
  - $y = \mu + v$ is the raw, unadjusted census statistic
  - μ is the truth
  - v is the pooled value of all of the aforementioned census errors
  - e is the DP error

Question 2

- 1-year data are protected by aggregation across geography
- 5-year data are protected by aggregation across time
- Both are protected by sampling
- PUMS data are protected by both geographic aggregation and sampling

Summary

- Balancing the tension is critical
- DP is an old tool recently dressed up a bit, which has attracted the interest and energy of the computer science community
- DP succeeds in some cases, i.e., protects privacy and delivers useful statistics
- DP fails in some cases, i.e., protects privacy and delivers worthless statistics
- Even when DP succeeds, it nearly always must be supplemented by the Census Bureau's standard tools of disclosure protection
- It isn't clear at this hour whether DP is even necessary
- **Communication, transparency, further research, and testing are key**

Thank You!

NORC at the UNIVERSITY of CHICAGO

insight for informed decisions™