The Formal Privacy Research Agenda for Complex Survey Statistics

John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau

Joint Statistical Meetings 2020
Invited Session 147: Private Data for the Public Good: Formal Privacy in Survey Organizations
Tuesday, August 4, 2020 10:00am

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

Privacy Protection out of the Shadows

- As I have argued since 2015, traditional statistical disclosure limitation is broken and must be fixed
- I am unaware of a single argument against using formal privacy methods that does not apply *a fortiori* to traditional SDL methods as well
- Privacy practices for many current statistical products depend upon obfuscation
- 2020 Census Disclosure Avoidance System is the most transparent view into Census Bureau privacy practices ever
- That system and all future systems depend on constant feedback and interaction with the user community
- Those feedback mechanisms must be built into the overall statistical design

Lessons learned to date

- It is far easier to implement modern disclosure limitation when the alternative is full suppression: insist on modern methods for new products
- Sampling does not imply privacy protection by itself, you have to bound the information leakage first
- Tacking formal privacy onto the current design of official statistical products usually fails
- Formal privacy requires careful definition of the estimand and modification of the traditional estimators, *but so do all statistical disclosure limitation methods if they are to be used reliably*
- Understanding how to incorporate SDL into the statistical workload and how to evaluate the resulting estimators is the first order problem
- My thoughts are largely based on Abowd and Schmutte (*Brookings Papers on Economic Activity*, Spring 2015, https://www.brookings.edu/wp-content/uploads/2015/03/AbowdText.pdf) which gives a full Bayesian analysis of the problem in the online appendix here: https://digitalcommons.ilr.cornell.edu/ldi/24/

Complete data likelihood function

$$\mathcal{L}_\theta\left(\theta_p,\theta_D\,|Y,R\right)=p_Y\left(Y\,|\theta_p\right)p_{R|Y}\left(R\,|Y,\theta_D\right)=p_{YR}\left(Y,R\,|\theta_p,\theta_D\right)$$

Y = complete data matrix N x K (index: i,j)
θ_p = process parameters
R = inclusion matrix N x K (r_{ij} = 1 if y_{ij} is included in the sample design; 0, otherwise)
θ_D = design parameters

Observed data likelihood function

$$\pounds_{\theta}^{(obs)}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) = p_{Y^{(obs)}R}\left(Y^{(obs)}, R\left|\theta_p, \theta_D\right.\right)$$
$$= \int p_{YR}\left(Y, R\left|\theta_p, \theta_D\right.\right) dY^{(mis)}$$

$Y^{(obs)}$ = observed data elements N x K (index: i,j)
$Y^{(mis)}$ = missing data elements N x K (index: i,j)
These correspond to $r_{ij} = 1$ and $r_{ij} = 0$, respectively

Inference and estimation without SDL

$$p_{\theta_p \theta_D | Y^{(obs)} R}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) \propto p_{\theta_D|\theta_p}\left(\theta_D \left|\theta_p\right.\right) p_{\theta_p}\left(\theta_p\right) p_{Y^{(obs)} R}\left(Y^{(obs)}, R \left|\theta_p, \theta_D\right.\right)$$

$$p_{\theta_P | Y^{(obs)} R}\left(\theta_p \left| Y^{(obs)}, R\right.\right) = \int p_{\theta | Y^{(obs)} R}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) d\theta_D \qquad (A.5)$$

$$\propto \int\int p_Y\left(Y \left|\theta_p\right.\right) p_{R|Y}\left(R \left|Y, \theta_D\right.\right) p_{\theta_D|\theta_p}\left(\theta_D \left|\theta_p\right.\right) p_{\theta_p}\left(\theta_p\right) dY^{(mis)} d\theta_D$$

The data inclusion model is *ignorable* if

$$p_{\theta_P | Y^{(obs)} R}\left(\theta_p \left| Y^{(obs)}, R\right.\right) \equiv p_{\theta_P | Y^{(obs)}}\left(\theta_p \left| Y^{(obs)}\right.\right). \qquad (A.6)$$

Ignorability here covers ignorable sampling and/or missing data.

Published data likelihood function

$$\pounds_{\theta}^{(pub)}\left(\theta_p, \theta_D, \theta_S \,|Z, R\right) = \int p_{Z|YR}\left(Z\,|Y, R, \theta_S\right) p_{YR}\left(Y, R\,|\theta_p, \theta_D\right) dY$$

$$= \int p_{Z|YR}\left(Z\,|Y, R, \theta_S\right) p_{R|Y}\left(R\,|Y, \theta_D\right) p_Y\left(Y\,|\theta_p\right) dY$$

Z = published data matrix N x K (index: i,j)

θ_S = SDL parameters

Inference and estimation including SDL

$$
\begin{aligned}
p_{\theta|ZR}\left(\theta_p, \theta_D, \theta_S \,|Z, R\right) &\propto \int p_{Z|YR}\left(Z\,|Y, R, \theta_S\right) p_{YR}\left(Y, R\,|\theta_p, \theta_D\right) p_\theta\left(\theta\right) dY \\
&= p_\theta\left(\theta\right) \pounds_\theta^{(pub)}\left(\theta_p, \theta_D, \theta_S \,|Z, R\right),
\end{aligned}
$$

$$
p_{\theta_P|ZR}\left(\theta_p \,|Z, R\right) = \int\int p_{\theta|ZR}\left(\theta_p, \theta_D, \theta_S \,|Z, R\right) d\theta_D d\theta_S
$$

$$
p_{\theta_P|ZR}\left(\theta_p \,|Z, R\right) = \int p_{\theta_P|Y^{(obs)}R}\left(\theta_p \,\middle|Y^{(obs)}, R\right) p_{Y^{(obs)}|ZR}\left(Y^{(obs)}\,|Z, R\right) dY^{(obs)}
$$

The inference or estimation should be conditioned on Z and R, which are not the same as the confidential observed data $Y^{(obs)}$.

Shape your future START HERE >

United States Census 2020

Inference and estimation including SDL-II

We define *ignorable statistical disclosure limitation* as

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)}=Z,R\right.\right)\equiv p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right)$$

for all $Y^{(obs)}$, Z, and R.

If the model possesses both ignorable inclusion and ignorable SDL then

$$p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}=Z\right.\right)\equiv p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right)$$

SDL-aware inference and estimation

$$\begin{aligned} p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) &= p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right) \\ &= \int p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right) p_{Y^{(obs)}|Z}\left(Y^{(obs)}\left|Z\right.\right) dY^{(obs)} \end{aligned} \qquad \text{(A.12)}$$

$$p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right) = p_{\theta_p|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right) \qquad \text{(A.13)}$$

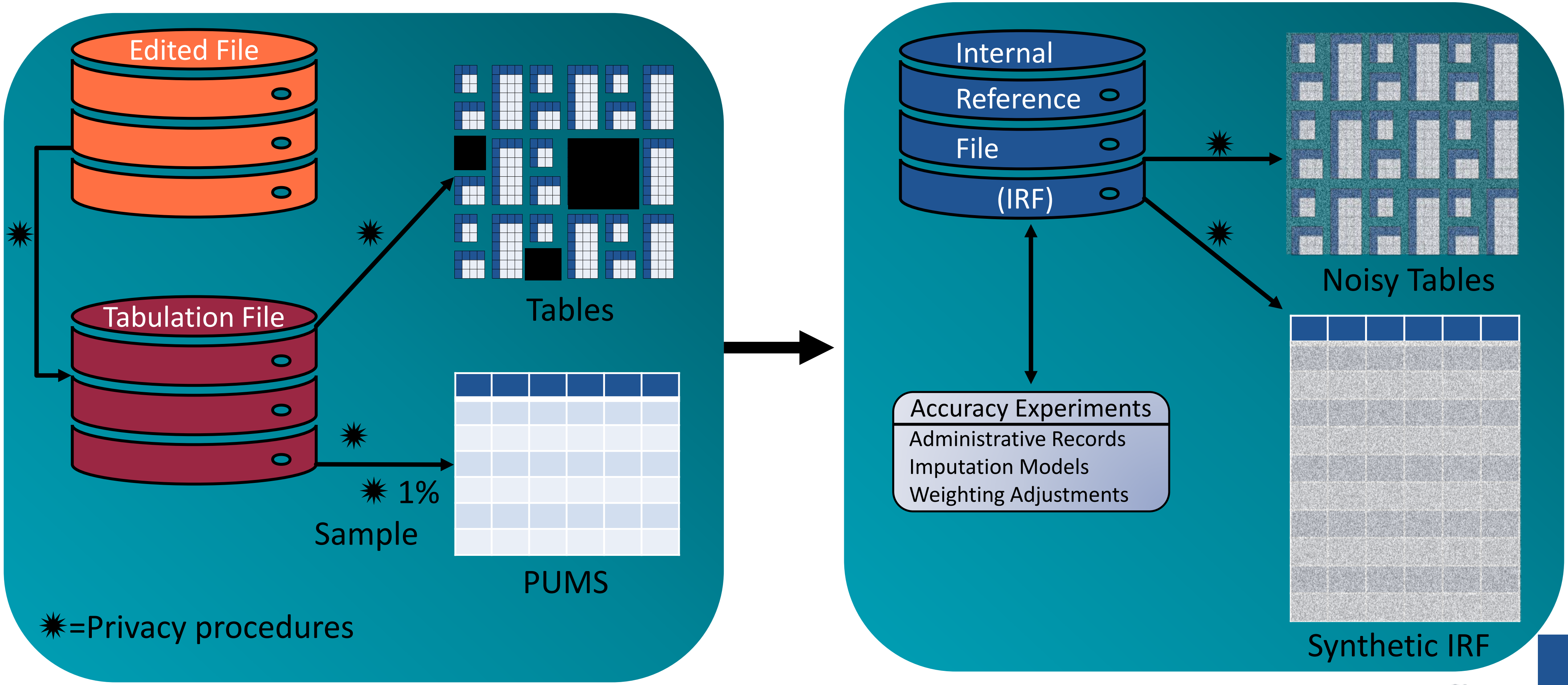
$$p_{\theta_S|ZR\theta_p\theta_D}\left(\theta_S\left|Z,R,\theta_p,\theta_D\right.\right) = p_{\theta_S|Z\theta_p}\left(\theta_S\left|Z,\theta_p\right.\right) \qquad \text{(A.14)}$$

and

$$p_{Y^{(obs)}|ZR\theta_p\theta_D\theta_S}\left(Y^{(obs)}\left|Z,R,\theta_p,\theta_D,\theta_S\right.\right) = p_{Y^{(obs)}|Z\theta_p\theta_S}\left(Y^{(obs)}\left|Z,\theta_p,\theta_S\right.\right). \qquad \text{(A.15)}$$

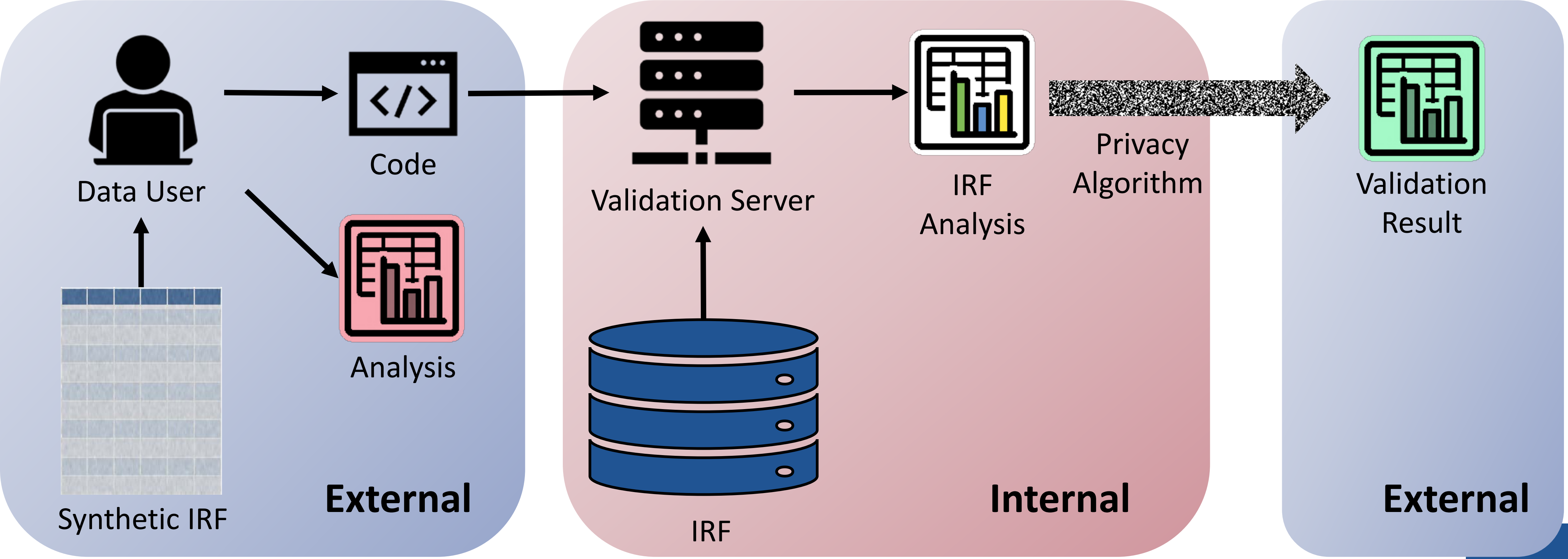
These MCMC equations implement non-ignorable SDL, assuming an ignorable, known inclusion mechanism (sampling probabilities are public).

Fewer privacy procedures allow for simpler and more adaptive workflows



Source: Rolando Rodríguez, U.S. Census Bureau, COPAFS/FCSM ACS Confidentiality and Data Access Webinar, July 23, 2020.

Data users will gain the ability to validate modeled output



Source: Rolando Rodríguez, U.S. Census Bureau, COPAFS/FCSM ACS Confidentiality and Data Access Webinar, July 23, 2020.

Thank you

John.Maron.Abowd@census.gov