Towards Good Statistical Inference from Differentially Private Data

Ruobin Gong

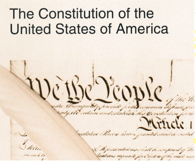
Rutgers University

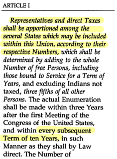
Department of Statistics
The University of Chicago

Nov 23, 2020

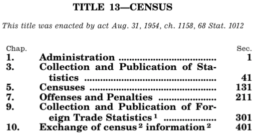
Modern data curators seek to meet two goals at once:

1. To **disclose** key statistics/use cases of the database, in accordance with its legal, policy, and/or ethical **mandates**.
2. To protect the **privacy** of individuals with provable guarantees.

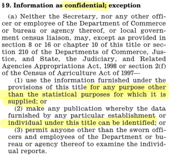



ARTICLE I

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, *three fifths of all other Persons.* The actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years, in such Manner as they shall by Law direct. The Number of


TITLE 13—CENSUS

This title was enacted by act Aug. 31, 1954, ch. 1158, 68 Stat. 1012


§ 9. Information as confidential; exception

(a) Neither the Secretary, nor any other officer or employee of the Department of Commerce or bureau or agency thereof, or local government census liaison, may, except as provided in section 8 or 16 or chapter 10 of this title or section 210 of the Departments of Commerce, Justice, and State, the Judiciary, and Related Agencies Appropriations Act, 1998 or section 2(f) of the Census of Agriculture Act of 1997—

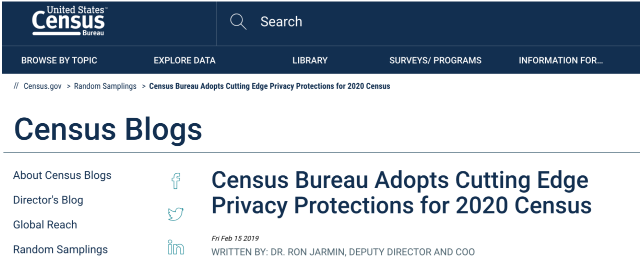
(1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(2) make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or

(3) permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports.

For example, the U.S. Census Bureau is constitutionally mandated to enumerate the population for apportionment every ten years. At the same time it is bound by Title 13 of U.S. Code to protect the confidentiality of respondents.

The U.S. Census Bureau Adopts Differential Privacy (Abowd, 2018)

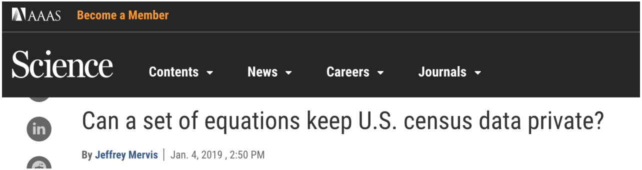


Census Blogs

Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census

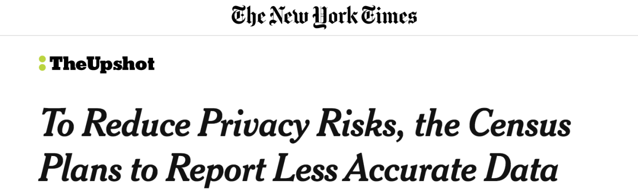
Fri Feb 15 2019

WRITTEN BY: DR. RON JARMIN, DEPUTY DIRECTOR AND COO



Can a set of equations keep U.S. census data private?

By **Jeffrey Mervis** | Jan. 4, 2019 , 2:50 PM



To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

- Statistical disclosure limitation mechanisms compliant with differential privacy guarantees protection with **provability**.

Definition. (Dwork et al., 2006). A random function M is *ϵ-differentially private* if for all neighboring datasets $\{(\mathcal{D}, \mathcal{D}') : d(\mathcal{D}, \mathcal{D}') = 1\}$,

$$Pr\left(M\left(\mathcal{D}'\right) \in A\right) \leq e^{\epsilon} Pr\left(M\left(\mathcal{D}\right) \in A\right),$$

for all $A \in \mathscr{B}(\mathbb{R}^p)$. Here, ϵ is called the *privacy loss budget.*

- The probabilistic specification of the mechanism M can be fully **transparent** (examples to follow), which makes good statistical inference possible.

But **possible** $\neq$ **easy**!

? Can data analysts adapt their statistical models to privatized data?

? Can data curators deliver the promised full transparency?

In their *Summary of the CNSTAT Workshop*, Hotz & Salvo (2020) wrote:

> *With some exceptions, applications demonstrated that the variability of small-area data (i.e., blocks, block groups, census tracts)* **compromised existing analyses***...[C]omparisons of 2010 Census data under the old DAS to 2020 Census data under DP may well show inexplicable trends.*

In addition, they noted:

> *...[U]nexpected issues with the* **post-processing** *of the proposed DAS, arising from the Bureau's need to hold some data cells* **invariant to change** *(e.g., total population at the state level).*

Do not ignore a privacy mechanism

Posit a simple linear model between two vector counts $(\boldsymbol{x}, \boldsymbol{y})$:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{e}.$$

OLS produces consistent estimators

$$\hat{\beta}_0 \longrightarrow \beta_0, \qquad \hat{\beta}_1 \longrightarrow \beta_1.$$

Protect the confidential counts with the ϵ-DP mechanism:

$$\boldsymbol{x}_{\text{dp}} = \boldsymbol{x} + \boldsymbol{u}, \qquad \boldsymbol{y}_{\text{dp}} = \boldsymbol{y} + \boldsymbol{v}, \quad \boldsymbol{u}, \boldsymbol{v} \sim Lap_n\left(\epsilon^{-1}\right)$$

Naïvely fitting the original model to differentially privatized data

$$\boldsymbol{y}_{\text{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\text{dp}} + \boldsymbol{e},$$

results in least squares estimates that are both biased (attenuated):

$$\hat{b}_0 \rightarrow \beta_0 \underbrace{+ \left(1 - \frac{\mathbb{V}(x)}{\mathbb{V}(x) + \sigma_u^2}\right) \mathbb{E}(x)\, \beta_1,} \qquad \hat{b}_1 \rightarrow \beta_1 \underbrace{- \frac{\sigma_u^2}{\mathbb{V}(x) + \sigma_u^2} \beta_1,}$$

with inflated regression residual variance

$$\mathbb{V}\left(\boldsymbol{y}_{\text{dp}} \mid \boldsymbol{x}_{\text{dp}}\right) = \sigma^2 \underbrace{+ \beta_1^2 \sigma_u^2 + \sigma_v^2}.$$

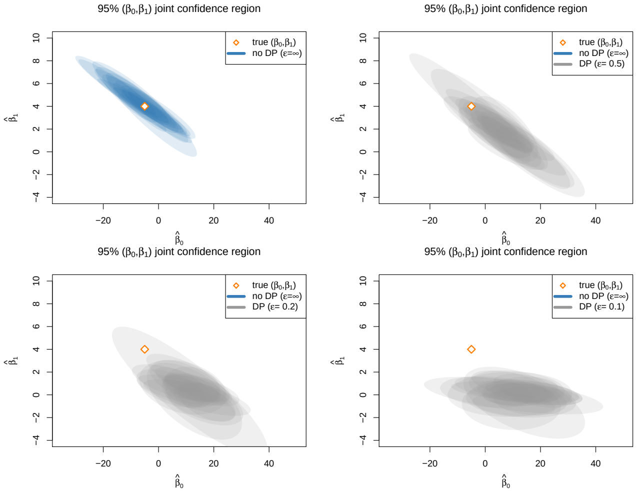


Figure: Naïve fitting with $\mathbf{x}_i \sim \mathit{Pois}(10)$, $\mathbf{y}_i = -5 + 4\mathbf{x}_i + \mathbf{e}_i$, $\mathbf{e}_i \sim N\left(0, 5^2\right)$, $n = 10$, at privacy budget levels $\epsilon = 0.5, 0.2, 0.1$, and ∞ (no privacy). Smaller ϵ induces more misguided confidence regions for (β_0, β_1). Each panel depicts 20 simulations.

Transparent privacy is principled privacy

A model adequate for the confidential data $\boldsymbol{s} = (\boldsymbol{x}, \boldsymbol{y})$ will almost certainly be inadequate if naïvely fitted to the privatized data $\boldsymbol{s}_{\text{dp}} = (\boldsymbol{x}_{\text{dp}}, \boldsymbol{y}_{\text{dp}})$:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{e} \qquad \not\Longrightarrow \qquad \boldsymbol{y}_{\text{dp}} = \beta_0 + \beta_1 \boldsymbol{x}_{\text{dp}} + \boldsymbol{e}.$$

Instead, augment the original model to include the privacy mechanism:

Likelihood for $\boldsymbol{\beta}$ based on privatized data $\boldsymbol{s}_{\text{dp}}$ is integrated over the confidential data $\boldsymbol{s}$, with respect to the privacy mechanism:

$$L(\boldsymbol{\beta}; \boldsymbol{s}_{\text{dp}}) = \int \underbrace{\eta_{\text{dp}}(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s})}_{\text{privacy mechanism}} \underbrace{\pi(\boldsymbol{s} \mid \boldsymbol{\beta})}_{\text{original model}} \partial \boldsymbol{s}$$

- The full transparency of $\eta_{\text{dp}}(\cdot \mid \cdot)$ is necessary to ensure unbiased inference for all questions for $\boldsymbol{\beta}$ using private data;
- Inference that respect $L(\boldsymbol{\beta}; \boldsymbol{s}_{\text{dp}})$ will be termed *exact*.

On behalf of data users...

How can data analysts adapt their statistical models to privatized data ?

- **Approximate computation** (ABC, Monte Carlo EM) can be systematically adapted to produce exact statistical inference with transparent privacy;

How can data curators deliver the promised full transparency ?

- If **invariants** are mandated, the congenial design of the privacy mechanism via **standard probabilistic conditioning** is a principled alternative to optimization-based post-processing.

Approximate Bayesian computation (ABC)

A Bayesian model is posited:

- prior: $\theta \sim \pi_0(\theta)$
- likelihood: $\boldsymbol{x} \mid \theta \sim \pi(\boldsymbol{x} \mid \theta)$
- posterior:

$$\pi\left(\theta \mid \boldsymbol{x}\right) \propto \pi_0\left(\theta\right) \pi\left(\boldsymbol{x} \mid \theta\right)$$

Sampling from the posterior via Monte Carlo requires that it at least can be evaluated. This is not the case for complex models (e.g. intractable or implicit likelihood).

Approximate Bayesian computation (ABC)

Algorithm 1

```
Input: observed data $\boldsymbol{x}_0$, integer $N > 0$;
Iterate: for $i = 1, \ldots, N$:
    step 1, simulate $\theta_i \sim \pi_0(\theta)$;
    step 2, simulate $\boldsymbol{x}_i \sim \pi(\boldsymbol{x} \mid \theta_i)$;
    step 3, accept $\theta_i$ if $\boldsymbol{x}_i = \boldsymbol{x}_0$, otherwise go to step 1;
Output: a set of parameter values $\{\theta_i\}_{i=1}^{N}$.
```

Algorithm 1 draws $\theta_i \sim \pi(\theta \mid \boldsymbol{x}_0)$, i.i.d.

Exact matching $\boldsymbol{x}_i = \boldsymbol{x}_0$ isn't practical, if $\boldsymbol{x}_0$ is high-dimensional or continuous.

Approximate Bayesian computation (ABC)

Algorithm 2

```
Input: observed summary data $\boldsymbol{s}_0 = \boldsymbol{s}(\boldsymbol{x}_0)$, integer $N > 0$,
    a kernel density $\eta$ with bandwidth $h > 0$;
Iterate: for $i = 1, \ldots, N$:
    step 1, simulate $\theta_i \sim \pi_0(\theta)$;
    step 2, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s}(\boldsymbol{x}) \mid \theta_i)$;
    step 3, accept $\theta_i$ with probability $c\eta\left((\boldsymbol{s}_i - \boldsymbol{s}_0)/h\right)$
        where $c^{-1} = \max\{\eta(\cdot)\}$, otherwise go to step 1;
Output: a set of parameter values $\{\theta_i\}_{i=1}^{N}$.
```

Two layers of approximation

$$\theta_i \sim \pi_{ABC}\left(\theta \mid \boldsymbol{s}_0\right) \xleftrightarrow{\eta(\cdot) \text{ and } h} \pi\left(\theta \mid \boldsymbol{s}_0\right) \xleftrightarrow{\boldsymbol{s}(\cdot)} \pi\left(\theta \mid \boldsymbol{x}_0\right)$$

Adapting a Bayesian model to DP queries

The Bayesian model is modified to:

- prior: $\theta \sim \pi_0(\theta)$
- confidential query likelihood: $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$
- privacy mechanism $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}, \cancel{\theta} \sim \eta_{\text{dp}}\left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}\right)$ ← ignorability

- observed/private posterior:

$$\pi\left(\theta \mid \boldsymbol{s}_{\text{dp}}\right) = \frac{\pi_0\left(\theta\right) \int \eta_{\text{dp}}(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s}}{\int \pi_0\left(\theta\right) \int \eta_{\text{dp}}(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s} d\theta}.$$

Adapting a Bayesian model to DP queries

The Bayesian model is modified to:

- prior: $\theta \sim \pi_0(\theta)$
- confidential query likelihood: $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$
- privacy mechanism (additive): $\boldsymbol{s}_{\mathrm{dp}} \mid \boldsymbol{s}, \cancel{\theta} \sim \eta\left(\left(\boldsymbol{s}_{\mathrm{dp}} - \boldsymbol{s}\right)/h\right)$

$$\boldsymbol{s}_{\mathrm{dp}}\left(\mathcal{D}\right) := \boldsymbol{s}\left(\mathcal{D}\right) + h\boldsymbol{u}, \qquad \boldsymbol{u} \sim \eta,\ \ h = h\left(\epsilon, \delta, \boldsymbol{s}\right) > 0$$

- observed/private posterior:

$$\pi\left(\theta \mid \boldsymbol{s}_{\mathrm{dp}}\right) = \frac{\pi_0\left(\theta\right) \int \eta\left(\left(\boldsymbol{s}_{\mathrm{dp}} - \boldsymbol{s}\right)/h\right) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s}}{\int \pi_0\left(\theta\right) \int \eta\left(\left(\boldsymbol{s}_{\mathrm{dp}} - \boldsymbol{s}\right)/h\right) \pi\left(\boldsymbol{s} \mid \theta\right) d\boldsymbol{s} d\theta}.$$

ABC produces exact posterior draws for DP data

Algorithm 3

```
Input: private query $\boldsymbol{s}_{\text{dp}}$, integer $N > 0$, perturbation
        mechanism w/ density $\eta$ and bandwidth $h(\epsilon, \delta, \boldsymbol{s}) > 0$;
Iterate: for $i = 1, \ldots, N$:
    step 1, simulate $\theta_i \sim \pi_0(\theta)$;
    step 2, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s} \mid \theta_i)$;
    step 3, accept $\theta_i$ with probability $c\eta\left((\boldsymbol{s}_{\text{dp}} - \boldsymbol{s}_i)/h\right)$
            where $c^{-1} = \max\{\eta(\cdot)\}$, otherwise go to step 1;
Output: a set of parameter values $\{\theta_i\}_{i=1}^{N}$.
```

Theorem (G. 2019)

Algorithm 3 draws $\theta_i \sim \pi(\theta \mid \boldsymbol{s}_{\text{dp}})$, i.i.d.

- Noisy ABC (Fearnhead & Prangle, 2012);
- ABC under the assumption of model error (Wilkinson, 2013).

Exact likelihood inference with Monte Carlo EM

Expectation-Maximization (Dempster et al., 1977) for private data:

- complete data is $(\boldsymbol{s}, \boldsymbol{s}_{\text{dp}})$;
- missing data is $\boldsymbol{s}$, where $\boldsymbol{s} \mid \theta \sim \pi(\boldsymbol{s} \mid \theta)$; ← data analyst
- observed data is $\boldsymbol{s}_{\text{dp}}$, where $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s} \sim \eta_{\text{dp}}(\cdot \mid \boldsymbol{s})$. ← data curator

Iterate till convergence:

- E-step:

$$\begin{aligned} Q(\theta; \theta^{(t)}) &= \mathbb{E}\left(\log L(\theta; \boldsymbol{s}, \boldsymbol{s}_{\text{dp}}) \mid \boldsymbol{s}_{\text{dp}}, \theta^{(t)}\right) \\ &= \mathbb{E}\left(\log \pi(\boldsymbol{s} \mid \theta) \mid \boldsymbol{s}_{\text{dp}}, \theta^{(t)}\right) + \text{const.} \end{aligned}$$

- M-step:

$$\theta^{(t+1)} = \operatorname{argmax}_\theta Q(\theta; \theta^{(t)}).$$

Exact likelihood inference with Monte Carlo EM

E-Step via Importance Sampling

```
Iterate: for $i = 1, \ldots, N$:
    step 1, simulate $\boldsymbol{s}_i \sim \pi(\boldsymbol{s} \mid \theta^{(t)})$;        ← data analyst
    step 2, assign weight $\omega_i = \eta_{\text{dp}}(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s}_i)$;        ← data curator
Output: a set of weighted samples $\{\boldsymbol{s}_i, \omega_i\}_{i=1}^N$.
```

$$\sum_{i=1}^{N} \omega_i b(\boldsymbol{s}_i) / \sum_{i=1}^{N} \omega_i \xrightarrow{p} \mathbb{E}\left(b(\boldsymbol{s}) \mid \boldsymbol{s}_{\text{dp}}, \theta^{(t)}\right), \qquad \text{as } N \to \infty.$$

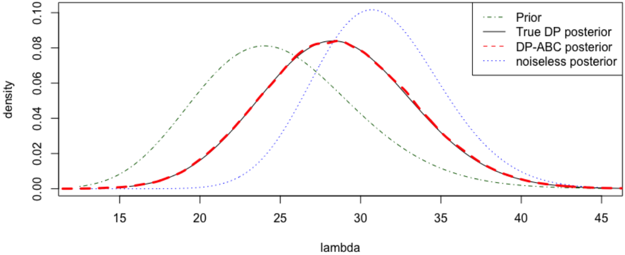
Take $b(\boldsymbol{s})$ to be:

- sufficient statistic for θ, if $\pi(\boldsymbol{s} \mid \theta)$ is exponential family;
- $\log \pi(\boldsymbol{s} \mid \theta)$ in general;
- $\nabla_\theta \log \pi(\boldsymbol{s} \mid \theta)$ and $\nabla_\theta^2 \log \pi(\boldsymbol{s} \mid \theta)$, towards estimating observed score function and Fisher information.

Numerical example: privatized count data

- Likelihood for confidential query $\boldsymbol{s} \mid \theta \sim \mathit{Pois}\,(\theta)$;
- Privacy mechanism $\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{s} \sim \epsilon^{-1} \mathit{Lap}(1)$ with $\epsilon = 0.2$;
- With prior $\theta \sim \mathit{Gamma}\,(25, 1)$, exact posterior is

$$\pi\left(\theta \mid \boldsymbol{s}_{\text{dp}}\right) \propto \theta^{\alpha-1} e^{-(\beta+1)\theta} \left[\frac{\Gamma\left(\left\lceil \boldsymbol{s}_{\text{dp}} \right\rceil, \theta_{\epsilon}^{+}\right)}{\Gamma\left(\left\lceil \boldsymbol{s}_{\text{dp}} \right\rceil\right)} e^{\theta_{\epsilon}^{+} - \epsilon \boldsymbol{s}_{\text{dp}}} + \frac{\gamma\left(\left\lceil \boldsymbol{s}_{\text{dp}} \right\rceil, \theta_{\epsilon}^{-}\right)}{\Gamma\left(\left\lceil \boldsymbol{s}_{\text{dp}} \right\rceil\right)} e^{\theta_{\epsilon}^{-} + \epsilon \boldsymbol{s}_{\text{dp}}} \right]$$



- Monte Carlo EM gives $\hat{\theta}_{\text{dp}} = 37.237$, $\hat{I}_{\text{dp}} = 1.582 \times 10^{-2}$;
- If naïvely treat $\boldsymbol{s}_{\text{dp}} = 37.4$ as confidential: $\hat{I} = 2.674 \times 10^{-2} \approx 169\% \times \hat{I}_{\text{dp}}$.

Takeaway

The analogy at play here:

$$\text{approximate computation on exact data}$$
$$\Updownarrow$$
$$\text{exact computation on approximate data}$$

The statistical tradeoff (utility vs privacy) becomes aligned with the computational tradeoff (approximation vs exactness).

. . .

Up next – Invariants: the roadblock to fully transparent privacy mechanisms

Privacy under mandated disclosure

- The data curator may be mandated to release privatized data congruent with **invariants**, a set of exact statistics computed from the confidential micro-data (Ashmead et al., 2019).
  - e.g. state population total, voting age population total, total housing units, non-negativity of counts, etc.
- Let $\mathcal{X}^* \subset \mathcal{X}$ be the set of dataset x's that obey the invariants. It restricts the value the query must satisfy to be

  $$\mathcal{S}^* = \left\{ s\left(x\right) \in \mathbb{R}^d : x \in \mathcal{X}^* \right\}.$$

  Note, $\mathcal{S}^* : \mathcal{X}^* \to \sigma\left(\mathbb{R}^d\right)$ is a *set-valued function* of the confidential dataset x^*.
- Challenge: how to find a mechanism, $\tilde{M}$, that satisfies the mandated disclosure:

  $$\tilde{M}\left(x\right) \in \mathcal{S}^*, \quad \forall x \in \mathcal{X}^*,$$

  while preserving the **provability** and **transparency** of differential privacy?

Curator’s post-processing may not be innocent post-processing

A common practice to construct invariant-respecting mechanisms is through **optimization-based post-processing** of an otherwise unconstrained ϵ-DP mechanism M:

$$f\left(M; \mathcal{S}^{*}\right) := \operatorname{argmin}_{s \in \mathcal{S}^{*}} \Delta\left(M, s\right),$$

for Δ some discrepancy measure.

This operation does **not** fall under the **post-processing theorem**!

Curator's post-processing may not be innocent post-processing

Theorem (Dwork & Roth, 2014)

If M is an ϵ-DP mechanism and g an arbitrary function, then $g \circ M$ is also ϵ-DP.

- Indeed for any g-measurable set B, the two events

$$g(M(x)) \in B \;\Leftrightarrow\; M(x) \in g^{-1}(B)$$

are equivalent, thus equiprobable according to M.

- However, $\mathcal{S}^* : \mathcal{X}^* \to \sigma\left(\mathbb{R}^d\right)$ is a *set-valued function* of the confidential data. For an f-measurable set B and any $x \in \mathcal{X}$, the equivalent events are

$$f\left(M(x); \mathcal{S}^*(x)\right) \in B \;\Leftrightarrow\; M(x) \in f^{-1}\left(B; \mathcal{S}^*(x)\right),$$

noting that the inverse map f^{-1} now depends on x.

- For x, x' neighboring datasets, the f-probability ratio

$$\frac{P\left(f(x) \in B\right)}{P\left(f(x') \in B\right)} = \frac{P\left(M(x) \in f^{-1}(B; \mathcal{S}^*(x))\right)}{P\left(M(x') \in f^{-1}(B; \mathcal{S}^*(x'))\right)}$$

may not be $\exp(\pm\epsilon)$-bounded.

Invariant-respecting DP design via standard probabilistic conditioning

Let M be an unconstrained ϵ-differentially private mechanism based on the deterministic query $s : \mathcal{X} \to \mathbb{R}^d$, and $\mathcal{S}^*$ be the set of query values conformal to the invariant $\mathcal{X}^*$. Propose an invariant-respecting mechanism M^*, such that

$$\boxed{M^*(x) \overset{L}{=} M(x) \mid M(x) \in \mathcal{S}^*.}$$

Theorem (G. and Meng, 2020)

For all $(x, x') \in \mathcal{X}^* \times \mathcal{X}^*$ such that $\mathrm{d}(x, x') = k$, there exists a real-valued $\gamma \in [-1, 1]$ such that for all $B \in \mathscr{B}(\mathbb{R}^d)$,

$$P(M^*(x) \in B) \leq \exp((1+\gamma)k\epsilon)\, P\left(M^*(x') \in B\right).$$

Note. The theorem always holds when taking $\gamma = 1$. For specific M and $\mathcal{S}^*$, it is possible to find $\gamma \in [-1, 1)$. Examples are available for which $\gamma = 0, -1/2, -1$.

Conditioning preserves statistical intelligibility

The meaning of unconditional DP. Let $\{M_\epsilon : \epsilon > 0\}$ be a class of ϵ-DP mechanisms. For every $B \in \mathscr{B}\left(\mathbb{R}^d\right)$ and every prior π the analyst harbors about x_i,

$$\pi\left(x_i = \omega \mid M_\epsilon\left(x\right) \in B\right) \in \left[\exp\left(-\epsilon\right)\pi\left(x_i = \omega\right),\ \exp\left(\epsilon\right)\pi\left(x_i = \omega\right)\right].$$

The meaning of conditional DP. Let $\{M^*_\epsilon : \epsilon > 0\}$ be a class of ϵ-DP mechanism that respect the invariant $\mathscr{X}^*$. For all $x \in \mathscr{X}^*$ such that $\exists x' \in \mathscr{X}^*$ so that $\mathrm{d}\left(x, x'\right) = 1$, and $\forall B \in \mathscr{B}\left(\mathcal{S}^*\right)$, the analyst's posterior probability

$$\pi\left(x_i = \omega \mid x \in \mathscr{X}^*, M^*_\epsilon\left(x\right) \in B\right) \in$$
$$\left[\exp\left(-\left(1+\gamma\right)\epsilon\right)\pi\left(x_i = \omega \mid x \in \mathscr{X}^*\right), \exp\left(\left(1+\gamma\right)\epsilon\right)\pi\left(x_i = \omega \mid x \in \mathscr{X}^*\right)\right].$$

The structural resemblance between the two statements is due to the conditional nature of M^*, which allows for statistical information from privacy mechanisms (constrained or otherwise) to be interpreted in the same – hence congenial – way.

A Monte Carlo implementation

We demonstrate a generic Metropolized Independent Sampler (MIS; Liu, 1996) for when the mandated invariants are expressed as a consistent system of linear equalities and inequalities

$$\mathcal{S}^* = \left\{ s \in \mathbb{R}^d : As = a, Bs \geq b \right\}.$$

The proposal is applicable to both discrete and continuous privatization schemes, and does not require knowing the normalizing constant.

Metropolized Independent Differentially Private Sampler with Invariants

```
Input: unconstrained privacy mechanism $p$,
       confidential query $s^*$, invariant parameters $(A, a, B, b)$,
       proposal distribution $q$, proposal index set $\mathcal{I}$,
       initial value $s^{(0)} \in \mathcal{S}^*$, integer nsim;
Iterate: for $t = 0, 1, \ldots, \text{nsim} - 1$, at $t + 1$:
    step 1, propose $\tilde{s}$:
        1-1. sample $\tilde{s}_{\mathcal{I}} \sim q$;
        1-2. solve for $\tilde{s}_{\mathcal{I}^c}$ in $A_{[\mathcal{I}]}\tilde{s}_{\mathcal{I}} + A_{[\mathcal{I}^c]}\tilde{s}_{\mathcal{I}^c} = a$;
        1-3. write $\tilde{s} = (\tilde{s}_{\mathcal{I}}, \tilde{s}_{\mathcal{I}^c})$;
    step 2, compute $\alpha(s^{(t)}, \tilde{s}) = \min\left\{1, \frac{p(\tilde{s})\mathbf{1}(B\tilde{s} \geq b)q\left(s_{\mathcal{I}}^{(t)}\right)}{p\left(s^{(t)}\right)q(\tilde{s}_{\mathcal{I}})}\right\}$;
    step 3, set $s^{(t+1)} = \tilde{s}$ with probability $\alpha(s^{(t)}, \tilde{s})$,
        otherwise set $s^{(t+1)} = s^{(t)}$.
Output: a set of draws $\{s^{(t)}\}$, $t = 1, \ldots, \text{nsim}$.
```

Example: sex-by-age contingency table with invariants

	< 5	6-10	11-15	16-17	18-19	20	21	22-24	25-29	30-34	35-39	40-44	45-49	50-54
Female	8	6	3	6	4	4	4	8	5	7	7	6	1	5
Male	3	4	5	8	6	4	5	5	5	6	10	7	3	2
					55-59	60-61	62-64	65-66	67-69	70-74	75-79	80-84	85+	Total
Female					4	4	9	6	2	8	8	8	7	**130**
Male					5	11	6	4	7	4	5	3	8	**126**
Voting				**43**									**213**	**256**

	< 5	6-10	11-15	16-17	18-19	20	21	22-24	25-29	30-34	35-39	40-44	45-49	50-54
Female	6	6	4	3	2	10	2	5	6	7	5	6	0	5
Male	9	4	5	6	3	4	5	5	3	4	7	8	5	3
					55-59	60-61	62-64	65-66	67-69	70-74	75-79	80-84	85+	Total
Female					4	5	10	8	3	8	8	12	5	**130**
Male					2	23	8	2	6	3	0	3	8	**126**
Voting				**43**									**213**	**256**

A confidential sex × age contingency table (top) and its corresponding constrained DP-release (bottom), subject to 1) total population, 2) proportion female population, 3) voting age population, and 4) nonnegative integer constraints. Constructed using a Double Geometric mechanism w/ $\epsilon = 0.5$ per cell.

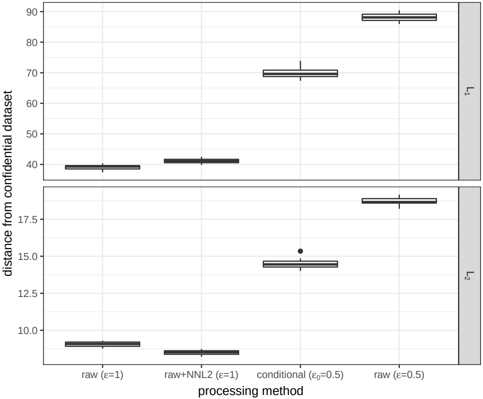


Figure: Average L_1 (top) and L_2 (bottom) distances of a simulated confidential dataset from its privatized releases using four processing methods: a) raw Double Geometric w/ $\epsilon = 1$; b) raw ($\epsilon = 1$) followed by nonnegative L_2 onto $\mathcal{S}^*$; c) proposed $\mathcal{S}^*$-conditional algorithm w/ $\epsilon_0 = 0.5$; d) raw Double Geometric w/ $\epsilon = 0.5$. Note that this comparison does not suggest relative accuracy between the nonnegative L_2 and the conditional mechanisms, as the effective privacy guarantee that either mechanism enjoys is undetermined.

Future Research

- Finding better γ: effective privacy determination for given M and M^*
- Combating computational complexity without compromising privacy, e.g. when non-perfect Markov chain Monte Carlo is employed
- The real dark side of invariants is its ability to *disrupt*, even *destroy*, the neighborhood structure of databases, i.e.

$$\left\{\left(x, x'\right) \in \mathcal{X}^* \times \mathcal{X}^* : \mathrm{d}\left(x, x'\right) = 1\right\} = \emptyset,$$

and

$$\left\{\left(x, x'\right) \in \mathcal{X}^* \times \mathcal{X}^* : \mathrm{d}\left(x, x'\right) \geq 1\right\} = \emptyset.$$

How to prevent privacy from becoming a (near)-vacuous promise?

Thank You

Paper/preprints associated with this talk are:

- Gong, R. (2019). Exact Inference with Approximate Computation for Differentially Private Data via Perturbations. *ArXiv:1909.12237*
- Gong, R. (2020). Transparent Privacy is Principled Privacy. *ArXiv:2006.08522*
- Gong, R. & Meng, X.-L. (2020). Congenial Differential Privacy under Mandated Disclosure. *ACM-IMS Foundations of Data Science Conference (FODS-2020)*

I gratefully acknowledge research support by the National Science Foundation.

Abowd, J. M. (2018). The us census bureau adopts differential privacy. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2867–2867).

Ashmead, R., Kifer, D., Leclerc, P., Machanavajjhala, A., & Sexton, W. (2019). *Effective privacy after adjusting for invariants with applications to the 2020 census* (Tech. Rep.). US Census Bureau.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265–284).

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474.

Hotz, V. J., & Salvo, J. (2020, 2 25). *Assessing the use of differential privacy for the 2020 Census: Summary of what we learned from the CNSTAT workshop* (Tech. Rep.). National Academies Committee on National Statistics (CNSTAT). (https://www.amstat.org/asa/files/pdfs/POL-CNSTAT_CensusDP_WorkshopLessonsLearnedSummary.pdf [Accessed: 04-08-2020])

Liu, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and computing*, *6*(2), 113–119.

Wilkinson, R. D. (2013). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *Statistical applications in genetics and molecular biology*, *12*(2), 129–141.

In case you ask...

Monte Carlo EM for simple linear regression

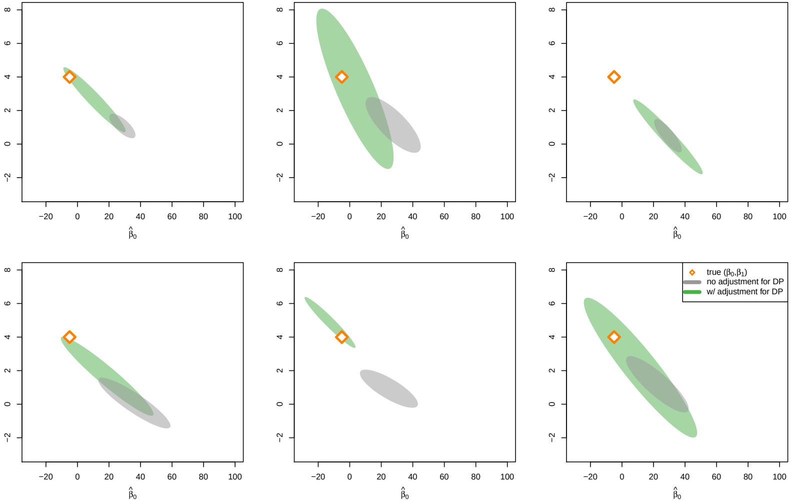
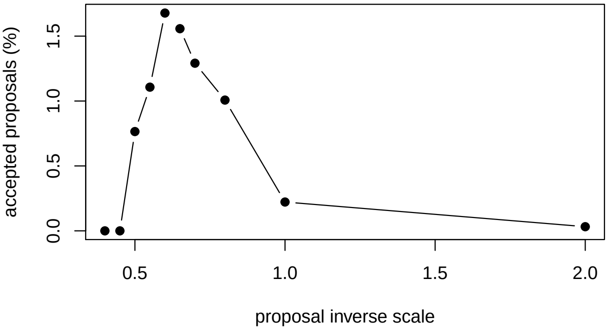


Figure: Correct model (green) fitted via Monte Carlo EM vs. naïve model (gray) on six instances of DP protected datasets ($\epsilon = 0.2$). Displayed 95% confidence ellipses are based on normal approximations at the MLE.

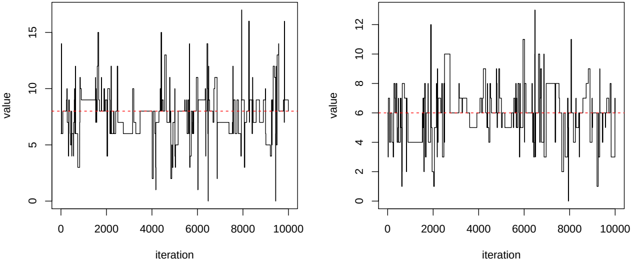
Performance Diagnostics of the MIS Algorithm

The acceptance rate is shown as a function of the proposal inverse scale parameter $\tilde{\epsilon}$. The acceptance rate is the highest in this example when $\tilde{\epsilon}$ is set to 0.6, just slightly larger than the privacy loss budget of the unconstrained privacy mechanism ($\epsilon = 0.5$ per cell). The acceptance rate achieved is about 1.68%.



Performance Diagnostics of the MIS Algorithm

Traceplots of $10,000$ draws from the Algorithm of respectively the second (in proposal index set $\mathcal{I}$) and the first (not in proposal index set $\mathcal{I}$) cells of the constrained differentially private contingency table, when $\tilde{\epsilon} = 0.6$.



A case of $\gamma = 0$: trivial invariants

Consider the trivial case where the set of invariants does not actually impose any restriction, i.e., $\mathcal{X}^* = \mathcal{X}$. It is then necessarily true that $\mathcal{S}^* = \mathcal{S}$, and the "constrained" differentially private mechanism is identical in distribution to the unconstrained one: $M^* \overset{L}{=} M$. In this case, $\gamma = 0$ and M^* is ϵ-differentially private.

A case of $\gamma = -1$: rounding

Let x be an indicator vector of length n, and the query is

$$s(x) = \lceil \sum x_i / 10 \rceil .$$

Suppose the invariant set mandated for disclosure is

$$\mathcal{X}^* = \left\{ (x_1, \ldots, x_n) \in \{0,1\}^n : \sum x_i \in [41, 50] \right\},$$

or equivalently, $\mathcal{S}^* = \{5\}$. For any M, the implied constrained privacy mechanism M^* is equivalent to a degenerate distribution: $P(M^*(x) = 5) = 1$ for all $x \in \mathcal{X}^*$. Furthermore, for all neighboring datasets $(x, x') \in \mathcal{X}^* \times \mathcal{X}^*$, and any B a measurable subset of $\mathbb{N}$,

$$P(M^*(x) \in B) = \exp(0)\, P\left(M^*\left(x'\right) \in B\right) = \begin{cases} 1 & \text{if } 5 \in B \\ 0 & \text{otherwise.} \end{cases}$$

Therefore in this particular instance, M^* is in fact 0-differentially private, corresponding to $\gamma = -1$, i.e. for those databases conformal to the invariant $\mathcal{X}^*$, M^* supplies no discriminatory information among them whatsoever.

A case of $\gamma = -1/2$: two-bin histogram with fixed sum

Suppose x is a binary vector, and the query $s = (s_1(x), s_2(x))$ tabulates the 0s and 1s in x. Employ the Laplace mechanism as the unconstrained mechanism, i.e.

$$M(x) = (m_1 = s_1 + u_1, m_2 = s_2 + u_2), \; u_i \overset{i.i.d.}{\sim} Lap(2\epsilon^{-1}),$$

expending a total of ϵ budget. Suppose that

$$\mathcal{S}^*(x) = \left\{(a_1, a_2) \in \mathbb{R}^2 : a_1 + a_2 = s_1(x) + s_2(x)\right\}.$$

The congenial $M^*(x)$ follows the conditional distribution $(s_1 + u_1, s_2 + u_2) \mid u_1 + u_2 = 0$, which has density

$$p(m_1 = m, m_2 = n - m) = \frac{\epsilon}{2} \exp\{-\epsilon |m - s_1|\},$$

which is equivalent to drawing a u from $Lap(\epsilon^{-1})$ and release $(m_1 = s_1 + u, m_2 = s_2 - u)$, thus maintaining the same ϵ guarantee as the unconstrained mechanism. However, when the sum is fixed, k must be set to 2 or greater because the nearest neighboring vectors (x, x') must have $\mathrm{d}(x, x') = 2$. Hence the ϵ privacy bound implies $k(1 + \gamma) = 2(1 + \gamma) = 1$, yielding $\gamma = -0.5$.