Data Augmentation MCMC for Bayesian Inference from Privatize Data[1]

Ruobin Gong

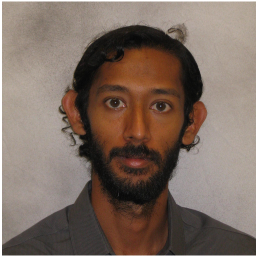
Rutgers University

July 25, 2022

Workshop on Differential Privacy and Statistical Data Analysis
Fields Institute

[1] Ju, Awan, G., & Rao. (2022). ArXiv:2206.00710.

Collaborators



Nianqiao (Phyllis) Ju
Purdue University

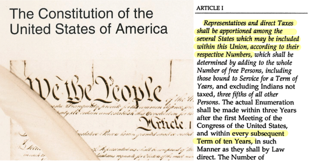
Jordan A. Awan
Purdue University

Vinayak Rao
Purdue University

Privacy: a challenge in modern data curation

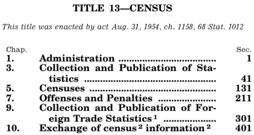
Modern data curators seek to meet two goals at once:

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 trust-worthy guarantees.


The Constitution of the United States of America

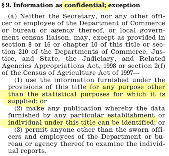
ARTICLE I

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

Chap.		Sec.
1.	Administration	1
3.	Collection and Publication of Statistics	41
5.	Censuses	131
7.	Offenses and Penalties	211
9.	Collection and Publication of Foreign Trade Statistics [1]	301
10.	Exchange of census [2] information [2]	401


§ 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—

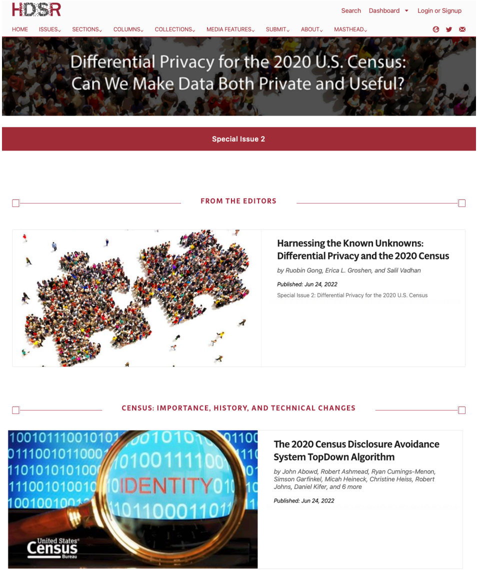
(1) use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or

(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.

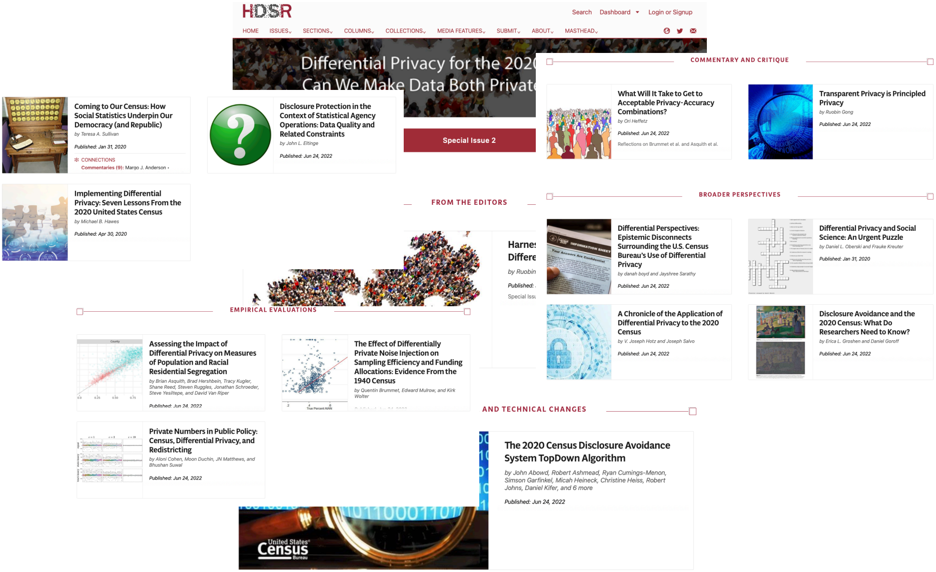
For example, the U.S. Census Bureau bears the constitutional mandate to enumerate the population every 10 years for apportionment. It is also bound by Title 13 of U.S. Code to protect respondent confidentiality.

The U.S. Census Bureau adopts differential privacy



Harvard Data Science Review (https://hdsr.mitpress.mit.edu)

The U.S. Census Bureau adopts differential privacy

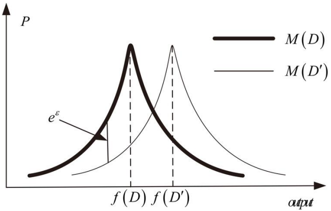


Harvard Data Science Review (https://hdsr.mitpress.mit.edu)

The mechanism of differential privacy

A random function $s_{\mathrm{dp}}(\boldsymbol{x}, \boldsymbol{r})$ is said to be ϵ-differentially private[2] if for all *neighboring* databases $(\boldsymbol{x}, \boldsymbol{x}')$ and all possible state a,

$$\frac{\Pr\left(s_{\mathrm{dp}}(\boldsymbol{x}, \boldsymbol{r}) = a \mid \boldsymbol{x}\right)}{\Pr\left(s_{\mathrm{dp}}(\boldsymbol{x}', \boldsymbol{r}) = a \mid \boldsymbol{x}'\right)} \leq \exp(\epsilon).$$



https://www.ons.gov.uk/peoplepopulationandcommunity

That is, differentially private mechanisms conceal the confidential data x by infusing crafted noise r into the data product s_{dp} for release:

$$\boldsymbol{x} \longrightarrow s_{\mathrm{dp}}(\boldsymbol{x}, \boldsymbol{r})$$

[2] Dwork et al. (2006). Calibrating noise to sensitivity in private data analysis. *TCC (pp 265-284)*

Differential privacy: benefits and challenges

- ✓ **Provability**: differential privacy guarantees are formal and verifiable;
- ✓ **Transparency**: The probabilistic specification of the privacy mechanism can publicized without sabotaging the privacy guarantee.
- ► Statistical implication: transparency is *necessary* for drawing principled inference from privatized data.[3]

> How do we leverage the privacy mechanism for statistical inference?

[3] G. (2022). Transparent Privacy is Principled Privacy. *HDSR*, Special Issue 2.

Situating our (statistical × privacy) framework

x is the truth

$$s_{\text{dp}} \mid x \sim \eta(s_{\text{dp}} \mid x)$$

- Infer x based on s_{dp};
- η is the only source of uncertainty.

x is a sample

$$x \mid \theta \sim f(x \mid \theta)$$

$$s_{\text{dp}} \mid x \sim \eta(s_{\text{dp}} \mid x)$$

- Infer θ based on s_{dp}.
- Uncertainty stems from η, f, and beyond

design framework

Choose the best mechanism (η) + estimator combo:[4]

$$\hat{\theta}_{\text{design}}\left(s^*_{\text{dp}}(x)\right)$$

adjustment framework

For a given mechanism (η), perform the best inference:

$$\hat{\theta}_{\text{adjust}}\left(s_{\text{dp}}(x)\right)$$

[4] Slavković & Seeman. (2022). Statistical Data Privacy: A Song of Privacy and Utility. ArXiv:2205.03336.

Situating our (statistical × privacy) framework

x is the truth

$$s_{dp} \mid x \sim \eta(s_{dp} \mid x)$$

- Infer x based on s_{dp};
- η is the only source of uncertainty.

x is a sample

$$x \mid \theta \sim f(x \mid \theta)$$
$$s_{dp} \mid x \sim \eta(s_{dp} \mid x)$$

- Infer θ based on s_{dp}.
- Uncertainty stems from η, f, and beyond

design framework

Choose the best mechanism (η) + estimator combo:[4]

$$\hat{\theta}_{\text{design}}\left(s^*_{dp}(x)\right)$$

adjustment framework

For a given mechanism (η), perform the best inference:

$$\hat{\theta}_{\text{adjust}}\left(s_{dp}(x)\right)$$

[4] Slavković & Seeman. (2022). Statistical Data Privacy: A Song of Privacy and Utility. ArXiv:2205.03336.

Statistical inference from privatized data

Without privacy:

- Likelihood inference: $\ell(\theta \mid x) = f(x \mid \theta)$;
- Bayesian inference: $p(\theta \mid x) \propto p(\theta) f(x \mid \theta)$.

With privacy:

- The **marginal likelihood** integrates over $\mathcal{X}$, the entire space of confidential databases:

$$\ell(\theta \mid s_{\mathrm{dp}}) = \int_{\mathcal{X}} \eta(s_{\mathrm{dp}} \mid x) f(x \mid \theta)\, dx.$$

- The (exact) Bayesian **posterior** distribution is

$$p(\theta \mid s_{\mathrm{dp}}) \propto p(\theta) \ell(\theta \mid s_{\mathrm{dp}}).$$

Statistical inference from privatized data

Without privacy:

- Likelihood inference: $\ell(\theta \mid x) = f(x \mid \theta)$;
- Bayesian inference: $p(\theta \mid x) \propto p(\theta) f(x \mid \theta)$.

With privacy:

- The **marginal likelihood** integrates over $\mathcal{X}$, the entire space of confidential databases:

$$\ell(\theta \mid s_{\text{dp}}) = \int_{\mathcal{X}} \eta(s_{\text{dp}} \mid x) f(x \mid \theta)\, dx.$$

- The (exact) Bayesian **posterior** distribution is

$$p(\theta \mid s_{\text{dp}}) \propto p(\theta) \ell(\theta \mid s_{\text{dp}}).$$

Statistical inference from privatized data

Without privacy:

- Likelihood inference: $\ell(\theta \mid x) = f(x \mid \theta)$;
- Bayesian inference: $p(\theta \mid x) \propto p(\theta) f(x \mid \theta)$.

With privacy:

- The **marginal likelihood** integrates over $\mathcal{X}$, the entire space of confidential databases:

$$\ell(\theta \mid s_{\text{dp}}) = \int_{\mathcal{X}} \eta(s_{\text{dp}} \mid x) f(x \mid \theta)\, dx.$$

- The (exact) Bayesian **posterior** distribution is

$$p(\theta \mid s_{\text{dp}}) \propto p(\theta) \ell(\theta \mid s_{\text{dp}}).$$

Statistical inference from privatized data

Without privacy:

- Likelihood inference: $\ell(\theta \mid x) = f(x \mid \theta)$;
- Bayesian inference: $p(\theta \mid x) \propto p(\theta) f(x \mid \theta)$.

With privacy:

- The **marginal likelihood** integrates over $\mathcal{X}$, the entire space of confidential databases:

$$\ell(\theta \mid s_{\text{dp}}) = \int_{\mathcal{X}} \eta(s_{\text{dp}} \mid x) f(x \mid \theta) \, dx.$$

- The (exact) Bayesian **posterior** distribution is

$$p(\theta \mid s_{\text{dp}}) \propto p(\theta) \ell(\theta \mid s_{\text{dp}}).$$

Existing solutions

Approximations:

- Asymptotic approximation (Wang et al., 2018)
- Variational approximation (Karwa et al., 2016)
- Parametric bootstrap (Ferrando et al., 2020)

Exact but limited:

- Integrate exactly (Awan & Slavković, 2018, 2020)
- MCMC with latent sufficient stat (Bernstein & Sheldon, 2018, 2019)
- Exact inference with approximate computation (Gong, 2019)

Our Goal

An **efficient** and **user-friendly** sampler for the **exact** posterior that works for **general** choices of the data model f and the prior p.

A traditional Gibbs sampler

Problem #1: If n individuals each contribute d features, then

$$\mathcal{X} = \mathbb{X}^{n \times d}.$$

The likelihood may be intractable, and the posterior *doubly* intractable.

$$\boxed{p(\theta \mid s_{\mathrm{dp}}) \propto p(\theta) \int_{\mathcal{X}} \eta(s_{\mathrm{dp}} \mid x) f(x \mid \theta)\, dx.}$$

A traditional Gibbs sampler

Problem #1: If n individuals each contribute d features, then

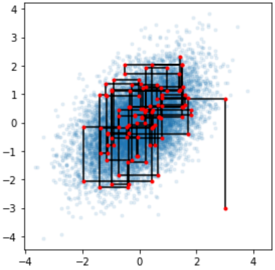
$$\mathcal{X} = \mathbb{X}^{n \times d}.$$

The likelihood may be intractable, and the posterior *doubly* intractable.

$$p(\theta \mid s_{\text{dp}}) \propto p(\theta) \int_{\mathcal{X}} \eta(s_{\text{dp}} \mid x) f(x \mid \theta)\, dx.$$

Data Augmentation (DA). Iterate the following:

1: sample $\theta \mid x, s_{\text{dp}} \overset{d}{=} \theta \mid x$
2: **for** $i = 1, \ldots, n$ **do**
3: sample $x_i \mid x_{-i}, \theta, s_{\text{dp}}$
4: **end for**



jessicastringham.net

A traditional Gibbs sampler

Problem #1: If n individuals each contribute d features, then

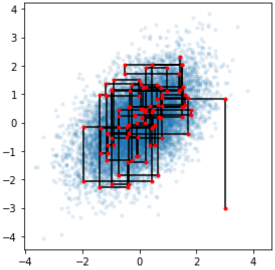
$$\mathcal{X} = \mathbb{X}^{n \times d}.$$

The likelihood may be intractable, and the posterior *doubly* intractable.

$$p(\theta \mid s_{\text{dp}}) \propto p(\theta) \int_{\mathcal{X}} \eta(s_{\text{dp}} \mid x) f(x \mid \theta)\, dx.$$

Data Augmentation (DA). Iterate the following:

sample $\theta \mid x, s_{\text{dp}} \stackrel{d}{=} \theta \mid x$
for $i = 1, \ldots, n$ **do**
  sample $x_i \mid x_{-i}, \theta, s_{\text{dp}}$
end for



jessicastringham.net

Problem #2: the conditional dist. $\boldsymbol{x} \mid \theta, s_{\text{dp}}$ is both f- and η-specific.

A general Metropolis-within Gibbs sampler

Solution. Propose $\boldsymbol{x} \mid \theta$ instead (or $x_i \mid \theta$ under conditional independence):

One Iteration of the privacy-aware Metropolis-within-Gibbs sampler

1: update $\theta \mid \boldsymbol{x}$
2: **for** $i = 1, \ldots, n$ **do**
3: propose $x_i^* \mid \theta$
4: accept x_i^* with probability

$$\alpha(x_i^\star \mid x_i, x_{-i}, \theta) = \min\left\{\frac{\eta(s_{\mathrm{dp}} \mid x_i^*, x_{-i})}{\eta(s_{\mathrm{dp}} \mid x_i, x_{-i})}, 1\right\}$$

5: **end for**

A general Metropolis-within Gibbs sampler

Solution. Propose $\boldsymbol{x} \mid \theta$ instead (or $x_i \mid \theta$ under conditional independence):

One Iteration of the privacy-aware Metropolis-within-Gibbs sampler

---

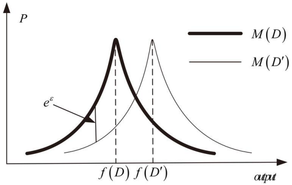
update $\theta \mid \boldsymbol{x}$
for $i = 1, \ldots, n$ **do**
    propose $x_i^* \mid \theta$
    accept x_i^* with probability

$$\alpha(x_i^\star \mid x_i, x_{-i}, \theta) = \min\left\{\frac{\eta(s_{\text{dp}} \mid x_i^*, x_{-i})}{\eta(s_{\text{dp}} \mid x_i, x_{-i})}, 1\right\}$$

end for

> If η is ϵ-DP, then for all θ, x_{-i}, x_i and x_i^*:
>
> $$\alpha(x_i^\star \mid x_i, x_{-i}, \theta) \geq \exp(-\epsilon).$$

As ϵ decreases (more privacy), acceptance rate α increases (more computational efficiency)!



e.g. for $\epsilon = 1$, $\alpha \geq 36.7\%$.

Requirements, run time, and efficiency

The proposed sampler requires:

- **Assumption 1**. The analyst knows how to sample the posterior if the data aren't privatized, i.e. she has a Markov kernel targeting $p(\theta \mid x)$.

Furthermore, if we have

- **Assumption 2** (*Record Additivity*). The privacy mechanism can be written as $\eta(s_{\mathrm{dp}} \mid x) = g\left(s_{\mathrm{dp}}, \sum_{i=1}^{n} t_i(x_i, s_{\mathrm{dp}})\right)$ for some known and tractable functions $g, t_1, \ldots, t_n$, then:

> The Gibbs sampler requires $O(n)$ number of operations to update the full latent database according to $p(x \mid \theta, s_{\mathrm{dp}})$.

Note:

- Even without privacy, one round of an MCMC procedure typically takes $O(n)$ time;
- Many commonly used DP mechanisms satisfy record additivity. e.g. additive, exponential mechanism, objective perturbation, etc.

Requirements, run time, and efficiency

The proposed sampler requires:

- **Assumption 1**. The analyst knows how to sample the posterior if the data aren't privatized, i.e. she has a Markov kernel targeting $p(\theta \mid x)$.

Furthermore, if we have

- **Assumption 2** (*Record Additivity*). The privacy mechanism can be written as $\eta(s_{\text{dp}} \mid x) = g\left(s_{\text{dp}}, \sum_{i=1}^{n} t_i(x_i, s_{\text{dp}})\right)$ for some known and tractable functions $g, t_1, \ldots, t_n$, then:

> The Gibbs sampler requires $O(n)$ number of operations to update the full latent database according to $p(x \mid \theta, s_{\text{dp}})$.

Note:

- Even without privacy, one round of an MCMC procedure typically takes $O(n)$ time;
- Many commonly used DP mechanisms satisfy record additivity. e.g. additive, exponential mechanism, objective perturbation, etc.

Ergodicity of the proposed sampler

Under the conditions

A1 the prior distribution is proper and $p(\theta) > 0$ for all θ in $\Theta = \{\theta \mid f_\theta(x) > 0 \text{ for some } x\}$;

A2 the model is such that the set $\{x : f(x \mid \theta) > 0\}$ does not depend on θ; and

A3 the privacy mechanism satisfies $\eta(s_{\text{dp}} \mid x) > 0$ for all $x \in \mathbb{X}^n$,

the Metropolis-within-Gibbs sampler on the joint space $(\mathbb{X}^n \times \Theta)$ is *ergodic* and it admits $p(x, \theta \mid s_{\text{dp}})$ as the unique limiting distribution.

Furthermore, if one can directly sample from $p(\theta \mid x)$, then the resulting (x, θ) chain as well as the marginal chains are *geometrically ergodic* if there exists $0 < a \leq b < \infty$ such that $a \leq f(x \mid \theta) \leq b$ for all θ and x.

Application: a naïve Bayes classifier

- $x = (x_1, \ldots, x_K)$ are *features*, each taking values in $\{1, \ldots, \mathcal{J}_K\}$;
- $y \in \{1, \ldots, I\}$ is the *class*;
- The non-private data consists of n i.i.d. copies of (x, y).
- Goal: predict the class given the features: $\Pr(y \mid x)$.
- The *naïve Bayes classifier* assumes $\Pr(x \mid y) = \prod_{k=1}^{K} \Pr(x_k \mid y)$;
- Release the noisy counts: $s_{\text{dp}} = \{n_{ijk} + \text{Laplace}(2K/\epsilon)\}_{ijk}$.

	X_1		X_2		…	X_K	
Y	1	2	1	2		1	2
1	n^1_{11}	n^1_{12}	n^2_{11}	n^2_{12}	…	n^K_{11}	n^K_{12}
2	n^1_{21}	n^1_{22}	n^2_{21}	n^2_{22}	…	n^K_{21}	n^K_{22}

TABLE 1

Sufficient statistics of the Naive Bayes model.

	X_1		X_2		…	X_K	
Y	1	2	1	2		1	2
1	p^1_{11}	p^1_{12}	p^2_{11}	p^2_{12}	…	p^K_{11}	p^K_{12}
2	p^1_{21}	p^1_{22}	p^2_{21}	p^2_{22}	…	p^K_{21}	p^K_{22}

TABLE 2

An example of the parameters of the Naive Bayes model for a $2 \times 2 \times K$ table.

Figure from Karwa et al. (2016)

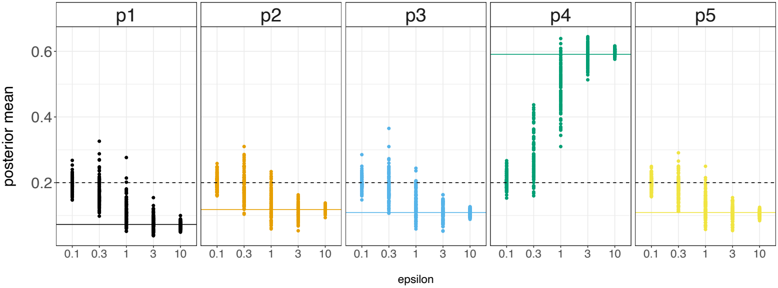
Simulation setup

For the simulation, set

- $N = 100$ (number of samples);
- $I = 5$ (number of classes);
- $K = 5$ (number of features);
- $\mathcal{J}_K = 3$ (number of options for each feature);
- $\epsilon \in \{.1, .3, 1, 3, 10\}$;
- Prior for all parameters: $\text{Dirichlet}(2, \ldots, 2)$.

Posterior mean

- Fix a confidential dataset;
- Create 100 privatized datasets at each ϵ value;
- Run one chain for each privatized dataset for 10,000 iterations;
- For each chain, calculate the posterior mean for $p_i = \Pr(y = i)$.



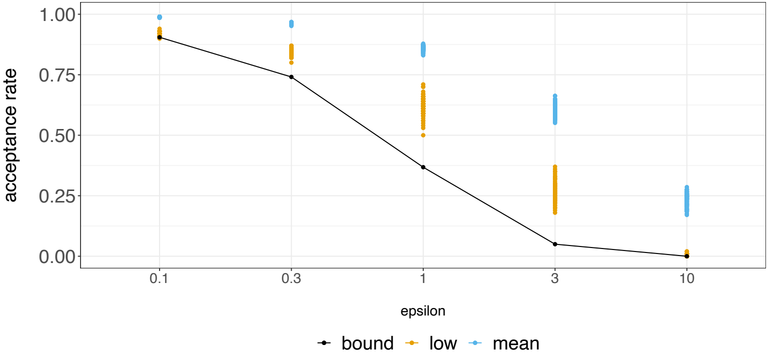
Frequentist coverage

Table. Frequentist coverage of a 90% credible interval for $p_i = \Pr(y = i)$ at different ϵ values. 100 replicates per ϵ value.

ϵ	$p_1 = .097$	$p_2 = .148$	$p_3 = .145$	$p_4 = .446$	$p_5 = .163$
.1	1	1	1	**.36**	1
.3	.97	1	1	**.59**	1
1	.94	.99	.97	**.83**	.98
3	.95	.91	.97	.89	.93
10	.92	.88	.94	.92	.9

Empirical acceptance rates

- 100 chains at each ϵ value;
- Each chain ran for 10,000 iterations;
- Minimum (orange) and mean (blue) acceptance rates for each chain.



Summary

An MCMC framework for Bayesian inference from privatized data:

- **Exact**: targets the correct posterior distribution;
- **General**: applicable to a wide range of statistical models and privacy mechanisms;
- **User-friendly**: mechanism independent, no (further) tuning parameters.

the privacy-efficiency alignment

Smaller ϵ implies higher acceptance rate: allowing the "**free exploitation**" of the privacy guarantee for computational efficiency.

Thank you

Awan, J., & Slavković, A. (2018). Differentially private uniformly most powerful tests for binomial data. In *Advances in neural information processing systems 31* (pp. 4208–4218). Curran Associates, Inc.

Awan, J., & Slavković, A. (2020). Differentially private inference for binomial data. *Journal of Privacy and Confidentiality*, *10*(1).

Bernstein, G., & Sheldon, D. R. (2018). Differentially private bayesian inference for exponential families. *Advances in Neural Information Processing Systems*, *31*.

Bernstein, G., & Sheldon, D. R. (2019). Differentially private bayesian linear regression. *Advances in Neural Information Processing Systems*, *32*.

Ferrando, C., Wang, S., & Sheldon, D. (2020). General-purpose differentially-private confidence intervals. *arXiv preprint arXiv:2006.07749*.

Gong, R. (2019). Exact inference with approximate computation for differentially private data via perturbations. *arXiv preprint arXiv:1909.12237*.

Karwa, V., Kifer, D., & Slavković, A. (2016). Private posterior distributions from variational approximations. *NIPS 2015 Workshop on Learning and Privacy with Incomplete Data and Weak Supervision*.

Wang, Y., Kifer, D., Lee, J., & Karwa, V. (2018). Statistical approximating distributions under differential privacy. *Journal of Privacy and Confidentiality*, *8*(1).

Ju, Awan, G., & Rao. (2022). Data Augmentation MCMC for Bayesian Inference from Privatize Data.[5] *ArXiv:2206.00710*.

[5] ruobin.gong@rutgers.edu. Gong and Rao thank the NSF for research support.