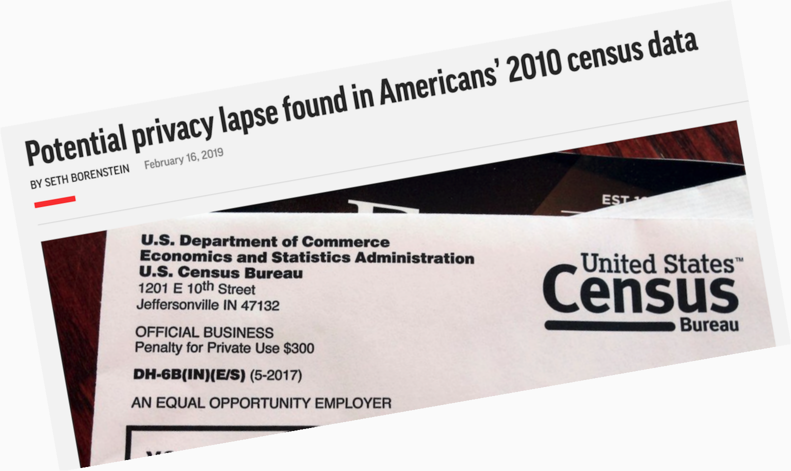
Differentially Private Bootstrap

New Privacy Analysis and Inference Strategies

Zhanyu Wang[1], Guang Cheng[2], Jordan Awan[1]

[1]Purdue University, Department of Statistics [2]UCLA, Department of Statistics

Privacy protection is hard



Privacy protection is hard

Potential privacy lapse found in Americans' 2010

BY SETH BORENSTEIN February 16, 2019

The New York Times

NEWS ANALYSIS

Poking Holes in Genetic Privacy

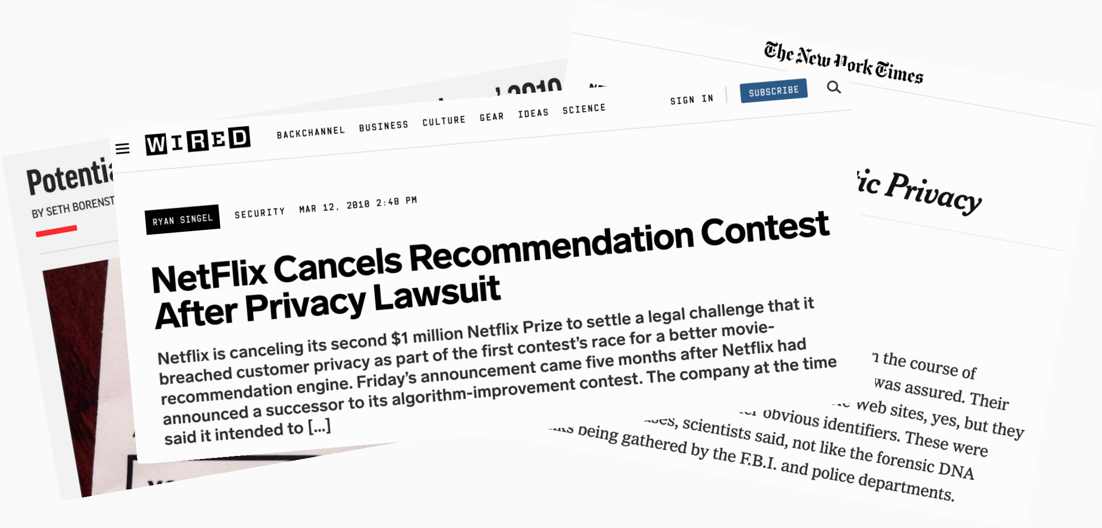
Give this article

By Gina Kolata

June 16, 2013

Not so long ago, people who provided DNA in the course of research studies were told that their privacy was assured. Their DNA sequences were on publicly available Web sites, yes, but they did not include names or other obvious identifiers. These were research databases, scientists said, not like the forensic DNA banks being gathered by the F.B.I. and police departments.

Privacy protection is hard



Potenti

BY SETH BORENST

WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE

RYAN SINGEL SECURITY MAR 12, 2010 2:48 PM

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine. Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to [...]

The New York Times

SIGN IN SUBSCRIBE

ic Privacy

n the course of
was assured. Their
web sites, yes, but they
obvious identifiers. These were
ses, scientists said, not like the forensic DNA
being gathered by the F.B.I. and police departments.

Privacy protection is hard

The New York Times

Your Data Were 'Anonymized'? These Scientists Can Still Identify You

Computer scientists have developed an algorithm that can pick out almost any American in databases supposedly stripped of personal information.

Differential Privacy (DP) is widely used

Figure 1: Differential Privacy is used in US Census 2020; Apple's study of diagnostic device, health and web browsing data; Google's Privacy Sandbox; Microsoft's analytics on app usage; Facebook's mobility data release during COVID-19; Amazon's AWS; Snapchat's machine learning models; Uber's detection of trends; Salesforce's reporting logs, etc.

Differential Privacy: state-of-the-art privacy protection measure

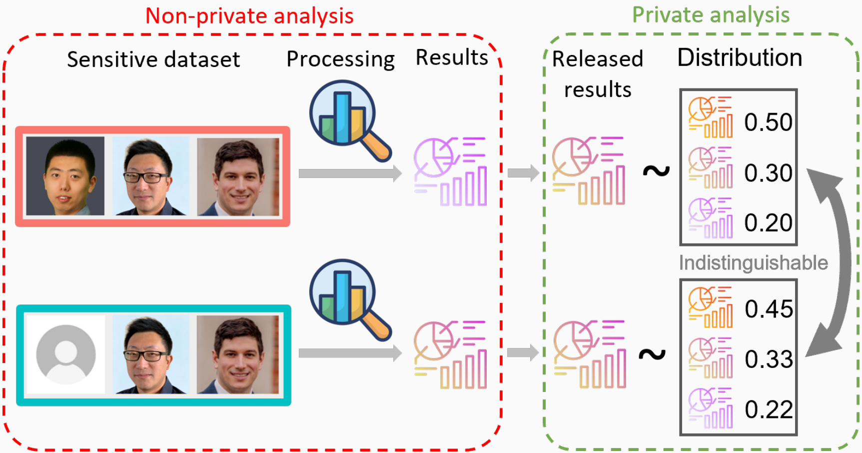
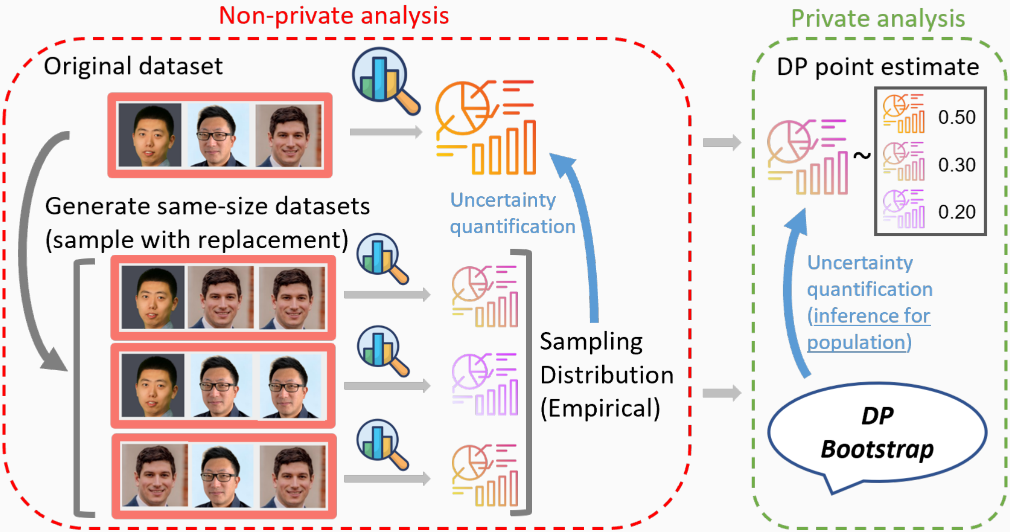
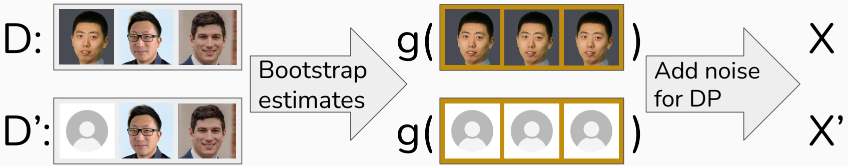


Figure 2: The output of the mechanism is roughly the same (approximately indistinguishable) when the input data is slightly changed. This is required for all datasets as input.

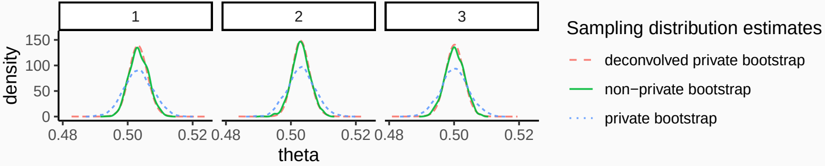
DP Bootstrap for private uncertainty quantification



DP Bootstrap: privacy analysis and implementation



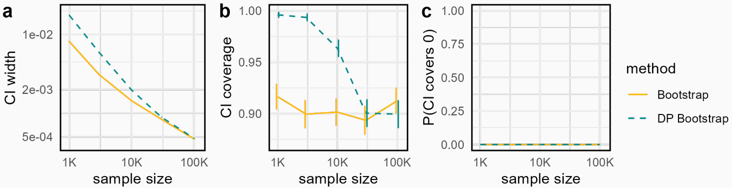
- If we use DP Bootstrap estimates for inference, its privacy cost is similar to releasing the same number of DP estimates based on the original dataset (uncertainty only from DP).
- The sampling distribution from DP Bootstrap is affected by the added DP noises. We use deconvolution to recover the non-private sampling distribution.



Private confidence intervals (CI) and its application

- We construct private CIs using quantiles of the deconvolved sampling distribution.
- Using the 2016 Canada Census Public Use Microdata, we build CI for the slope parameter in the quantile regression between market income and shelter cost.
- To the best of our knowledge, we are the first to do private inference in quantile regression.

Figure 3: Results of CIs for the slope parameter. The confidence level is 90%, and the privacy guarantee is 1-Gaussian DP. We have 2000 replicates to evaluate the performance of our CI:
(a) DP Bootstrap has a slightly larger CI width compared to non-private Bootstrap,
(b) The coverage is satisfactory (always above 90%; close to 90% for large sample size),
(c) The CI never contains 0, which means there is dependence between market income and shelter cost.



Thank you!

Our paper is on arXiv:

https://arxiv.org/abs/2210.06140

Contact me:

wang4094@purdue.edu

An example for de-anonymization

Data Considered for Sharing

Age	Zip Code	Gender	Diagnosis
15	00000	Male	Diabetes
21	00001	Female	Influenza
36	10000	Male	Broken Arm
91	10001	Female	Acid Reflux

Voter Registration Records (Identified Resource)

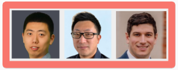
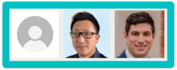
Birthdate	Zip Code	Gender	Name
2/2/1989	00001	Female	Alice Smith
3/3/1974	10000	Male	Bob Jones
4/4/1919	10001	Female	Charlie Doe

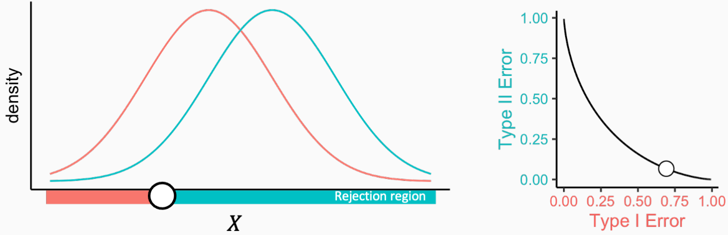
Linking two data sources to identity diagnoses.

Figure 4: The Guidance on De-identification of Protected Health Information. `hhs.gov`. Dataset on the left is released without Name. But using another public dataset on the right, we can recover the names in the anonymized dataset.

Hypothesis Testing, Trade-off Function, and f-DP

- The trade-off function maps the Type I error to the optimal corresponding Type II error.

$H_0 : X \sim M($  $)$ $H_1 : X \sim M($  $)$



- If $\mathrm{HammingDistance}(D, D') = 1$, we denote $D \cong D'$ and call them neighboring datasets.
- Differential privacy (DP) ensures the test is hard for any neighboring datasets in hypotheses.
- A random algorithm $\mathcal{M}$ is f-DP if the trade-off function between $\mathcal{M}(D)$ and $\mathcal{M}(D')$ for any $D \cong D'$ is lower bounded by f; it is μ-Gaussian DP (GDP) if the f is the trade-off function between $\mathcal{N}(0, 1)$ and $\mathcal{N}(\mu, 1)$.

Existing results and their problems

- It was mistakenly claimed that one could obtain the standard error of the output without additional privacy cost using bootstrap[1].
- There was also a wrong analysis of the DP bootstrap using the privacy loss distribution[2].
- The correct DP analysis of subsampling with replacement was only given in (ϵ, δ)-DP[3]. They did not consider the composition of the subsampling results and ignored the application on bootstrap methods.

---

[1]Thomas Brawner and James Honaker (2018). "Bootstrap inference and differential privacy: Standard errors for free." In: *Unpublished Manuscript*.

[2]Antti Koskela, Joonas Jälkö, and Antti Honkela (2020). "Computing tight differential privacy guarantees using fft." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2560–2569.

[3]Borja Balle, Gilles Barthe, and Marco Gaboardi (2018). "Privacy amplification by subsampling: Tight analyses via couplings and divergences." In: *Advances in Neural Information Processing Systems* 31.

DP guarantee with single bootstrap estimate

Theorem

Let $\underline{f} = (f_1, \ldots, f_n)$ be a sequence of tradeoff functions and $\underline{p} = (p_1, \ldots, p_m)$ be a vector of probability mass. Assume $\mathcal{M}$ satisfies $T_{\mathcal{M}(D),\mathcal{M}(D')} \geq f_i$ for any $\mathrm{HammingDist}(D, D') = i$.

For any given $\lambda \in (-\infty, 0]$, we can find α_i such that $f_i'(\alpha_i) = \lambda$. For $\sum_{i=1}^m p_i = 1$ and $\alpha = \sum_{i=1}^m p_i \alpha_i$, define $\mathrm{mix}(\underline{p}, \underline{f}) : \alpha \mapsto \sum_{i=1}^m p_i f_i(\alpha_i)$.

1. *The mapping $\mathrm{mix}(\underline{p}, \underline{f})$ is well-defined.*
2. *Let $p_0 = (1 - 1/n)^n$, $p_i = \frac{1}{p_0}\binom{n}{i}(1/n)^i(1 - 1/n)^{n-i}$, $f_0(\alpha) = 1 - \alpha$.*
   Then $\mathcal{M} \circ \texttt{boot}$ is $f_{\texttt{boot}}$-DP where $f_{\texttt{boot}} := \mathrm{mix}\left((p_0, \underline{p}), (f_0, \underline{f})\right)$;
   In addition, a stronger result is $f_{\texttt{boot}} := \mathrm{Symm}(p_0 f_0 + (1 - p_0)\mathrm{mix}(\underline{p}, \underline{f}))$ and $\mathrm{Symm}(\cdot)$ maps asymmetric tradeoff functions to symmetric ones (w.r.t. the line $y = x$).

DP guarantee with multiple bootstrap estimates

- As the bootstrap method estimates the sampling distribution with the empirical distribution of multiple bootstrap estimates, we provide DP analysis for the mechanism outputting multiple DP Bootstrap estimates.

Theorem

Assume $\mathcal{M}_i$ satisfies μ_B-GDP. If $\lim_{B\to\infty} \mu_B\sqrt{(2-2/\mathrm{e})B} \to \mu$ and we let $\mathcal{M}'_i = \mathcal{M}_i \circ \texttt{boot}$, $\mathcal{M}^B_{\texttt{boot}} = \{\mathcal{M}'_1, \ldots, \mathcal{M}'_B\}$, then $\mathcal{M}^B_{\texttt{boot}}$ asymptotically satisfies μ-GDP.

- Although the trade-off function for $\mathcal{M} \circ \texttt{boot}$ is not in the form of GDP, the nature of bootstrap method allows us to assume the composition number is large and the asymptotic privacy analysis can be a good approximation.

Private confidence intervals (CI) and its application

- We construct private CIs using quantiles of the deconvolved sampling distribution.
- We conduct real-world experiments on the 2016 Census Public Use Microdata Files.
- First, we build CIs for the population mean of the individual's market income in Ontario. We use DP Bootstrap with the Gaussian mechanism and compare our results with NoisyVar[4]. The confidence level is 90%, and the privacy guarantee is 1-GDP.

Table 1: Results of CIs for the mean income. ($n = 200,000$.)

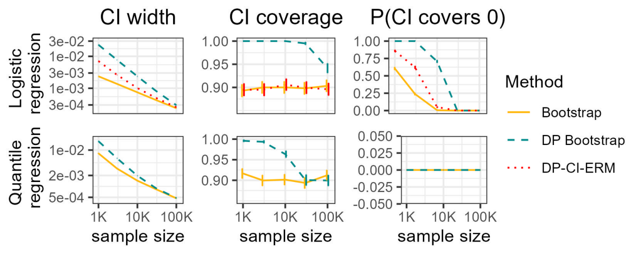
Method	CI Coverage	CI Width
Bootstrap	0.910 (0.006)	279.4 (0.54)
DP Bootstrap	0.905 (0.007)	291.0 (0.54)
NoisyVar	0.857 (0.008)	253.6 (0.16)

[4] Wenxin Du et al. (2020). "Differentially private confidence intervals." In: *arXiv preprint arXiv:2001.02285*.

Private confidence intervals (CI) and its application

- Then we build CIs for the slope parameter in the logistic regression and quantile regression between the market income and shelter cost. We use DP Bootstrap with the output perturbation mechanism (built on empirical risk minimization) and compare our results with DP-CI-ERM[5]. The confidence level is 90%, and the privacy guarantee is 1-GDP.

Figure 5: Results of CIs for the slope parameter. DP-CI-ERM cannot be used on quantile regression since it is based on the Hessian of the loss, which is 0 for $\rho_\tau(z) = (\tau - \mathbb{1}(z \leq 0))z$, $z = y - x^\intercal\theta$.



[5] Yue Wang, Daniel Kifer, and Jaewoo Lee (2019). "Differentially Private Confidence Intervals for Empirical Risk Minimization." In: *Journal of Privacy and Confidentiality* 9.1.

References

- Balle, Borja, Gilles Barthe, and Marco Gaboardi (2018). "Privacy amplification by subsampling: Tight analyses via couplings and divergences." In: *Advances in Neural Information Processing Systems* 31.
- Brawner, Thomas and James Honaker (2018). "Bootstrap inference and differential privacy: Standard errors for free." In: *Unpublished Manuscript*.
- Du, Wenxin et al. (2020). "Differentially private confidence intervals." In: *arXiv preprint arXiv:2001.02285*.
- Koskela, Antti, Joonas Jälkö, and Antti Honkela (2020). "Computing tight differential privacy guarantees using fft." In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2560–2569.
- Wang, Yue, Daniel Kifer, and Jaewoo Lee (2019). "Differentially Private Confidence Intervals for Empirical Risk Minimization." In: *Journal of Privacy and Confidentiality* 9.1.