Hypothesis testing under local differential privacy

Tom Berrett
University of Warwick

One World YoungStatS Webinar

November 9th, 2022

Personal data

The collection and use of personal data is increasingly common in modern society.

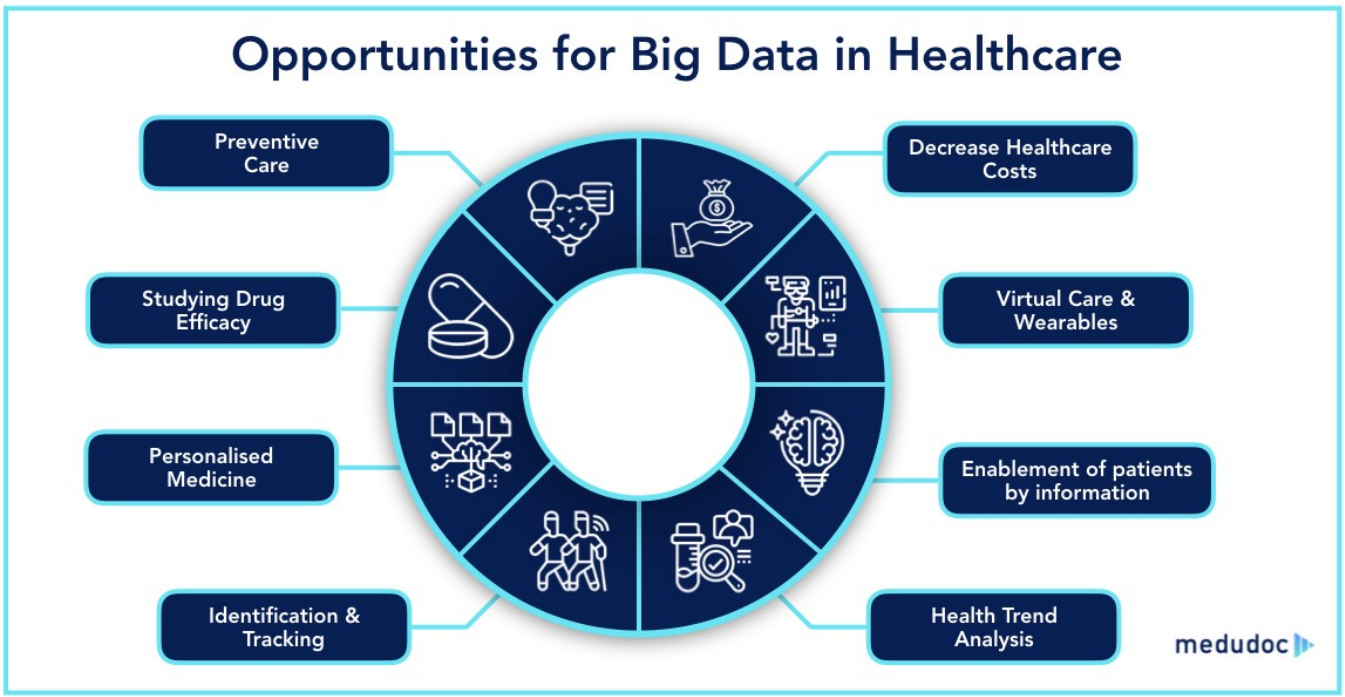
Souce: Paris Marx, `medium.com`

Data protection laws and bad publicity drive organisations to demonstrate respect for individuals' privacy.

Sensitive information

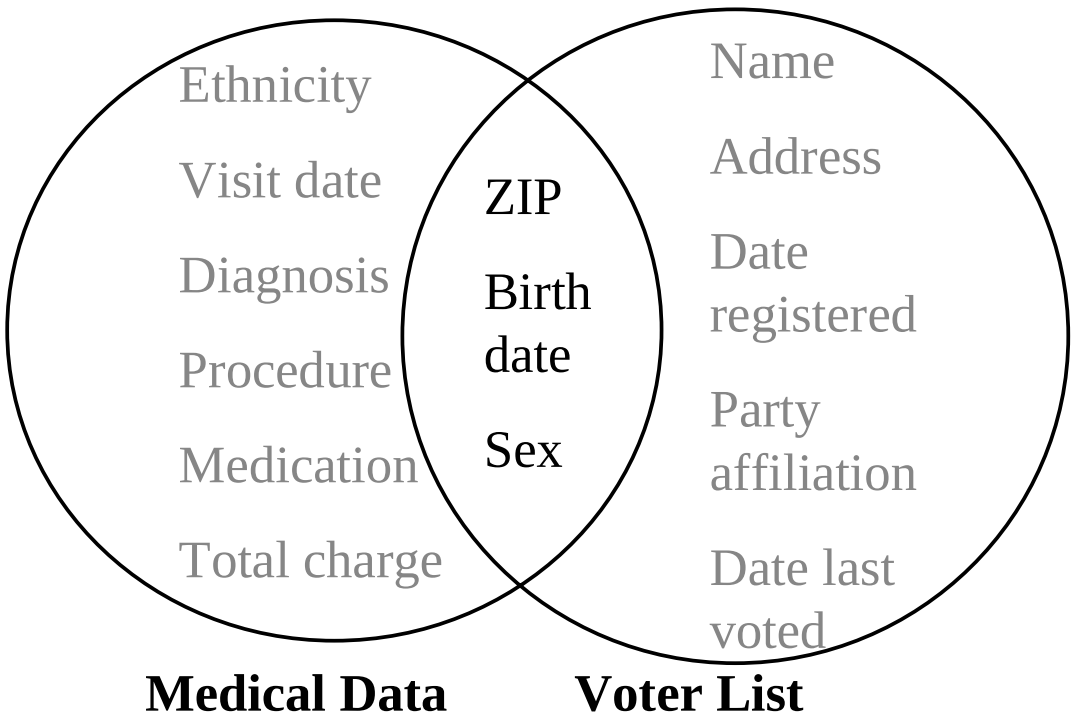
Many classical application areas also involve large amounts of sensitive information. For example:

- Medicine and public health;
- Census;
- Finance.



Traditional anonymisation is not enough

Removing names/addresses is insufficient to prevent re-identification.

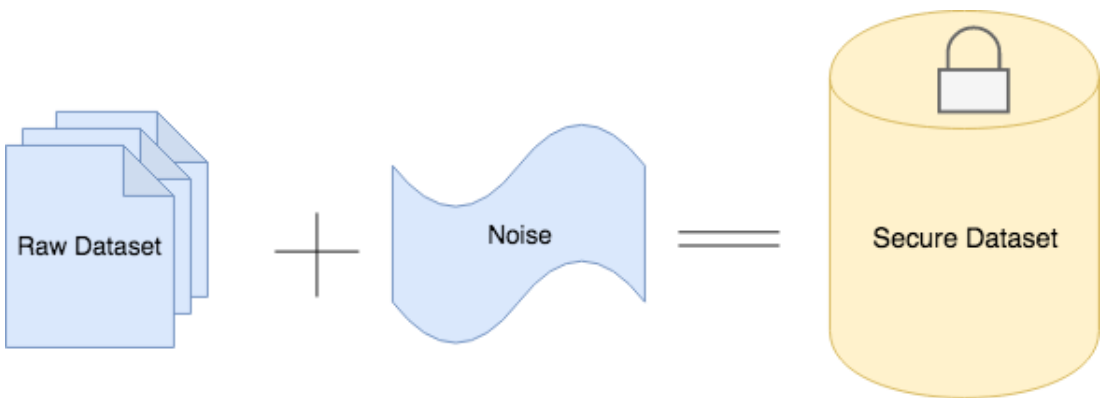


Through 'anonymised' state medical records and publicly available voter registration lists, Sweeney (2002) was able to find the medical records of the governor of Massachusetts.

Privacy mechanisms

A *privacy mechanism* is a randomised algorithm taking an input dataset $\mathsf{X} = (X_1, \ldots, X_n)$ in $\mathcal{X}^n$ and producing publishable data Z. Formally, it is a collection of conditional distributions $Q = \{Q(\cdot|\mathsf{x}) : \mathsf{x} \in \mathcal{X}\}$ such that

$$\mathsf{Z}|\{\mathsf{X} = \mathsf{x}\} \sim Q(\cdot|\mathsf{x}).$$



Source: Abhishek Tandon, `medium.com`

How much noise should we add? What type of noise?

Differential privacy

Privacy mechanism Q is called α-*differentially private* (Dwork et al., 2006) if

$$\sup_A \frac{Q(A|\mathrm{x})}{Q(A|\mathrm{x}')} \leq e^{\alpha}$$

for all x, x' such that $d(\mathrm{x}, \mathrm{x}') := \sum_{i=1}^{n} \mathbb{1}_{x_i \neq x_i'} \leq 1$.

Differential privacy

Privacy mechanism Q is called α-*differentially private* (Dwork et al., 2006) if

$$\sup_A \frac{Q(A|\mathrm{x})}{Q(A|\mathrm{x}')} \leq e^{\alpha}$$

for all x, x' such that $d(\mathrm{x}, \mathrm{x}') := \sum_{i=1}^{n} \mathbb{1}_{x_i \neq x_i'} \leq 1$.

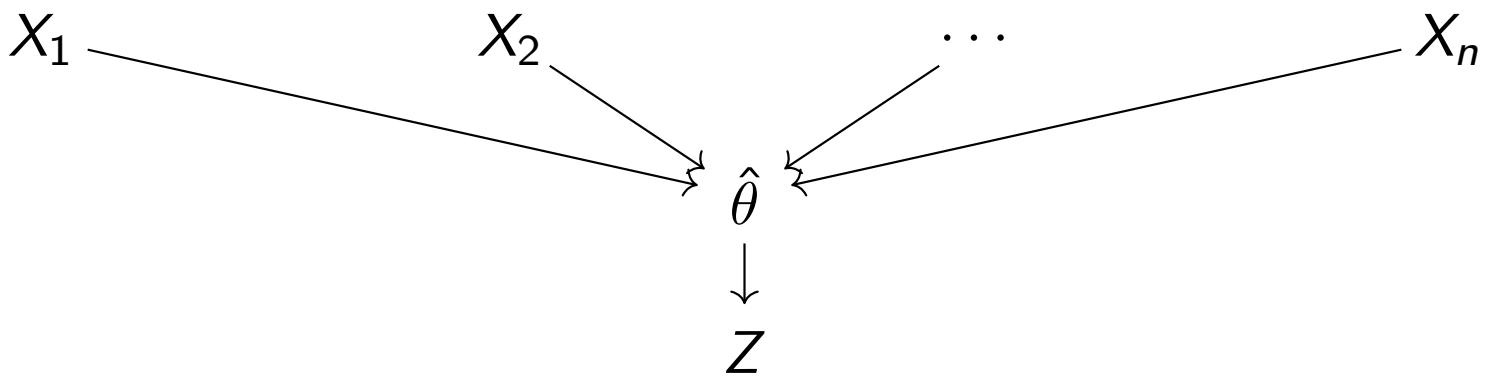
Differential privacy provides a rigorous framework to control the amount of personal information in published data. Large scale applications include

- Google Chrome (Erlingsson, Pihur and Korolova, 2014);
- Apple in iOS and macOS (Tang et al., 2017);
- Microsoft (Ding, Kulkarni and Yekhanin, 2017);
- Uber (Near, 2018);
- US Census (Machanavajjhala et al., 2008; Dwork, 2019).

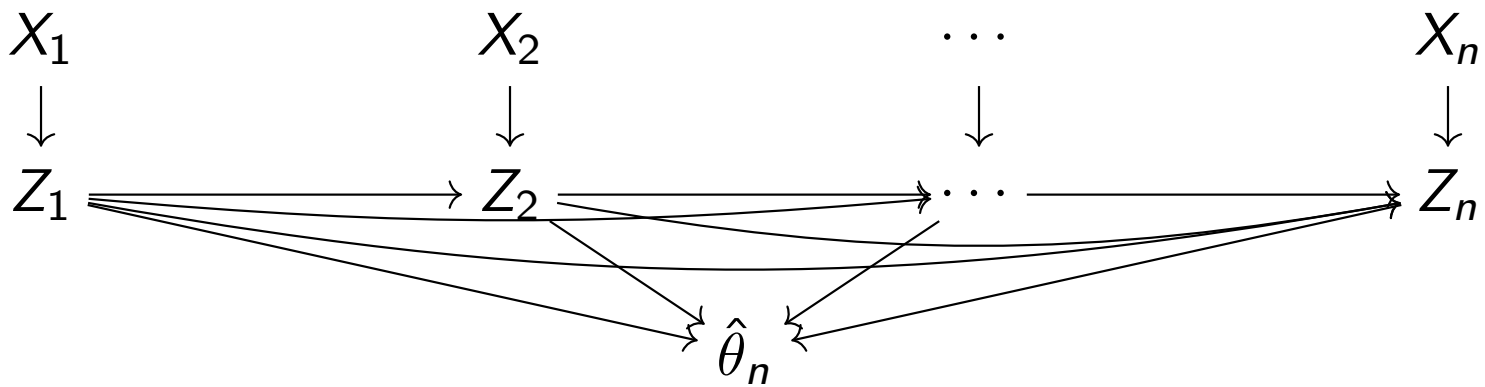
Can also be used to demonstrate GDPR compliance (Cohen and Nissim, 2020).

(Central) Differential privacy

The earliest work (Dwork et al., 2006) assumes a trusted data curator.



We consider the local model (e.g. Duchi et al., 2013):



Simple hypothesis testing

Simple hypothesis testing

Consider the simple hypothesis testing problem

$$H_0 : P = P_0 \quad \text{vs.} \quad H_1 : P = P_0$$

for fixed distributions P_0, P_1, given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$.

The classical LR statistic $\prod_{i=1}^{n} \frac{dP_1}{dP_0}(X_i)$ is difficult to privatise, but we can use ideas from *robust statistics* (e.g. Chen et al., 2016; Gopi et al., 2020).

In the non-private setting the Scheffé test rejects H_0 if and only if

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A^c\}} > \frac{1}{2}\{P_0(A) + P_1(A)\},$$

where A is such that $P_0(A) - P_1(A) = \sup_S \{P_0(S) - P_1(S)\}$.

Simple hypothesis testing

This can be applied to the output of the randomised response mechanism (Warner, 1965; Gopi et al., 2020)

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in A^c\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in A^c\}}, & \text{otherwise.} \end{cases}$$

Reject if and only if

$$\frac{e^\alpha + 1}{n(e^\alpha - 1)} \sum_{i=1}^{n} \left(Z_i - \frac{1}{e^\alpha + 1} \right) > \frac{1}{2}\{P_0(A) + P_1(A)\}.$$

Simple hypothesis testing

This can be applied to the output of the randomised response mechanism (Warner, 1965; Gopi et al., 2020)

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in A^c\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in A^c\}}, & \text{otherwise.} \end{cases}$$

Reject if and only if

$$\frac{e^\alpha + 1}{n(e^\alpha - 1)} \sum_{i=1}^{n} \left(Z_i - \frac{1}{e^\alpha + 1} \right) > \frac{1}{2}\{P_0(A) + P_1(A)\}.$$

Analysing the risk of this test shows that

$$\mathcal{R}_{n,\alpha} := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\phi \in \Phi_Q} \left\{ \mathbb{E}_{P_0,Q}(\phi) + \mathbb{E}_{P_1,Q}(1-\phi) \right\} \leq 2\exp[-Cn\alpha^2 \mathrm{TV}(P_0, P_1)^2].$$

Simple hypothesis testing

This can be applied to the output of the randomised response mechanism (Warner, 1965; Gopi et al., 2020)

$$Z_i = \begin{cases} \mathbb{1}_{\{X_i \in A^c\}}, & \text{w.pr. } e^\alpha/(1+e^\alpha), \\ 1 - \mathbb{1}_{\{X_i \in A^c\}}, & \text{otherwise.} \end{cases}$$

Reject if and only if

$$\frac{e^\alpha + 1}{n(e^\alpha - 1)} \sum_{i=1}^{n} \left(Z_i - \frac{1}{e^\alpha + 1} \right) > \frac{1}{2}\{P_0(A) + P_1(A)\}.$$

Analysing the risk of this test shows that

$$\mathcal{R}_{n,\alpha} := \inf_{Q \in \mathcal{Q}_\alpha} \inf_{\phi \in \Phi_Q} \left\{ \mathbb{E}_{P_0,Q}(\phi) + \mathbb{E}_{P_1,Q}(1-\phi) \right\} \leq 2\exp[-Cn\alpha^2 \mathrm{TV}(P_0, P_1)^2].$$

There is a lower bound to match:

$$\mathcal{R}_{n,\alpha} \geq (1/2)\exp[-16n\alpha^2 \mathrm{TV}(P_0, P_1)^2].$$

Simple hypothesis testing

We combine private and robust analyses to show that, under ε-Huber contamination ($X_i \sim (1-\varepsilon)P + \varepsilon G$), the minimax risk satisfies

$$(1/2)\exp[-16n\alpha^2\{\mathrm{TV}(P_0, P_1) - \varepsilon/(1-\varepsilon)\}_+^2] \leq \mathcal{R}_{n,\alpha}(\varepsilon) \leq 2\exp[-Cn\alpha^2\{\mathrm{TV}(P_0, P_1) - \varepsilon/(1-\varepsilon)\}_+^2]$$

For combined error rate ≤ 0.1 we require:

- Classical model: $\mathrm{H}(P_0, P_1) \gtrsim 1/\sqrt{n}$;
- ε-Huber with $n = \infty$: $\mathrm{TV}(P_0, P_1) > \varepsilon/(1-\varepsilon)$ (e.g. Chen et al., 2016);
- α-LDP: $\mathrm{TV}(P_0, P_1) \gtrsim 1/\sqrt{n\alpha^2}$ (e.g. Gopi et al., 2020);
- α-LDP and ε-Huber: $\mathrm{TV}(P_0, P_1) \gtrsim \varepsilon + 1/\sqrt{n\alpha^2}$ (Li et al., 2022).

Simple hypothesis testing

We combine private and robust analyses to show that, under ε-Huber contamination ($X_i \sim (1-\varepsilon)P + \varepsilon G$), the minimax risk satisfies

$$(1/2)\exp[-16n\alpha^2\{\mathrm{TV}(P_0, P_1) - \varepsilon/(1-\varepsilon)\}_+^2] \leq \mathcal{R}_{n,\alpha}(\varepsilon) \leq 2\exp[-Cn\alpha^2\{\mathrm{TV}(P_0, P_1) - \varepsilon/(1-\varepsilon)\}_+^2]$$

For combined error rate ≤ 0.1 we require:

- Classical model: $\mathrm{H}(P_0, P_1) \gtrsim 1/\sqrt{n}$;
- ε-Huber with $n = \infty$: $\mathrm{TV}(P_0, P_1) > \varepsilon/(1-\varepsilon)$ (e.g. Chen et al., 2016);
- α-LDP: $\mathrm{TV}(P_0, P_1) \gtrsim 1/\sqrt{n\alpha^2}$ (e.g. Gopi et al., 2020);
- α-LDP and ε-Huber: $\mathrm{TV}(P_0, P_1) \gtrsim \varepsilon + 1/\sqrt{n\alpha^2}$ (Li et al., 2022).

There are deep connections between robust statistics and (local) differential privacy (e.g. Dwork and Lei, 2009; Avella-Medina, 2021; Li et al., 2022).

Goodness-of-fit testing

Goodness-of-fit testing

Raw data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$, a discrete distribution on $\mathbb{N}$.

Want to test

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1(\delta, \mathbb{L}_r) : \|p - p_0\|_r \geq \delta$$

for $r = 1, 2$.

Goodness-of-fit testing

Raw data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$, a discrete distribution on $\mathbb{N}$.

Want to test

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1(\delta, \mathbb{L}_r) : \|p - p_0\|_r \geq \delta$$

for $r = 1, 2$.

In the non-private problem with $r = 1$ we may have

$$\delta \approx \sqrt{\frac{\|p_0\|_{2/3}}{n}}$$

and still have non-trivial power (Valiant and Valiant, 2014, 2017; Balakrishnan and Wasserman, 2019).

Goodness-of-fit testing

Raw data $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$, a discrete distribution on $\mathbb{N}$.

Want to test

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1(\delta, \mathbb{L}_r) : \|p - p_0\|_r \geq \delta$$

for $r = 1, 2$.

In the non-private problem with $r = 1$ we may have

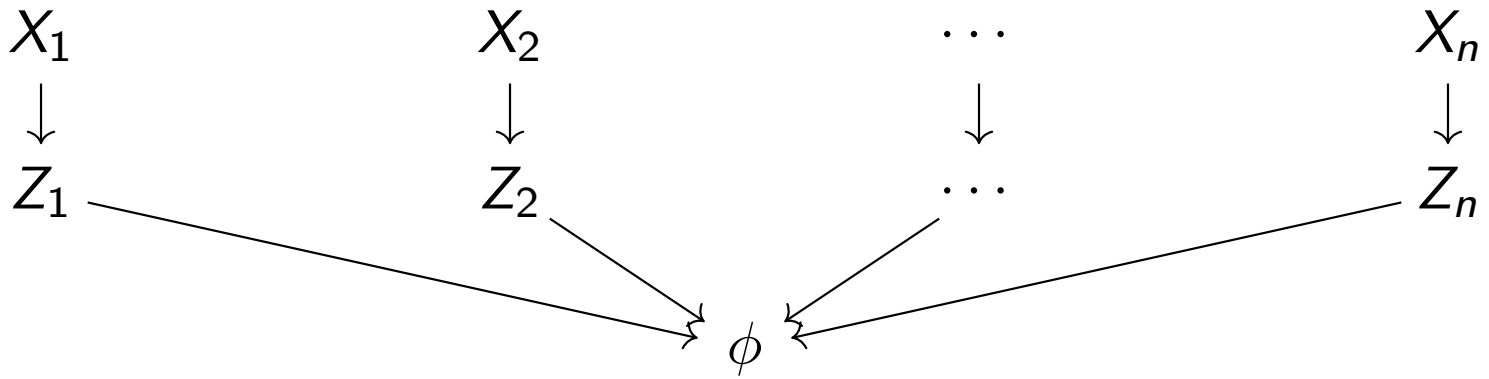
$$\delta \approx \sqrt{\frac{\|p_0\|_{2/3}}{n}}$$

and still have non-trivial power (Valiant and Valiant, 2014, 2017; Balakrishnan and Wasserman, 2019).

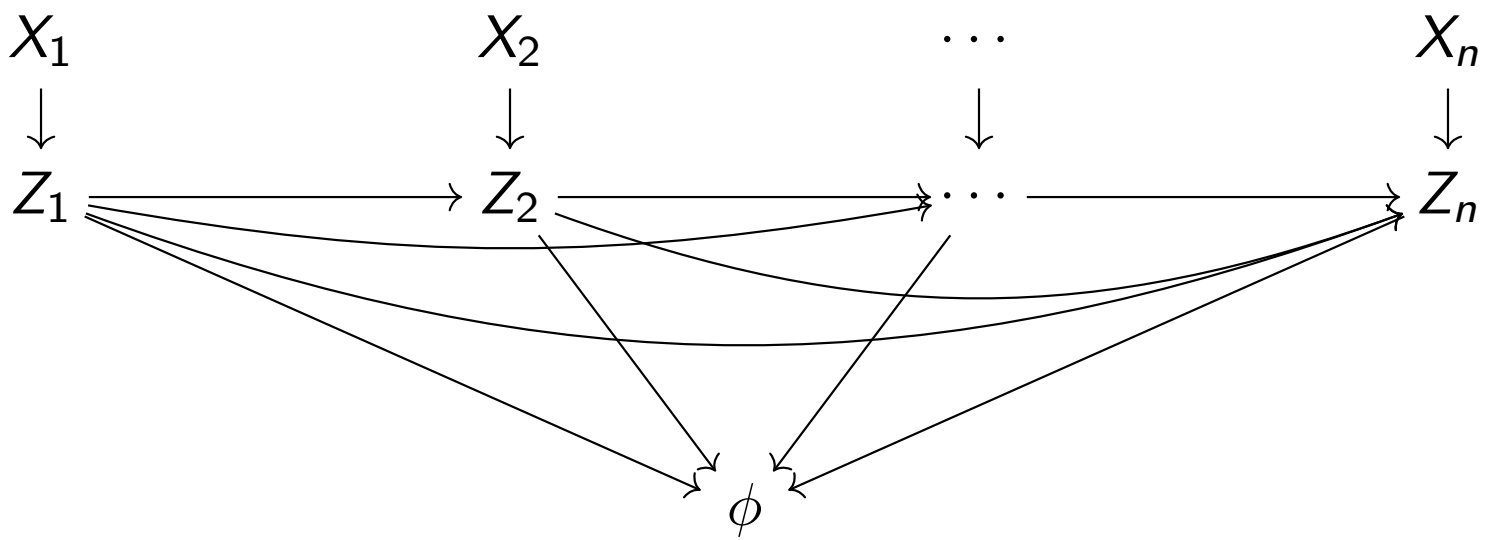
Q: How are local testing rates affected by local differential privacy?

Interactive vs. Non-interactive

For simple hypothesis testing a *non-interactive* method was optimal.



In this problem we see that *sequentially interactive* methods can do better.



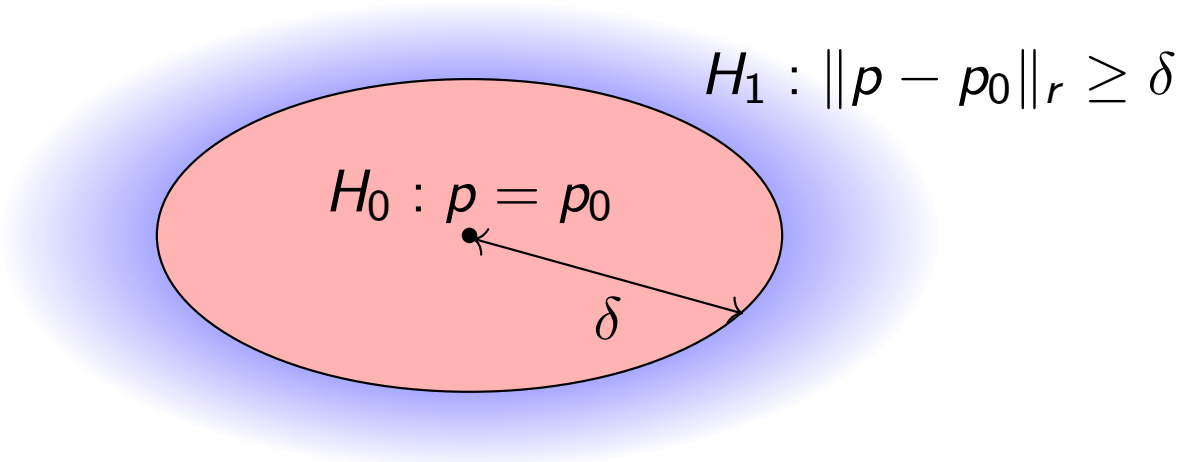
Minimax separation rate

We measure performance through the minimax separation rate: the smallest δ for which we have non-trivial power. Given mechanism Q and $\gamma > 0$ this is

$$\mathcal{E}_n(Q, p_0, \mathbb{L}_r) = \inf\left\{\delta > 0 : \inf_{\phi \in \Phi_Q} \sup_{p \in H_1(\delta, \mathbb{L}_r)} \left\{\mathbb{E}_{p_0}(\phi) + \mathbb{E}_p(1-\phi)\right\} \leq \gamma\right\}$$

We also want to find the best mechanism in our classes

$$\mathcal{E}^{\mathrm{NI}}_{n,\alpha}(p_0, \mathbb{L}_r) = \inf_{Q \in \mathcal{Q}^{\mathrm{NI}}_\alpha} \mathcal{E}_n(Q, p_0, \mathbb{L}_r), \quad \mathcal{E}^{\mathrm{I}}_{n,\alpha}(p_0, \mathbb{L}_r) = \inf_{Q \in \mathcal{Q}^{\mathrm{I}}_\alpha} \mathcal{E}_n(Q, p_0, \mathbb{L}_r).$$



Non-interactive rates

Optimal rate for $p_0 = \mathrm{Unif}([d])$ in $\mathbb{L}_1$ is $\frac{d^{3/4}}{\sqrt{n\alpha^2}}$ (Acharya et al., 2019).

Theorem (B. and Butucea (2020))

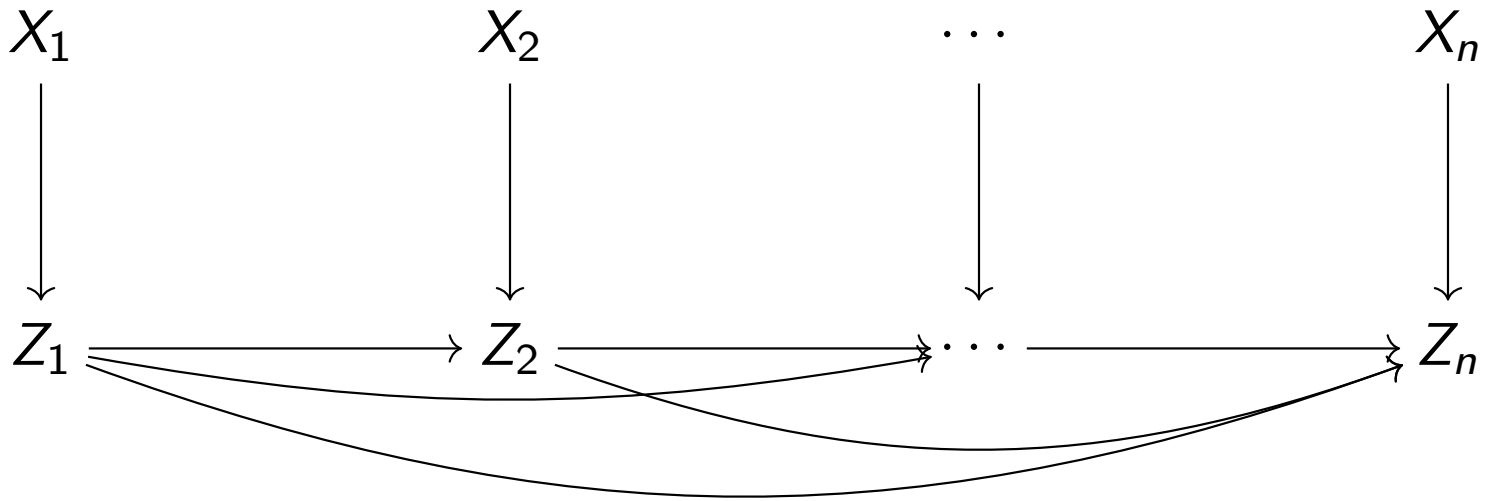
$$\mathcal{E}^{\mathrm{NI}}_{n,\alpha}(p_0, \mathbb{L}_1) \lesssim \frac{j_*^{3/4}}{\sqrt{n\alpha^2}} \quad \textit{and} \quad \mathcal{E}^{\mathrm{NI}}_{n,\alpha}(p_0, \mathbb{L}_2) \lesssim \frac{j_{**}^{1/4}}{\sqrt{n\alpha^2}},$$

with (nearly) matching lower bound for $\mathbb{L}_1$. Here j_*, j_{**} are 'effective support sizes', e.g.

$$j_* = j_*(n\alpha^2, p_0, \mathbb{L}_1) := \min\left\{ j \in \mathbb{N} : \frac{j^{3/4}}{(n\alpha^2)^{1/2}} \geq \sum_{j'=j+1}^{\infty} p_0(j') \right\}.$$

Interactive rates

Going back to the interactive setting:



Theorem

We have

$$\mathcal{E}^{\mathrm{NI}}_{n,\alpha}(p_0, \mathbb{L}_1) \lesssim \frac{\tilde{j}^{1/2}}{\sqrt{n\alpha^2}} \quad \textit{and} \quad \mathcal{E}^{\mathrm{I}}_{n,\alpha}(p_0, \mathbb{L}_2) \lesssim \frac{1}{\sqrt{n\alpha^2}},$$

with (nearly) matching lower bounds, where $\tilde{j}$ is another 'effective support size'.

Examples

Table: Separation rates (up to log factors) for testing discrete distributions on $\mathbb{N}$.

	Noninteractive		Interactive	
p_0	$\mathbb{L}_1$	$\mathbb{L}_2$	$\mathbb{L}_1$	$\mathbb{L}_2$
Unif$[d]$	$\frac{d^{3/4}}{\sqrt{n\alpha^2}}$	$\leq \frac{d^{1/4}}{\sqrt{n\alpha^2}}$ $\gtrsim \frac{d^{1/4}}{\sqrt{n\alpha^2}} \wedge \frac{1}{\sqrt{d}}$	$\frac{d^{1/2}}{\sqrt{n\alpha^2}}$	$\frac{1}{\sqrt{n\alpha^2}}$
$\propto j^{-1-\beta}$	$(n\alpha^2)^{-\frac{2\beta}{4\beta+3}}$	$\leq (n\alpha^2)^{-\frac{2\beta}{4\beta+1}}$ $\gtrsim (n\alpha^2)^{-\frac{2\beta+1}{4\beta+3}}$	$(n\alpha^2)^{-\frac{2\beta}{4\beta+2}}$	$\frac{1}{\sqrt{n\alpha^2}}$
$\propto j^{\eta} e^{-cj^{\beta}}$	$\frac{\log^{3/(4\beta)}(n\alpha^2)}{\sqrt{n\alpha^2}}$	$\frac{\log^{1/(4\beta)}(n\alpha^2)}{\sqrt{n\alpha^2}}$	$\frac{\log^{2/(4\beta)}(n\alpha^2)}{\sqrt{n\alpha^2}}$	$\frac{1}{\sqrt{n\alpha^2}}$

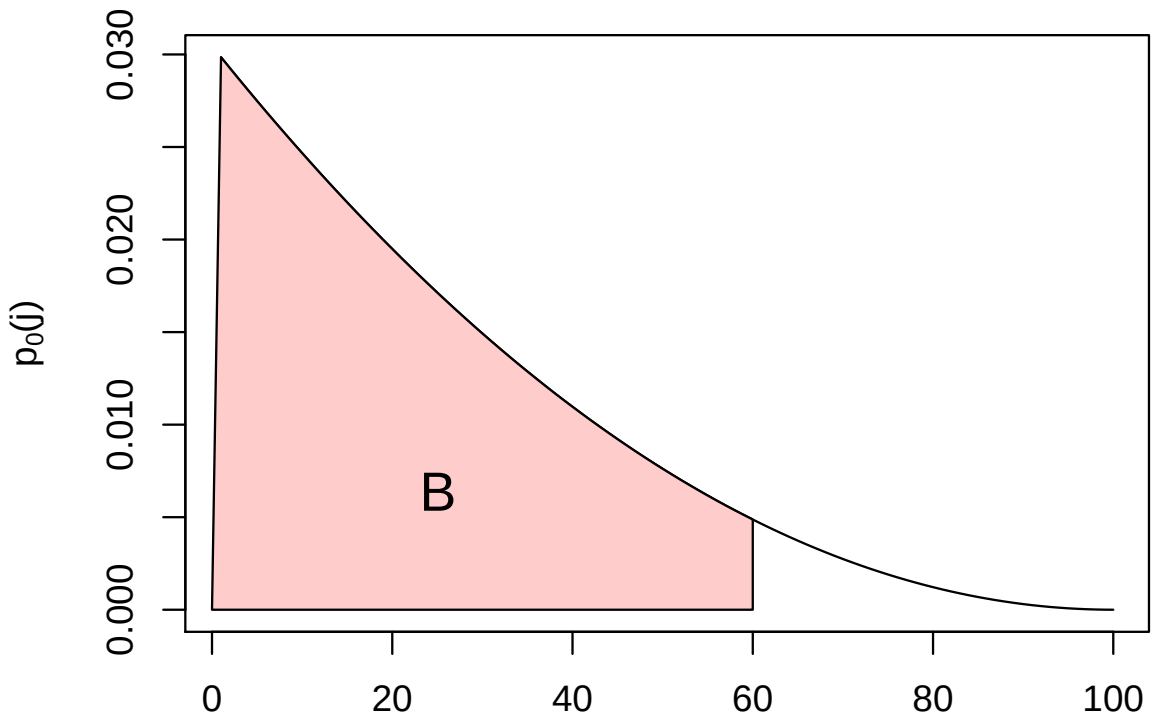
Non-interactive procedure

Let $(W_{ij}) \overset{\text{i.i.d.}}{\sim} \text{Laplace}$. Given $B \subset \mathbb{N}$, with first half of data generate

$$Z_{ij} = \mathbb{1}_{\{X_i = j\}} + \frac{2}{\alpha} W_{ij}, \qquad j \in B$$

and find

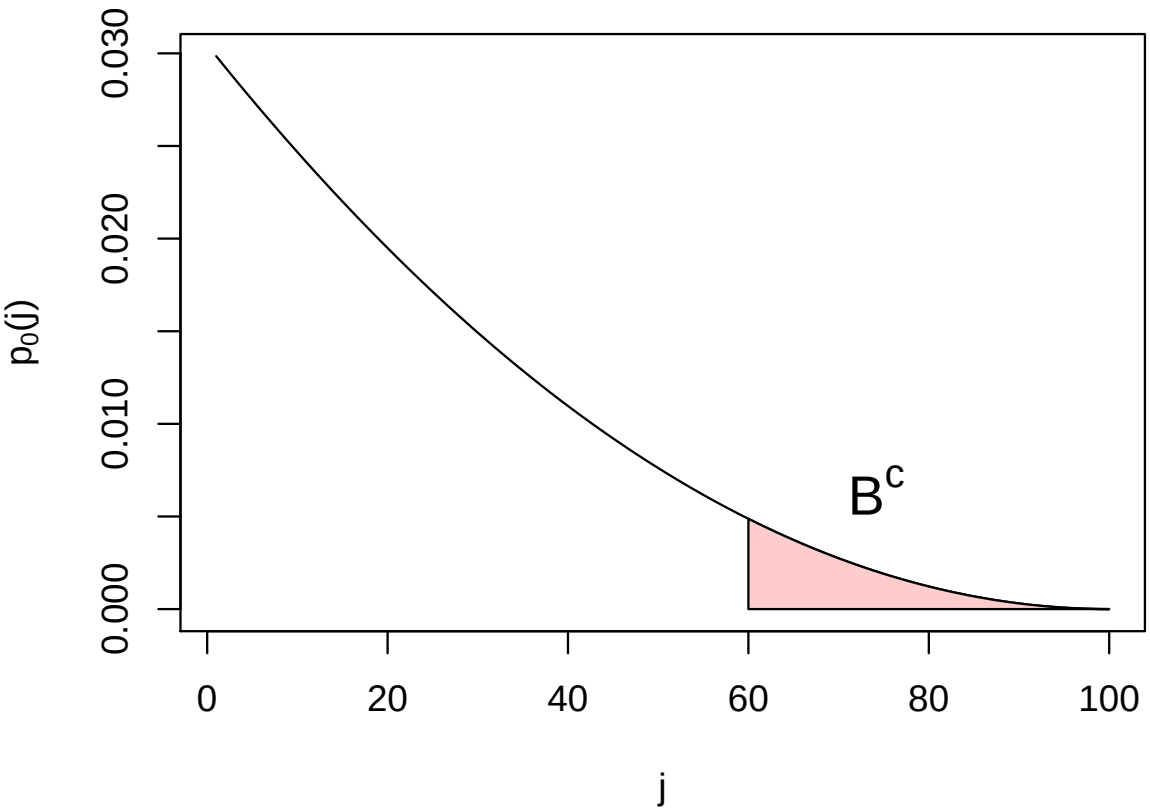
$$S_B = \sum_{j \in B} \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \{Z_{i_1 j} - p_0(j)\}\{Z_{i_2 j} - p_0(j)\}.$$



Non-interactive procedure

With second half:

$$Z_i = \mathbb{1}_{\{X_i \notin B\}} + \frac{2}{\alpha} W_{i1}, \quad T_B = \frac{1}{n} \sum_{i=n+1}^{2n} \{Z_i - p_0(B^c)\}.$$



Reject if $\max(S_B, T_B)$ is large to show $\mathcal{E}^{\mathrm{NI}}_{n,\alpha}(p_0, \mathbb{L}_1) \lesssim \frac{|B|^{3/4}}{\sqrt{n\alpha^2}} \vee p_0(B^c)$.

Interactive procedure

We use T_B to deal with the tail of p_0 as before, but estimate

$$\sum_{j \in B} \{p(j) - p_0(j)\}^2 = \sum_{j \in B} p(j)\{p(j) - p_0(j)\} - \sum_{j \in B} p_0(j)\{p(j) - p_0(j)\}$$

differently.

Two-steps: find some $\hat{p}_j$ then estimate the linear functional of p,

$$\sum_{j \in B} p(j)\{\hat{p}_j - p_0(j)\},$$

using optimal linear functional estimators of Rohde and Steinberger (2020). See also Butucea, Rohde and Steinberger (2020).

Interactive procedure

With first half of sample generate

$$Z_{ij} = \mathbb{1}_{\{X_i = j\}} + \frac{2}{\alpha} W_{ij}, \qquad j \in B$$

and calculate $\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} Z_{ij}$.

With second half, set $c_\alpha = \frac{e^\alpha + 1}{e^\alpha - 1}$ and $\tau = (n\alpha^2)^{-1/2}$ and generate Z_i in $\{-c_\alpha \cdot \tau, c_\alpha \cdot \tau\}$ with

$$\mathbb{P}(Z_i = c_\alpha \cdot \tau | X_i = j) = \frac{1}{2}\left(1 + \frac{[\hat{p}_j - p_0(j)]_{-\tau}^{\tau}}{c_\alpha \cdot \tau}\right),$$

where $[v]_{-\tau}^{\tau} = (-\tau) \vee v \wedge \tau$. Reject if T_B is large or if

$$D_B = \frac{1}{n} \sum_{i=n+1}^{2n} Z_i - \sum_{j=1}^{d} p_0(j)[\hat{p}_j - p_0(j)]_{-\tau}^{\tau}$$

is large.

Continuous case

Methods/results from discrete GoF testing can be extended to case of continuous distributions with Hölder smooth densities (Dubois et al., 2022).

f_0	Non-interactive	Interactive	Non-private
$\mathcal{U}([a,b])$	$(n\alpha^2)^{-2/7}$	$(n\alpha^2)^{-1/3}$	$n^{-2/5}$
$\mathcal{N}(0,1)$	$(n\alpha^2)^{-2/7}$	$(n\alpha^2)^{-1/3}$	$n^{-2/5}$
Beta(a,b)	$(n\alpha^2)^{-2/7}$	$(n\alpha^2)^{-1/3}$	$n^{-2/5}$
Cauchy$(0,a)$	$(n\alpha^2)^{-2/13}$	$(n\alpha^2)^{-1/5}$	$n^{-2/5}$
Pareto(a,k)	$(n\alpha^2)^{-2k/(7k+6)}$	$(n\alpha^2)^{-k/(3k+2)}$	$n^{-2k/(2+3k)}$

Table: Examples of $\mathbb{L}_1$ testing rates (up to log factors) for Lipschitz densities. The non-private rates can be found in Balakrishnan and Wasserman (2019).

Conclusion

By considering simple hypothesis testing we see links between robust statistics and LDP (also in mean/median estimation, density estimation...)

LDP constraints reduce the effective sample size and can change rates of convergence.

With more complex problems there can be a gap between non-interactive and sequentially interactive rates.

Thank you!

Li, M., B. and Yu, Y. (2022+) On robustness and local differential privacy. `arXiv:2201.00751`.

B. and Butucea, C. (2020) Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *NeurIPS 34*.

Dubois, A., B. and Butucea, C. (2022) Goodness-of-fit testing for Hölder continuous densities under local differential privacy. *Foundations of Modern Statistics – Festschrift in Honor of Vladimir Spokoiny*

References

Acharya, J., Canonne, C. L., Freitag, C. and Tyagi, H. (2019a) Test without trust: Optimal locally private distribution testing. *AISTATS 2019*.

Avella-Medina, M. (2021) Privacy-preserving parametric inference: a case for robust statistics. *J. Amer. Statist. Assoc.*, **116**, 969–983.

Balakrishnan, S. and Wasserman, L. (2019) Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.*, **47**, 1893–1927.

B. and Butucea, C. (2020) Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. *NeurIPS 34*.

Butucea, C., Rohde, A. and Steinberger, L. (2020) Interactive versus non-interactive locally, differentially private estimation: Two elbows for the quadratic functional. *Available at:* `arXiv:2003.04773`.

Cohen, A. and Nassim, K. (2020) Towards formalizing the GDPR's notion of singling out. *PNAS*, **117**, 8344–8352.

Chen, M., Gao, C. and Ren, Z. (2016) A general decision theory for Huber's ε-contamination model. *Electronic Journal of Statistics*, **10**, 3752–3774.

Ding, B., Kulkarni, J., and Yekhanin S. (2017) Collecting telemetry data privately. *NeurIPS*, 3571–3580.

References

Duchi, J. C., Jordan, M. I. and Wainwright, M. J. (2013) Local privacy and minimax bounds: Sharp rates for probability estimation. *NeurIPS*, 1529–1537.

Dubois, A., B., Butucea, C. (2022) Goodness-of-fit testing for Hölder continuous densities under local differential privacy. *Foundations of Modern Statistics – Festschrift in Honor of Vladimir Spokoiny*

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 265–284.

Dwork, C. and Lei, J. (2009) Differential privacy and robust statistics. *Annual ACM Symposium on Theory of Computing*, 371–380.

Dwork, C. (2019) Differential privacy and the US census. *PODS*.

Erlingsson, U., Pihur, V. and Korolova, A. (2014) Rappor: Randomized aggregatable privacy-preserving ordinal response. *Proc. 2014 ACM SIGSAC conference on computer and communications security*, 1054–1067.

Gopi, S., Kamath, G., Kulkarni, J., Nikolov, A., Wu, Z. S. and Zhang, H. (2020) Locally private hypothesis selection. *Conference on Learning Theory*, 1785–1816.

Li, M., B. and Yu, Y. (2022) On robustness and local differential privacy. *Available at* `arXiv:2201.00751`.

References

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008) Privacy: Theory meets practice on the map. *IEEE 24th international conference on data engineering*, 277–286.

Near, J. (2018) Differential privacy at scale: Uber and Berkeley collaboration. *Enigma 2018*.

Rohde, A. and Steinberger, L. (2020) Geometrizing rates of convergence under local differential privacy constraints. *Ann. Statist.*, **48**, 2646–2670.

Sweeney, L. (2002) k-anonymity: A model for protecting privacy. *Fuzziness and Knowledge-Based Systems*, **10**, 557–570.

Tang J., Korolova, A., Bai, X., Wang, X. and Wang X. (2017) Privacy loss in Apple's implementation of differential privacy on macOS 10.12. *Available at* `arXiv:1709.02753`.

Valiant, G. and Valiant, P. (2014) An automatic inequality prover and instance optimal identity testing. *FOCS 2014*.

Valiant, G. and Valiant, P. (2017) An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, **46**, 429–455.

Warner, S. L. (1965) Randomized sesponse: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, **60**, 63–69.