Differentially Private Linear Regression

Daniel Alabi

Joint work with *Audra McMillan (BU/Northeastern/Apple), Jayshree Sarathy, Adam Smith (BU), and Salil Vadhan*

Main paper: https://arxiv.org/abs/2007.05157

To learn more about differential privacy: http://people.seas.harvard.edu/~salil/cs208/

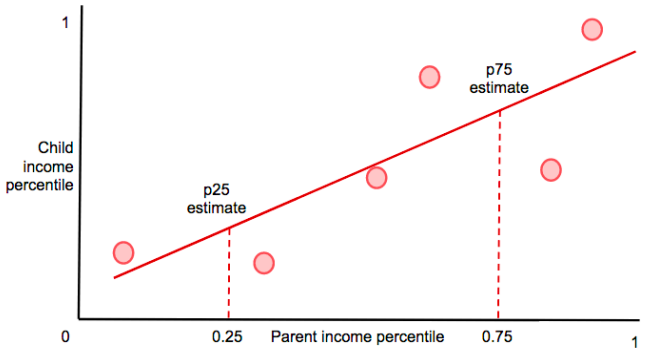
My email: alabid@g.harvard.edu

Quick Overview of Terms

- Differential Privacy (DP): mathematically rigorous definition of individual-level data privacy

- In statistics/ML/CS, linear regression (LR): linear approach to modeling the relationship between a scalar/dependent variable and one or more explanatory/independent variables

- Simple linear regression: LR for 1 explanatory variable



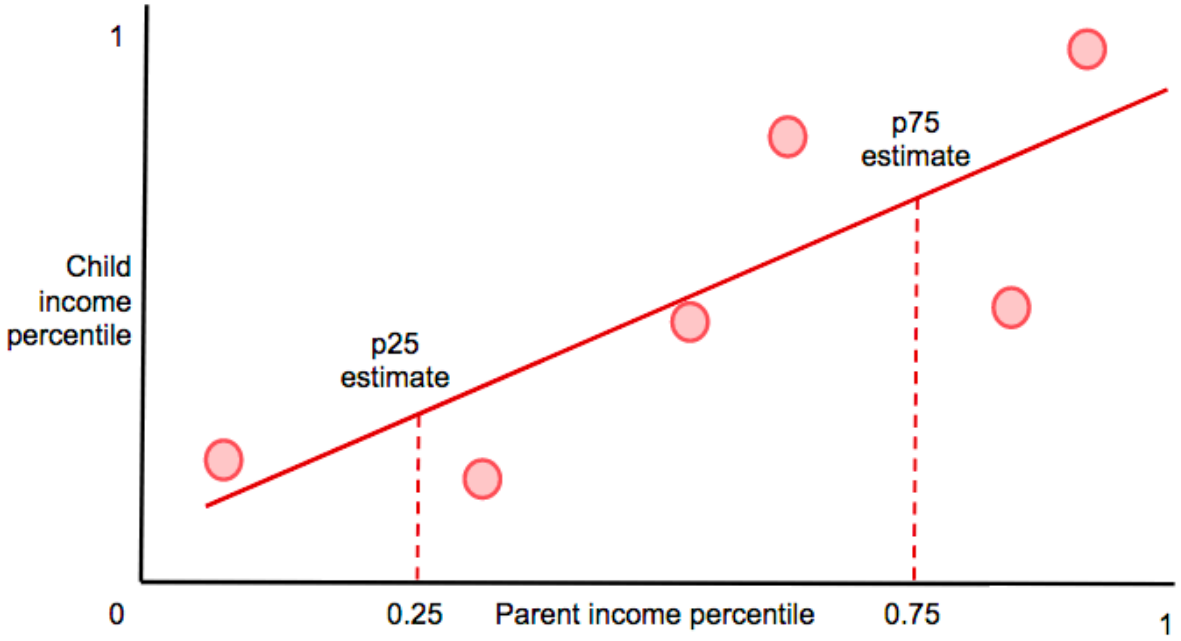
Main Question

Is it possible to design DP linear regression algorithms where the distortion added by the private algorithm is less than the standard error, even for small datasets?

Motivating Application: Opportunity Atlas

Opportunity Insights Application

- Neighborhood-level predictions of social/economic mobility via simple linear regression



OI team provide noise infusion algorithm (not formally private) [Chetty-Friedman ‘19] with sufficient accuracy (i.e., error due to privacy less than standard error). We provide formally private algorithms for this problem based on robust linear regression estimators (rather than OLS).

Table of Contents

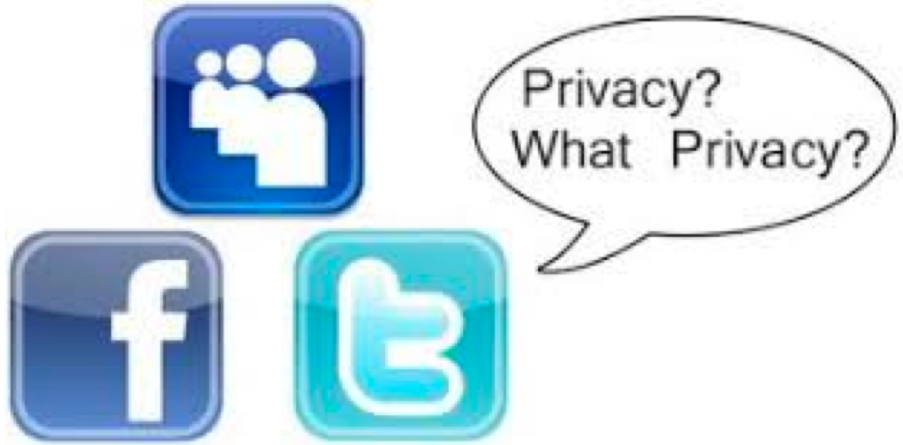
The General Problem

We have a dataset with sensitive information, such as:

1. Health records (e.g., reveals which disease a patient has)
2. Census data (e.g., reveals income range)
3. Social network activity (e.g., which pages you like)

How can we:

1. Allow the use of the data?
2. Protect the privacy of the data subjects?
3. Achieve both (1) and (2)?



Some Approaches to Solve the Problem

"Anonymize the Data": Are we happy with this solution? Why or why not?

Name	Sex	Blood	…	HIV?
James	M	B	…	N
Peter	M	O	…	Y
…	…	…	…	…
Paul	M	A	…	N
Eve	F	B	…	Y

Name	Sex	Blood	…	HIV?
XXXXX	M	B	…	N
XXXXX	M	O	…	Y
…	…	…	…	…
XXXXX	M	A	…	N
XXXXX	F	B	…	Y

Some Approaches to Solve the Problem

"Anonymize the Data": Not sufficient because of linkage attacks!

87% of US population have unique date of birth, gender, and postal code!

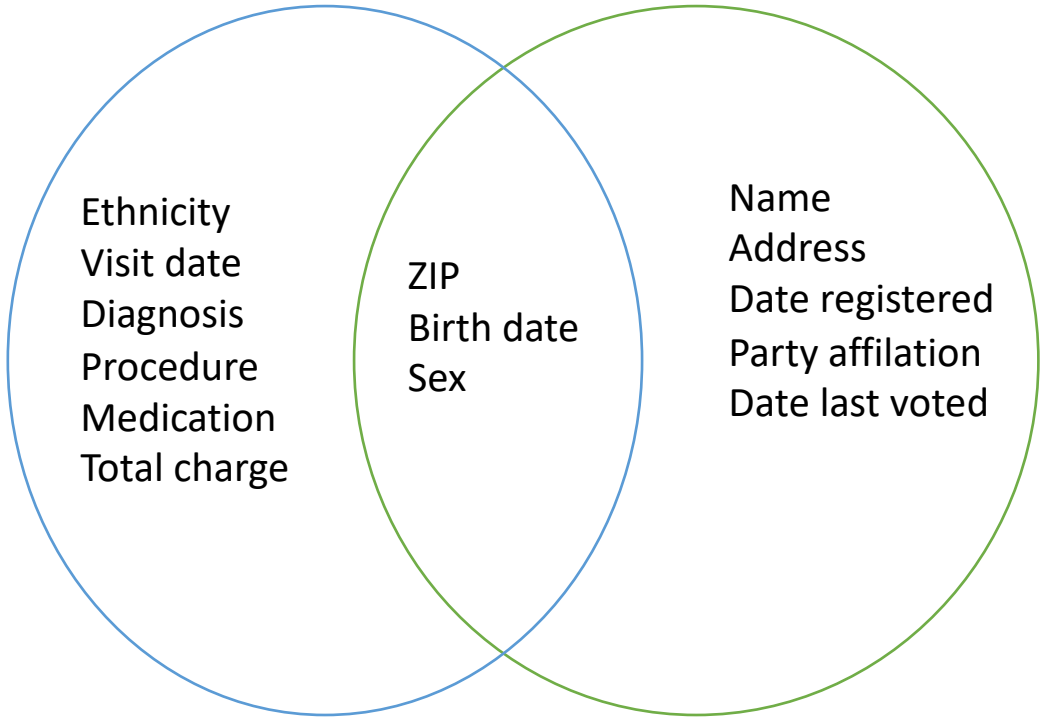
[Golle and Partridge '09]

Some Approaches to Solve the Problem

"Anonymize the Data": Reidentification via Linkage

Can uniquely identify > 60% of the U.S. population [Sweeney '00, Golle '06, Sweeney '97]

Name	Sex	Blood	…	HIV?
XXXXX	M	B	…	N
XXXXX	M	O	…	Y
…	…	…	…	…
XXXXX	M	A	…	N
XXXXX	F	B	…	Y



The story so far

- Motivations
  - Reidentification via Linkage Attacks
  - Reconstruction and Inference Attacks

Reconstruction attack: If we have dataset $x \in \{0,1\}^n$ and person i has sensitive bit x_i and attacker/adversary gets $q_S(x) = \sum_{i \in S} x_i$ for $O(n)$ random $S \subseteq [n]$.

The story so far

- Motivations
  - Reidentification via Linkage Attacks
  - Reconstruction and Inference Attacks

Reconstruction attack: If we have dataset $x \in \{0,1\}^n$ and person i has sensitive bit x_i and attacker/adversary gets $q_S(x) = \sum_{i \in S} x_i$ for $O(n)$ random $S \subseteq [n]$.

[Dinur-Nissim '03]: With high probability, adversary can reconstruct 0.99 fraction of the dataset $x \in \{0,1\}^n$ if noise added to each query is less than $o(\sqrt{n})$.

The story so far

- Motivations
  - Reidentification via Linkage Attacks
  - Reconstruction and Inference Attacks

Reconstruction attack: If we have dataset $x \in \{0,1\}^n$ and person i has sensitive bit x_i and attacker/adversary gets $q_S(x) = \sum_{i \in S} x_i$ for $O(n)$ random $S \subseteq [n]$.
[Dinur-Nissim '03]: With high probability, adversary can reconstruct 0.99 fraction of the dataset $x \in \{0,1\}^n$ if noise added to each query is less than $o(\sqrt{n})$.

Inference attack: Attacker gets $\tilde{O}(n^2)$ count queries with noise $o(n)$ and needs to know if someone is in the dataset or not.

The story so far

Message

Releasing too many statistics with too much accuracy can lead to a reconstruction of the entire dataset or inference attacks

Some Approaches to Solve the Problem

So now what?

- Encryption doesn’t work
- Anonymization doesn’t work
- Even adding insufficient noise to an attacker’s query is not good enough

Some Approaches to Solve the Problem

So now what?

- Encryption doesn’t work
- Anonymization doesn’t work
- Even adding insufficient noise to an attacker’s query is not good enough

Possible Responses:

- Privacy is an illusion!

Some Approaches to Solve the Problem

So now what?

- Encryption doesn't work
- Anonymization doesn't work
- Even adding insufficient noise to an attacker's query is not good enough

Possible Responses:

- Privacy is an illusion!
- In the long run, it's better to use data for research!

  Ignore privacy!

Some Approaches to Solve the Problem

So now what?

- Encryption doesn’t work
- Anonymization doesn’t work
- Even adding insufficient noise to an attacker’s query is not good enough

Possible Responses:

- Privacy is an illusion!
- In the long run, it’s better to use data for research! Ignore privacy!
- Never release statistics about any dataset!

Some Approaches to Solve the Problem

So now what?

- Encryption doesn’t work
- Anonymization doesn’t work
- Even adding insufficient noise to an attacker’s query is not good enough

Possible Responses:

- Privacy is an illusion!
- In the long run, it’s better to use data for research! Ignore privacy!
- Never release statistics about any dataset!
- Is there a way to add enough noise to queries and still allow for usefulness?

Main Message of this Talk

Yes, there is a way to add enough noise to queries and still allow for usefulness!

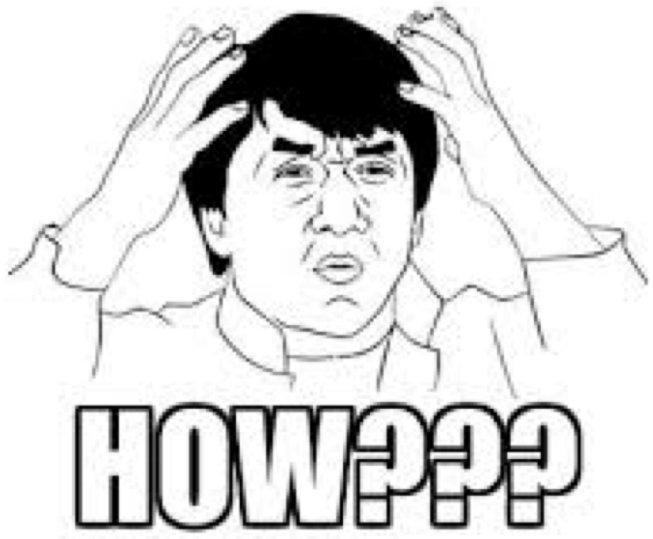


Table of Contents

Differential Privacy

- Utility
- Privacy
- Definition

Differential Privacy

- Utility: enable “statistical analysis” on datasets
  - Can release (noisy) statistics such as means, sums, regression estimates, etc.
  - Predictions from trained machine learning models and statistical models

Differential Privacy

- Utility: enable “statistical analysis” on datasets
  - Can release (noisy) statistics such as means, sums, regression estimates, etc.
  - Predictions from trained machine learning models and statistical models
- Privacy: protect each individual in dataset against all possible attack strategies
  - Now and in the future!
  - Even with use of auxiliary information or datasets!
  - Group privacy also allowed!

Differential Privacy

- Utility: enable “statistical analysis” on datasets
  - Can release (noisy) statistics such as means, sums, regression estimates, etc.
  - Predictions from trained machine learning models and statistical models
- Privacy: protect each individual in the dataset against all possible attack strategies
  - Now and in the future!
  - Even with use of auxiliary information or datasets!
  - Group privacy also allowed!
- Definition: pure and approximate

Differential Privacy: Definition

[Dwork-McSherry-Nissim-Smith ‘06]

Other references:

Motivated from and based off of work in

[Dinur-Nissim ‘03, Dwork-Nissim ‘04, Blum-Dwork-McSherry-Nissim ‘05]

Differential Privacy: Definition

[Dwork-McSherry-Nissim-Smith ‘06]

Intuition: for a statistic, the effect of each individual (whether in the dataset or not) should be close to nothing.

Differential Privacy: Definition

[Dwork-McSherry-Nissim-Smith ‘06]

Intuition: for a statistic, the effect of each individual (whether in the dataset or not) should be close to nothing.

Worst-case notion: protects against all possible adversaries and any kind of individual.

Differential Privacy: Definition

[Dwork-McSherry-Nissim-Smith ‘06]

For any algorithm $\mathcal{A}$, it satisfies ϵ differential privacy if

For all datasets D, D' differing in exactly one row all queries q

Distribution of $\mathcal{A}(D, q)$ is at most ϵ away from $\mathcal{A}(D', q)$

The smaller ϵ is, the more privacy is ensured!

Differential Privacy: Definition

[Dwork-McSherry-Nissim-Smith ‘06]

For any algorithm $\mathcal{A}$, it satisfies ϵ differential privacy if

For all datasets D, D' differing in exactly one row all queries q

Distribution of $\mathcal{A}(D, q)$ is at most ϵ away from $\mathcal{A}(D', q)$

For all sets T,

$\Pr[\mathcal{A}(D, q) \in T] \leq (1 + \epsilon) \Pr[\mathcal{A}(D', q) \in T]$

Simple Linear Regression

Assume that:

1) $\forall\, i \in [n],\ y_i = \alpha \cdot x_i + \beta + e_i,$ e_i are error terms

2) $\forall i \in [n],\quad x_i \in \mathbb{R}$

Simple Linear Regression

Assume that:

1) $\forall\, i \in [n],\ y_i = \alpha \cdot x_i + \beta + e_i,$ e_i are error terms

2) $\forall i \in [n],$ $x_i \in \mathbb{R}$

$\hat{\alpha}, \hat{\beta}$ are non-DP estimates of α, β

The goal is to calculate and release DP estimates of:

$\hat{\alpha}, \hat{\beta}$ or $\hat{p}_{25} = 0.25 \cdot \hat{\alpha} + \hat{\beta}, \hat{p}_{75} = 0.75 \cdot \hat{\alpha} + \hat{\beta}.$

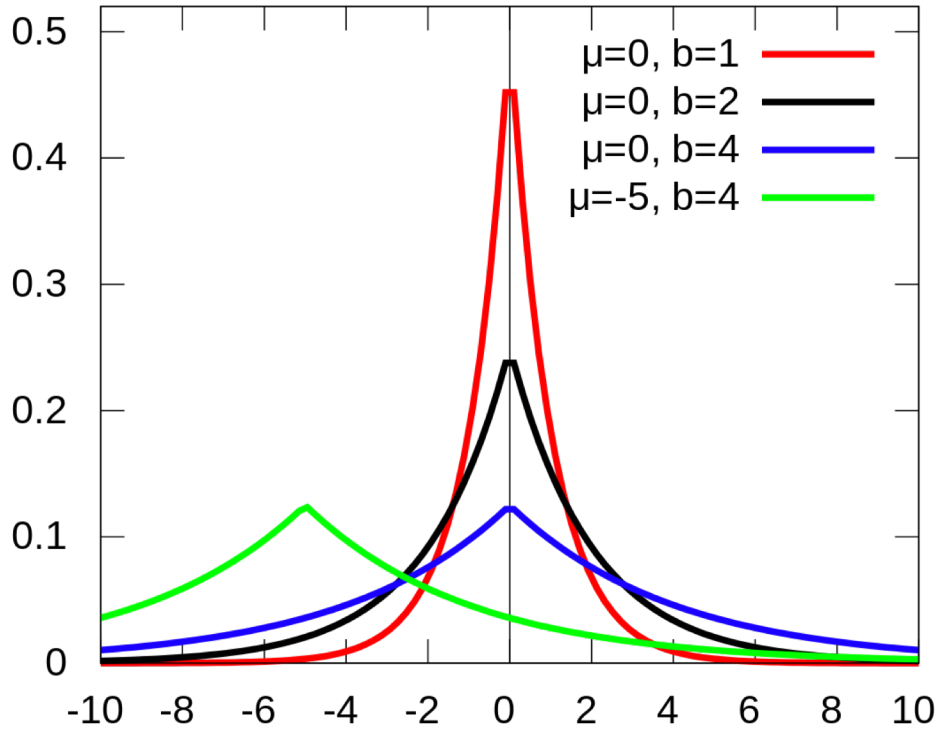
Table of Contents

Warmup: Laplace Mechanism

$\mathrm{Lap}(\mu, b)$ is the Laplace Distribution with scale b and mean μ.

Some properties:

- Has standard deviation $\sqrt{2} \cdot b$
- It's a "double-exponential" distribution



Warmup:
Laplace Mechanism for Sum and Average

1. $\mathcal{A}(x) = \sum_i x_i + \text{Lap}(0, \frac{1}{\epsilon})$

where $x_i \in [0, 1]$ for all $i \in [n]$.

2. $\mathcal{A}(x) = \frac{1}{n} \cdot \sum_i x_i + \text{Lap}(0, \frac{1}{n \cdot \epsilon})$

where $x_i \in [0, 1]$ for all $i \in [n]$.

Scales of Laplace distribution calculated based on how much a datapoint can affect a statistic.

DPSuffStats (closest to OLS estimator)

$$X = (x_1, \ldots, x_n)^T, Y = (y_1, \ldots, y_n)^T$$

OLS estimator

$$\hat{\alpha}^{OLS} = ncov(X, Y)/nvar(X)$$

where the sufficient statistics are

- $ncov(X, Y) = (X - \bar{X})^T (Y - \bar{Y})$
- $nvar(X) = (X - \bar{X})^T (X - \bar{X})$

To make DP, add Laplace noise to the sufficient statistics.

Advantages: as efficient as OLS; can release sufficient statistics (for other tasks); geometric interpretation

DPGradDescent

Convex optimization problem that defines OLS:

Minimize $||Y - (\alpha \cdot X + \beta)||^2$

Solve using gradient descent.

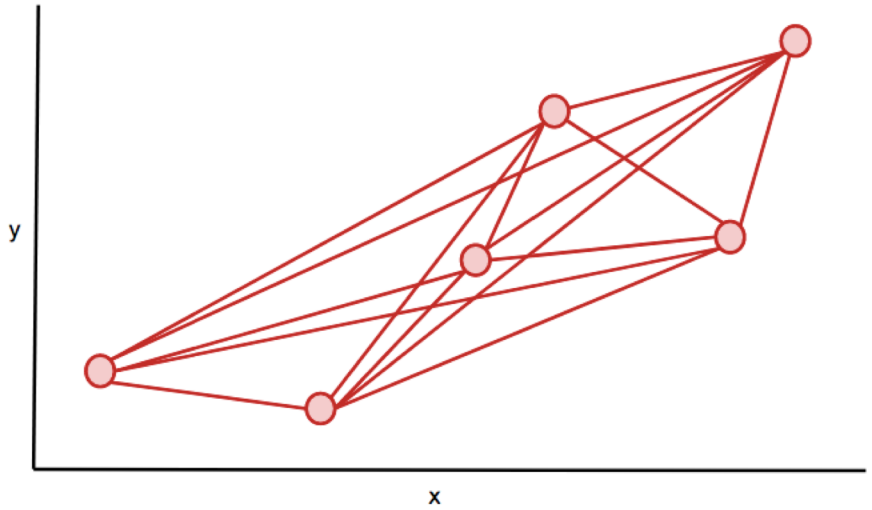
To make DP, add noise to the gradient computation process.

Advantages: inherits most benefits of non-private gradient descent (e.g., parallelizability, fine-tuning of optimization steps)

DPTheilSen

Theil-Sen estimator (Theil 50, Sen 68):
1) For $i \neq j \in [n], X_i \neq X_j$, compute slopes of pair of points as follows:
$Z_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$.
2) Compute median of the Z's.

To make DP, make median computation DP.

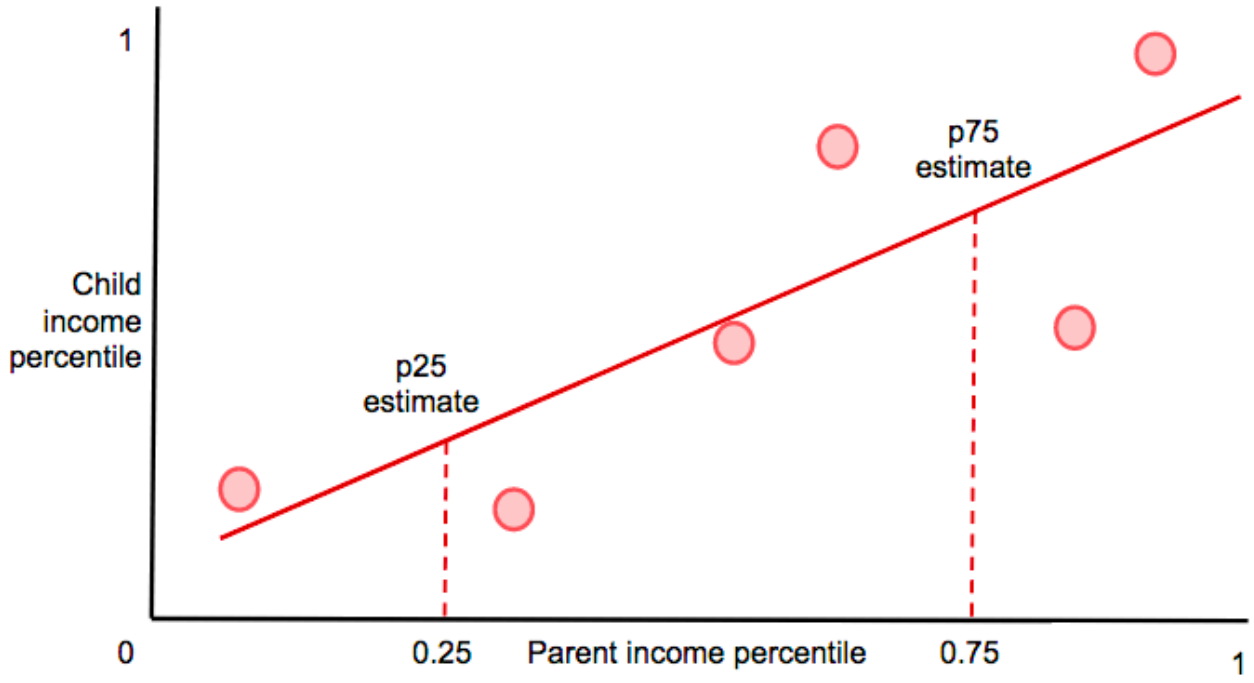


Advantages: performs best on smallest datasets (e.g., tens or hundreds); as DP median computation gets better (i.e., more accurate, faster), DPTheilSen gets better.

Table of Contents

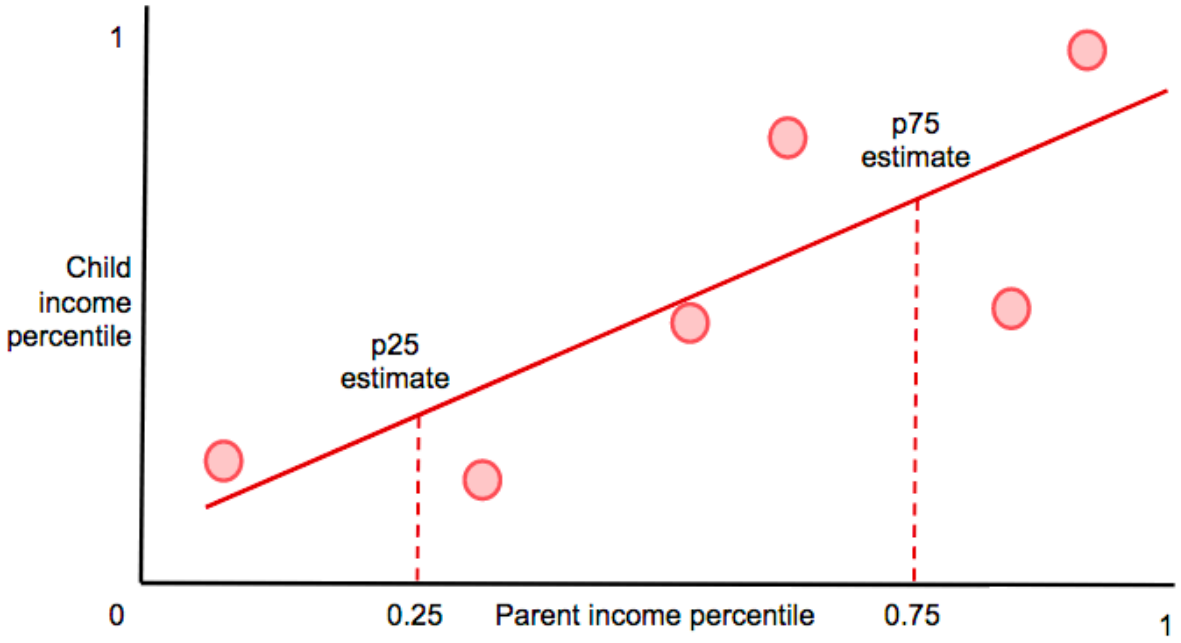
Opportunity Insights Application

- Neighborhood-level predictions of social mobility via simple linear regression



Opportunity Insights Application

- Neighborhood-level predictions of social mobility via simple linear regression



They provide noise infusion algorithm (not formally private) [Chetty-Friedman ‘19] with sufficient accuracy (i.e., error due to privacy less than standard error).

Error Metrics

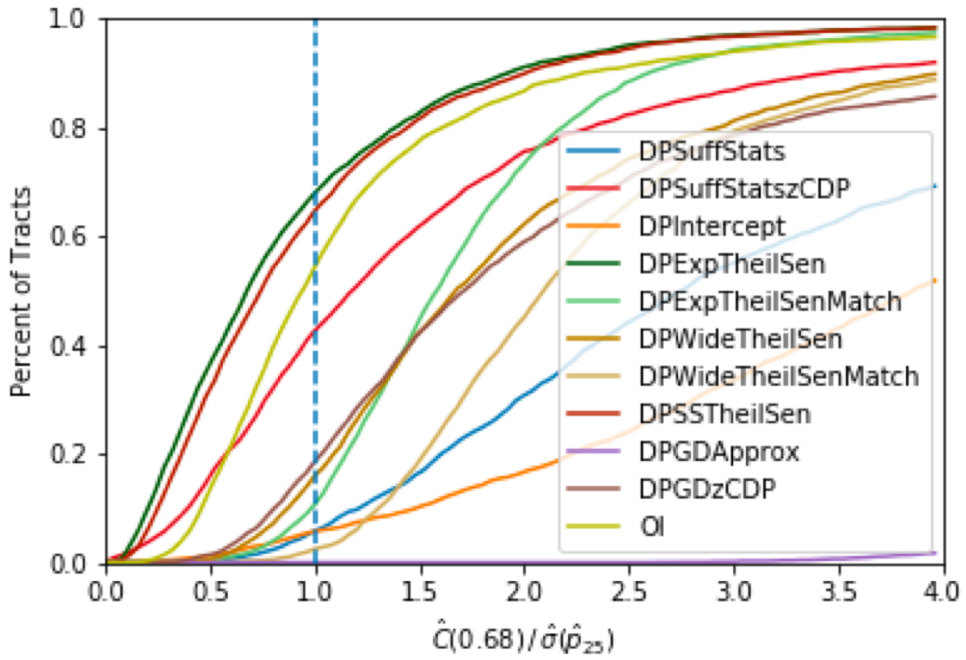
Recall that the goal is to calculate and release DP estimates of:

$\hat{\alpha}, \hat{\beta}$ or $\hat{p}_{25} = 0.25 \cdot \hat{\alpha} + \hat{\beta}, \hat{p}_{75} = 0.75 \cdot \hat{\alpha} + \hat{\beta}$.

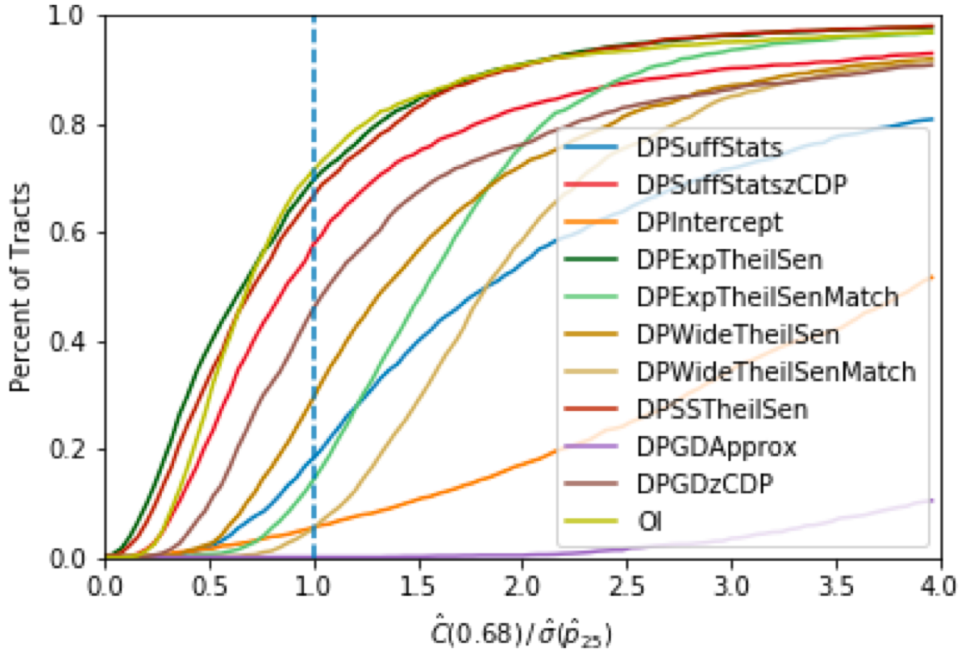
We use tildes to denote private estimates and hats for non-private estimates.

- $C(q) = \min\{c : \mathbb{P}(|\tilde{p}_{25} - \hat{p}_{25}| \leq c) \geq q\}$ for any $q \in [0, 1]$
- $\hat{C}(q) = \min\{c : \geq q\ fraction\ of\ trials\ have\ error \leq c\}$ for any $q \in [0, 1]$
- $\sigma(p) =$ standard deviation of estimator p under a noise model (i.e., error terms gaussian)
- $\hat{\sigma}(p) =$ standard error of estimator p = empirical estimate of $\sigma(p)$

Opportunity Insights (OI) Application (IL, NC)



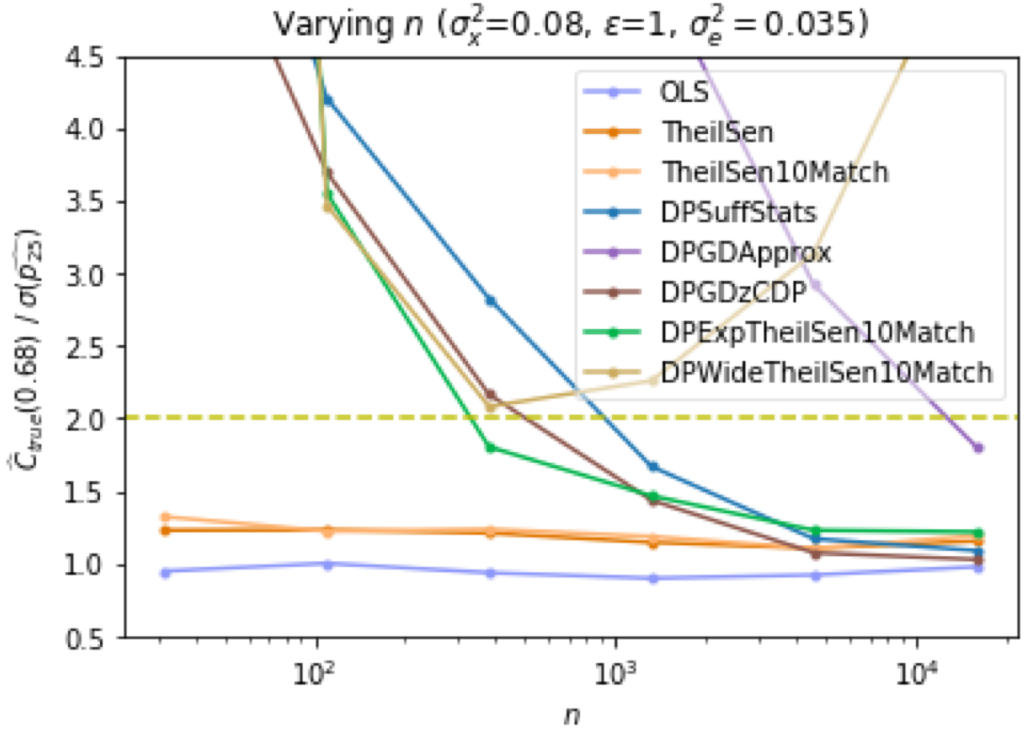
Illinois

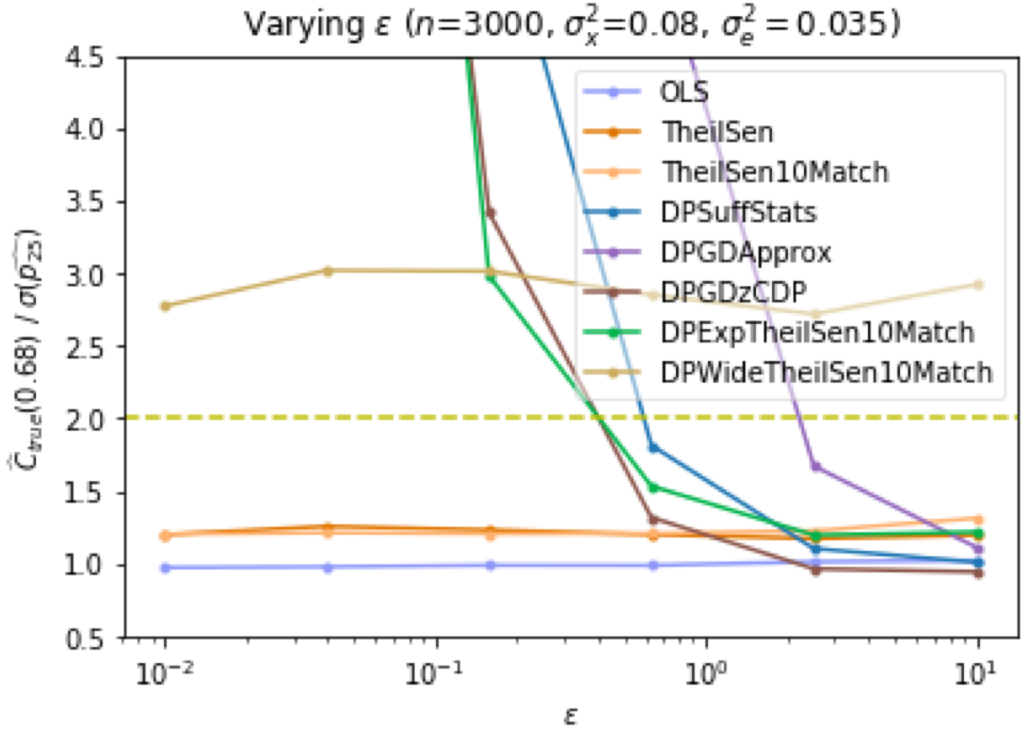


North Carolina

Main Takeaway: DPExpTheilSen outperforms or matches the OI method

Synthetic Datasets





Main Takeaway: Private robust methods (e.g., DPExpTheilSen) approach non-private robust methods; Private non-robust methods (e.g., DPSuffStats) approach non-private non-robust methods; there's a cross-over between private robust/non-robust methods

Conclusion

- Differential Privacy is a mathematically rigorous definition of individual data privacy.
- It’s being used by the U.S. Census Bureau (for the 2020 Decennial Census), Google, Apple, Facebook.
- It is possible to design DP simple linear regression algorithms where the distortion added by the private algorithm is less than the standard error, even for small datasets.

Future Work (with you?). Any Questions?

	[DL09]	[ZZX+12]	[DJW13]	[BST14]	[She15]	[She17]
Uses Bayesian Approach?	No	No	No	No	No	No
Point Estimates?	Yes	Yes	Yes	Yes	Yes	Yes
Uncertainty Estimates?	No	No	No	No	No	Yes
Multiple Linear Regression?	Yes	Yes	Yes	Yes	Yes	Yes
Small Dataset (e.g., ≤ 500)	Yes	No	No	No	No	No
Ridge/OLS/Robust Estimator	Robust	OLS	OLS	OLS	Ridge	OLS/Ridge
Distributed?	No	No	Yes	No	No	No

	[STU17]	[Wan18]	[CKS+19]	[BS19]	[CF19]	[AMS+20]
Uses Bayesian Approach?	No	No	No	Yes	No	No
Point Estimates?	Yes	Yes	Yes	Yes	Yes	Yes
Uncertainty Estimates?	No	No	Yes	No	No	No
Multiple Linear Regression?	Yes	Yes	Yes	Yes	No	No
Small Dataset (e.g., ≤ 500)	No	No	No	No	Yes	Yes
Ridge/OLS/Robust Estimator	OLS	Ridge	OLS/Robust	OLS	OLS	OLS/Robust
Distributed?	Yes	No	No	No	No	No