Formal privacy protection for data products combining individual and employer frames

Samuel Haney

October 4, 2015

Joint work with Ashwin Machanavajjhala, Mark Kutzbach, Matthew Graham, John Abowd, and Lars Vilhuber

Disclaimer & Acknowledgements

- Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

- Machanavajjhala acknowledges support from NSF grants 1253327, 1408982, and 1443014.
- Abowd acknowledges direct support from NSF grants BCS-0941226 and TC-1012593.
- Abowd and Vilhuber acknowledge support from NSF grant SES-1131848.
- This paper was written while Abowd was visiting the Center for Labor Economics at UC-Berkeley.

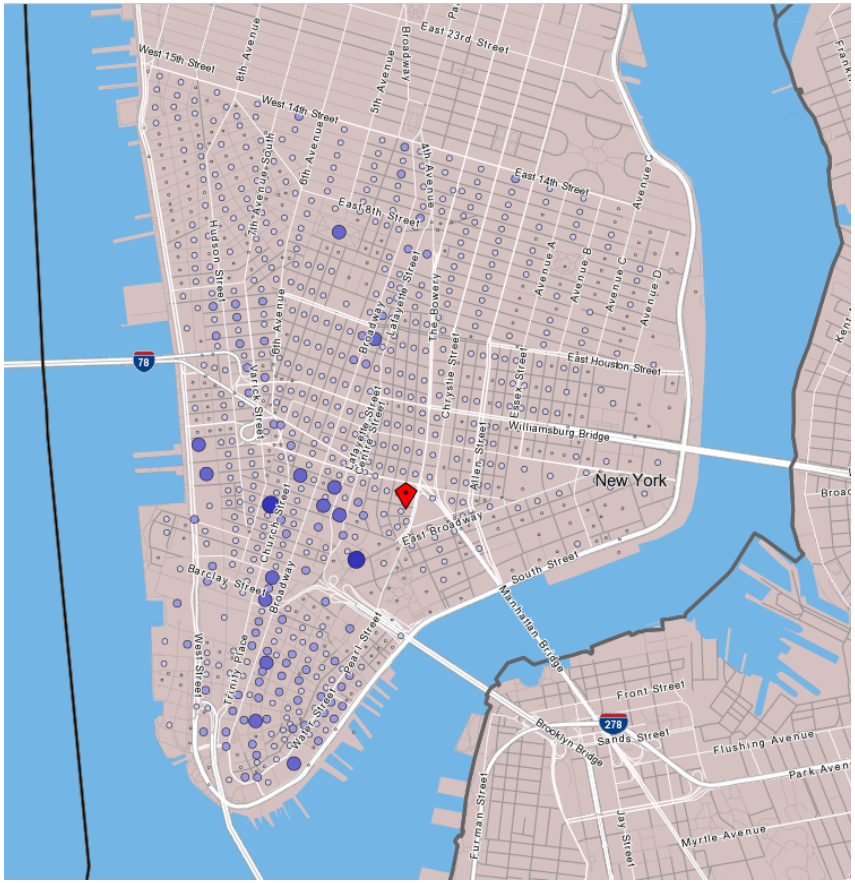
Background

- LEHD Origin-Destination Employment Statistics (LODES) are a public-use database of jobs (Worker-Firm connection) released by the U.S. Census Bureau.
- In the most recent data year, LODES constitutes:
  - 128 million jobs for 119 million workers at 6.2 million employer firms with 7.6 million establishments in the U.S.
  - 2 million census blocks with employment
  - 5.5 million census blocks with workers residing in them
- Firm characteristics: Location (2×10^6), industry (20), ownership (2)
- Person characteristics: Home location (5.5×10^6), age (3), sex (2), race (6), ethnicity (2), educational attainment (4), and job earnings (3)

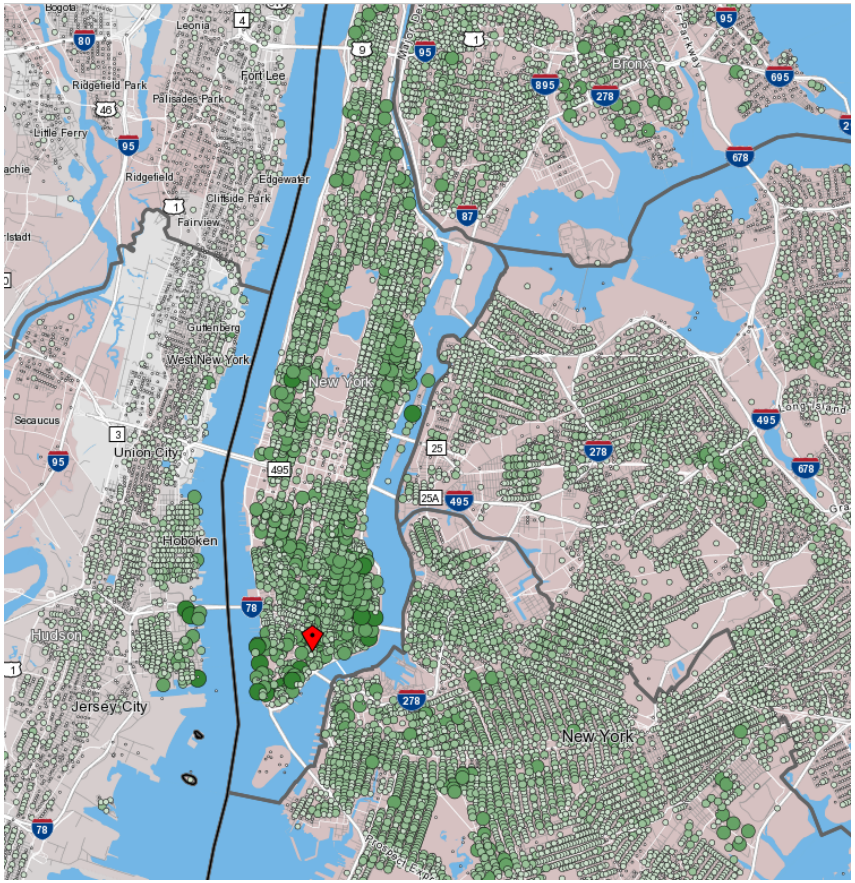
Existing Protections

- Under existing U.S. law, some characteristics are public and some are private.
  - LODES must protect the size and characteristics of a firm's workforce.
  - LODES must protect the existence of any particular job and its characteristics.
  - Not protected are the existence of an employer in a location, an industry, and an ownership class.
- LODES uses a permanent multiplicative noise distortion factor at the employer and establishment levels (plus synthetic methods for small cells) to protect employment counts. (Abowd et al, 2006)
- Residence location is protected by synthetic data methods using probabilistic differential privacy. (Machanavajjhala et al, 2008)

Existing LODES data in OnTheMap application



Employment in Lower Manhattan



Residences of Workers Employed in Lower Manhattan

Available at `http://onthemap.ces.census.gov/`.

Goals

- Answer marginal queries over individual characteristics (e.g. sex, age, salary), and public workplace characteristics (e.g. geographic location).

Goals

- Answer marginal queries over individual characteristics (e.g. sex, age, salary), and public workplace characteristics (e.g. geographic location).
- Give algorithms with a provable privacy guarantee for both individuals and employer businesses.

Goals

- Answer marginal queries over individual characteristics (e.g. sex, age, salary), and public workplace characteristics (e.g. geographic location).
- Give algorithms with a provable privacy guarantee for both individuals and employer businesses.
- Algorithms should perform comparably to the current protection system.

Goals

- Answer marginal queries over individual characteristics (e.g. sex, age, salary), and public workplace characteristics (e.g. geographic location).
- Give algorithms with a provable privacy guarantee for both individuals and employer businesses.
- Algorithms should perform comparably to the current protection system.

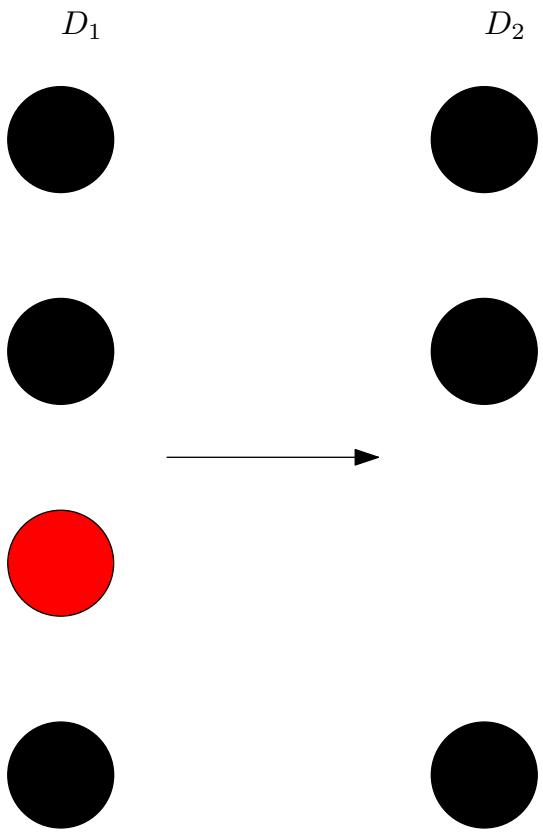
(For the purposes of this talk, assume queries are just total employment over some geographic region.)

Differential Privacy Guarantee

The output of an algorithm should be insensitive to the addition or removal of a single individual (or establishment).

Neighbors

Neighboring datasets differ in one entry.

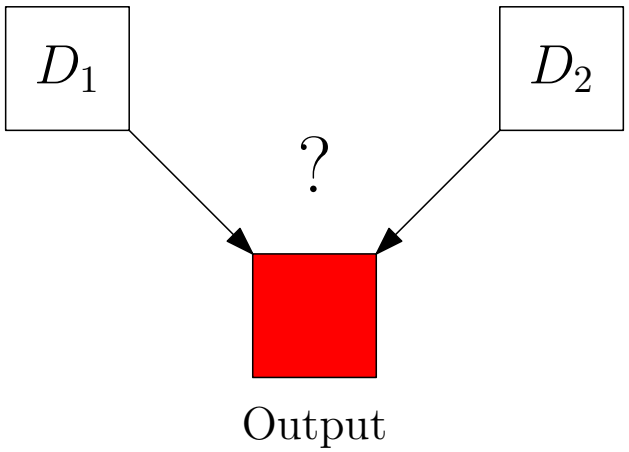


Differential Privacy

Definition (ϵ-Differential Privacy)

A mechanism $\mathcal{M}$ satisfies ϵ-differential privacy if for all outputs $S \subseteq range(\mathcal{M})$, and for all neighbors D_1 and D_2,

$$\Pr[\mathcal{M}(D_1) \in S] \leq e^{\epsilon} \cdot \Pr[\mathcal{M}(D_2) \in S]$$



Semantics

- Suppose an adversary believes that record r being in the dataset is c times as likely as r not in the dataset.
- After seeing the output, the adversary's new belief c' is bounded by $e^{-\epsilon}c \leq c' \leq e^{\epsilon}c$.

Sensitivity

Definition

Let N denote the set of pairs of neighboring datasets. The L_1 sensitivity of a query q is:

$$\Delta_q = \max_{(\mathbf{x},\mathbf{x}')\in N} |q(\mathbf{x}) - q(\mathbf{x}')|$$

Ensuring differential privacy generally requires adding noise proportional to the sensitivity.

Sensitivity

Example: The query that returns the total size (number of records) has sensitivity 1.

Laplace Mechanism

The Laplace mechanism is ϵ-differentially private.

$$\mathcal{L}(q, \mathbf{x}) = q(\mathbf{x}) + \mathrm{Lap}(\sigma)$$

where $\sigma = \Delta_q / \epsilon$.

Laplace Mechanism

The Laplace mechanism is ϵ-differentially private.

$$\mathcal{L}(q, \mathbf{x}) = q(\mathbf{x}) + \mathrm{Lap}(\sigma)$$

where $\sigma = \Delta_q / \epsilon$.

Is this approach any good?

Error

We can judge a mechanism $\mathcal{M}$ by the error it produces.

The per query L_1 error is

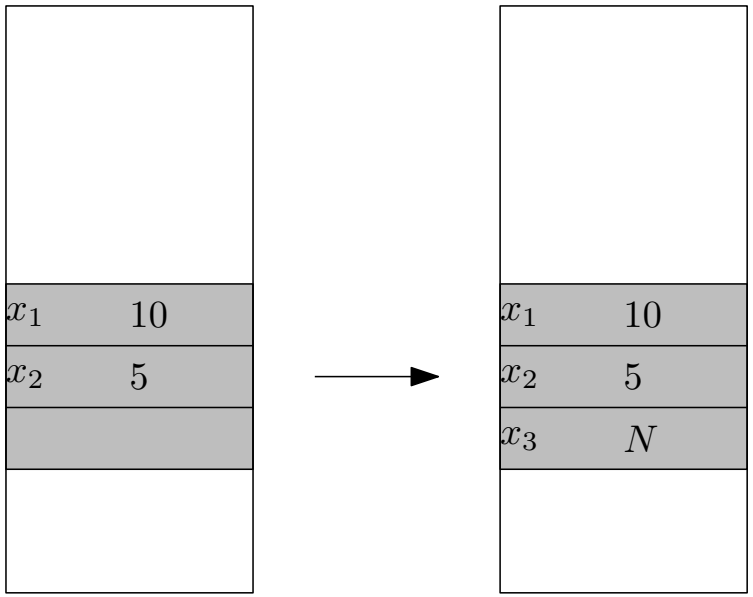
$$\mathbb{E}\left[|\text{true answer} - \text{noisy answer}|\right]$$

Error

For Laplace mechanism, error is Δ_q/ϵ.

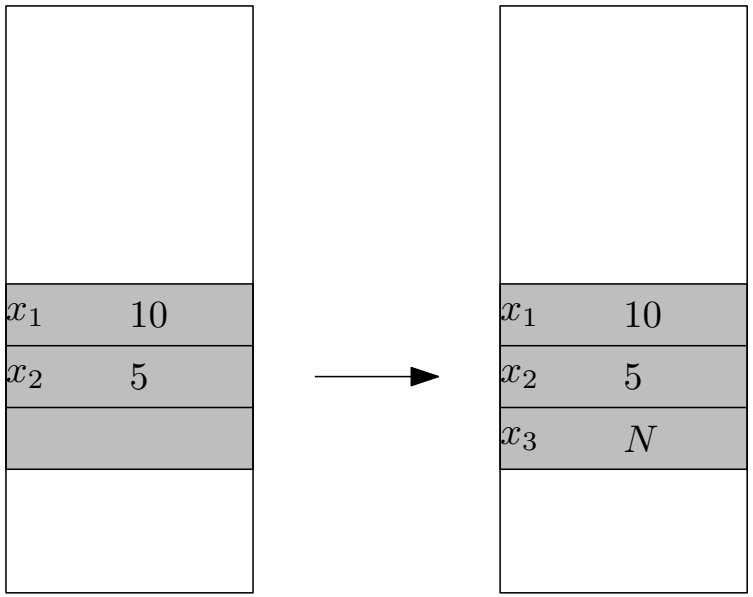
Error

For Laplace mechanism, error is Δ_q/ϵ.



Error

For Laplace mechanism, error is Δ_q/ϵ.



For our queries, Δ_q is the maximum allowable employment size!

Problem & Solution

Problem:

- The sensitivity is the maximum allowable employment.
- The error incurred will almost always dominate the count.

Problem & Solution

Problem:

- The sensitivity is the maximum allowable employment.
- The error incurred will almost always dominate the count.

Solution:

1. Provide a weaker (but still provable) privacy guarantee by re-defining neighboring databases.
2. Use the *local sensitivity* rather than the standard sensitivity (called global sensitivity).

Problem & Solution

Problem:

- The sensitivity is the maximum allowable employment.
- The error incurred will almost always dominate the count.

Solution:

1. **Provide a weaker (but still provable) privacy guarantee by re-defining neighboring databases.**
2. Use the *local sensitivity* rather than the standard sensitivity (called global sensitivity).

Employer protection under differential privacy is *too* strong:

- Existence of establishment need not be protected.
- We can assume it is common knowledge whether the employment is very large or very small.

Privacy Guarantee Relaxation

The output of a mechanism is not significantly changed due to the presence/absence of a single establishment.

$\longrightarrow$

The output of a mechanism is not significantly changed due to a change in employment (of a single establishment) from C to $C' \in [C - \alpha C, C + \alpha C]$.

Semantics

Suppose $C' \in [C - \alpha C, C + \alpha C]$ where C is the true employment of an establishment.

- Suppose an adversary believes that the employment of the establishment being C is p times as likely as C'.
- After seeing the output, the adversary's new belief p' is bounded by $e^{-\epsilon} p \leq p' \leq e^{\epsilon} p$.

Privacy Guarantee Relaxation

Definition (simplified)

$\mathbf{x}$ and $\mathbf{y}$ are neighbors if they satisfying the following:

- $\mathbf{x}$ and $\mathbf{y}$ differ in the employment of one establishment (let $C_{\mathbf{x}}, C_{\mathbf{y}}$ be those employments).
- $C_{\mathbf{x}} - \alpha C_{\mathbf{x}} \leq C_{\mathbf{y}} \leq C_{\mathbf{x}} + \alpha C_{\mathbf{x}}$.

Problem & Solution

Problem:

- The sensitivity is the maximum allowable employment.
- The error incurred will almost always dominate the count.

Solution:

1. Provide a weaker (but still provable) privacy guarantee by re-defining neighboring databases.
2. **Use the local sensitivity rather than the standard sensitivity (called global sensitivity).**

Local vs. Global sensitivity

Local sensitivity depends on the input database:

Definition

Let $N(\mathbf{x})$ denote neighbors of $\mathbf{x}$. The L_1 local sensitivity of a query q with respect to $\mathbf{x}$ is:

$$\Delta_q(\mathbf{x}) = \max_{\mathbf{x}' \in N(\mathbf{x})} |q(\mathbf{x}) - q(\mathbf{x}')|$$

Local vs. Global sensitivity

- Local sensitivity can vary greatly depending on database.

Local vs. Global sensitivity

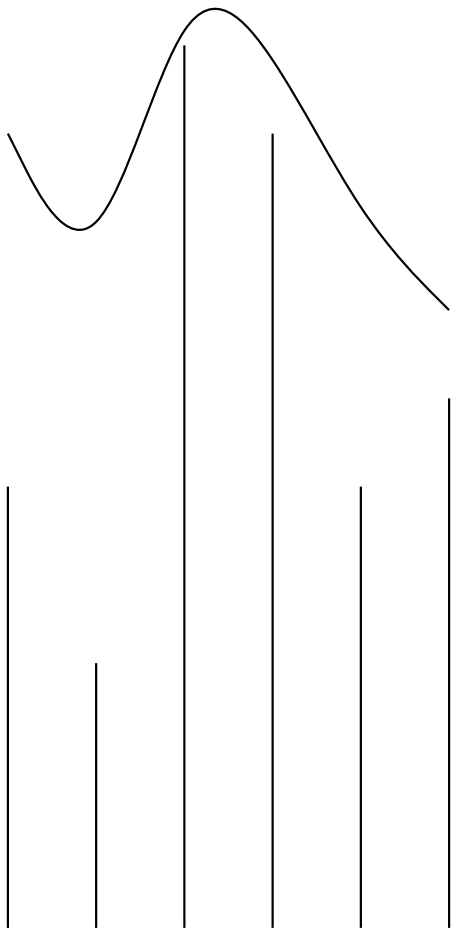
- Local sensitivity can vary greatly depending on database.
- Consider $median(0, 50, 100)$ versus $median(1, 2, 3)$.

Local vs. Global sensitivity

- Local sensitivity can vary greatly depending on database.
- Consider $median(0, 50, 100)$ versus $median(1, 2, 3)$.
- Local sensitivity is related to the largest employment in the query region, rather than the largest total employment.

Local vs. Global sensitivity

We add noise according to a smoothed upper bound on this sensitivity (Nissim et al).



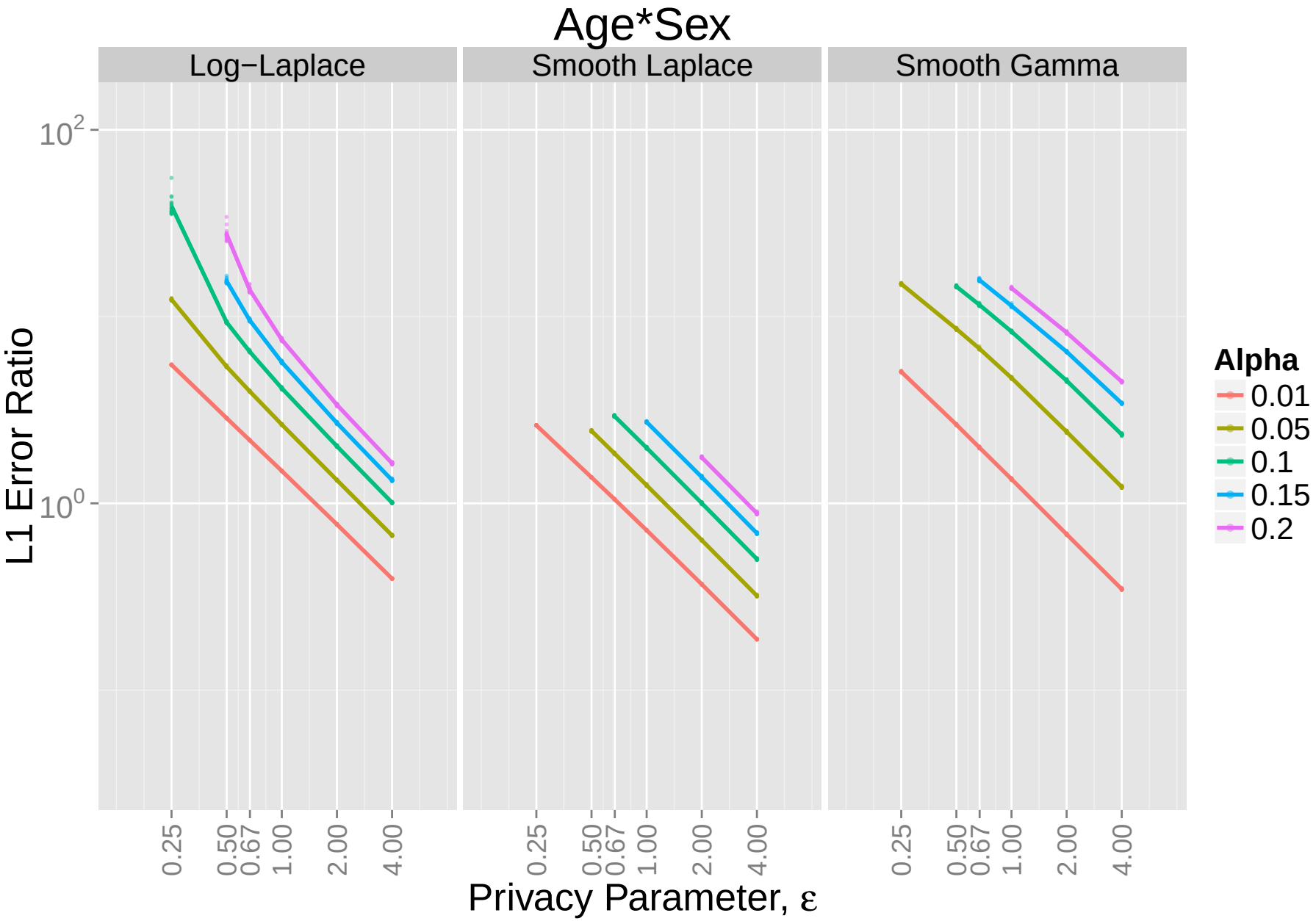
Extensions (see paper)

- Allow for queries on more attributes.

Extensions (see paper)

- Allow for queries on more attributes.
- Protect individuals under conventional differential privacy.

Experiments



Summary

We answer marginal queries over establishment and individual attributes, while providing provable protection to both establishments and individuals. Our algorithms have comparable error to the current protection system.

Questions?