Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics

October 25, 2016

Outline

- Preliminaries
- Dataset
- Privacy requirements
- Current SDL protection
- Applying differential privacy
- Employee-Employer privacy
- Algorithms and results

Combined employer-employee data

- whether or not a specific individual is employed
- the count of employees in a specific workplace

Preliminaries

- D is a table of records with schema $(A_1, ..., A_k)$
- $dom(A_i)$ is the domain of attribute A_i
- $dom(V)$ is the multidimensional domain $\times_{A \in V} dom(A)$ for the set of attributes $V = \{A_{i_1}, ..., A_{i_m}\}$
- for each record t in the table $t[A_i] \in dom(A_i)$ is a value of attribute A_i
- $n = |D|$ is the size of the table
- database with schema $(S_1, ..., S_m)$ is a collection of tables $D_1, ..., D_m$, where D_i has schema S_i

Preliminaries

- D is a table of records with schema $(A_1, ..., A_k)$
- $dom(A_i)$ is the domain of attribute A_i
- $dom(V)$ is the multidimensional domain $\times_{A \in V} dom(A)$ for the set of attributes $V = \{A_{i_1}, ..., A_{i_m}\}$
- for each record t in the table $t[A_i] \in dom(A_i)$ is a value of attribute A_i
- $n = |D|$ is the size of the table
- database with schema $(S_1, ..., S_m)$ is a collection of tables $D_1, ..., D_m$, where D_i has schema S_i

Preliminaries

- D is a table of records with schema $(A_1, ..., A_k)$
- $dom(A_i)$ is the domain of attribute A_i
- $dom(V)$ is the multidimensional domain $\times_{A \in V} dom(A)$ for the set of attributes $V = \{A_{i_1}, ..., A_{i_m}\}$
- for each record t in the table $t[A_i] \in dom(A_i)$ is a value of attribute A_i
- $n = |D|$ is the size of the table
- database with schema $(S_1, ..., S_m)$ is a collection of tables $D_1, ..., D_m$, where D_i has schema S_i

Marginal Query

Let $V = \{A_{i_1}, ..., A_{i_m}\}$ denote a subset of attributes chosen from D. Let $dom(V) = \times_{A \in V} dom(A)$. The marginal query $q_V(D)$ is defined as a vector of $|dom(V)|$ counts, one for each cell $\mathbf{v} = (v_1, ..., v_m) \in dom(V)$. The count corresponding to cell $\mathbf{v}$, denoted by $q_V(D, \mathbf{v})$ is

$$|\{t \in D | t[A_{i_1}] = v_1 \wedge ... \wedge t[A_{i_m}] = v_m\}| \tag{1}$$

$q_\emptyset(D)$ returns a single cell whose count is the size of the table.

SQL:

Select Count(*)
From D
Group By $A_{i_1}, ..., A_{i_m}$

Marginal Query

Let $V = \{A_{i_1}, ..., A_{i_m}\}$ denote a subset of attributes chosen from D. Let $dom(V) = \times_{A \in V} dom(A)$. The marginal query $q_V(D)$ is defined as a vector of $|dom(V)|$ counts, one for each cell $\mathbf{v} = (v_1, ..., v_m) \in dom(V)$. The count corresponding to cell $\mathbf{v}$, denoted by $q_V(D, \mathbf{v})$ is

$$|\{t \in D | t[A_{i_1}] = v_1 \wedge ... \wedge t[A_{i_m}] = v_m\}| \tag{1}$$

$q_\emptyset(D)$ returns a single cell whose count is the size of the table.

SQL:

Select Count(*)
From D
Group By $A_{i_1}, ..., A_{i_m}$

$(\epsilon, \delta) - \textit{DifferentialPrivacy}$

Let $\mathcal{M}$ be a randomized algorithm. Let the tables D and D' be a neighbors with the same schema. Then $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all D and D' and for all $S \subset range(\mathcal{M})$

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S] + \delta \tag{2}$$

Sensitivity

Let $\mathcal{I}$ denote the set of all tables with a given schema. let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query function on that table that outputs a vector of d real numbers. The sensitivity of q, denoted Δ_q, is

$$\Delta_q = max_{D,D' neighbors} \left\| q(D) - q(D') \right\|_1$$

$(\epsilon, \delta) - DifferentialPrivacy$

Let $\mathcal{M}$ be a randomized algorithm. Let the tables D and D' be a neighbors with the same schema. Then $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all D and D' and for all $S \subset range(\mathcal{M})$

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S] + \delta \tag{2}$$

Sensitivity

Let $\mathcal{I}$ denote the set of all tables with a given schema. let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query function on that table that outputs a vector of d real numbers. The sensitivity of q, denoted Δ_q, is

$$\Delta_q = max_{D,D' neighbors} \left\| q(D) - q(D') \right\|_1$$

Laplace Mechanism

Let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query on a table and let $\eta \sim Lap(\lambda)$. The algorithm which returns $\tilde{q}(D) = q(D) + \eta^d$ satisfies ϵ-differential privacy, where η^d is a vector of d independently drawn Laplace random variables.

Expected L_p Error

Let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query over a table and $\tilde{q}(D)$ be a noisy answer returned by an algorithm. The expected L_p error of the algorithm is

$$E(\|q(D) - \tilde{q}(D)\|_p)$$

where $\|x\|_p$ is the L_p norm, and expectation is over the randomness of the algorithm.

Laplace Mechanism

Let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query on a table and let $\eta \sim Lap(\lambda)$. The algorithm which returns $\tilde{q}(D) = q(D) + \eta^d$ satisfies ϵ-differential privacy, where η^d is a vector of d independently drawn Laplace random variables.

Expected L_p Error

Let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query over a table and $\tilde{q}(D)$ be a noisy answer returned by an algorithm. The expected L_p error of the algorithm is

$$E(\|q(D) - \tilde{q}(D)\|_p)$$

where $\|x\|_p$ is the L_p norm, and expectation is over the randomness of the algorithm.

Database structure

Workplace

workplace id	industry	ownership	geography

Worker

worker id	age	sex	race	ethnicity	education

Job

worker id	workplace id

*** *Assume each worker has exactly one job*

Privacy Requirements

- The existence of a job held by a particular individual is confidential
- The existence of an employer business as well as its type and location is not confidential
- Characteristics of an establishment's workforce must be protected

Formal Privacy

Uninformed attackers: do not possess detailed background knowledge about specific individuals and establishments in the data

Informed attackers: possess specific background knowledge about specific employers and employees in the data

Assumptions

- Adversary knows the set of all establishments $\mathcal{E}$ and their public attributes
- Adversary knows the universe U of all employees
- Each worker $w \in U$ has a set of private attributes $A_1, ..., A_k$
- For each emplyee w the attacker's belief is defined as π_w, a probability distribution over all values in $\mathcal{T}$
- Adversary's belief about all employees in U: $\theta = \prod_{w \in U} \pi_w$
- The set of all possible adversarial beliefs $\Theta = \{\theta\}$
- We distinguish subset of attackers $\Theta_{weak} \subset \Theta$ with a prior for each worker $\pi_w = \pi_{1,e} \times \pi_{2,w}$

Assumptions

- Adversary knows the set of all establishments $\mathcal{E}$ and their public attributes
- Adversary knows the universe U of all employees
- Each worker $w \in U$ has a set of private attributes $A_1, ..., A_k$
- For each emplyee w the attacker's belief is defined as π_w, a probability distribution over all values in $\mathcal{T}$
- Adversary's belief about all employees in U: $\theta = \prod_{w \in U} \pi_w$
- The set of all possible adversarial beliefs $\Theta = \{\theta\}$
- We distinguish subset of attackers $\Theta_{weak} \subset \Theta$ with a prior for each worker $\pi_w = \pi_{1,e} \times \pi_{2,w}$

No re-identification of individuals:

Employee Privacy Requirement

For randomized algorithm A, if for some $\epsilon \in (0, \infty)$, and for every employee $w \in U$, for every adversary $\theta \in \Theta$, for every $a, b \in \mathcal{T}$ such that $Pr_\theta[w = a] > 0$ and $Pr_\theta[w = b] > 0$, and for every output $\omega \in range(A)$:

$$log\left(\frac{Pr_{\theta,A}[w = a|A(D) = \omega]}{Pr_{\theta,A}[w = b|A(D) = \omega]} \Big/ \frac{Pr_\theta[w = a]}{Pr_\theta[w = b]}\right) \leq \epsilon \tag{3}$$

Then the algorithm protects employees against informed attackers at privacy level ϵ.

No precise inference of establishment size

Employer Size Requirement

A randomized algorithm A protects establishment size against an informed attacker at privacy level (ϵ, α) if, for every informed attacker $\theta \in \Theta$, for every pair of numbers x, y, and for every output $\omega \in range(A)$:

$$\left| log \left(\frac{Pr_{\theta,A}[|e| = x | A(D) = \omega]}{Pr_{\theta,A}[|e| = y | A(D) = \omega]} \Big/ \frac{Pr_{\theta}[|e| = x]}{Pr_{\theta}[|e| = y]} \right) \right| \leq \epsilon \qquad (4)$$

whenever $x \leq y \leq \lceil (1+\alpha)x \rceil$ and $Pr_\theta[|e| = x]$, $Pr_\theta[|e| = y] > 0$. We say that an algorithm weakly protects establishments against an informed attacker if the condition above holds for all $\theta \in \Theta_{weak}$.

No precise inference of establishment shape

Employer Shape Requirement

Let $e_{\mathcal{X}}$ denote the subset of employees working at e who have values in $\mathcal{X} \subset A_1 \times ... \times A_k$. A randomized algorithm A protects establishment shape against an informed attacker at privacy level (ϵ, α) if, for every informed attacker $\theta \in \Theta$, for every property of a worker record $\mathcal{X} \subset A_1 \times ... \times A_k$, for every pair of numbers $0 < p \leq q \leq min(1, (1+\alpha)p)$, and for every output $\omega \in range(A)$ and for every number z,

$$\left| log \left(\frac{Pr_{\theta,A}[|e_{\mathcal{X}}|/|e| = p, |e| = z | A(D) = \omega]}{Pr_{\theta,A}[|e_{\mathcal{X}}|/|e| = q, |e| = z | A(D) = \omega]} \Big/ \frac{Pr_{\theta}[|e_{\mathcal{X}}|/|e| = p, |e| = z]}{Pr_{\theta}[|e_{\mathcal{X}}|/|e| = q, |e| = z]} \right) \right| \leq \epsilon$$

whenever $Pr_{\theta}[|e_{\mathcal{X}}|/|e| = p, |e| = z], Pr_{\theta}[|e_{\mathcal{X}}|/|e| = q, |e| = z] > 0$.

Current SDL Protection

- Database is perturbed before answering queries
- Every establishment w is assigned a unique, time-invariant, confidential distortion factor $f_w \in [1-\beta, 1-\alpha] \cup [1+\alpha, 1+\beta]$
- Zero values are kept unmodified
- Additional output perturbation to limit re-identification of inndividual workes.

Differential Privacy

Bipartite graph

- edge differential privacy: not strong enough
- node differential privacy: too strong

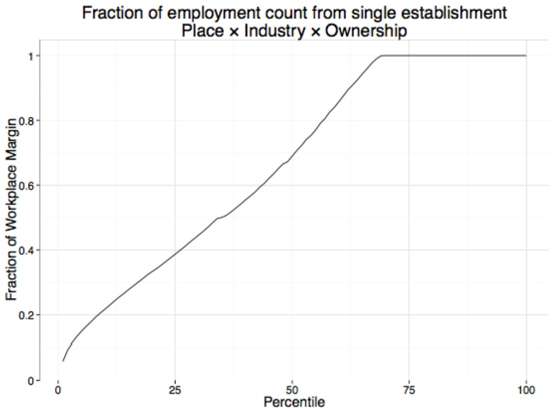

Figure 2: Employment counts from a single establishment

DEFINITION 7.1 (STRONG α-NEIGHBORS). *Let D and D' be two employer-employee tables such that they differ in the employment attribute of exactly one record (say corresponding to establishment e). Let E denote the set of workers employed at e in D, and E' denote the set of workers employed at e in D'. Then D and D' are neighbors if $E \subseteq E'$, and $|E| \leq |E'| \leq \max((1+\alpha)|E|, |E|+1)$*

DEFINITION 7.2 ((α, ϵ)-EMPLOYEE-EMPLOYER PRIVACY). *A randomized algorithm $\mathcal{M}$ is said to satisfy (α, ϵ)-Employee-Employer Privacy, if for every set of outputs $S \subseteq range(M)$, and every pair of strong α-Neighbors D and D', we have*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S]$$

THEOREM 7.1. *Let $\mathcal{M}$ be an algorithm satisfying (α, ϵ)-employer-employee privacy. Then, $\mathcal{M}$ satisfies the individual privacy requirement at privacy level ϵ, and the establishment size and shape requirements at privacy level (ϵ, α).*

DEFINITION 7.3 (WEAK α-NEIGHBORS). *Let D and D' be two employer-employee tables such that they differ in the employment attribute of exactly one record (say corresponding to establishment e). Let $\phi : U \to \{0,1\}$ be any property of a worker record, and for any $E \subset U$, let $\phi(E) = \sum_{r \in E} \phi(r)$. Let E denote the set of workers employed at e in D, and E' denote the set of workers employed at e in D'. D and D' are called weak α-neighbors if for every ϕ*

$$\phi(E) \leq \phi(E') \leq \max((1+\alpha)\phi(E), 1) \tag{7}$$

DEFINITION 7.4 (WEAK (α, ϵ)-EMPLOYEE-EMPLOYER PRIVACY) *A randomized algorithm $\mathcal{M}$ is said to satisfy weak (α, ϵ)-Employee-Employer Privacy, if for every set of outputs $S \subseteq range(M)$, and every pair of weak α-Neighbors D and D', we have*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S]$$

THEOREM 7.2. *Let A be an algorithm satisfying weak (α, ϵ)-employer-employee privacy. Then, A satisfies the individual privacy requirement at privacy level ϵ and the establishment shape requirement at level (ϵ, α). A satisfies the establishment size requirement at level (ϵ, α) for weak adversaries.*

). THEOREM 7.3. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be (α, ϵ_1)- and (α, ϵ_2)-employer-employee private algorithms. Releasing the outputs of $\mathcal{M}_1(D)$ and $\mathcal{M}_2(D)$ results in $(\alpha, \epsilon_1 + \epsilon_2)$-employer-employee privacy. The same holds for weak (α, ϵ)-employer-employee privacy.*

THEOREM 7.4. *Let D_1 and D_2 represent subsets of records from the employer-employee dataset that pertain to distinct sets of establishments. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be (α, ϵ)- and (α, ϵ)-employer-employee private algorithms. Releasing the outputs of $\mathcal{M}_1(D_1)$ and $\mathcal{M}_2(D_2)$ results in (α, ϵ)-employer-employee privacy. The same holds for weak (α, ϵ)-employer-employee privacy.*

Composition

THEOREM 7.5. *Let D_1 and D_2 represent subsets of records from the employer-employee dataset that pertain to distinct workers, but have records that arise from the same establishment. Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be (α, ϵ)- and (α, ϵ)-employer-employee private algorithms. Releasing the outputs of $\mathcal{M}_1(D_1)$ and $\mathcal{M}_2(D_2)$ results in (α, ϵ)-employer-employee privacy. The same* does not *hold for weak (α, ϵ)-employer-employee privacy.*

Intuition for the Proof:

$$d(A \cap D_1, B \cap D_1) + d(A \cap D_2, B \cap D_2) \leq d(A, B),$$

Log-Laplace Algorithm

THEOREM 8.1. *Suppose q_v is a query over only establishment attributes. Then, releasing q_v using Algorithm 1 satisfies (α, ϵ)-employer-employee privacy.*

Algorithm 1 Log-Laplace Mechanism

Require: : n : the sum of employment counts for a set of cells, α, ϵ: privacy parameters

Ensure: : $\tilde{n}$: the noisy employment count

Set $\gamma \leftarrow 1/\alpha$

$\ell \leftarrow \ln(n + \gamma)$

Sample $\eta \sim \mathit{Laplace}(2\ln(1+\alpha)/\epsilon)$

$\tilde{n} \leftarrow e^{\ell+\eta} - \gamma$

Suppose q_v is a query over both establishment attributes and employee attributes. Then, releasing q_v using Algorithm 1 satisfies weak (α, ϵ)-employer-employee privacy.

Log-Laplace Algorithm

LEMMA 8.2. *Let x denote a real number, and $\tilde{x}$ the random variable denoting the output of the Log-Laplace mechanism. Let $\lambda = 2\ln(\alpha+1)/\epsilon$. Then, when $\lambda < 1$, $E[\tilde{x}] + \gamma = (x+\gamma)/(1-\lambda^2)$. When $\lambda \geq 1$, $E[\tilde{x}]$ is unbounded.*

THEOREM 8.3. *The expected squared relative error of the Log-Laplace mechanism for q_v is bounded when $\lambda = 2\ln(\alpha+1)/\epsilon$ is less than 1, and is given by:*

$$\begin{aligned} \mathcal{E}_{rel}(q_v) &= \max_D \left(\frac{|q_v(D) - \mathcal{M}(D)|}{q_v(D)} \right) \\ &\leq (1+\gamma)^2 \frac{2\lambda^2 + 4\lambda^4}{(1-4\lambda^2)(1-\lambda/^2)} \end{aligned} \tag{9}$$

Smooth Sensitivity-based Algorithm

DEFINITION 8.1 (LOCAL SENSITIVITY). *Let q be a query, and $\mathcal{I}$ be a domain of datasets. The local sensitivity of query q for a dataset $x \in \mathcal{I}$ is*

$$LS_q(x) = \max_{y: y \in nbrs(x)} \|q(x) - q(y)\|_1$$

DEFINITION 8.2. *Let q be a query and b a smoothing parameter. Let $\mathcal{I}$ denote the universe of all datasets. The b-smooth sensitivity of query q with respect to database x is defined as*

$$S^*_{q,b}(x) = \max_j e^{-jb} A^{(j)}(x),$$

$$\text{where } A^{(j)}(x) = \max_{y \in \mathcal{I}: d(x,y) \leq j} LS_q(y),$$

and $d(x, y)$ *is the smaller integer* ℓ *such that there exist databases* $x = x_0, x_1, \ldots, x_\ell = y$, *such that for all i,* x_{i-1} *and* x_i *are neighbors according to either Definition 7.1 or 7.3.*

Smooth Sensitivity-based Algorithm

DEFINITION 8.3 ([26]). *A probability distribution h is (a, b)-*admissible, *where a and b are functions of ϵ and δ, if $\forall \lambda \in \mathbb{R}, \Delta \in \mathbb{R}^d$ with $|\lambda| \leq b$ and $\|\Delta\|_1 \leq a$, and $\forall S \subseteq \mathbb{R}^d$,*

$$\Pr_{Z\sim h}[Z \in S] \leq e^{\epsilon/2} \Pr_{Z\sim h}[Z \in S + \Delta] + \frac{\delta}{2}, \text{ and} \tag{10}$$

$$\Pr_{Z\sim h}[Z \in S] \leq e^{\epsilon/2} \Pr_{Z\sim h}\left[Z \in S \cdot e^{\lambda}\right] + \frac{\delta}{2}. \tag{11}$$

THEOREM 8.4. *Suppose h is an (a, b)-admissible probability distribution with $\delta = 0$, and $Z \sim h$. For query q, let $S(x)$ be a b-smooth upper bound on the local sensitivity of q. Then, the algorithm $\mathcal{M}(x) = q(x) + \frac{S(x)}{a} \cdot Z$ satisfies (α, ϵ)-Employer-Employee privacy.*

LEMMA 8.5. *Let q_v be a query on x. Let x_v be the maximum number of workers belonging to a single workplace and matching the conditions in v. Then, the b-smooth sensitivity of x, $S^*_{v,b}(x)$, is*

$$S^*_{v,b}(x) = \begin{cases} \max(x_v \cdot \alpha, 1) & \text{if } e^b \geq (1+\alpha), \\ \text{unbounded} & \text{otherwise.} \end{cases} \tag{12}$$

LEMMA 8.6 ([26]). *$h(z) \propto \frac{1}{(1+|z|^{\gamma})}$ is $(\epsilon/4\gamma, \epsilon/\gamma)$-admissible for $\gamma > 0$ ($\delta = 0$).*

Smooth Sensitivity-based Algorithm

Algorithm 2 Smooth Gamma

Require: : n : true count, α, ϵ: privacy parameters, $\alpha + 1 \leq e^{\epsilon/4}$
Ensure: : $\tilde{n}$: noisy count
Sample $\eta \sim \frac{1}{(1+|z|^4)}$
$\tilde{n} \leftarrow n + \frac{S^*_{v,\epsilon/4}(x)}{\epsilon/16}\eta$,

LEMMA 8.7. *Suppose q_v is a query over only establishment attributes. Then releasing q_v using Algorithm 2 satisfies (α, ϵ)-Employer-Employee privacy.*

Suppose q_v is a query over both establishment and individual attributes. Then releasing q_v using Algorithm 2 satisfies weak (α, ϵ)-Employer-Employee privacy.

LEMMA 8.8. *Algorithm 2 is unbiased and has expected L_1 error of $O(\frac{x_v \cdot \alpha}{\epsilon} + 1)$.*

DEFINITION 9.1 ((α, ϵ)-EMPLOYEE-EMPLOYER PRIVACY). *A randomized algorithm $\mathcal{M}$ is said to satisfy $(\alpha, \epsilon, \delta)$-Employee-Employer Privacy, if for every set of outputs $S \subseteq range(M)$, and every pair of strong α-Neighbors D and D', we have*

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} Pr[\mathcal{M}(D') \in S] + \delta$$

LEMMA 9.1 ([26]). *The Laplace distribution, $h(z) \propto \frac{1}{2} \epsilon^{-|z|}$, is $(\epsilon/2, \frac{\epsilon}{2\ln(1/\delta)})$-admissible.*

Approximating Privacy

Algorithm 3 Smooth Laplace

Require: : n : true count, α, ϵ: privacy parameters, $\alpha + 1 \leq e^{\frac{\epsilon}{2\ln(1/\delta)}}$.
Ensure: : $\tilde{n}$: noisy count

Sample $\eta \sim Laplace(1)$

$$\tilde{n} \leftarrow n + \frac{S^*_{v,\frac{\epsilon}{2\ln(1/\delta)}}(x)}{\epsilon/2}\eta,$$

LEMMA 9.2. *Suppose q_v is a query over only establishment attributes. Then releasing q_v using Algorithm 3 satisfies $(\alpha, \epsilon, \delta)$-employer employee privacy.*

Suppose q_v is a query over both establishment and individual attributes. Then releasing q_v using Algorithm 3 satisfies weak $(\alpha, \epsilon, \delta)$-employer employee privacy.

LEMMA 9.3. *Algorithm 3 is unbiased and expected L_1 error is* $O(\frac{x_v \cdot \alpha}{} + 1)$

Queries and Quality Measures: We use three types of query work-loads to evaluate our algorithms.

- **Workload 1** A marginal over all establishment characteristics: industry sector, ownership, and location at the resolution of places (e.g., cities and towns).
- **Workload 2** Single queries over all establishment attributes, and over the worker attributes of sex and education.
- **Workload 3** The marginal over all establishment attributes, and sex and education.

Empirical Evaluation and Findings

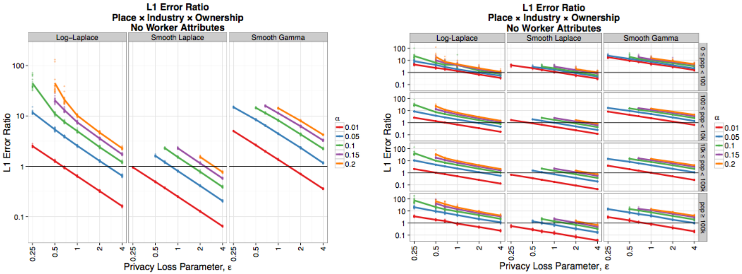

Figure 3: Average L_1 error of releasing place by industry sector by ownership marginal compared to the current system.

Empirical Evaluation and Findings

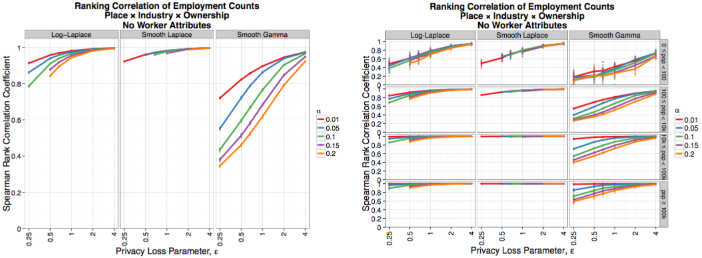

Figure 4: Spearman correlation between tested model and input noise infusion on the count of total workers ranked by place by industry section by ownership.

Empirical Evaluation and Findings

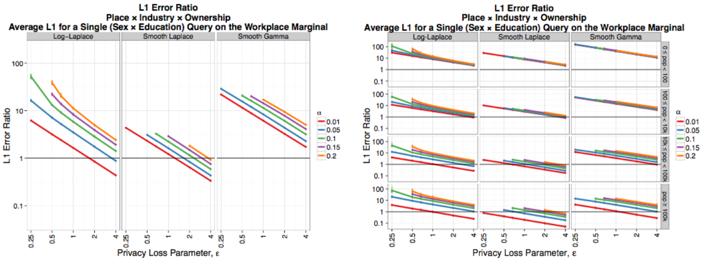

Figure 5: Average L_1 error of releasing single queries in the place by industry sector by ownership by sex by education marginal, compared to the current system.