Formal privacy protection for data products combining individual and employer frames[1]

Samuel Haney*, Ashwin Machanavajjhala*, Mark Kutzbach†, Matthew Graham†, John Abowd**, Lars Vilhuber**

* Department of Computer Science, Duke University, USA, {shaney, ashwin}@cs.duke.edu

† LEHD Program, US Census Bureau, USA, {matthew.graham, mark.j.kutzbach}@census.gov

** School of Industrial and Labor Relations, Cornell University, USA, {john.abowd, lars.vilhuber}@cornell.edu

Abstract. Published tabular summaries of linked employer-employee data usually use a job frame (statutory employer linked to a specific employee) but include characteristics of both the individual (employee) and workplace (employer establishment). Formal privacy protection of these characteristics requires defining the sensitivity of the published statistic to variation in a single individual or a single workplace (establishment). We propose a model that simultaneously protects individuals and establishments using parameters that control the conventional differential privacy for individuals and a generalization that provides a similar privacy guarantee for the employment magnitudes associated with an employer establishment. We implement our model using three alternative noise distributions. We present results for cross-sectional employment summaries for combinations of employer industry, geography, and ownership; and employee sex, age, race, ethnicity, and education. The system is illustrated using the LEHD Origin-Destination Employment Statistics (LODES) database displayed in the U.S. Census Bureau's OnTheMap application.

1 Introduction

In this paper we present a case study in applying provably private algorithms for publishing tabular summaries of linked employer-employee data. The data we consider are provided by the U.S. Census Bureaus Longitudinal Employer-Household Dynamics Program (LEHD) [2]. LEHD combines confidential survey and administrative data on jobs and releases a public-use dataset, known as the LEHD Origin-Destination Employment Statistics (LODES), that is widely used by government, businesses, and academics for regional planning, investment, and research. LODES tabulates over 130 million jobs in a year by workplace and residence location, commuting flows, as well as by a set of job, employer, and worker characteristics [1]. The combination of detailed geography and characteristics results in many cells having a small number of workers or employers.

[1] Disclaimer: 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.

The existing scheme for protecting the privacy of entities represented in the LODES data uses a combination of ad-hoc and provably private techniques. Workplace characteristics and job counts are protected using a combination of noise infusion techniques with no provable guarantees [3], while worker residence locations are protected using an algorithm that provably ensure probabilistic differential privacy [6].

Differential privacy [4] has quickly risen to become the gold standard for privacy protection. Algorithms that ensure differential privacy take as input a table, and their outputs do not change much on *neighboring inputs* that differ in the presence or absence of a single record in the input table. However, applying differential privacy to the LODES data, especially to protect workplace characteristics, poses a number of challenges. First, all work in differential privacy assumes data formatted as a single table, and the unit to be protected a single row in the table. However, the LODES data is a collection of tables describing properties of workers, workplaces and jobs. Next, there are multiple entities, each of which need different kinds of privacy protection. The presence or absence of workers in the data must be kept secret. The employment count and the distribution of the employees must be kept secret for the workplaces. Finally, there are certain attributes that do not need any protection, e.g., industry, ownership and geography of a workplace. Thus, directly adapting existing differentially private algorithms to this data is not possible.

Recent work has proposed a novel privacy framework, called Blowfish [5], that generalizes differential privacy and allows defining more expressive notions of what must be kept secret in a dataset. Blowfish accomplishes this by specifying a *policy graph* that redefines the notion of neighboring databases. In addition the Blowfish framework also possesses all the nice provable properties of differential privacy. In this paper, we show how to apply Blowfish to the protection of LODES data, and propose new algorithms for publishing tabular summaries over the data. More specifically, our contributions are:

- Present a Blowfish privacy policy for protecting workers and workplaces in the LODES data. The policy accounts for (a) data laid out as multiple tables, (b) protection of aggregate properties of workplaces, and (c) allowing certain information to be released without protection.
- Develop three novel algorithms for the release of marginal employment counts. All three algorithms provably satisfy privacy according to the specification in the Blowfish policy graph. We also prove analytical bounds on error for them.
- We empirically evaluate the algorithms on the LODES data and present the cost of provable protection by comparing the ratio of the error under the new algorithms to the error under the existing protection scheme (that has no provable privacy properties).

Organization Preliminaries are described in Section 2. We describe the LODES data and privacy challenges in Section 3. We review the Blowfish privacy framework and describe the specific privacy policy used for the LODES data in Section 4. Our algorithms are presented in Section 5 and evaluated in Section 6.

2 Preliminaries

Database and Queries Let D be a table of records with schema $(A_1, \ldots, A_k)$. The domain of each attribute A_i is denoted $dom(A_i)$. For each record t in the table, we let $t[A_i] \in dom(A_i)$ be value of attribute A_i. Let $n = |D|$ denote the size of the table; i.e. D has n records. A database with schema $(S_1, \ldots, S_m)$ is a collection of tables $(D_1, \ldots, D_m)$, where D_i has schema S_i.

We will consider *marginal queries* over tables in this paper.

Definition 2.1 (Marginal Query) *A marginal query q_V is defined by a tuple $V = (v_1, \ldots, v_k)$. Each $v_i \in dom(A_i) \cup \{*\}$, where $*$ is a wildcard which matches any value of A_i. q_V is answered on table D as follows:*

$$q_V(D) \ := \ |\{t \in D | t[A_i] = v_i \forall i\}| \,. \tag{1}$$

For example, the query $(*, \ldots, *)$ returns the size of the table.

Differential Privacy Differential privacy is usually defined on single tables (rather than an entire database). A mechanism or algorithm is differentially private if its output is not significantly affected by presence or absence of a single record from the input table.

Definition 2.2 ((ϵ, δ)-Differential Privacy [4]) *Let $\mathcal{M}$ be a randomized algorithm. Let D and D' be tables that differ in the presence of a single record; i.e., $|(D \backslash D') \cup (D' \backslash D)| = 1$. $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all $S \subseteq range(\mathcal{M})$,*

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

δ allows for the ratio of probabilities to be unbounded with a small failure probability. To avoid algorithms that disclose individual records, δ should be set smaller than $1/n$. When $\delta = 0$, we refer to the condition as ϵ-differential privacy. We call tables that differ in the presence of a single record *neighbors*.

Queries over tables can be answered while satisfying differential privacy by adding noise that is related to the *sensitivity* of a query.

Definition 2.3 (Sensitivity) *Let $\mathcal{I}$ denote the set of all tables with a given schema. Let $q : \mathcal{I} \to \mathbb{R}^d$ be a query function on that table that outputs a vector of d real numbers. The sensitivity of q, denoted Δ_q, is*

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

The Laplace mechanism is a commonly used ϵ-differentially private technique.

Definition 2.4 (Laplace Mechanism [4]) *Let $q : \mathcal{I} \to \mathbb{R}^d$ be a query on a table. Let $\eta \sim Lap(\lambda)$ denote a random variable drawn from the Laplace distribution with pdf $Pr[\eta = x] \propto e^{-|x|/\lambda}$. 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.*

Definition 2.5 (Expected L_p Error) *Let $q : \mathcal{I} \rightarrow \mathbb{R}^d$ be a query over a table, and $\tilde{q}(D)$ be the noisy answer returned by an algorithm. The expected L_p error of the algorithm is:*

$$\mathbb{E}\left(||q(D) - \tilde{q}(D)||_p\right) \tag{2}$$

where $||x||_p$ is the L_p norm, and expectation is over the randomness of the algorithm. We use L_1 and L_2 error in this paper.

3 Challenges in Applying Differential Privacy to Linked Employer-Employee Data

3.1 LODES Data and Queries

The LODES data are produced from an extract of the LEHD Infrastructure Files, which are composed of administrative records, census and survey data focused on the labor market, worker, and firm statistics. State unemployment insurance reporting and account information and federal worker earnings records provide information on employment location for jobs and residential information for workers, which form the basis of the LODES data product. LODES are published as an annual cross-section of jobs held on April 1 of each year from 2002 onwards.

The LODES data is organized as a relation with three database tables – Job, Worker and Workplace. The Workplace table contains one record per establishment and describes the following attributes – NAICS code (denoting industry the establishment belongs to), Ownership (Public/Private), geography (or the Census block that the establishment is located at). The Worker table contains one record for each individual working in an establishment at that point of time. Worker attributes include Age, Sex, Race, Ethnicity, and Education. Finally, the Job table contains pairs (w, i) of worker and workplace ids denoting that worker i works at establishment w. We assume each worker has exactly one job. We leave out some of the attributes that do not feature in our queries.

We will be interested in releasing marginal counts of employment over workplaces. More specifically, let V_I and V_W denote vectors over the workplace and worker attributes (along with wildcards) respectively, as in Definition 2.1. Let D_I denote the worker table, D_W denote the workplace table, and D_J the job table. Let V be the concatenation of marginals V_I and V_W. The marginal count of interest is the number of workers matching the criteria in V_I who work in establishments matching the criteria in V_W:

$$\begin{aligned} q_V \quad := \quad & |\{(t_I, t_W) \in D_I \times D_W | (t_I[id], t_W[id]) \in D_J, \\ & t_I[A_i] = v_i \forall v_i \in V_I, t_W[A_j] = v_j \forall v_j \in V_W\}| \end{aligned}$$

3.2 Privacy Desiderata

Protecting the data described above poses a number of novel challenges.

- The tabular summary of interest is not defined on a single table, but on a collection of tables. Prior work on differential privacy only considers a single table.

- A number of attributes about establishments are public attributes that do not need protection, including industry, ownership and geography. Prior work on differential privacy assumes that all attributes are private.
- Aggregate properties of establishments like the employment count must be protected. Just relying on the differentially private protection of individuals is insufficient – adding or removing an individual only changes the employment count of an establishment by 1, and adding noise to hide this change would still allow the employment count of an establishment to be inferred accurately. On the other hand, requiring differential privacy at the level of establishments (where neighbors differ in the presence or absence of an establishment) is too strict and can destroy the utility of the data. For instance, if half the population in a town works in a hospital, adding noise to hide the presence of the hospital would result in data with no utility for that town.

4 Privacy Policy for Linked Employer-Employee Data

In this section, we present the privacy definition that will be used for satisfying the privacy requirements stated in the previous section. We begin by reviewing a recent proposal to extend differential privacy using privacy policies, called Blowfish (Section 4.1). We can use Blowfish to explicitly define what information must be kept secret, thus potentially increasing utility by relaxing privacy. We then present a privacy policy to customize differential privacy for the tasks of releasing marginals on the LODES data (Section 4.2). We discuss privacy semantics implied by the resulting privacy definition in Section 4.3.

4.1 Blowfish Privacy

He et al [5] propose a class of privacy definitions, called *Blowfish*, that extend differential privacy by generalizing the notion of neighboring databases. Data owners can thus specify what information should be protected by choosing neighboring databases. Neighboring databases are succinctly encoded using a *policy graph* defined below.

Definition 4.1 (Policy graph) *A policy graph is a graph $G = (V, E)$ with $V \subseteq \mathcal{T} \cup \{\perp\}$, where $\perp$ is the name of a special vertex, and $E \subseteq (\mathcal{T} \cup \{\perp\}) \times (\mathcal{T} \cup \{\perp\})$.*

Intuitively, the policy graph defines pairs of domain values that an adversary should not distinguish between. An edge $(u, v) \in E$ means that an adversary can't tell whether an individual's value is u or v. $\perp$ is a dummy value not in $\mathcal{T}$, and an edge $(u, \perp) \in E$ means that an adversary should not be able to distinguish between the presence of a tuple with value u or the absence of the tuple from the database.

Definition 4.2 (Blowfish neighbors) *We consider a policy graph $G = (V, E)$. Let D and D' be two databases. D and D' are neighbors, denoted $(D, D') \in \mathcal{N}(G)$, iff exactly one of the following is true:*

- *D and D' differ in the value of exactly one entry such that $(u, v) \in E$, where u is the value of the entry in D and v is the value of the entry in D';*
- *$D' = D \cup \{u\}$, or $D = D' \cup \{u\}$, such that $(u, \perp) \in E$.*

Definition 4.3 ((ϵ, G)-Blowfish Privacy) *Let G be a policy graph. A mechanism $\mathcal{M}$ satisfies (ϵ, G)-Blowfish privacy if for any subset of outputs $S \subseteq range(\mathcal{M})$, and for any pair of neighboring databases $(D, D') \in \mathcal{N}(G)$,*

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

Differential privacy is a special case of Blowfish with the following policy graph:

$$G_{DP} = (V, E) \text{ such that } E = \{(u, \perp) \mid u \in \mathcal{T}\},$$

Blowfish policies guarantee weaker privacy semantics than differential privacy in the following way. Consider two tables that differ in one tuple: $D_1 = D \cup \{u\}$ and $D_2 = D \cup \{v\}$. For mechanism $\mathcal{M}$ satisfying (ϵ, G)-Blowfish privacy (Defs 4.2-4.3), we have,

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon \cdot \mathrm{dist}_G(u,v)} \cdot \Pr[\mathcal{M}(D') \in S], \tag{3}$$

where $\mathrm{dist}_G(u, v)$ is the length of the shortest path between u and v in G. Values u and v that are farther away in G can be distinguished more easily than values that are closer in G. In differential privacy, all pair (u, v) have the same distance (equal to 2) under G_{DP}, and are equally protected.

4.2 Customizing Privacy for LODES

We now show how to customize differential privacy using Blowfish privacy to overcome the challenges outlined in Section 3. Rather than considering the data as multiple tables, we represent it as a single table as follows. Each row corresponds to an establishment. Apart from the attributes of an establishment that are present in the Workplace table, each row has an additional attribute *employment*. If U is the universe of all worker records (i.e., the worker table $D_I \subset U$), the value of the employment attribute is $S \subset U$. Each element in S is a worker record (complete with all the attributes). Note that this view of the data can be constructed by joining the three tables. We denote by $\mathcal{P}$ and $\mathcal{NP}$ the sets of private and public attributes, respectively, associated with an establishment.

We define neighbors using the following Blowfish graph $G = (V, E)$. Each node in V is a triple (S, P, NP), where $S \subset U$ is the employment set of an establishment, and P, NP are vectors of attribute values for the private and public attributes for an establishment, respectively. There is an edge between two nodes (S, P, NP) and (S', P', NP') if one of the following holds:

$$NP = NP' \ \wedge \ P = P' \ \wedge \ S \neq S' \ \wedge \ (|S' \setminus S| = 1 \ \vee \ S' \approx_\alpha S)\text{, or} \tag{4}$$

$$NP = NP' \ \wedge \ S = S' \ \wedge \ P \neq P' \tag{5}$$

The neighborhood of sets of employees S and S' is defined in terms of α-*closeness*. Let $\mathcal{S}$ denote the set of multisets whose elements are drawn from a domain $\mathcal{T}$. Let $h : \mathcal{S} \rightarrow \mathbb{N}^{|\mathcal{T}|}$ denote the histogram function that takes as input a multiset $S \in \mathcal{S}$ and outputs the number of times each domain element appears in S. Let $h(S)[v]$ denote the number of times domain element v appears in S. Two multisets $S, S' \in \mathcal{S}$ are said to be α-close, denoted by $S' \approx_\alpha S$, if for every domain element $v \in \mathcal{T}$, $h(S')[v] \leq (1+\alpha)h(S)[v]$.

The first condition (Eq 4) defines neighborhood based on employment, and the second condition (Eq 5) defines neighborhood based on the non-private attributes. All neighbors have the same values for the non-private attributes. The first part of the *or* condition in Eq 4 is required to protect workers (adding or removing a worker changes the employment of some establishment by 1). The second part of the *or* condition is required to protect both the employment counts of workplaces as well as the distribution of employees, and states that an adversary should not be able to tell the exact employment count of an establishment and the distribution of the employees within a multiplicative factor of α.

4.3 Privacy Semantics

Using the Blowfish policy helps address the privacy challenges raised in Section 3.

Protecting Workers: Eq 4 requires neighboring databases that differ in the presence of one individual row in the Worker table ($|S' \setminus S| = 1$). Thus, workers are afforded as much privacy protection as under differential privacy.

Protecting Workplaces: Eq 4 requires neighboring databases differ in the employment of an establishment by a multiplicative factor of α. Thus if there are two databases, one where an establishment has employment count x and another has employment count y, then $y \leq (1+\alpha)^k x$ implies that the ratio of probabilities that an output was generated from these databases is bounded by $e^{k\cdot\epsilon}$. Moreover, the distribution of the employees (ages, race, sex, etc) is also protected by ensuring $S' \approx_\alpha S$ in neighbors. This is similar in spirit to the current ad-hoc protection mechanism for workplace characteristics [3]. Additionally, Eq 5 ensures that private attribute values are protected, since neighbors differ in the value of these attributes.

Multiple Tables: Both Worker and Workplace tables are protected by our Blowfish policy. Additionally, our Blowfish policy ensures that adding/removing an individual from the Worker table also adds/removes a job from the Job table. This ensures that there are no *referential integrity violations* when constructing a neighboring database.

Public Attributes: Finally, neighboring databases do not differ in the public attributes NP of the Workplace table. Thus given two possible values np and np' for the public attributes, the distance in the policy graph $d_G(np, np') = \infty$. Thus, these attribute values are not protected under our policy, and algorithms satisfying this policy are allowed to disclose the values of these attributes exactly. This helps us since we can disclose the exact number of establishments (but not their employments) with a given combination of industry, ownership and geography without adding any noise. In particular, we can disclose whether or not a cell has no establishments, and do not need to add noise to cells with no establishments

5 Algorithms

In this section, we give algorithms for answering queries under policy graph G. Each algorithm answers a single marginal query. Like differential privacy, Blowfish privacy also satisfies *parallel composition* [5] that allows us to answer multiple marginal queries without any loss in privacy, as long as the queries filter on disjoint attribute values. That is, we can release the number of employees in some Census tract of age 25-30, and the number of employees in that tract of age 30-35 each with a privacy parameter ϵ, and parallel composition will ensure that these multiple releases together ensure privacy with parameter ϵ. Our marginal queries in this section can only specify private individual values and public firm values, not private firm values (the marginal must have a '$*$' for these values). We defer all proofs to the appendix.

5.1 Log-Laplace Algorithm

Algorithm 1 Log-Laplace Mechanism

```
Require: : $n$ : the sum of employment counts for a set of cells, $\epsilon$: privacy parameter, $\alpha, \Delta$:
    multiplicative and additive protection factors
Ensure: : $\tilde{n}$: the noisy employment count
    Set $\gamma \leftarrow \Delta/\alpha$
    $\ell \leftarrow \ln(n + \gamma)$
    Sample $\eta \sim Laplace(2\ln(1 + \alpha)/\epsilon)$
    $\tilde{n} \leftarrow e^{\ell+\eta} - \gamma$
```

Algorithm 1 describes the log-laplace mechanism. The two parameters α and Δ represent the multiplicative and additive protection factors. α ensures that adversaries can't distinguish whether a firm's employment count is x or $(1 + \alpha)x$, while Δ ensures that adding or removing Δ workers does not change the output significantly. In our experiments, we use $\Delta = 1$. Note that the algorithm's privacy properties follow directly from the fact that the sensitivity of $\log q_V$ is $\max(\ln(\alpha + 1), \ln \Delta)$.

Theorem 5.1 *Let G be the Blowfish policy graph defined in Section 4.2. Then releasing q_V using Algorithm 1 satisfies (ϵ, G)-Blowfish privacy.*

While the original laplace mechanism is unbiased (the expectation of the noisy sum equals the true sum), the log laplace mechanism is not. In particular we can show:

Lemma 5.2 *Let x denote a real number, and $\tilde{x}$ the random variable denoting the output of the multiplicative 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 5.3 *The expected squared relative error of the multiplicative mechanism for* q_V *is bounded when* $\lambda = 2\ln(\alpha + 1)/\epsilon$ *is less than* 1*, and is given by:*

$$\mathcal{E}_{rel}(q_V) = \max_{D} \left(\frac{|q_V(D) - \mathcal{M}(D)|}{q_V(D)} \right) \leq \ (1+\gamma)^2 \frac{2\lambda^2 + 4\lambda^4}{(1-4\lambda^2)(1-\lambda/^2)} \tag{6}$$

Discussion: One could use the original Laplace mechanism with sensitivity M, where M is the maximum employment size of an establishment. The relative error of the Laplace mechanism is $M^2/x^2\epsilon^2$. For small x, the relative error will be very large. In contrast, the relative error of the log laplace mechanism is *independent* of q_V's true answer.

5.2 An Algorithm Based on Smooth Sensitivity

We can provide a mechanism using the smooth sensitivity framework [7]. The smooth sensitivity framework aims to add noise based on *local sensitivity* of the input database rather than global sensitivity across all databases. While local sensitivity can be much smaller than global sensitivity, adding noise proportional to local sensitivity does not ensure differential privacy, and hence, local sensitivity must be "smoothed".

Definition 5.1 (Local Sensitivity) *Let* q *be a query, and* $\mathcal{I}$ *be a domain of datasets. The local sensitivity of query* q *a dataset* $x \in \mathcal{I}$ *is*

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

Note that the global sensitivity is the maximum local sensitivity over all databases. The *smooth sensitivity* depends on the local sensitivity. Nissim et al[7] show that is is enough to add noise proportional to the smooth sensitivity. We adapt these results to our setting with alternate neighbors.

Definition 5.2 (Smooth Sensitivity [7]) *Let* q *be a query and* β *be a smoothing parameter. Recall that* $\mathcal{I}$ *is the domain of datasets. The* (β, G)*-smooth sensitivity with respect to database* x *of query* q *is defined as*

$$S^*_{q,\beta}(x) = \max_{j} e^{-j\beta} A^{(j)}(x), \text{ where } A^{(j)}(x) = \max_{y \in \mathcal{I} : d(x,y) \leq j} LS_q(y),$$

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 Blowfish policy G.*

Let $N(S)$ denote $\bigcup_{x \in S} nbrs(x)$, and $c \cdot S$ denote $\{c \cdot s | s \in S\}$.

Definition 5.3 ([7]) *A probability distribution* h *is* (a, β)-admissible*, where* a *and* β *are functions of* ϵ *and* δ*, if* $\forall \lambda \in \mathbb{R}, \Delta \in \mathbb{R}^d$ *with* $|\lambda| \leq \beta$ *and* $\|\Delta\|_1 \leq a$*, and* $\forall S \subseteq \mathbb{R}^d$*,*

$$\Pr_{Z \sim h}[Z \in S] \leq e^{\epsilon/2} \Pr_{Z \sim h}[Z \in S + \Delta] + \frac{\delta}{2}, \text{ and} \tag{7}$$

$$\Pr_{Z \sim h}[Z \in S] \leq e^{\epsilon/2} \Pr_{Z \sim h}\left[Z \in S \cdot e^{\lambda}\right] + \frac{\delta}{2}. \tag{8}$$

We can now adapt the following theorem from [7] to show that adding noise from admissible distributions lead to Blowfish private algorithms.

Theorem 5.4 *Suppose h is an (a, β)-admissible probability distribution, and $Z \sim h$. For query q, let $S(x)$ be an upper bound on the (β, G)-smooth sensitivity of q. Then the algorithm $\mathcal{M}(x) = q(x) + \frac{S(x)}{a} \cdot Z$ is (ϵ, δ, G)-Blowfish private.*

We now compute the (β, G)-smooth sensitivity of our queries and describe two admissible distributions.

Lemma 5.5 *Let G be the policy graph from Section 4.2. Let q_V be a query. Let*

$$x_V = \max_{t_w \in D_w : \forall v_i \in V_I, t_i[A_i] = v_i} \|t_I \in D_I | (t_I[id], t_W[id]) \in D_J \textit{ and } t_w[A_j] = v_j \forall v_j \in V_W\| . \tag{9}$$

That is, x_V is the maximum number of workers matching the conditions in V all belonging to the same workplace. Then the (β, G)-smooth sensitivity of x, $S^*_{V,\beta}(x)$, is

$$S^*_{V,\beta}(x) = \begin{cases} x_V \cdot \alpha & \textit{if } e^\beta \geq (1 + \alpha), \\ \textit{unbounded} & \textit{otherwise}. \end{cases} \tag{10}$$

Lemma 5.6 ([7]) *The following distributions are admissible.*

- *The Laplace distribution, $h(z) \propto \frac{1}{2} \cdot \epsilon^{-|z|}$, is $(\epsilon/2, \frac{\epsilon}{2\ln(1/\delta)})$-admissible.*
- *$h(z) \propto \frac{1}{(1+|z|^\gamma)}$ is $(\epsilon/4\gamma, \epsilon/\gamma)$-admissible for $\gamma > 0$ ($\delta = 0$). We will use $\gamma = 4$.*

Using these two distributions gives us the Smooth Laplace (Alg 2) and the Smooth Gamma (Alg 3) algorithms. While the former satisfies (ϵ, δ, G)-Blowfish privacy, the latter satisfies the stronger (ϵ, G)-Blowfish condition. The two algorithms are very similar so we give Algorithm 2 here, and defer Algorithm 3 to the appendix.

Algorithm 2 Smooth Laplace

```
Require: : $n$ : the sum of employment counts for a set of cells, $\epsilon$: privacy parameter, $\alpha$: multiplicative protection factor, $\alpha + 1 \leq e^{\frac{\epsilon}{2\ln(1/\delta)}}$.
Ensure: : $\tilde{n}$: the noisy employment count
  Sample $\eta \sim Laplace(1)$
  Sample $\eta' \sim Laplace(1/\epsilon)$
  $\tilde{n} \leftarrow n + \frac{S^*_{V,\frac{\epsilon}{2\ln(1/\delta)}}(x)}{\epsilon/2}\eta + \eta'$,
```

Lemma 5.7 *Algorithms 2 and 3 are unbiased and have bounded expected error.*

6 Empirical Evaluation

Dataset: The data used for these experiments were a 3-state sample from the LODES data infrastructure, to which standard edits and imputations had already been applied. The sample was taken from the 2011 snapshot, in which Quarter 2 (April-June) is the reference period. To be included jobs had to qualify as beginning-of-quarter jobs, which means that the job (person-firm relationship) has positive earnings in the reference quarter (Q2) as well as the previous quarter (Q1). Then the assumption is that the person was employed in the job on the first day of Q2. The count of jobs in the sample was 10.9 million jobs in about 527,000 establishments.

Queries: We compute marginal queries over public workplace characteristics industry, ownership and location (census tract), and private worker characteristics age, sex and race. We report errors on 4 marginals – the 3-way marginal with only workplace characteristics (94,363 cells), 4-way marginal including age (283,089 cells), 5-way marginal including age and sex (566,178 cells), and a 6-way marginal with all the attributes (3,397,068 cells). L_1 errors reported are averages over all the cells, over 20 independent trials. Error is reported as a ratio to the error of the current protection scheme for workplace counts [3].

Algorithms: We compare the Log-laplace, Smooth Gamma and Smooth Laplace algorithms. While the first two satisfy (ϵ, G)-Blowfish privacy, Smooth Laplace algorithm is (ϵ, δ, G)-Blowfish private. We do not vary δ as part of our evaluation, since δ does not affect the amount of noise added. After fixing ϵ and α, the value of δ that is achieved is computed from $\ln(\alpha + 1) = \frac{\epsilon}{2\ln(1/\delta)}$. This is the lowest value of δ which still satisfies (ϵ, δ, G)-Blowfish privacy. Smooth Laplace, and Smooth Gamma, require that certain conditions on ϵ, δ, and α are met. Points are plotted only if these conditions are met (see Algorithms 2, and 3 for more details). Similarly, we do not plot errors for the Log-laplace mechanism when the expectation of the noisy count is unbounded (see Lemma 5.2). We present results for $\epsilon \in \{0.25, 0.5, 0.67, 1.0, 2.0, 4.0\}$.

Results: Our key findings (Figure 1) are:

- The logarithm of the error is inversely proportional to the logarithm of the privacy parameter (note the log scale on both the x and y axes in Figure 1). In fact, for the smooth sensitivity based mechanisms, L_1 error is linear in the privacy parameter ϵ.
- For larger epsilon values there is no cost of provable privacy protection. We see that the error of our algorithms are smaller than the error of existing privacy protection. This can be perhaps explained due to the fact that the existing perturbation algorithm is a *local algorithm*; i.e., perturbation is performed on the input database on each record. On the other hand, our algorithms work on the entire dataset and on the output of the query. An interesting open question is empirically comparing the privacy protections of the existing perturbation algorithm and our new provably private algorithms.
- As α increases, the privacy protection increases and so does the error for all three algorithms.

- The Log-laplace mechanism has lower expected error when the data are aggregated coarsely. On the other hand, the smooth gamma algorithm has lower error when the data are disaggregated.
- Allowing a small failure probability in the privacy protection helps improve error as shown by the smooth Laplace algorithm.

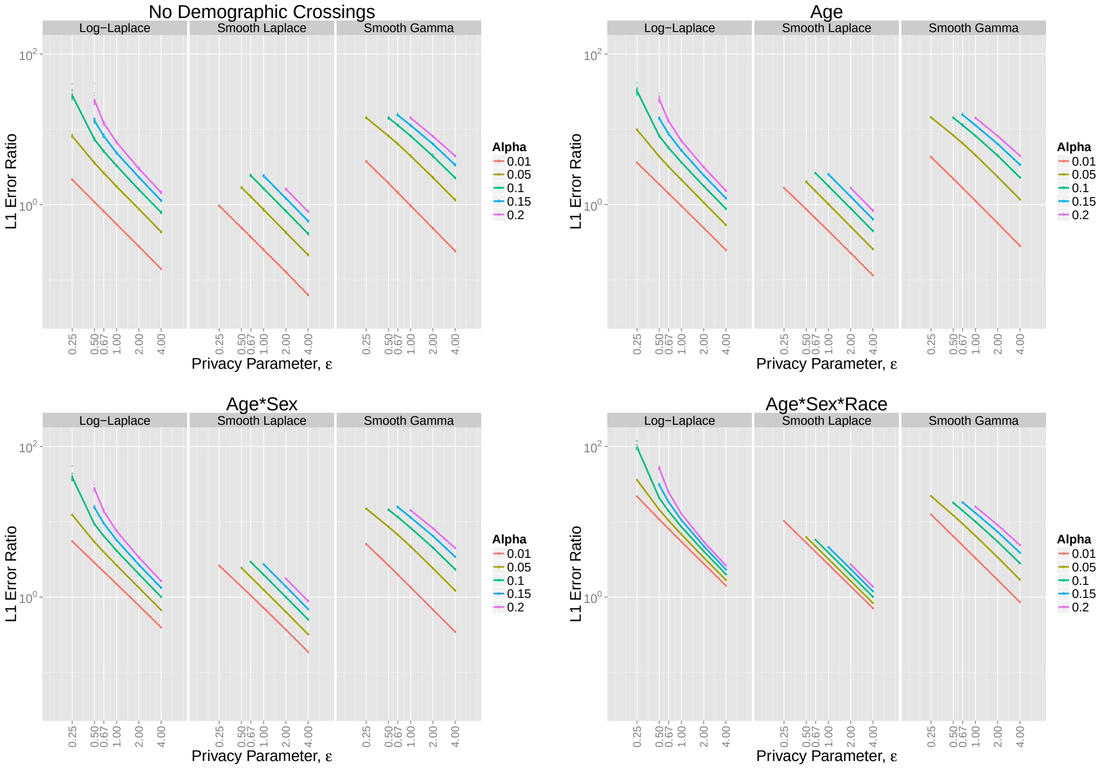

Figure 1: Empirical comparison of proposed algorithms for releasing marginals

7 Conclusions

We considered the problem of applying ϵ-differential privacy to releasing marginal counts from a dataset collected by the US Census. We found a number of challenges in directly applying existing differentially private techniques including handling multiple tables, and protecting aggregate properties of entities. We used the Blowfish framework to specify a privacy policy that fit the privacy requirements of these data and developed algorithms with provable privacy guarantees. We found many privacy settings where our provably private algorithms incurred lesser error than the error of the existing protection mechanism (that has no privacy guarantees), suggesting that provable privacy does not always incur a cost in terms of utility.

References

[1] Lehd origin-destination employment statistics (lodes) technical document. http://lehd.ces.census.gov/data/lodes/LODES7/LODESTechDoc7.1.pdf.

[2] U.s. census bureaus longitudinal employer-household dynamics program. http://lehd.ces.census.gov/.

[3] J. M. Abowd, B. E. Stephens, and L. Vilhuber. Confidentiality protection in the census bureaus quarterly workforce indicators. Technical Report TP-2006-02, U.S. Census Bureau, LEHD Program, December 2006.

[4] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3,4):211–407, Aug. 2014.

[5] X. He, A. Machanavajjhala, and B. Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1447–1458, 2014.

[6] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE*, pages 277–286, 2008.

[7] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84. ACM, 2007.

A Smooth Gamma Algorithm

Algorithm 3 Smooth Gamma

```
Require: : $n$ : the sum of employment counts for a set of cells, $\epsilon$: privacy parameter, $\alpha$:
    multiplicative protection factor, $\alpha + 1 \leq e^{\epsilon/4}$
Ensure: : $\tilde{n}$: the noisy employment count
    Sample $\eta \sim \frac{1}{(1+|z|^4)}$
    Sample $\eta' \sim Laplace(1/\epsilon)$
    $\tilde{n} \leftarrow n + \frac{S^*_{V,\epsilon/4}(x)}{\epsilon/16}\eta + \eta'$,
```

B Proof of Theorem 5.1

Proof: Consider two neighboring datasets (in the form described in Section 4.2) that differ in one tuple. Let (S, P, NP) and (S', P', NP') denote the values of the entry in which the two databases differ, and n_- denote the sum of all other counts. Note that we are disallowing queries with $P \neq P'$, so we just consider the case where $S \subseteq S'$. Let

$x = |S|$ and $y = |S'|$. We need to ensure privacy for two cases: (i) $x = \alpha \cdot y$, and (ii) $x = y + \Delta$. In case (i),

$$\begin{aligned}
\frac{P(M(D_1) = o)}{P(M(D_2) = o)} &= \frac{P(M(x + n_-) = o)}{P(M(y + n_-) = o)} \\
&= \frac{P(\eta = \ln(o + \gamma) - \ln(x + n_- + \gamma)}{P(\eta = \ln(o + \gamma) - \ln(y + n_- + \gamma)} \\
&\leq exp\left(\frac{\epsilon}{\ln \alpha} \ln\left(\frac{\alpha \cdot y + n_- + \gamma}{y + n_- + \gamma}\right)\right) \\
&\leq exp\left(\frac{\epsilon}{\ln \alpha} \cdot \ln(\alpha)\right) = e^{\epsilon}
\end{aligned}$$

In case (ii),

$$\begin{aligned}
\frac{P(M(D_1) = o)}{P(M(D_2) = o)} &= \frac{P(M(n_-) = o)}{P(M(1 + n_-) = o)} \\
&\leq exp\left(\frac{\epsilon}{\ln \alpha} \ln\left(\frac{\Delta + n_- + \gamma}{n_- + \gamma}\right)\right) \\
&\leq exp\left(\frac{\epsilon}{\ln \alpha} \ln(\alpha)\right) = e^{\epsilon}
\end{aligned}$$

□

C Proof of Theorem 5.2

Proof:

$$E[\tilde{x}] = -\gamma + (x + \gamma) \cdot E[e^{\eta}]$$

where $\eta \sim Laplace(\lambda)$. $E[e^{\eta}]$ corresponds to the value of the moment generating function $M_{\eta}(1)$.

$$M_{\eta}(1) = E[e^{\eta}] = 1 + \sum_{n=1}^{\infty} E[\eta^n]/n!$$

Since $Laplace(\lambda)$ is an even distribution, for all odd i, $E[\eta] = 0$. Moreover, $E[\eta^{2n}] = 2n!\lambda^{2n}$. Therefore, when $\lambda < 1$

$$\begin{aligned}
E[\tilde{x}] &= -\gamma + (x + \gamma) \cdot \sum_{n=1}^{\infty} \lambda^{2n} \\
&= -\gamma + (x + \gamma)/(1 - \lambda^2)
\end{aligned}$$

When λ is not bounded by 1, then the expected value is not bounded. Thus this mechanism is good only when $\lambda < 1$. □

D Proof of Theorem 5.3

Proof: Let y denote $x+\gamma$, where $q_V(D)=x$ is the true sum. Similarly, let $\tilde{y}$ denote $\tilde{x}+\gamma$, where $\tilde{x}$ is the output of the multiplicative laplace mechanism. We will show that

$$E((\frac{y-\tilde{y}}{y})^2) = \frac{2\lambda^2+4\lambda^4}{(1-4\lambda^2)(1-\lambda/^2)}$$

The result in the theorem directly follows.

$$\begin{aligned} E((y-\tilde{y})^2/y^2) &= E(\tilde{y}^2)/y^2 - 2E(\tilde{y})/y + 1 \\ &= E(\tilde{y}^2)/y^2 - 2/(1-\lambda^2) + 1 \end{aligned}$$

$E(\tilde{y}^2)/y^2 = E[e^{2\cdot\eta}]$, where $\eta \sim Laplace(\lambda)$. $E[e^{2\cdot\eta}]$ corresponds to the value of the moment generating function $M_\eta(2)$.

$$\begin{aligned} E(\tilde{y}^2)/y^2 &= E[e^{2\cdot\eta}] = M_\eta(2) = 1 + \sum_{n=1}^{\infty} 2^n E[\eta^n]/n! \\ &= 1 + \sum_{n=1}^{\infty} (2\lambda)^{2n} \\ &= 1/(1-4\lambda^2) \qquad \text{when } \lambda < 1/2 \end{aligned}$$

Therefore, we have:

$$\begin{aligned} E((y-\tilde{y})^2/y^2) &= 1/(1-4\lambda^2) - 2/(1-\lambda^2) + 1 \\ &= \frac{2\lambda^2+4\lambda^4}{(1-4\lambda^2)(1-\lambda/^2)} \end{aligned}$$

□

E Proof of Theorem 5.4

Proof: For $y \in nbrs(x)$, we must show that

$$\Pr[\mathcal{M}(x) \in S] \le e^\epsilon \cdot \Pr[\mathcal{M}(y) \in S]. \tag{11}$$

We have

$$\begin{aligned} \Pr[\mathcal{M}(x) \in S] &= \Pr_{Z\sim h}\left[Z \in \frac{S-q(x)}{S(x)/\alpha}\right] & (12) \\ &\le \Pr_{Z\sim h}\left[Z \in \frac{S-q(y)}{S(x)/\alpha}\right] \cdot e^{\epsilon/2} + \frac{\delta}{2} & (13) \\ &\le \Pr_{Z\sim h}\left[Z \in \frac{S-q(y)}{S(y)/\alpha}\right] \cdot e^{\epsilon} + \delta & (14) \\ &\le \Pr[\mathcal{M}(y) \in S] \cdot e^\epsilon + \delta. & (15) \end{aligned}$$

(13) holds by the first property of Definition 5.3. (14) holds by the second property of Definition 5.3. □

F Proof of Lemma 5.5

Proof: The local sensitivity of q_V with respect to x is the maximum amount by which any firm's (matching the criteria in V) count of employees (matching the criteria in V) can change. Note that x_V is the largest such count, and therefore

$$LS_{q_V}(x) = x_V \cdot (1+\alpha) - x_V = x_V \cdot \alpha \tag{16}$$

Then, we have

$$A^{(j)}(x) = \max_{y \in D: d(x,y) \leq j} y_V \cdot \alpha.$$

This value is maximized by maximizing y_V. The maximum value for y_V is $x_v(1+\alpha)^j$. Therefore,

$$A^{(j)}(x) = x_V \cdot \alpha(1+\alpha)^j.$$

Our smooth sensitivity is therefore

$$S^*_{q,\beta}(x) = \max_j (\frac{1+\alpha}{e^\beta})^j x_V \alpha.$$

Our databases do not have a fixed size, so j can be any positive integer, and therefore the smooth sensitivity is not necessarily bounded. When $e^\beta < (1+\alpha)$,

$$S^*_{q,\beta}(x) = \max_j (\frac{1+\alpha}{e^\beta})^k x_V \alpha = \lim_{j \to \infty} (\frac{1+\alpha}{e^\beta})^j x_V \alpha,$$

which is unbounded. When $e^\beta \geq (1+\alpha)$, this limit is bounded, and in this case

$$S^*_{q,\beta}(x) = x_V \cdot \alpha.$$

□

G Proof of Lemma 5.7

Proof: It is well known that the Laplace mechanism is unbiased with Laplace(λ) having mean squared error of λ^2. We prove here that $h(z) \propto \frac{1}{(1+|z|^\gamma)}$ is unbiased with bounded error for $\gamma = 4$. $h(z)$ is unbiased since

$$\mathbb{E}[h(z)] = \int \frac{z}{1+|z|^4} dz = 0.$$

The variance of $h(z)$ is given by

$$\int \frac{z^2}{1+|z|^4} dz = \frac{\pi}{\sqrt{2}} \approx 2.2.$$

The error mean squared error is therefore 2.2^2. $\square$