Regression Model Fitting under Differential Privacy and Model Inversion Attack

Yue Wang
University of North Carolina at Charlotte
Charlotte, NC, USA
ywang91@uncc.edu

Cheng Si
University of Arkansas
Fayetteville, AR, USA
cxs040@uark.edu

Xintao Wu
University of Arkansas
Fayetteville, AR, USA
xintaowu@uark.edu

Abstract

Differential privacy preserving regression models guarantee protection against attempts to infer whether a subject was included in the training set used to derive a model. It is not designed to protect attribute privacy of a target individual when model inversion attacks are launched. In model inversion attacks, an adversary uses the released model to make predictions of sensitive attributes (used as input to the model) of a target individual when some background information about the target individual is available. Previous research showed that existing differential privacy mechanisms cannot effectively prevent model inversion attacks while retaining model efficacy. In this paper, we develop a novel approach which leverages the functional mechanism to perturb coefficients of the polynomial representation of the objective function but effectively balances the privacy budget for sensitive and non-sensitive attributes in learning the differential privacy preserving regression model. Theoretical analysis and empirical evaluations demonstrate our approach can effectively prevent model inversion attacks and retain model utility.

1 Introduction

Privacy-preserving data mining is an important research area. In many applications, sensitive datasets such as financial transactions, medical records, or genetic information about individuals are often only disclosed to authorized users, yet models learned from them are made public. The released models may be exploited by an adversary to breach privacy of both participant individuals in the datasets and regular non-participant individuals.

Differential privacy [Dwork *et al.*, 2006] has been developed and shown as an effective mechanism to protect privacy of participant individuals. Simply speaking, differential privacy is a paradigm of post-processing the models such that the inclusion or exclusion of a single individual from the dataset makes no statistical difference to the results found. In other words, differential privacy aims to achieve the goal, i.e., the risk to one's privacy should not substantially increase as a result of participating in a database when models built from the database are released to public. On the contrary, the models (or even the perturbed models which preserve differential privacy of participants) may be exploited to breach attribute privacy of regular individuals who are not necessarily in the dataset. In [Fredrikson *et al.*, 2014], the authors developed a model inversion attack where an adversary uses the released model to make predictions of sensitive attributes (used as input to the model) of a target individual when some background information about the target individual is available. Fredrikson et al. showed that differential privacy mechanisms prevent model inversion attacks only when the privacy budget is very small. However, for privacy budgets effective at preventing attacks, the model utility in terms of performing simulated clinical trials is significantly lost.

Hence it is imperative to develop mechanisms to achieve differential privacy protection for participants and prevent attribute privacy disclosure from model inversion attacks while retaining the utility of the released models. In this paper, we focus on regression models which have been widely applied in many applications. Regression studies often involve continuous data (e.g., blood lipid levels or heights) in addition to categorical attributes (e.g., gender, race and disease). Various regression models including linear regression, logistic regression, and lasso models have been developed. There are generally two approaches to derive differential privacy preserving regression models. The first approach is to directly perturb the output coefficients of the regression models. However, this approach requires an explicit sensitivity analysis of the regression models, which is often infeasible. The second approach is to add noise to the objective function used to derive regression models [Chaudhuri and Monteleoni, 2008]. Recently the authors [Zhang *et al.*, 2012] developed the functional mechanism which adds noise to the coefficients of polynomial representation of the objective function, and showed that deriving a bound on the amount of noise needed for the functional mechanism involves a fairly simple calculation on the object function.

Differential privacy preserving regression models [Chaudhuri and Monteleoni, 2008; Zhang *et al.*, 2012] guarantee protection against attempts to infer whether a subject was included in the training set used to derive a model. It is not effective to protect attribute privacy, which is the target of the model inversion attacks. This is because the mechanism perturbs coefficients equally no matter whether they correspond to sensi-

Table 1: Notations

Symbol	Definition
$t_i = (\mathbf{x}_i, y_i)$	the i-th tuple
ω	the parameter vector of the regression model
$\rho(\omega)$	the released regression mode
$f(t_i, \omega)$	the cost function on tuple t_i
$f_D(\omega)$	$f_D(\omega) = \sum_{t_i \in D} f(t_i, \omega)$
ω^*	$\omega^* = \arg\min_\omega f_D(\omega)$
$\phi(\omega)$	a product of elements in $\omega_1, \omega_2, ..., \omega_d$
Φ_j	the set of all possible ϕ of order j
$\lambda_{\phi t_i}$	the polynomial coefficient of ϕ in $f(t_i, \omega)$
ϵ	privacy budget for the regression model
ϵ_s, ϵ_n	privacy for sensitive, non-sensitive attributes

tive or non-sensitive attributes. However, model inversion attacks seek to exploit correlations between the target sensitive attributes, known non-sensitive attributes and the model output. In this paper, we aim to develop a new approach to learn differential privacy preserving regression models which effectively prevent model inversion attacks and retain the model utility. Our approach leverages the functional mechanism but effectively balances the privacy budget for sensitive and non-sensitive attributes in learning the differential privacy preserving regression models.

1.1 Problem Formalization

Let D be a data set that contains n tuples $t_1, t_2, ..., t_n$ regarding d explanatory attributes $X_1, X_2, \cdots, X_d$ and one response attribute Y. The explanatory attributes can be divided into two groups: non-sensitive attributes and sensitive attributes. For simplicity, we consider there is only one sensitive attribute X_s and all remaining ones are non-sensitive. Our analysis can be straightforwardly extended to multiple sensitive attributes. For each explanatory attribute X_i, without loss of generality, we assume its domain $\mathcal{X}_i$ in the range of $[-1, 1]$. Similarly, we denote $\mathcal{Y}$ as the domain of the response attribute Y, which could be $[-1, 1]$ for linear regression or $\{0, 1\}$ for logistic regression. We denote each tuple t_i as $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{id})$. Throughout this paper, we use bold lower-case variables, e.g., $\mathbf{x}_i$, to represent vectors; upper-case alphabets, e.g., X_i, to denote an attribute; calligraphic upper-case alphabets, e.g., $\mathcal{X}_i$, to denote the domain of attribute X_i. $\mathbf{x}^T$ refers the transpose of vector $\mathbf{x}$. Table 1 summarizes the notations used in this paper.

The data mining task is to release a regression model from D to predict the attribute value of Y given the corresponding attribute value of $X_1, ..., X_d$. That is to say, we are to release a regression function ρ parameterized with a real number vector $\omega = (\omega_1, ..., \omega_d)$. The model takes $\mathbf{x}_i$ as input and output the corresponding prediction for y_i as $\hat{y}_i = \rho(\mathbf{x}_i, \omega)$. Most regression analytical methods often iteratively optimize some objective functions with various constraints. A cost function f is often chosen to measure the difference between the original and predicted values based on specific ω. The optimal model parameter ω^* is defined as the one that minimizes the cost function.

$$\omega^* = \arg\min_\omega f_D(\omega) = \arg\min_\omega \sum_{i=1}^{n} f(t_i, \omega). \quad (1)$$

In our paper, we consider two commonly used regression models, linear regression and logistic regression.

Definition 1. *(Linear Regression) Assume without loss of generality that Y has a domain of $[-1, 1]$. A linear regression on D returns a prediction function $\hat{y}_i = \rho(\mathbf{x}_i, \omega^*) = \mathbf{x}_i^T \omega^*$, where ω^* is a d-dimensional real vector that minimizes the following cost function.*

$$\omega^* = \arg\min_\omega f_D(\omega) = \arg\min_\omega \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \omega)^2. \quad (2)$$

Definition 2. *(Logistic Regression) Assume Y has a domain of $\{0, 1\}$. A logistic regression on D returns a prediction function which returns $\hat{y}_i = 1$ with the probability $P(\hat{y}_i = 1 | \mathbf{x}_i, \omega^*) = exp(\mathbf{x}_i^T \omega^*)/(1 + exp(\mathbf{x}_i^T \omega^*))$, where ω^* is a d-dimensional real vector that minimizes the following cost function.*

$$\begin{aligned} \omega^* &= \arg\min_\omega f_D(\omega) \\ &= \arg\min_\omega \sum_{i=1}^{n} (log(1 + exp(\mathbf{x}_i^T \omega)) - y_i \mathbf{x}_i^T \omega). \end{aligned} \quad (3)$$

Releasing the regression model under differential privacy requires noise injection to the model parameter ω^*. Adding noise to ω^* involves the derivation of the sensitivity of ω^*, which is rather challenging. In this paper, we apply the *functional mechanism* proposed in [Zhang *et al.*, 2012], which perturbs the objective function of the regression model. However, the release model parameter ω^* or its perturbed one $\bar{\omega}$ could be exploited by the adversary to predict the value of sensitive input attributes $x_{\alpha s}$ for a target individual α when some background information about the target individual is available. Formally, the adversary has access to the regression model with parameters ω^*, the domain value and marginal probability of each attribute, accuracy metrics of the model like the confusion matrix, in addition to some background knowledge of the target individual including the value of a subset of non-sensitive input attributes and the value of output attribute of the model y_α.

Our research problem is how to derive the perturbed regression model parameter $\bar{\omega}$ such that we achieve differential privacy protection for participants and prevent attribute privacy disclosure from model inversion attacks on regular individuals while retaining the utility of the regression model.

2 Background

2.1 Differential Privacy

We revisit the formal definition and the classic mechanism of differential privacy. In prior work on differential privacy, a database is treated as a collection of *rows*, with each row corresponding to the data of a different individual. Differential privacy ensures that the inclusion or exclusion of one individual's record makes no statistical difference on the output.

Definition 3. *(Differential Privacy[Dwork* et al.*, 2006]) A randomized function A gives ϵ-differential privacy if for all data sets D and D' differing at most one row, and all $S \subseteq Range(A)$*

$$Pr[A(D) \in S] \leq e^{\epsilon} \cdot Pr[A(D') \in S] \tag{4}$$

The privacy parameter ϵ controls the amount by which the distributions induced by two neighboring data sets may differ (smaller values enforce a stronger privacy guarantee).

A general method for computing an approximation to any function f while preserving ϵ-differential privacy is given in [Dwork *et al.*, 2006]. It computes the sum of the true answer and random noise generated from a Laplace distribution. The magnitude of the noise distribution is determined by the sensitivity of the computation and the privacy parameter specified by the data owner. The sensitivity of a computation bounds the possible change in the computation output over any two neighboring data sets (differing at most one record).

Definition 4. *(Global Sensitivity[Dwork* et al.*, 2006]) The global sensitivity of a function $f : D^n \rightarrow \boldsymbol{R}^d$*

$$GS_f(D) := \max_{D,D' s.t. D' \in \Gamma(D)} ||f(D) - f(D')||_1 \tag{5}$$

Theorem 1. *(Laplace Mechanism[Dwork* et al.*, 2006]) An algorithm A takes as input a data set D, and some $\epsilon > 0$, a query Q with computing function $f : D^n \rightarrow \boldsymbol{R}^d$, and outputs*

$$A(D) = f(D) + (Y_1, ..., Y_d) \tag{6}$$

where the Y_i are drawn i.i.d from $Lap(GS_f(D)/\epsilon)$. The Algorithm satisfies ϵ-differential privacy.

2.2 Model Inversion

Model inversion attack [Fredrikson *et al.*, 2014] leverages the released regression model $y = \rho(\mathbf{x}, \omega^*)$ trained from a dataset D which contains a sensitive attribute X_s. An adversary then exploits the released model to predict the sensitive input attribute value of the target individual based on some of the target individual's background (values of some non-sensitive input attributes, e.g. demographic information, for the model) and the observed response attribute value.

The model inversion attack algorithm works as follows. The adversary has access to the regression model ρ with parameter ω^* trained over a dataset D drawn i.i.d from an unknown prior distribution p. Recall that D has input domain $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ and output domain $\mathcal{Y}$. The target individual is represented by $t_\alpha = (\mathbf{x}_\alpha, y_\alpha)$. The adversary is assumed to know values of some (or all) input attributes of the target individual except the sensitive one, i.e., $S \subseteq X \backslash X_s$, and the output value y_α. The sensitive attribute value the adversary wants to learn is referred to as $x_{\alpha s}$. Note that the target individual t_α is not necessarily in D. In addition to the released model $\rho(\mathbf{x}, \omega^*)$, the adversary also has access to marginal $p_1, ..., p_d, p_y$ of the joint prior p, the input domain and the output domain, the information π about the model prediction performance where $\pi(y, y') = Pr(y_i = y | \rho(x_i, \omega^*) = y')$. The algorithm makes prediction by estimating the probability of a potential target attribute value given the available information of the target individual and the model.

- Find the feasible set $\hat{\mathcal{X}} \subseteq \mathcal{X}$, i.e., for $\forall x \in \hat{\mathcal{X}}$, x matches $\mathbf{x}_\alpha$ on each known attribute in S.
- if $\hat{\mathcal{X}} = 0$, return null; otherwise, return $\hat{x}_{\alpha s} = z$ that maximizes $\sum_{x \in \hat{\mathcal{X}}: x_s = z} \pi(y_\alpha, \rho(x_\alpha, \omega)) \prod_{1 \leq j \leq d} p_j(x_j)$.

In step 1, the algorithm filters the domain space using the known attribute values of the target individual. In step 2, the algorithm calculates weight to each candidate row in the domain space based on known priors and how well the model's output on that row coincides with the target individual's model output value. It then returns the value of the target sensitive attribute with the largest weight computed by marginalizing the other attributes. The model inversion algorithm is optimal as it minimizes the expected misclassification rate on the maximum-entropy prior given the model and marginals. It was demonstrated in [Fredrikson *et al.*, 2014] that the value of the sensitive attribute is predicted with significantly better accuracy than guessing based on marginal distributions. It was also concluded that differential privacy mechanisms can prevent model inversion attacks only when privacy budget is very small, but in those cases, the private model usually does not simultaneously retain desirable efficacy. In clinical trials, such lack of efficacy may put patients in increased risk of health problems.

3 Our Approach

We propose a new approach to provide regression models under differential privacy and against model inversion attacks. Our approach aims to improve privacy specifically for sensitive attributes while retaining the efficacy of the released regression model by balancing the privacy budget for sensitive and non-sensitive attributes. Our approach leverages the *functional mechanism* proposed in [Zhang *et al.*, 2012] but perturbs the polynomial coefficients of the objective function with different magnitudes of noise.

3.1 Functional Mechanism Revisited

Functional mechanism achieves ϵ-differential privacy by perturbing the objective function $f_D(\omega)$ and then releasing the model parameter $\bar{\omega}$ that minimizes the perturbed objective function $\bar{f}_D(\omega)$ instead of the original one. Because $f_D(\omega)$ is a complicated function of ω, the functional mechanism exploits the polynomial representation of $f_D(\omega)$.

The model parameter ω is a vector that contains d values $\omega_1, \omega_2, ..., \omega_d$. Let $\phi(\omega)$ denote a product of $\omega_1, \omega_2, ..., \omega_d$, namely, $\phi(\omega) = \omega_1^{c_1} \cdot \omega_2^{c_2} ... \omega_d^{c_d}$ for some $c_1, c_2, ..., c_d \in \mathbf{N}$. Let $\Phi_j (j \in \mathbf{N})$ denote the set of all products of $\omega_1, \omega_2, ..., \omega_d$ with degree j, i.e.,

$$\Phi_j = \{\omega_1^{c_1} \cdot \omega_2^{c_2} ... \omega_d^{c_d} | \sum_{l=1}^{d} c_l = j\}. \tag{7}$$

By the Stone-Weierstrass Theorem, any continuous and differentiable $f(t_i, \omega)$ can always be written as a polynomial of $\omega_1, \omega_2, ..., \omega_d$, i.e., for some $J \in [0, \infty]$, we have

$$f(t_i, \omega) = \sum_{j=0}^{J} \sum_{\phi \in \Phi_j} \lambda_{\phi t_i} \phi(\omega), \tag{8}$$

where $\lambda_{\phi t_i} \in R$ denotes the coefficient of $\phi(\omega)$ in the polynomial. Similarly, $f_D(\omega)$ can also be expressed as a polynomial of $\omega_1, ..., \omega_d$.

For example, the polynomial expression of the linear regression is as follows.

$$\begin{aligned} f_D(\omega) &= \sum_{t_i \in D} (y_i - \mathbf{x}_i^T \omega)^2 \\ &= \sum_{t_i \in D} y_i^2 - \sum_{j=1}^{d} (2 \sum_{t_i \in D} y_i x_{ij}) \omega_j \\ &+ \sum_{1 \leq j,l \leq d} (\sum_{t_i \in D} x_{ij} x_{il}) \omega_j \omega_l \end{aligned} \tag{9}$$

We can see that $f_D(\omega)$ only involves monomials in $\Phi_0 = \{1\}$, $\Phi_1 = \{\omega_1, \omega_2, ..., \omega_d\}$, and $\Phi_2 = \{\omega_i \omega_j | i, j \in [1, d]\}$. Each $\phi(\omega)$ has its own coefficient, e.g., for ω_j, its polynomial coefficient $\lambda_{\phi_{t_i}} = -2 y_i x_{ij}$.

$f_D(\omega)$ is perturbed by injecting Laplace noise into its polynomial coefficients λ_ϕ, and then the model parameter $\bar{\omega}$ is derived to minimize the perturbed function $\bar{f}_D(\omega)$. Each polynomial coefficient λ_ϕ is perturbed by adding Laplace noise $Lap(\frac{\Delta}{\epsilon})$, where $\Delta = 2max_t \sum_{j=1}^{J} \sum_{\phi \in \Phi_j} ||\lambda_{\phi t}||_1$, according to the following Lemma 1.

Lemma 1. *[Zhang* et al.*, 2012] Let D and D' be any two neighboring datasets. Let $f_D(\omega)$ and $f_{D'}(\omega)$ be the objective functions of regression analysis on D and D', respectively, and denote their polynomial representations as follows:*

$$f_D(\omega) = \sum_{j=1}^{\mathrm{J}} \sum_{\phi \in \Phi_j} \sum_{t_i \in D} \lambda_{\phi t_i} \phi(\omega),$$

$$f_{D'}(\omega) = \sum_{j=1}^{\mathrm{J}} \sum_{\phi \in \Phi_j} \sum_{t'_i \in D'} \lambda_{\phi t'_i} \phi(\omega).$$

Then, we have the following inequality

$$\sum_{j=1}^{\mathrm{J}} \sum_{\phi \in \Phi_j} || \sum_{t_i \in D} \lambda_{\phi t_i} - \sum_{t'_i \in D'} \lambda_{\phi t'_i} ||_1 \leq 2 \max_t \sum_{j=1}^{\mathrm{J}} \sum_{\phi \in \Phi_j} ||\lambda_{\phi t}||_1. \tag{10}$$

where t_i, t'_i or t is an arbitrary tuple.

When the polynomial form of an objective function (e.g., logistic regression objective function) contains terms with unbounded degrees, [Zhang *et al.*, 2012] developed an approximation polynomial form based on Taylor expansion. The perturbation method based on the functional mechanism [Zhang *et al.*, 2012] also removes the requirements, i.e., the convexity of the objective function, from the original function perturbation approach [Chaudhuri and Monteleoni, 2008].

3.2 Improved Perturbation of Objective Function

To better optimize the balancing between privacy and the regression model' efficacy, we propose a new algorithm based on the functional mechanism to improve privacy specifically for sensitive attributes. To improve the privacy on X_s, we aim to weaken the correlation between X_s and the model output Y by perturbing the corresponding ω_s more intensely. In other words, we need to add noise with larger magnitude to the coefficients of the monomials involving ω_s and add noise with smaller magnitude to the other coefficients. As a result, we expect to retain the utility of the released regression model while achieving differential privacy for participants and preventing model inversion attacks.

In general, a database can contain more than one sensitive attribute. We allocate privacy budget ϵ_n for non-sensitive attributes and ϵ_s for sensitive ones. We introduce a ratio parameter, γ such that $\epsilon_s = \gamma \epsilon_n$ and $0 < \gamma \leq 1$. The smaller the γ, the more noise added to the sensitive attributes.

Algorithm 1 Functional Mechanism with Different Perturbation of Coefficients

Input: Database D, objective function $f_D(\omega)$, privacy threshold ϵ, privacy budget ratio γ

Output: $\bar{\omega}$

```
1: Set Φ_n = {}, Φ_s = {};
2: for each 1 ≤ j ≤ J do
3:     for each φ ∈ Φ_j do
4:         if φ does not contain ω_s for any sensitive attribute then
5:             Add φ into Φ_n;
6:         else
7:             Add φ into Φ_s;
8:         end if
9:     end for
10: end for
11: Set Δ = 2max_t Σ_{j=1}^{J} Σ_{φ∈Φ_j} ||λ_φt||_1;
12: Set β_1 = 2max_t Σ_{φ∈Φ_n} ||λ_φt||_1/Δ;
13: Set β_2 = 2max_t Σ_{φ∈Φ_s} ||λ_φt||_1/Δ; (β_1 + β_2 = 1)
14: Set ε_n = 1/(β_1 + γβ_2) ε, ε_s = γ/(β_1 + γβ_2) ε;
15: for each 1 ≤ j ≤ J do
16:     for each φ ∈ Φ_j do
17:         if φ ∈ Φ_n then
18:             set λ_φ = Σ_{t_i∈D} λ_φt_i + Lap(Δ/ε_n);
19:         else
20:             set λ_φ = Σ_{t_i∈D} λ_φt_i + Lap(Δ/ε_s);
21:         end if
22:     end for
23: end for
24: Let f̄_D(ω) = Σ_{j=1}^{J} Σ_{φ∈Φ_j} λ_φ φ(ω);
25: Compute ω̄ = argmin_ω f̄_D(ω);
26: return ω̄;
```

Specifically, we split all ϕs into two subsets Φ_n and Φ_s based on whether they involve any sensitive attribute, as shown in Lines 1-10 of Algorithm 1. Secondly, we determine the privacy budget according to the given ϵ and the privacy budget ratio γ. In Line 11, we set Δ based on the maximum value of all the coefficients $\lambda_{\phi t}$ of $\phi(\omega)$ in the polynomial. Accordingly, In Lines 12-13, β_1 and β_2 can be considered as the fraction of contributions to Δ from coefficients corresponding to elements in Φ_n and that in Φ_s. We will derive formula of Δ, β_1 and β_2 for linear regression and logistic regression and show they do not disclose any private information about dataset D in Results 1 and 2, respectively. Thirdly, we add noise to polynomial coefficients of $\phi \in \Phi_n$ with ϵ_n and to those of $\phi \in \Phi_s$ with ϵ_s, to derive the differentially private objective function $\bar{f}_D(\omega)$. Finally, we calculate and output the optimized $\bar{\omega}$ according to $\bar{f}_D(\omega)$. Next we show our algorithm achieves ϵ-differential privacy.

Theorem 2. *Algorithm 1 satisfies ϵ-differential privacy.*

Proof. Assume D and D' are two neighbouring datasets. Without loss of generality, D and D' differ in the last row t_n and $t_n^{'}$. Δ is calculated as Line 11 of Algorithm 1, and $\bar{f}(\omega)$ is the output of Line 24. Φ_s (Φ_n) denotes the set of ϕ that does (does not) contain sensitive attribute ω_s. The maximum difference of the objective function on D and D' is then the maximum difference of coefficients introduced by t_n and $t_n^{'}$. Adding Laplace noise to coefficients would produce the differentially private objective function. Specifically, we can add different magnitudes of noise to coefficients corresponding to $\phi \in \Phi_n$ and those corresponding to $\phi \in \Phi_s$. γ is pre-determined as the ratio of such difference of noise magnitude. Formally, we have

$$
\begin{aligned}
Pr(\bar{f}(\omega|D)) = \prod_{\phi\in\Phi_n} exp(\frac{\epsilon_n || \sum\limits_{t_i\in D} \lambda_{\phi t_i} - \lambda_\phi ||_1}{\Delta}) \\
\prod_{\phi\in\Phi_s} exp(\frac{\epsilon_s || \sum\limits_{t_i\in D} \lambda_{\phi t_i} - \lambda_\phi ||_1}{\Delta})
\end{aligned} \tag{11}
$$

Similarly, we have the formula for $Pr(\bar{f}(\omega|D'))$.

$$
\begin{aligned}
&\frac{Pr(\bar{f}(\omega|D))}{Pr(\bar{f}(\omega|D'))} \\
&\le \frac{\prod\limits_{\phi\in\Phi_n} exp(\frac{\epsilon_n}{\Delta} || \sum\limits_{t_i\in D} \lambda_{\phi t_i} - \sum\limits_{t_i^{'}\in D'} \lambda_{\phi t_i^{'}} ||_1)}{\prod\limits_{\phi\in\Phi_s} exp(\frac{\epsilon_s}{\Delta} || \sum\limits_{t_i\in D} \lambda_{\phi t_i} - \sum\limits_{t_i^{'}\in D'} \lambda_{\phi t_i^{'}} ||_1)} \\
&= \frac{\prod\limits_{\phi\in\Phi_n} exp(\frac{\epsilon_n}{\Delta} ||\lambda_{\phi t_n} - \lambda_{\phi t_n^{'}}||_1)}{\prod\limits_{\phi\in\Phi_s} exp(\frac{\epsilon_s}{\Delta} ||\lambda_{\phi t_n} - \lambda_{\phi t_n^{'}}||_1)} \\
&\le \prod_{\phi\in\Phi_n} exp(\frac{\epsilon_n}{\Delta} 2\max_t ||\lambda_{\phi t}||_1) \prod_{\phi\in\Phi_s} exp(\frac{\epsilon_s}{\Delta} 2\max_t ||\lambda_{\phi t}||_1) \\
&= exp(\epsilon_n \beta_1 + \epsilon_s \beta_2) \\
&= exp(\frac{\beta_1}{\beta_1 + \gamma\beta_2}\epsilon + \frac{\gamma\beta_2}{\beta_1 + \gamma\beta_2}\epsilon) = exp(\epsilon)
\end{aligned} \tag{12}
$$

□

Our algorithm needs to derive Δ, β_1 and β_2 to add noise with different magnitudes to the polynomial coefficients of sensitive attributes and non-sensitive attributes. Result 1 shows their derived formulas for linear regression and Result 2 shows for logistic regression. We can see they only involve the number of attributes d and the number of sensitive attributes k. As a result, they do not disclose any private information of the dataset D, which guarantees the rigorous ϵ-differential privacy. Due to space limitations, we only give the proof for linear regression in Result 1 and skip the proof for logistic regression in Result 2.

Result 1. *For linear regression defined in Definition 1, assume there are k sensitive attributes among the total d input attributes. We have in Algorithm 1, $\Delta = 2(d^2 + 2d)$; $\beta_1 = \frac{d-k}{d}$ and $\beta_2 = \frac{k}{d}$.*

Proof. According to Equation 9, we have

$$
\begin{aligned}
\Delta &= 2 \max_{t=(\mathbf{x},y)} \sum_{j=1}^{J} \sum_{\phi\in\Phi_j} ||\lambda_{\phi t}||_1 \\
&\le 2 \max_{t=(\mathbf{x},y)} (2\sum_{j=1}^{d} yx_{(j)} + \sum_{1\le j,l\le d} x_{(j)}x_{(l)}) \\
&= 2(2d + d^2)
\end{aligned} \tag{13}
$$

where $x_{(j)}$ denotes the jth entry in vector $\mathbf{x}$, which satisfies $|x_{(j)}| \le 1$. Similarly, for the coefficients related to k sensitive attributes, we have

$$
2 \max_{t=(\mathbf{x},y)} \sum_{j=1}^{J} \sum_{\phi\in\Phi_s} ||\lambda_{\phi t}||_1 = 2(2k + kd) \tag{14}
$$

Thus $\beta_2 = \frac{2(2k+kd)}{2(2d+d^2)} = \frac{k}{d}$. Similarly we have $\beta_1 = \frac{d-k}{d}$. □

Result 2. *For logistic regression defined in Definition 2, assume there are k sensitive attributes among the total d input attributes. Algorithm 1 can be applied with the approximate objective function $\sum_{i=1}^{n} \frac{1}{8}(x_i^T\omega)^2 + \sum_{i=1}^{n}(\frac{1}{2} - y_i)x_i^T\omega$ based on Taylor expansion. We have $\Delta = \frac{d^2}{4} + 3d$ and $\beta_1 = \frac{d-k}{d}, \beta_2 = \frac{k}{d}$.*

4 Evaluation

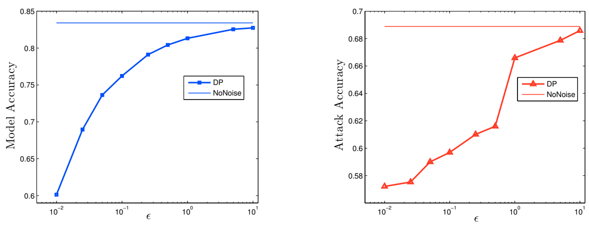
In our experiments, we mainly focus on the problem of releasing the logistic regression model under differential privacy against model inversion attacks. We use the Adult dataset [Lichman, 2013] to evaluate the performance of Algorithm 1 and apply five-fold cross validation for all the accuracy calculation. The Adult dataset contains census information of 30,175 individuals with 14 attributes such as *age, workclass, education, marital-status, hours-per-week* and so on. The regression task is to predict whether the *income* of an individual is greater than 50K. Among the 13 input attributes, we pick *marital status* as the sensitive attribute which the model inversion attack would target.

Figure 1 shows how both the accuracy of the released model and the accuracy of the model inversion attack are affected by different ϵ values varying from 0.01 to 10. In this experiment, we do not differentiate the privacy budget for sensitive attribute and non-sensitive attribute. We can see from Figure 1(a) that the prediction accuracy of the regression model increases as ϵ increases and from Figure 1(b) that the accuracy of the model inversion attack on *marital status* also increases as as ϵ increases. This is not surprising because the larger ϵ is, the less noise introduced in the released model. Hence, the

Table 2: Privacy Budget for $\epsilon = 1$

γ	0.5	0.25	0.1	0.05	0.025	0.01
ϵ_s	0.52	0.265	0.107	0.054	0.027	0.011
ϵ_n	1.04	1.061	1.074	1.079	1.081	1.082

model has high utility but also incurs high risk under model inversion attacks. We can see even with small ϵ values such as 0.01, the model inversion attack still outperform random guessing based on the marginal probability (0.57 vs. 0.53) although the model utility is significantly lost.



(a) Model Accuracy (b) Attack Accuracy

Figure 1: Accuracy of logistic regression model and that of model inversion attack vs. varying ϵ

In the second experiment, we set the privacy threshold $\epsilon = 1$ and change the privacy budget ratio γ from $\{1, 0.5, 0.25, 0.1, 0.05, 0.025, 0.01\}$. Table 2 shows the corresponding ϵ_s and ϵ_n values under each γ. Note that these values can be easily derived from Result 2.

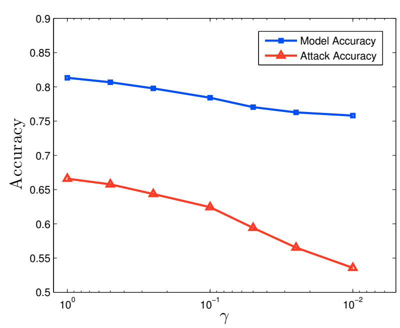


Figure 2: Accuracy of logistic regression and that of model inversion attack vs. varying γ when $\epsilon = 1$

Figure 2 shows the accuracy trend of the model prediction task and the model inversion attack. We can see that the prediction accuracy of the released model generally stays stable. However, the accuracy of the model inversion attack shows a clear decreasing trend as γ goes small. For example, when γ is 0.01 or 0.025 , the model inversion attack would fail to beat random guessing based on the marginal probability; while the prediction accuracy of regression model can still stay higher than 75%. This result shows that our approach can significantly decrease the attribute privacy disclosure risk due to the model inversion attack while retaining the utility of the released regression model.

5 Related Work

Differential privacy research has been significantly studied from the theoretical perspective, e.g., [Chaudhuri and Monteleoni, 2008; Hay *et al.*, 2010; Kifer and Machanavajjhala, 2011; Lee and Clifton, 2012; Ying *et al.*, 2013]. There are also studies on the applicability of enforcing differential privacy in real world applications, e.g., collaborative recommendation [McSherry and Mironov, 2009], logistic regression [Chaudhuri and Monteleoni, 2008; Zhang *et al.*, 2012], publishing contingency tables [Xiao *et al.*, 2010; Barak *et al.*, 2007] or data cubes [Ding *et al.*, 2011], privacy preserving integrated queries [McSherry, 2009], synthetic graph generation [Wang and Wu, 2013; Mir and Wright, 2009; Sala *et al.*, 2011], computing graph properties such as degree distributions [Hay *et al.*, 2009] and clustering coefficient [Rastogi *et al.*, 2009; Wang *et al.*, 2012], and spectral graph analysis [Wang *et al.*, 2013] in social network analysis. The mechanisms of achieving differential privacy mainly include the classic approach of adding Laplace noise [Dwork *et al.*, 2006], the exponential mechanism based on the smooth sensitivity [McSherry and Mironov, 2009], and the functional perturbation approach [Chaudhuri and Monteleoni, 2008; Zhang *et al.*, 2012]. Privacy preserving models based on differential privacy guarantee protection against attempts to infer whether a subject was included in the training set used to derive models. However, they are not designed to protect attribute privacy of a target individual when model inversion attacks are launched. In this paper, we have studied how to effectively prevent model inversion attacks while retaining model efficacy.

There are several studies that showed differential privacy still could leak various type of private information. In [Kifer and Machanavajjhala, 2011], the authors showed that when rows in a database are correlated, or when previous exact statistics for a dataset have been released, differential privacy cannot achieve the ultimate privacy goal – nearly all evidence of an individual's participation should be removed. The authors in [Cormode, 2011] showed that if one is allowed to pose certain queries relating sensitive attributes to quasi-identifiers, it is possible to build a differentially-private Naive Bayes classifier that accurately predicts the sensitive attribute. The authors [Dankar and El Emam, 2012] examined the various tradeoffs between interactive and non-interactive mechanisms and the limitation of utility guarantees in differential privacy. Another notable work [Lee and Clifton, 2012] studied the relationship of ϵ to the relative nature of differential privacy.

6 Conclusion and Future Work

Recent work [Fredrikson *et al.*, 2014] showed that the existing differential privacy mechanisms cannot prevent model inversion attacks while retaining desirable model efficacy. In this paper, we have developed an effective approach which simultaneously protects differential privacy of participants and prevents sensitive attribute disclosure of regular individuals from model inversion attacks while retaining the efficacy of released regression models. Leveraging the functional mechanism [Zhang *et al.*, 2012], our approach rewrites the objective function in its polynomial representation and adds more (less) noise to the polynomial coefficients with (w/o) sensitive attributes. Our approach can effectively weaken the correlation between the sensitive attributes with the output to prevent model inversion attacks whereas retaining the utility of the released model by decreasing the perturbation effect on non-sensitive attributes. As a result, we still achieve ϵ-differential privacy for participants. In our future work, we will evaluate our research on real world applications such as clinical study which involves genetic privacy. We plan to theoretically analyze applicability of model inversion attacks under different background knowledge. We will explore other perturbation strategies to decrease utility loss under differential privacy and potential model inversion attacks.

Acknowledgments

The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported in part by U.S. National Institute of Health (1R01GM103309).

References

[Barak *et al.*, 2007] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282. ACM, 2007.

[Chaudhuri and Monteleoni, 2008] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, pages 289–296. Citeseer, 2008.

[Cormode, 2011] Graham Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *KDD*, pages 1253–1261. ACM, 2011.

[Dankar and El Emam, 2012] Fida Kamal Dankar and Khaled El Emam. The application of differential privacy to health data. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 158–166. ACM, 2012.

[Ding *et al.*, 2011] Bolin Ding, Marianne Winslett, Jiawei Han, and Zhenhui Li. Differentially private data cubes: optimizing noise sources and consistency. In *SIGMOD*, pages 217–228, 2011.

[Dwork *et al.*, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, pages 265–284, 2006.

[Fredrikson *et al.*, 2014] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, pages 17–32, 2014.

[Hay *et al.*, 2009] Michael Hay, Chao Li, Gerome Miklau, and David Jensen. Accurate estimation of the degree distribution of private networks. In *ICDM*, pages 169–178. IEEE, 2009.

[Hay *et al.*, 2010] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the Accuracy of Differentially Private Histograms Through Consistency. *Proceedings of the VLDB Endowment*, 3(1), 2010.

[Kifer and Machanavajjhala, 2011] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, pages 193–204, 2011.

[Lee and Clifton, 2012] Jaewoo Lee and Chris Clifton. Differential identifiability. In *KDD*, pages 1041–1049, 2012.

[Lichman, 2013] M. Lichman. UCI machine learning repository, 2013.

[McSherry and Mironov, 2009] Frank McSherry and Ilya Mironov. Differentially Private Recommender Systems. In *KDD*. ACM, 2009.

[McSherry, 2009] Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *SIGMOD*, pages 19–30. ACM, 2009.

[Mir and Wright, 2009] Darakhshan Mir and Rebecca Wright. A differentially private graph estimator. In *ICDMW*, pages 122–129. IEEE, 2009.

[Rastogi *et al.*, 2009] Vibhor Rastogi, Miachel Hay, Gerome Miklau, and Dan Suciu. Relationship privacy: Output perturbation for queries with joins. In *PODS*, pages 107–116. ACM, 2009.

[Sala *et al.*, 2011] Alessandra Sala, Xiaohan Zhao, Christo Wilson, Haitao Zheng, and Ben Zhao. Sharing graphs using differentially private graph models. In *ACM SIGCOMM*, pages 81–98. ACM, 2011.

[Wang and Wu, 2013] Yue Wang and Xintao Wu. Preserving differential privacy in degree-correlation based graph generation. *Transactions on Data Privacy*, 6:127–145, 2013.

[Wang *et al.*, 2012] Yue Wang, Xintao Wu, Jun Zhu, and Yang Xiang. On learning cluster coefficient of private networks. In *ASONAM*, 2012.

[Wang *et al.*, 2013] Yue Wang, Xintao Wu, and Leting Wu. Differential privacy preserving spectral graph analysis. In *PAKDD (2)*, pages 329–340, 2013.

[Xiao *et al.*, 2010] Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Differential privacy via wavelet transforms. In *ICDE*, pages 225–236. IEEE, 2010.

[Ying *et al.*, 2013] Xiaowei Ying, Xintao Wu, and Yue Wang. On linear refinement of differential privacy-preserving query answering. In *PAKDD (2)*, pages 353–364, 2013.

[Zhang *et al.*, 2012] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5(11):1364–1375, 2012.