Contract Design for Purchasing Private Data Using a Biased Differentially Private Algorithm*

Mohammad Mahdi Khalili†
EECS, University of Michigan
khalili@umich.edu

Xueru Zhang†
EECS, University of Michigan
xueru@umich.edu

Mingyan Liu
EECS, University of Michigan
mingyan@umich.edu

ABSTRACT

Personal information and other types of private data are valuable for both data owners and institutions interested in providing targeted and customized services that require analyzing such data. In this context, privacy is sometimes seen as a commodity: institutions (data buyers) pay individuals (or data sellers) in exchange for private data. In this study, we examine the problem of designing such data contracts, through which a buyer aims to minimize his payment to the sellers for a desired level of data quality, while the latter aim to obtain adequate compensation for giving up a certain amount of privacy. Specifically, we use the concept of differential privacy and examine a model of linear and nonlinear queries on private data. We show that conventional algorithms that introduce differential privacy via zero-mean noise fall short for the purpose of such transactions as they do not provide sufficient degree of freedom for the contract designer to negotiate between the competing interests of the buyer and the sellers. Instead, we propose a biased differentially private algorithm which allows us to customize the privacy-accuracy tradeoff for each individual. We use a contract design approach to find the optimal contracts when using this biased algorithm to provide privacy, and show that under this combination the buyer can achieve the same level of accuracy with a lower payment as compared to using the unbiased algorithms, while incurring lower privacy loss for the sellers.

CCS CONCEPTS

• **Security and privacy → Economics of security and privacy**.

1 INTRODUCTION

Advances in data centers have enabled storing large amounts of data containing private information of individuals/firms. These data have value for institutions interested in analyzing them for a variety of purposes such as targeted advertising. Individuals are not willing to share their data due to privacy concerns; even when they are not concerned with how institutions use their respective data, they can still be reluctant to share due to the possibility of data breaches. Within this context, privacy has become a commodity that institutions often have to pay monetary or non-monetary compensation for using it. For instance, Datacoup is a new startup offering monthly payment in return for the access to users' online accounts and credit card transactions. While Datacoup protects users' identities as well as credit card numbers, it provides aggregated and/or de-identified information about the users to any third party, including advertisers, data buyers and analytics partners [1].

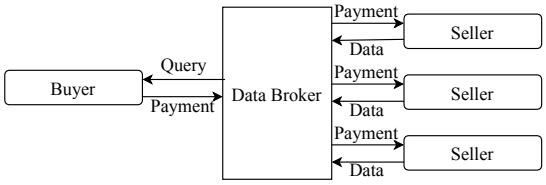

Figure 1: Interaction of buyer and sellers.

Studies of privacy as a commodity include arbitrage-free privacy-preserving pricing mechanisms, see e.g., [11], designing contracts for data privacy and utility [14], auctions and direct mechanisms for selling privacy [8, 9], as well as dynamic privacy pricing [15]. A more detailed literature review is given in Section 2.

In this paper, we consider a single buyer, whose goal is to minimize his payment to data owners, also referred to as sellers, provided that the purchased data satisfy a certain level of accuracy. The sellers value their privacy, but are willing to sell their data provided the cost of their privacy loss, measured by the concept of differential privacy [6], is adequately compensated by the payment.

The transaction takes place as follows. The buyer announces his desired accuracy level of a certain computational output, e.g., in the form of a query over certain types of data, to a trusted third party, also referred as the *data broker*. The data broker collects relevant data from different individuals/sellers and generates such an output, which he then releases to the buyer. The buyer pays each individual, through the broker, an amount commensurate with the privacy loss the individual experiences as a result of the release of the computational output. Figure 1 illustrates these interactions. A data contract among these parties stipulates the payment amount and quantifies accuracy as well as privacy guarantees associated with the payment. In the current model the broker is assumed to be a neural, non-profit entity, but our analysis and conclusions hold if the broker charges a fixed processing fee. A key component of this framework is a differentially private algorithm that preserves the privacy of the input data and returns a differentially private output for the query. Toward this end, we propose a randomized algorithm that, in contrast to most commonly used algorithms that add a zero-mean noise to the data, see e.g., [11], adds not only a zero-mean noise to the private data, but also a bias. As we will show, the introduction of this bias allows the broker to add less noise to

*This work is supported by the NSF under grants CNS-1616575, CNS-1646019, CNS-1739517. † indicates equal contribution.

NetEcon'19, June 28, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6837-7/19/06…$15.00
https://doi.org/10.1145/3338506.3340273

the data and increase the accuracy of the output simultaneously. Furthermore, it provides an additional degree of freedom that the broker can use to judiciously determine individual privacy losses based on individual privacy valuations. As a result, we show that by choosing the bias term carefully, a contract can be designed for the buyer to obtain the desired accuracy level at a lower cost, as compared to when an unbiased algorithm is used, while at the same time the sellers experience less privacy loss. In other words, both buyer and sellers benefit from using this algorithm. It is worth noting that [9] also introduces a biased differentially private algorithm for linear queries and one-dimensional data which offers only a single privacy level to the participant sellers. We generalize the algorithm introduced by [9] in the following aspects: i) Our algorithm is able to assign different privacy loss to the sellers, ii) We also extend our algorithm for nonlinear queries and multidimensional data.

Our main contribution is two-fold. Firstly, we present a new algorithm for generating differentially private estimates of a family of linear and nonlinear queries, and show that this algorithm allows the data broker to assign different privacy losses to different individuals. Secondly, we use a contract design approach to derive optimal data contracts that minimize the buyer's payment while satisfying his accuracy requirement and the seller's privacy constraint. This is done under two scenarios, one with full information, where the data broker knows the sellers' privacy valuation, and one with information asymmetry, where the broker does not know their privacy valuation but its distribution. We show in both scenarios, the broker can leverage the proposed algorithm to guarantee a lower privacy loss to an individual with higher privacy valuation.

The remainder of the paper is organized as follows. We present related work in Section 2 and preliminaries on differential privacy and query in Section 3. A biased differentially private algorithm is introduced in Section 4. In Section 5 we analyze the contract design problem between one buyer and one seller, and between one buyer and multiple sellers under full information. We discuss the contract design problem for purchasing private data under information asymmetry in Section 6. Section 7 concludes the paper, and proofs of the theorems and generalization of our algorithm for non linear queries as well as multi dimensional data are provided in online Appendix [2].

2 RELATED WORK

Studies most relevant to the present paper are [5, 8, 9, 14]. In [14], contracts are designed for a data market where data utility and privacy are considered, with the main conclusion that when the data collector requires a large amount of data, it is better to purchase from those who care the least about their privacy. It, however, does not provide any algorithm or mechanism to ensure privacy. A truthful mechanism for purchasing privacy is proposed in [5, 8, 9]. Gosh and Roth [9] introduce a fixed price auction mechanism using a biased algorithm which offers only a single privacy level to the sellers participating in the mechanism. This work was extended in [8], where the cost of privacy loss is correlated with the private data. Cummings *et.al* [5] design a truthful mechanism for a data aggregation problem where a buyer collects unbiased estimate of each individual's data and makes a payment based on the variance of the estimate. The buyer then calculates the average of all unbiased estimates to find a better estimate. It is worth noting that this mechanism is only applicable when the expected values of individuals' data are the same.

Privacy preserving mechanisms have also been studied in the context of data aggregation and task bidding in crowd sensing [10, 13], and in the context of security information exchange [12].

3 PRELIMINARIES

In this section, we review the notion of differential privacy proposed in [6, 7], which is widely used in the machine learning and optimization literature [3, 16, 17] to quantify privacy leakage. Then we will introduce a type of linear query.

We consider n individuals indexed by $\{1, 2, \cdots, n\}$. Let $d_i \in X$ be individual i's private data where X is a subset of real numbers. Extension to higher dimensional data is discussed in online Appendix [2]. An individual incurs a cost if his privacy is violated.

Differential privacy and accuracy: Consider database $D = (d_1, d_2, \cdots, d_n) \in X^n$, the collection of n individuals' data. Database $D = (d_1, d_2, \cdots, d_n)$ and $D^{(i)} = (d_1^{(i)}, d_2^{(i)}, \cdots, d_n^{(i)})$ are said to be neighbors if $d_j = d_j^{(i)}$ for all $j \neq i$ and $d_i \neq d_i^{(i)}$. In other words, D and $D^{(i)}$ are neighbors if and only if individual i's data is different in D and $D^{(i)}$.

Definition 3.1 (ϵ-Differential Privacy [6, 7]). An algorithm $A : X^n \to R$ is ϵ_i-differentially private with respect to individual i, if for all neighboring databases $D \in X^n$ and $D^{(i)} \in X^n$ differing only in element i, and for any $S \subset R$ we have, $\frac{Pr\{A(D)\in S\}}{Pr\{A(D^{(i)})\in S\}} \leq \exp\{\epsilon_i\}$.

This suggests that $A(.)$ is in general a randomized algorithm. A common method for making an algorithm ϵ_i-differentially private is adding Laplace noise to its output. Let $N(b)$ be the symmetric Laplacian noise with zero mean and parameter b. Then $N(b)$ has a variance of $2b^2$ and a distribution given by: $f(x) = \frac{1}{2b} \exp\{-\frac{|x|}{b}\}$.

Definition 3.2 (Accuracy). We say algorithm $A(.)$ is K-accurate for query $Q(D)$ if $E\left[(A(D) - Q(D))^2\right] \leq K,\ \forall D \in X^n$, i.e., algorithm A is K-accurate if its Mean Squared Error (MSE) is at most K. Smaller K indicates a more accurate algorithm.

There are other definitions for accuracy (e.g., see [7]), but the above choice does not affect the applicability of our methodology and main conclusions.

A type of linear query:

Definition 3.3 (Linear Query). A linear Query $Q : X^n \to R$ over the database $D = (d_1, d_2, \cdots, d_n)$ is a linear function evaluated as follows: $Q(D) = \sum_{i=1}^n q_i \cdot d_i$, where $q_i \in R$ are constants.

Without loss of generality, we will assume that $X = [0, 1]$ and $q_i = 1, \forall i$. Note that if $q_i \neq 1$, then we can assume that $d_i \in [0, q_i]$ and $Q(D)$ is the summation of d_i's. The generality of a summation form of query lies in the fact that it is sufficient to implement many machine learning algorithms in a differentially private manner [4]. Extension to non-linear queries is discussed in Appendix [2].

We next examine the relationship between accuracy K and privacy loss ϵ in this type of linear query. Intuitively, we expect an algorithm with high accuracy to also have high privacy loss. Below we find a lower bound on the total privacy loss $\sum_{i=1}^n \epsilon_i$ as a function of K.

Theorem 3.4 (Lower Bound on Total Privacy Loss). *If algorithm $A(D)$ is K-accurate and $K < (n/2)^2$,[1] then the total privacy loss is at least $\ln \frac{(n-\sqrt{K})^2}{K}$. Moreover, if $K < (m/2)^2$, then at least $n-m+1$ individuals experience non-zero privacy loss.*

Theorem 3.4 implies that as $K \to 0$, privacy loss approaches infinity logarithmically. We will introduce an algorithm in Section 4 under which the total privacy loss is close to the lower bound when K is close to $(n/2)^2$.

4 UNBIASED AND BIASED ALGORITHMS

As mentioned, a common way to provide differential privacy to an algorithm is to add zero-mean noise.

Theorem 4.1 (An unbiased algorithm). *Let $A_u(D) = Q(D) + N(b)$. Then $A_u(D)$ is $\frac{1}{b}$-differentially private with respect to each individual. Moreover, $A_u(D)$ is $2b^2$-accurate.*

$A_u(D) = Q(D) + N(b)$ is an unbiased algorithm, as $E[A_u(D) - Q(D)] = 0$. We next introduce a biased estimate $A_{new}(D)$ of $Q(D)$ such that $E[A_{new}(D)] \neq Q(D)$.

Theorem 4.2 (A biased algorithm). *Let $A_{new}(D) = \sum_{i=1}^n a_i \cdot d_i + \sum_{i=1}^n \frac{1-a_i}{2} + N(b)$ where $0 \le a_i \le 1$, $\forall i$. Then $A_{new}(D)$ is $\left[\left(\sum_{i=1}^n \frac{1-a_i}{2}\right)^2 + 2b^2\right]$-accurate. Moreover, the algorithm is $\frac{a_i}{b}$-differentially private with respect to individual i.*

Algorithm $A_{new}(D)$ is a biased algorithm with the following bound on the bias:

$$|E[A_{new}(D) - Q(D)]| = |\sum_{i=1}^n (a_i - 1)d_i + \frac{1-a_i}{2}| \le \sum_{i=1}^n \frac{1-a_i}{2} \quad (1)$$

Therefore, increase in a_i decreases the algorithm's bias, improves its accuracy, and increases its privacy loss. Note that the bias does not depend on parameter b, and that $A_{new}(D)$ reduces to $A_u(D)$ by setting $a_i = 1, \forall i$.

5 CONTRACT DESIGN UNDER FULL INFORMATION

5.1 A single buyer and a single seller

We begin by presenting a model of a single buyer and a single seller: the individual has data $D = (d)$ and the buyer wants to find an estimate of d.

The individual has cost function $c(v, .) : R_+ \to R_+$, where v is the seller's *type* or his valuation of privacy; this is also referred to as his *privacy attitude*. The seller incurs a cost of $c(v, \epsilon)$ if he experiences privacy loss ϵ. We assume that $c(v, \epsilon)$ is increasing in privacy valuation v and privacy loss ϵ, and the cost of revealing data is zero if there is zero privacy loss, i.e., $c(v, 0) = 0$, $\forall i$.

The data transaction is facilitated by a contract (p, ϵ, K), whereby by accepting it the seller receives payment p and reports actual data d to the data broker, while the broker uses an algorithm to find an estimate of $Q(D)$ which is ϵ-differentially private and K-accurate; this estimate is then reported to the buyer.

[1] In the next sections, we will show that If $K > (n/2)^2$, there exists algorithm $A(D)$ which is K-accurate and 0-differentially private with respect to each individual. More precisely, $A(D)$ could be pure noise if $K > (n/2)^2$.

Under the full information scenario, we assume the seller's privacy attitude v is public information. To ensure the seller accepts contract (p, ϵ, K), the contract has to satisfy the Individual Rationality (IR) constraint that the payment it receives sufficiently compensates for its privacy cost, i.e., $(IR): \quad p \ge c(v, \epsilon)$.

The goal is to find a K-accurate estimate of $Q(D) = (d)$ with the minimum amount of payment. If algorithm $A_{new}(.)$ is used to find an estimate of $Q(D)$, then the contract design problem can be written as follows:

$$\min_{\{0 \le a \le 1, b > 0, p, \epsilon = a/b\}} p$$
$$\text{s.t.} \quad (IR)\ p \ge c(v, \epsilon), \ \ (AC)\ ((1-a)/2)^2 + 2b^2 \le K \quad (2)$$

where AC denotes the accuracy constraint, following the privacy and accuracy property of $A_{new}(.)$ derived earlier in Thm 4.2. Note that in this case minimizing p equals to minimizing privacy loss ϵ.

If we apply the unbiased algorithm $A_u(.)$ (setting $a = 1$), then the corresponding optimization problem becomes:

$$\min_{\{b>0, p\}} p \quad \text{s.t. } (IR)\ p \ge c(v, 1/b), \ \ (AC)\ 2b^2 \le K \quad (3)$$

Note that both IR and AC constraints are binding in the above problems, otherwise one can always increase b and decrease p. As a result, the optimal solution (p^*, b^*) to (3) is given by $b^* = \sqrt{K/2}$, $p^* = c(v, \sqrt{2/K})$, while the optimal solution to (2) can be found using the following theorem.

Theorem 5.1. *The optimal solution (p^*, a^*, b^*) to (2) is as follows.*
1) If $K > 1/4$, then $p^* = 0$, $a^* = 0$, and $b^* = \sqrt{(4K-1)/8}$.
2) If $K < 1/4$, then $a^* = 1 - 4K$, $b^* = \sqrt{(K - 4K^2)/2}$, and $p^* = c(v, \frac{a^*}{b^*} = \sqrt{(2/K) - 8})$.
3) If $K = 1/4$, then there is no solution to (2).

Thm 5.1 implies that at sufficiently low accuracy levels ($K \ge 1/4$) the optimal strategy for the seller is to not provide any data, or alternatively, for the data broker to report simply the noise. Thm 5.1 also leads to the following result.

1) $K > (1/4)$: Privacy loss under $A_u(.)$ is $\sqrt{(2/K)}$ while under $A_{new}(.)$ it is zero. Thus $A_{new}(.)$ decreases the cost from $c(v, \sqrt{2/K})$ to zero.

2) $K < (1/4)$: Privacy loss under $A_u(.)$ is $\sqrt{(2/K)}$ while under $A_{new}(.)$ is $\sqrt{(2/K) - 8}$; thus again $A_{new}(.)$ reduces privacy loss and the resulting cost. Notice that as the IR constraint is binding, minimizing p is equivalent to minimizing ϵ. Therefore, $\epsilon = \sqrt{(2/K) - 8}$ is the minimum privacy loss that we can obtain under algorithm A_{new} and subject to accuracy K.

As stated earlier, Thm 3.4 suggests that a K-accurate estimate of $Q(D)$ has privacy loss at least $2\ln(1 - \sqrt{K}) - \ln K$. The privacy loss $\sqrt{(2/K) - 8}$ under $A_{new}(.)$ approaches this lower bound as $K \to 1/4$. Figure 2 compares the minimum privacy loss using algorithms $A_u(.)$ and $A_{new}(.)$. Clearly $A_{new}(.)$ outperforms $A_u(.)$ in terms of the cost/privacy-accuracy tradeoff: by introducing a bias, $A_{new}(.)$ uses less noise (as compared to $A_u(.)$) to reach a given privacy loss which improves accuracy.

5.2 A model of N sellers

We now consider the scenario with n sellers and a single buyer with query on database $D = (d_1, d_2, \cdots, d_n)$, where data d_i belongs to

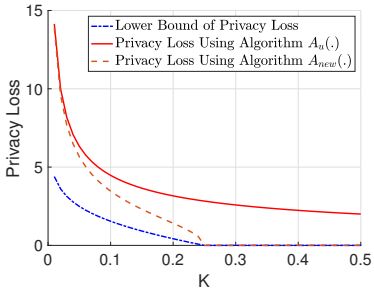

Figure 2: Minimum privacy loss under different algorithms.

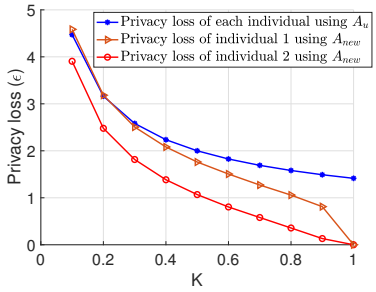

Figure 3: Privacy loss v.s. accuracy, under full information. Using A_{new} individuals experience less privacy loss than using A_u.

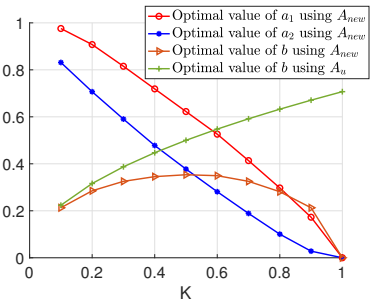

Figure 4: Optimal values of a_1, a_2, b v.s. accuracy, under full information. As $v_2 > v_1$, the optimal contract using A_{new} sets $a_2^* < a_1^*$.

seller i. Moreover, seller i has privacy valuation v_i. Without loss of generality we assume $v_1 \le v_2 \le \cdots \le v_n$. Similar as before, we assume the individual privacy cost function $c(v_i, \epsilon_i)$ is increasing in i's type v_i and privacy loss ϵ_i, and $c(v_i, 0) = 0$. The buyer wishes to obtain an estimate for query $Q(D) = \sum_{i=1}^n d_i$, with accuracy K and minimum payment.

If the broker uses algorithm $A_{new}(.)$ to get an estimate of $Q(D)$, we have $A_{new}(D) = \sum_{i=1}^n (a_i \cdot d_i + \frac{1-a_i}{2}) + N(b)$. With this algorithm, individual i experiences privacy loss $\epsilon_i = (a_i/b)$. Similar to the optimization problem (2), we can write the problem for finding contracts (p_i, ϵ_i, K) as follows:

$$\min_{\{0\le a_i\le 1,\ p_i\ge 0,\ \epsilon_i = a_i/b,\ b>0\}} \quad \sum_{i=1}^n p_i$$
$$\text{s.t.} \quad (IR) \qquad p_i \ge c(v_i, \epsilon_i)\ i \in \{1, 2, \cdots, n\}$$
$$(AC) \qquad \left(\sum_{i=1}^n \frac{1-a_i}{2}\right)^2 + 2b^2 \le K. \qquad (4)$$

It is easy to verify that the (IR) constraint and (AC) constraint in (4) are binding and optimization problem (4) can be written as follows,

$$\min_{\{0\le a_i\le 1, b>0, \epsilon_i = \frac{a_i}{b}, i=1,2,\cdots,n\}} \quad \sum_{i=1}^n c(v_i, \epsilon_i)$$
$$\text{s.t.} \quad (AC) \qquad \left(\sum_{i=1}^n \frac{1-a_i}{2}\right)^2 + 2b^2 \le K \ (5)$$

If $v_1 = v_2 = \cdots = v_n = 1$ and $c(v, \epsilon) = v \cdot \epsilon$, then optimization problem (5) is equivalent to minimizing total privacy loss under algorithm $A_{new}(.)$ subject to accuracy K.

A closed form solution is not easy to find in general and depends on the form of the cost function. We next provide an example to highlight the salient features of the biased algorithm in the context of the contract design problem.

Example 5.2. Consider a case of two sellers with privacy attitudes $v_1 = 5$ and $v_2 = 10$ respectively, and the following cost function: $c(v, \epsilon) = v \cdot (e^\epsilon - 1)$.

Using algorithm $A_{new}(.)$, the broker offers the contract $(p_1, \epsilon_1 = a_1/b, K)$ to seller 1 and $(p_2, \epsilon_2 = a_2/b, K)$ to seller 2. Figure 3 plots the privacy loss of each seller as a function of K. It shows that both individuals experience less loss as compared to using $A_u(.)$. Moreover, we see that $A_{new}(.)$ allows the broker to take advantage of the full information and assign different privacy loss to different individuals to minimize the cost to the sellers (lower loss for those with higher privacy valuation). Figure 4 shows the optimal values for parameter a_1, a_2, b; it suggests that $A_{new}(.)$ adds less noise to the output as compared to algorithm $A_u(.)$. This example helps highlight the two reasons why $A_{new}(.)$ outperforms $A_u(.)$:

1) Under $A_{new}(.)$, the broker is able to assign different privacy losses to the two individuals. To minimize the total cost, an individual with higher privacy valuation is afforded lower privacy loss in the optimal contract.

2) Under $A_{new}(.)$, the broker uses less noise (as compared to $A_u(.)$) to provide the same privacy guarantee, which in turn increases accuracy. In other words, as in the case of a single seller, $A_{new}(.)$ improves privacy-accuracy tradeoff.

Next we solve (4) under a linear cost model.

Theorem 5.3. *Let $c(v, \epsilon) = v \cdot \epsilon$, $K < (\frac{n}{2})^2$ and $s_{i+1} = (n-i) - 4 \cdot K \cdot \frac{v_{i+1}}{(n-i)\cdot v_{i+1} + \sum_{j=1}^i v_j}$, $\forall i \ge n - 2\sqrt{K}$, $i \le n-1$. Let $m+1$ be the first index where $s_{m+1} \le 0$ (if $s_i \ge 0$, $\forall i$, then set $m = n$). Then the solution to problem (4) is given by:*

$$a_1^* = a_2^* = \cdots = a_{m-1}^* = 1,\ a_m^* = \min\{s_m, 1\},\ a_{m+1} = \cdots = a_n = 0$$
$$b^* = \sqrt{\frac{1}{2}(K - (\frac{2K \cdot v_m}{(n-m+1)\cdot v_m + \sum_{j=1}^{m-1} v_j})^2)},\ p_i^* = c(v_i, \frac{a_i^*}{b^*})$$

Note that if $K > (n/2)^2$, then $a_1 = a_2 = \cdots = a_n = 0$ and $b = \sqrt{\frac{K-(n/2)^2}{2}}$ give a feasible solution to (4). This point is optimal because its objective value is zero. Thus, if $K > (n/2)^2$, then the output will be a pure noise.

6 CONTRACT DESIGN UNDER INFORMATION ASYMMETRY

We now turn to the scenarios where the sellers' privacy attitudes are their private information unknown to the buyer or the broker, and focus on the case of two sellers.[2] We will assume the broker knows that the privacy attitudes come from a binary (high and low types) distribution: $v_i = \begin{cases} v_H & \text{w.p. } \pi \\ v_L & \text{w.p. } 1-\pi \end{cases}$, $i = 1, 2$, where

[2]This is for simplicity of presentation; results obtained in this section remain valid for more than two sellers.

$v_H \geq v_L$. Moreover, we assume that v_1 and v_2 are independent. In this case the optimal thing to do for the broker is to design a menu of contracts $\{(p_H, \epsilon_H, K), (p_L, \epsilon_L, K)\}$, one for each of the two types, such that a seller of a certain type will choose the contract designed for his type, i.e., he will select (p_t, ϵ_t, K) if he is of type v_t, where $t \in \{L, H\}$. The IR constraint remains the same: $p_t \geq c(v_t, \epsilon_t), t \in \{H, L\}$. An additional constraint in this case is Incentive Compatibility (IC), which ensures that a seller does not increase his utility by selecting the contract of the opposing type (i.e., misrepresenting his own type):

$$\begin{array}{ll} (IC) & p_H - c(v_H, \epsilon_H) \geq p_L - c(v_H, \epsilon_L)\,, \\ (IC) & p_L - c(v_L, \epsilon_L) \geq p_H - c(v_L, \epsilon_H)\,. \end{array} \tag{6}$$

The broker has two options, offering the same menu to both sellers, or offering the menu to one of the sellers and not using data from the other seller. Which option to invoke depends on which one results in lower payment, given the problem parameters. We next examine the contract design problem under each option in detail.

6.1 Broker offers both sellers the menu of contracts

Given the constraint (6), by accepting a contract the seller essentially reveals his type. Algorithm $A_{new}(.)$ is then used by the broker to obtain an estimate of $Q(D)$. Due to the uncertainty in the seller types, $A_{new}(.)$ returns the following possibilities:

$$A_{new}(D) = \begin{cases} a_H d_1 + a_H d_2 + \frac{1-a_H}{2} + \frac{1-a_H}{2} + N(b) \text{ w.p. } \pi^2 \\ a_H d_1 + a_L d_2 + \frac{1-a_H}{2} + \frac{1-a_L}{2} + N(b) \text{ w.p. } \pi(1-\pi) \\ a_L d_1 + a_H d_2 + \frac{1-a_L}{2} + \frac{1-a_H}{2} + N(b) \text{ w.p. } \pi(1-\pi) \\ a_L d_1 + a_L d_2 + \frac{1-a_L}{2} + \frac{1-a_L}{2} + N(b) \text{ w.p. } (1-\pi)^2 \end{cases}$$

and we have $\epsilon_H = (a_H/b),\ \epsilon_L = (a_L/b)$.

The goal of the data broker is to provide *expected* accuracy K. Under algorithm $A_{new}(.)$ the expected accuracy is given by: $e(a_L, a_H, b) = \pi^2 \cdot (2b^2 + (1-a_H)^2) + (1-\pi)^2 \cdot (2b^2 + (1-a_L)^2) + 2\pi \cdot (1-\pi) \cdot (2b^2 + ((1-a_H)/2 + (1-a_L)/2)^2)$.
Accordingly, the contract design problem can be written as follows:

$$\begin{array}{lll} \min\limits_{\{p_i, a_i, b\}, i \in \{H,L\}} & \pi^2 \cdot (2p_H) + (1-\pi)^2 \cdot (2p_L) + 2\pi(1-\pi) \cdot (p_H + p_L) & \\ \text{s.t.} & (IR) & p_i \geq c(v_i, a_i/b),\ i \in \{H, L\} \\ & (IC) & p_i - c(v_i, a_i/b) \geq p_j - c(v_i, a_j/b),\ i, j \in \{H, L\} \\ & (AC) & e(a_L, a_H, b) \leq K,\ i \in \{H, L\} \\ & & 0 \leq a_i \leq 1, p_i \geq 0, b > 0,\ i \in \{H, L\} \end{array} \tag{7}$$

To solve this problem we use the following lemma.

Lemma 6.1. *The following holds for the optimization problem* (7)*:*
1) Constraint $p_H \geq c(v_H, a_H/b)$ is binding.
2) Constraint $p_L \geq c(v_L, a_L/b)$ is redundant.
3) Constraint $p_L - c(v_L, a_L/b) \geq p_H - c(v_L, a_H/b)$ is binding.

It follows that the solution to (7) satisfies the followings:

$$p_H = c(v_H, \frac{a_H}{b}), \quad p_L - c(v_L, a_L/b) = p_H - c(v_L, a_H/b)$$
$$\implies p_L = c(v_H, \frac{a_H}{b}) + c(v_L, a_L/b) - c(v_L, a_H/b) \tag{8}$$
$$p_H - c(v_H, a_H/b) \geq p_L - c(v_H, a_L/b) \implies$$
$$c(v_H, \frac{a_L}{b}) - c(v_H, \frac{a_H}{b}) \geq c(v_L, \frac{a_L}{b}) - c(v_L, \frac{a_H}{b}) \tag{9}$$

Using (8) and (9) we can remove p_H and p_L from problem (7) and rewrite (7) as follows:

$$\begin{array}{lll} \min\limits_{0 \leq a_i \leq 1,\ b>0,\ \ i \in \{H,L\}} & \pi^2 \cdot (2c(v_H, a_H/b)) & \\ & +2 \cdot (1-\pi)^2 \cdot (c(v_H, a_H/b) + c(v_L, a_L/b) - c(v_L, a_H/b)) & \\ & +2\pi(1-\pi) \cdot (2c(v_H, a_H/b) + c(v_L, a_L/b) - c(v_L, a_H/b)) & \\ \text{s.t.} & c(v_H, a_L/b) - c(v_H, a_H/b) \geq c(v_L, a_L/b) - c(v_L, a_H/b) & \\ & (AC)\ \ e(a_L, a_H, b) \leq K,\ i \in \{H, L\} \end{array} \tag{10}$$

Notice that if $\frac{\partial c(v_H, \epsilon)}{\partial \epsilon} \geq \frac{\partial c(v_L, \epsilon)}{\partial \epsilon}$ (marginal cost increasing in privacy valuation), then the constraint $c(v_H, \frac{a_L}{b}) - c(v_H, \frac{a_H}{b}) \geq c(v_L, \frac{a_L}{b}) - c(v_L, \frac{a_H}{b})$ implies $a_H \leq a_L$, i.e., a seller with higher privacy attitude experiences lower privacy loss. It is also worth noting that if $a_H = 0$ is the solution to (10), then the broker offers only a single contract (instead of a menu) and a seller with lower privacy attitude accepts that.

6.2 Broker offers only one seller the menu of contracts

As in the previous case, by selecting from the menu the seller reveals his type. The broker then uses $A_{new}(.)$ according to the revealed type, which returns the following:

$$A_{new}(D) = \begin{cases} a_H d_1 + \frac{1-a_H}{2} + \frac{1}{2} + N(b_H) & \text{w.p. } \pi \\ a_L d_1 + \frac{1-a_L}{2} + \frac{1}{2} + N(b_L) & \text{w.p. } 1-\pi \end{cases},$$

and $\epsilon_H = a_H/b_H,\ \epsilon_L = a_L/b_L$.

Note in this case the noise parameter b is type-dependent since only one individual contributes to the input.

The expected accuracy in this case is given by[3]:

$$\begin{aligned} e(a_L, a_H, b_L, b_H) \quad &= \quad \pi \cdot (2b_H^2 + ((2-a_H)/2)^2) \\ &+ \quad (1-\pi) \cdot (2b_L^2 + ((2-a_L)/2)^2)\,. \end{aligned} \tag{11}$$

Accordingly, the contract design problem is given by:

$$\begin{array}{lll} \min\limits_{\{0 \leq a_i \leq 1, p_i \geq 0, b_i > 0,\ i \in \{H,L\}\}} & \pi \cdot (p_H) + (1-\pi) \cdot (p_L) & \\ \text{s.t.} & (IR)\ \ p_i \geq c(v_i, a_i/b_i),\ i \in \{H, L\} & \\ & (IC)\ \ p_i - c(v_i, a_i/b_i) \geq p_j - c(v_i, a_j/b_i),\ i, j \in \{H, L\} & \\ & (AC)\ \ e(a_L, a_H, b_L, b_H) \leq K,\ i \in \{H, L\} \end{array} \tag{12}$$

Lemma 6.1 also holds for optimization problem (12).

6.3 Broker's decision

Given problem parameters K, v_H, v_L, π, the broker compares the solutions to optimization problems (10) and (12), and chooses an option that offers the lower payment and the associated menu of contracts. These solutions are generally found numerically.

As a comparison, under algorithm $A_u(.)$ the data broker can only offer a single contract (as $A_u(.)$ requires $a_H = a_L = 1$) and the corresponding optimal contract under algorithm $A_u(.)$ is given by: $(AC):\ b^* = \sqrt{K/2},\ \epsilon^* = 1/b^*, \quad (IR):\ p^* = c(v_H, 1/b^*)$.

We next present an example to highlight the comparison: $c(v, \epsilon) = v \cdot \epsilon,\ v_H = 5,\ v_L = 1,\ \pi = 0.5$.

Figure 5 illustrates the total payment under algorithm $A_{new}(.)$ and $A_u(.)$; here $A_{new}(.)$ denotes the optimal choice between the

[3]Note that $e(a_L, a_H, b_L, b_H) \geq \frac{1}{4}$; thus if $K < \frac{1}{4}$ then this is not a viable option, and the broker should offer the menu to both sellers.

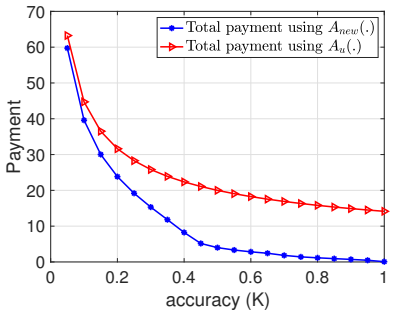

Figure 5: Total payment v.s. accuracy, under information asymmetry. The proposed method results in much lower cost.

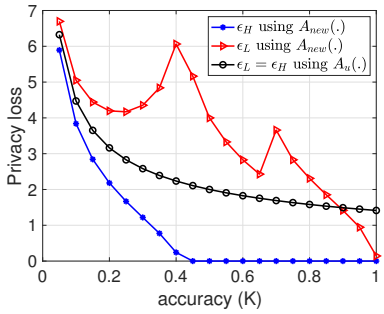

Figure 6: Privacy loss v.s. accuracy, under information asymmetry. The proposed method is able to differentiate heterogenous privacy valuations.

two versions of the algorithm presented in the previous two subsections, as determined by whether the menu is offered to both sellers (equation (10)) or only one of them (equation (12)). The results show that the payment is significantly lower by using $A_{new}(.)$. Figure 6 further illustrates the privacy loss of each seller as a function of K. As seen, the contract designed for the lower privacy type carries higher privacy loss ($\epsilon_H < \epsilon_L$) in order to decrease the total payment; by contrast, under $A_u(.)$ the broker is not able to differentiate between the two types. The two peaks in the ϵ_L curve under $A_{new}(.)$ are due to the following reasons. For $K \leq 0.4$, it is optimal for the broker to offer both sellers a menu of contracts, one for the high privacy type and one for the low privacy type, and a seller will select the right one. In the region $0.4 < K \leq 0.65$, it is optimal for the broker to offer each agent a single contract (of the low privacy type) such that a seller (if of a low type) accepts it, or (if of a high type) rejects it and walks away. This accounts for the discontinuity at $K = 0.4$. In the region $K > 0.65$, the broker offers a single contract (of the low privacy type) to only one of the sellers by random selection, and that seller accepts or rejects it depending on his type, while the other seller is not offered anything. This accounts for the discontinuity at $K = 0.65$.

7 CONCLUSION

In this study, we considered a data contract problem concerning the purchasing of private data between a single buyer and multiple sellers. We proposed a biased differentially private algorithm which allows a data broker to assign different privacy losses to different individuals depending on their privacy valuations. Using a contract design approach, we found the optimal pricing mechanism to minimize the cost of obtaining a K-accurate estimate of linear and nonlinear queries. We showed that the broker can take advantage of our proposed algorithm under both full information and information asymmetric cases, and afford lower privacy loss to individuals with higher privacy valuations. As a result, the cost to the buyer is lower and individuals experience lower privacy loss as compared to using a common unbiased algorithm.

REFERENCES

[1] *Datacoup*. http://datacoup.com/.
[2] *Online Appendix*. Available at http://bit.ly/2Fi4k9w.
[3] M. Abadi, Ú. Erlingsson, I. Goodfellow, H. B. McMahan, I. Mironov, N. Papernot, K. Talwar, and L. Zhang. 2017. On the Protection of Private Information in Machine Learning Systems: Two Recent Approches. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. 1–6.
[4] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. 2005. Practical privacy: the SuLQ framework. In *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 128–138.
[5] Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani. 2015. Accuracy for Sale: Aggregating Data with a Variance Constraint. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science (ITCS '15)*. ACM, New York, NY, USA, 317–324.
[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284.
[7] Cynthia Dwork, Aaron Roth, and others. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
[8] Lisa K. Fleischer and Yu-Han Lyu. 2012. Approximately Optimal Auctions for Selling Privacy when Costs Are Correlated with Data. In *Proceedings of the 13th ACM Conference on Electronic Commerce (EC '12)*. ACM, New York, NY, USA, 568–585.
[9] Arpita Ghosh and Aaron Roth. 2011. Selling Privacy at Auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC '11)*. ACM, New York, NY, USA, 199–208.
[10] Haiming Jin, Lu Su, Bolin Ding, Klara Nahrstedt, and Nikita Borisov. 2016. Enabling privacy-preserving incentives for mobile crowd sensing systems. In *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 344–353.
[11] Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. 2014. A Theory of Pricing Private Data. *ACM Trans. Database Syst.* 39, 4, Article 34 (Dec. 2014), 28 pages.
[12] Iman Vakilinia, Deepak K Tosh, and Shamik Sengupta. 2017. Privacy-preserving cybersecurity information exchange mechanism. In *2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*. IEEE, 1–7.
[13] Iman Vakilinia, Jiajun Xin, Ming Li, and Linke Guo. 2016. Privacy-preserving data aggregation over incomplete data for crowdsensing. In *2016 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 1–6.
[14] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. R. Liu. 2015. Privacy or Utility in Data Collection? A Contract Theoretic Approach. *IEEE Journal of Selected Topics in Signal Processing* 9, 7 (Oct 2015), 1256–1269.
[15] L. Xu, C. Jiang, Y. Qian, Y. Zhao, J. Li, and Y. Ren. 2017. Dynamic Privacy Pricing: A Multi-Armed Bandit Approach With Time-Variant Rewards. *IEEE Transactions on Information Forensics and Security* 12, 2 (Feb 2017), 271–285.
[16] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. 2018. Improving the Privacy and Accuracy of ADMM-Based Distributed Algorithms. *arXiv preprint arXiv:1806.02246* (2018).
[17] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. 2018. Recycled admm: Improve privacy and accuracy with less computation in distributed algorithms. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 959–965.