Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

Designing Contracts for Trading Private and Heterogeneous Data Using a Biased Differentially Private Algorithm

MOHAMMAD MAHDI KHALILI[1], XUERU ZHANG [2], AND MINGYAN LIU [2] (Fellow, IEEE),

[1]Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716 USA (e-mail: khalili@udel.edu)

[2]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: {xueru,mingyan}@umich.edu)

Corresponding author: Mohammad Mahdi Khalili (e-mail: khalili@udel.edu).

Mohammad Mahdi Khalili and Xueru Zhang contributed to this manuscript equally. This work is supported by the NSF under grants CNS-1616575, CNS-1646019, CNS-1739517. A preliminary version of this work [1] appeared in the 14th Workshop on the Economics of Networks, Systems and Computation (NetEcon 2019). In addition to a better exposition of our work by including proofs and technical analysis, this work extends our previous work by considering a scenario where sellers' privacy valuations are drawn from an unknown probability distribution (Section VI), nonlinear queries (Section VIII), and multidimensional data (Section IX).

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 randomized algorithm to generate differentially private output and show that this algorithm 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 conventional, unbiased algorithms, while at the same time incurring lower privacy loss for the sellers.

INDEX TERMS Contract Design, Differential Privacy, Information Asymmetry

I. INTRODUCTION

Advances in technology and data centers have enabled storing large amounts of data containing private information of individuals or firms. These data have value for institutions interested in analyzing them for a variety of purposes such as targeted advertising. Individuals are typically 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 which offers 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 purchasers, and analytics partners [2].

Studies of privacy as a commodity include arbitrage-free privacy-preserving pricing mechanisms, see e.g., [3], designing contracts for data privacy and utility [4], auctions and direct mechanisms for selling privacy [5], [6], as well as dynamic privacy pricing [7]. A more detailed literature review is given in Section II.

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

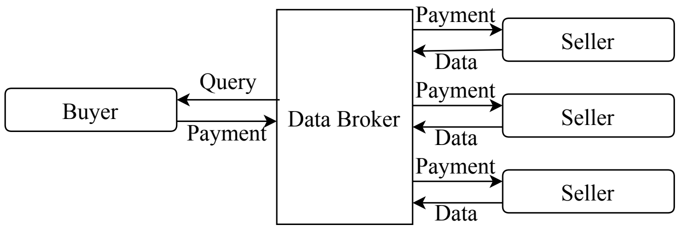


FIGURE 1: Interaction of buyer and sellers.

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 [8], 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 generated output satisfies the data minimization principle [21] imposed by a law/regulator. Under this principle, the least amount of information must be used for generating such an output. As a result of the release of the computational output, the buyer pays each individual, through the broker, an amount commensurate with the privacy loss the individual experiences. 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. The broker is assumed to be a neutral, non-profit entity in the current model, 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., [3], 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 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 [5] also introduces a biased differentially private algorithm for linear queries and one-dimensional data, but it offers only a single privacy level to the participating sellers. The present paper generalizes the algorithm introduced in [5] in the following aspects: i) our algorithm is able to afford different privacy protection/losses to different sellers, and ii) our algorithm can be extended to 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. We show that in both scenarios, the broker can leverage the proposed algorithm to guarantee a lower privacy loss for the sellers and a lower payment for the buyer.

The preliminary version of this work appeared in [1] where the proposed differentially private algorithm and contract design method were only applicable to linear queries and one-dimensional data. In addition to a better exposition of our previous work by including proofs and technical analysis in Section XI, the present paper extends our previous work in the following aspects,

- The proposed data contract under information asymmetry in [1] is only applicable to a scenario with two sellers whose privacy valuations come from a Bernoulli distribution. In the present paper, we consider a more general setting and propose a new data contract in Section VI for a scenario with n sellers whose privacy valuations are drawn from an unknown probability distribution.
- We introduce a biased differentially private algorithm for non-linear queries in Section VIII. This algorithm improves the privacy-accuracy tradeoff as compared to the unbiased algorithm, and allows data broker to assign different privacy losses to individuals when the buyer requests a non-linear query.
- We extend our biased differential private algorithm to multidimensional data in Section IX, and show that our methodology and results for one-dimensional data are equally applicable to the multi-dimensional case.

The remainder of the paper is organized as follows. We present related work in Section II and preliminaries on differential privacy and query in Section III. We introduce our randomized differentially private biased algorithm in Section IV. In Section V, we analyze the contract design problem between one buyer and multiple sellers under full information. We discuss the contract design problem for purchasing private data under information asymmetry in Section VI. We provide numerical examples in Section VII and generalize our algorithm for non-linear queries as well as multi-dimensional data in Section VIII and Section IX, respectively. Finally, Section X concludes the paper.

II. RELATED WORK

The literature in data market mechanisms to some degree parallels that in general market mechanism design, with some of them considering privacy preservation as an added element in the mechanism.

For instance, arbitrage-free mechanism is a pricing mechanism where buyers are not able to pay less for their true target by purchasing and combining substitute targets [9]. In other words, the arbitrage-free pricing mechanism does not allow the buyer to cheat the market/seller. The problem of arbitrage-free mechanisms arises in many different markets such as the energy market [10], the financial market [11]. Arbitrage-free pricing mechanisms for the data market was studied in [12], [13], [14], but privacy leakage throughout the process was not considered. Similar mechanisms were studied in [3] for linear queries, where a random noise was added to the actual query to preserve privacy and it is assumed that all individuals have the same privacy valuation.

Of the literature on data markets, the most relevant to the present paper are [4], [15], [5], [6]. In [4], 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. Gosh and Roth [5] 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 [6], where the cost of privacy loss is correlated with the private data. Cummings *et.al* [15] also design a truthful mechanism for the data aggregation problem where a buyer collects unbiased estimate of each individual's data and make a payment based on the variance of the estimate. Then, the buyer calculates the average of all unbiased estimates to find a better estimate. It is worth noting that this work is only applicable to a scenario where 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, see e.g., [16], [17], as well as in the context of security information exchange, see e.g., [18].

III. PRELIMINARIES

In this section, we review the notion of differential privacy first proposed in [8], [19] which we will use to quantify privacy leakage, and then introduce a type of linear query. Extension to any type of nonlinear query is discussed in Section VIII. 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. Extensions to higher dimensional data is discussed in Section IX. An individual incurs a cost if his privacy is violated.

A. 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 1 (ϵ-Differential Privacy [8], [19]): 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. Using Definition 1, it is easy to see that if $A(.)$ is ϵ_i-differentially private w.r.t. individual i, $i = 1, \cdots, n$, and if $D = (d_1, \cdots, d_n)$ and $D' = (d'_1, \cdots, d'_n)$ differ in more than one element, then we have

$$\frac{Pr\{A(D) \in S\}}{Pr\{A(D') \in S\}} \leq \exp\{\sum_{i \in I} \epsilon_i\}, \ \forall S,$$

where $d_j = d'_j$ if $j \notin I$ and $d_j \neq d'_j$ if $j \in I$.

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:

$$f(x) = \frac{1}{2b} \exp\{-\frac{|x|}{b}\} . \tag{1}$$

Definition 2 (Accuracy): We say algorithm $A(.)$ is K-accurate for query $Q(D)$ if

$$E\left[(A(D) - Q(D))^2\right] \leq K, \ \forall D \in X^n , \tag{2}$$

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 [19]), but the above choice does not affect the applicability of our methodology and main conclusions.

B. A TYPE OF LINEAR QUERY

Definition 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:

$$Q(D) = \sum_{i=1}^{n} q_i \cdot d_i , \tag{3}$$

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 [20]. Extension to nonlinear queries is discussed in Section VIII.

We next examine the relationship between accuracy K and privacy loss ϵ in this type of linear queries. 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 1 (Lower Bound on Total Privacy Loss): If algorithm $A(D)$ is K-accurate and $K < (\frac{n}{2})^2$,[1] then the total privacy loss is at least $\ln \frac{(n-\sqrt{K})^2}{K}$. Moreover, if $K < (\frac{m}{2})^2$, then at least $n - m + 1$ individuals experience non-zero privacy loss.

Theorem 1 implies that as $K \to 0$, privacy loss approaches infinity logarithmically. We will introduce an algorithm in Section IV under which the total privacy loss is close to the lower bound when K is close to $(\frac{n}{2})^2$.

IV. UNBIASED AND BIASED ALGORITHMS

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

Theorem 2 (An unbiased algorithm [19]): 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 $E\left[A_u(D) - Q(D)\right] = 0$. We next introduce a biased estimate $A_{new}(D)$ of $Q(D)$ such that $E\left[A_{new}(D)\right] \neq Q(D)$.

Theorem 3 (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 $0 \leq a_i \leq 1$, $\forall i$. 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.

Proof. We first derive the accuracy of $A_{new}(D)$.

$$
\begin{aligned}
(A_{new}(D) - Q(D))^2 &= \left(\sum_{i=1}^{n}((a_i - 1) \cdot d_i + \frac{1-a_i}{2})\right)^2 \\
&\quad +2\sum_{i=1}^{n}\left((a_i - 1) \cdot d_i + \frac{1-a_i}{2}\right) \cdot N(b) + N(b)^2 \\
&\leq \left(\sum_{i=1}^{n} \frac{1-a_i}{2}\right)^2 + N(b)^2 \\
&\quad +2(\sum_{i=1}^{n}(a_i - 1)d_i + \frac{1-a_i}{2})N(b) \,,
\end{aligned}
$$

where the inequality holds because $0 \leq d_i \leq 1$. Continuing,

$$
\begin{aligned}
E\left[(A_{new}(D) - Q(D))^2\right] &\leq \left(\sum_{i=1}^{n} \frac{1-a_i}{2}\right)^2 + E(N(b)^2) \\
&\quad +2\left(\sum_{i=1}^{n}(a_i - 1)d_i + \frac{1-a_i}{2}\right) E(N(b)) \\
&= \left(\sum_{i=1}^{n} \frac{1-a_i}{2}\right)^2 + 2b^2 \,.
\end{aligned}
$$

We next derive its privacy. Let $D = (d_1, d_2, \cdots, d_n)$ and $D' = (d'_1, d_2, d_3, \cdots, d_n)$ be two neighboring databases and let $s = \sum_{i=1}^{n} a_i \cdot d_i + \frac{1-a_i}{2}$ and $s' = a_1 d'_1 + \frac{1-a_1}{2} + \sum_{i=2}^{n} a_i \cdot d_i + \frac{1-a_i}{2}$. We then have

$$
\begin{aligned}
Pr\left\{A_{new}(D) \in S\right\} &= \int_{x \in S-s} \frac{1}{2b} e^{-\frac{|x|}{b}} dx \\
&= \int_{x \in S-s'} \frac{1}{2b} e^{-\frac{|x + a_1 \cdot d_1 - a_1 \cdot d'_1|}{b}} dx \\
&\leq e^{\frac{a_1 \cdot |d_1 - d'_1|}{b}} \int_{x \in S-s'} \frac{1}{2b} e^{-\frac{|x|}{b}} dx \\
&\leq e^{\frac{a_1}{b}} Pr(A_s(D') \in S) \,,
\end{aligned}
$$

where the notation $S - s := \{x - s | x \in S\}$. Therefore, $A_{new}(D)$ is $\frac{a_1}{b}$-differentially private with respect to individual 1. Similarly, we can show that $A_{new}(D)$ 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:

$$
\begin{aligned}
E\left[A_{new}(D) - Q(D)\right] &= \sum_{i=1}^{n}(a_i - 1) \cdot d_i + \frac{1-a_i}{2} \\
\implies \sum_{i=1}^{n} \frac{-1 + a_i}{2} &\leq E\left[A_{new}(D) - Q(D)\right] \leq \sum_{i=1}^{n} \frac{1-a_i}{2} \\
\implies \left|E\left[A_{new}(D) - Q(D)\right]\right| &\leq \sum_{i=1}^{n} \frac{1-a_i}{2} \,. \qquad (4)
\end{aligned}
$$

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 $a_i = 1, \forall i$.

V. PROBLEM FORMULATION

We consider a scenario with n sellers/individuals indexed by $\mathcal{N} = \{1, \ldots, n\}$ and a single buyer who is interested in obtaining a query on database $D = (d_1, d_2, \cdots, d_n)$, where data d_i belongs to seller i. Let $\boldsymbol{v} = [v_1, \ldots, v_n]$ be a vector of individuals' privacy valuations, where v_i is the *type* or the *privacy valuation* of individual i; this is also referred to as his *privacy attitude*. Individual i has cost function $c(v_i, .) : R_+ \cup \{0\} \to R_+ \cup \{0\}$. He incurs a cost of $c(v, \epsilon_i)$ if he experiences privacy loss ϵ_i. We assume that $c(v_i, \epsilon_i)$ is increasing in ϵ_i, and $c(v', \epsilon) \geq c(v, \epsilon)$ if and only if $v' \geq v$, i.e., a higher type implies higher privacy cost, and the cost of revealing data is zero if there is zero privacy loss, i.e., $c(v_i, 0) = 0$, $\forall i$. In this section we assume that the data broker knows the sellers' privacy valuations, i.e., $v_i, \forall i \in \mathcal{N}$ is common knowledge.

The buyer wishes to obtain a K-accurate estimate of $Q(D)$ with minimum payment. The data transaction between the sellers and the buyer is facilitated by a contract $(p_i, \epsilon_i, K)_{i \in \mathcal{N}}$, which stipulates that by accepting it seller i receives payment p_i and reports actual data d_i to the data broker, while the broker uses an algorithm to find an estimate

[1] Our problem is interesting if $K < (\frac{n}{2})^2$. In the next sections, we will show that If $K > (\frac{n}{2})^2$, 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 $K > (\frac{n}{2})^2$.

of $Q(D)$ which is K-accurate and ϵ_i-differentially private with respect to individual i. We assume that the individuals' data have to be used under a certain privacy principle. We will consider two such principles.

- **Principle 1** The total privacy loss experienced by the individuals has to be minimized. Moreover, individuals have to experience the same privacy loss.
- **Principle 2** The total privacy cost incurred by the sellers has to be minimized.

Based on the US Privacy Act of 1974 [21], each agency must follow data minimization principles and collect the least amount of information for its purposes. These two principles are meant to satisfy this privacy law. However, our methodology is more generally applicable. It is worth mentioning that Principle 1 not only imposes restriction on total privacy loss but also ensures that the sellers are treated equally. This is compatible with the General Data Protection Regulation (GDPR) Lawfulness, Fairness, and Transparency Principle [22].

A. OPTIMAL CONTRACT UNDER PRINCIPLE 1

Under Principle 1, the data broker has to assign the same privacy loss to each individual and minimize total privacy loss for finding a K-accurate estimate of $Q(D)$. In order to do so, the broker solves the following optimization problem to find the minimum required privacy loss using algorithm $A_{new}(D)$.

$$\begin{aligned} \min_{a,b,\epsilon} \quad & \epsilon \\ \text{s.t.} \quad (AC) \quad & \left(\frac{n}{2} - n \cdot \frac{a}{2}\right)^2 + 2b^2 = K \\ & \epsilon = \frac{a}{b},\ b > 0,\ 0 \le a \le 1 \end{aligned} \tag{5}$$

Theorem 4: The solution to optimization (5) is given by,

$$\hat{a} = \frac{n^2 - 4K}{n^2},\ \hat{b} = \sqrt{\frac{K}{2} - \frac{2K^2}{n^2}}, \tag{6}$$

$$\hat{\epsilon} = \frac{1}{n}\sqrt{(2n^2 - 8K)/K} \tag{7}$$

Earlier Theorem 1 suggests that a K-accurate estimate of $Q(D)$ has total privacy loss at least $2\ln(n - \sqrt{K}) - \ln K$. The minimum total privacy loss $\sqrt{\frac{2n^2}{K} - 8}$ under $A_{new}(.)$ approaches this lower bound as $K \to \frac{n^2}{4}$. Figure 2 compares the minimum total privacy loss using algorithms $A_u(.)$ and $A_{new}(.)$ for a scenario with $n = 10$ individuals. Clearly $A_{new}(.)$ outperforms $A_u(.)$ in terms of the 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.

Under Principle 1, contract $(p_i, \epsilon_i = \hat{\epsilon})$ is offered to individual i. To ensure individual i accepts contract $(p_i, \epsilon_i = \hat{\epsilon})$, the contract has to satisfy the Individual Rationality (IR) constraint which implies that the payment to each individual should sufficiently compensate for its privacy cost, i.e.,

$$(IR): \quad p_i \ge c(v_i, \epsilon_i)\ \forall i \in \mathcal{N}. \tag{8}$$

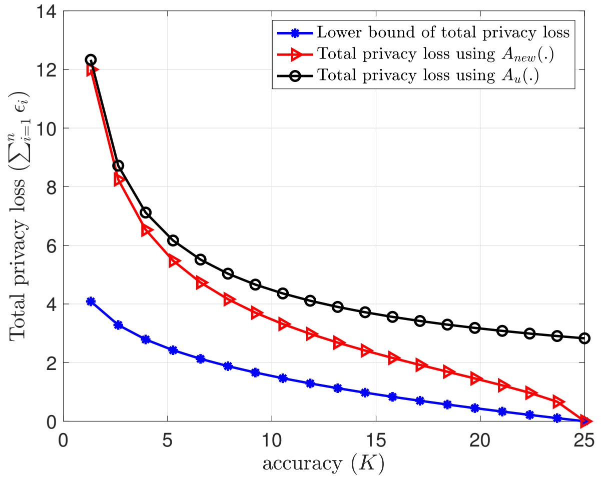


FIGURE 2: Minimum privacy loss under different algorithms.

Since v_i is common information, $p_i = c(v_i, \hat{\epsilon})$ would minimize the total payment made by the buyer. Therefore, *optimal contract* $\{\hat{p}_i, \hat{\epsilon}_i\}_{i \in \mathcal{N}}$, which implements principle 1, is given by,

$$\hat{p}_i = c(v_i, \hat{\epsilon}),\ \hat{\epsilon}_i = \hat{\epsilon},\ \forall i \in \mathcal{N}. \tag{9}$$

B. OPTIMAL CONTRACT UNDER PRINCIPLE 2

Under Principle 2, the total privacy cost incurred by the individuals must be minimized. In this case, the privacy loss assigned to each individual using algorithm $A_{new}(D)$ can be obtained by the following optimization problem,

$$\begin{aligned} \min_{\{a_i, b, \epsilon_i\}} \quad & \sum_{i=1}^{n} c(v_i, \epsilon_i) \\ \text{s.t.} \quad (AC) \quad & \left(\sum_{j=1}^{n} \frac{1 - a_j}{2}\right)^2 + 2b^2 = K \\ & 0 \le a_i \le 1,\ \epsilon_i = \frac{a_i}{b},\ b > 0,\ i \in \{1, \dots, n\}. \end{aligned} \tag{10}$$

A closed form solution to optimization problem (10) is not easy to find in general and depends on the form of the cost function. Below we solve (10) under a linear cost model.

Theorem 5: Let $c(v, \epsilon) = v \cdot \epsilon$, $K < (\frac{n}{2})^2$, $v_1 \le v_2 \le \ldots \le v_n$ and $s_{i+1} = (n - i) - 4 \cdot K \cdot \frac{v_{i+1}}{(n-i) \cdot v_{i+1} + \sum_{j \le i} v_j}, \forall i \ge \left\lceil n - 2\sqrt{K} \right\rceil$, $i \le n - 1$, where $[x]$ is the largest integer less than or equal to x. 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 (10) is given by:

$$
\begin{aligned}
a_1^* &= a_2^* = \ldots = a_{m-1}^* = 1,\ a_m^* = \min\{s_m, 1\}, \\
a_{m+1}^* &= a_{m+2}^* = \ldots = a_n^* = 0, \\
b^* &= \sqrt{\frac{1}{2}(K - (\frac{2K \cdot v_m}{(n-m+1)\cdot v_m + \sum_{j=1}^{m-1} v_j})^2)} \\
\epsilon_i^* &= \frac{a_i^*}{b^*}. \qquad (11)
\end{aligned}
$$

Proof. See Appendix. ■

Note that if $K > (\frac{n}{2})^2$, then $a_1 = a_2 = \cdots = a_n = 0$ and $b = \sqrt{\frac{K-(n/2)^2}{2}}$ give a feasible solution to (10). This point is optimal because its objective value is zero. Thus, if $K > (\frac{n}{2})^2$, the output of algorithm $A_{new}(D)$ will be a pure noise. In addition to a linear cost model, a closed form solution to optimization problem (10) can be calculated using Algorithm 1 if the cost function has the following form: $c(v, \epsilon) = v\cdot(\epsilon)^r$, where $r > 1$ is a constant.

Theorem 6: Let $c(v, \epsilon) = v \cdot (\epsilon)^r$ and $r > 1$. Then, Algorithm 1 finds the optimal solution to optimization problem (10).

For notational convenience, let $h_i(\boldsymbol{v})$ be the privacy loss of individual i obtained from optimization problem (10), and $h(\boldsymbol{v}) = [h_1(\boldsymbol{v}), \ldots, h_n(\boldsymbol{v})]$.

Under Principle 2, the broker offers contract $(p_i, \epsilon_i = h_i(\boldsymbol{v}))$. The contract has to satisfy the (IR) constraint defined in (8). Under full information, $p_i = c(v_i, h_i(\boldsymbol{v}))$ satisfies the (IR) constraint and minimizes the total payment. Therefore, under Principle 2 and algorithm $A_{new}(D)$, the optimal contract is given by,

$$p_i^* = c(v_i, h_i(\boldsymbol{v})),\ \epsilon_i^* = h_i(\boldsymbol{v}), \forall i \in \mathcal{N}. \qquad (12)$$

C. COMPARISON OF THE OPTIMAL CONTRACT UNDER ALGORITHM $A_{new}(D)$ AND $A_u(D)$

So far, we have used biased algorithm $A_{new}(D)$ to find the optimal contract. In this section, we study the contract design problem using $A_u(D)$ and compare it with the contract design problem using $A_{new}(D)$.

As we mentioned in Section IV, $A_u(D) = Q(D) + N(b)$ has only one degree of freedom and is K-accurate if $b = \sqrt{K/2}$. Therefore, the optimal contract which minimizes the total payment using $A_u(D)$ is given by,

$$\overline{p}_i = c(v_i, \sqrt{2/K}),\ \overline{\epsilon}_i = \sqrt{2/K} \qquad (13)$$

We make two observation here. First, the individuals experience privacy loss $\overline{\epsilon}_i = \sqrt{2/K}$ under algorithm $A_u(D)$, while their privacy loss is $\hat{\epsilon}_i = \frac{1}{n}\sqrt{(2n^2 - 8K)/K}$ under Principle 1 and algorithm $A_{new}(D)$. Therefore, $A_{new}(D)$ under Principle 1 is able to decrease the total privacy leakage as compared to $A_u(D)$. Second, $A_{new}(D)$ under Principle 2 decreases the total privacy cost as compared to $A_u(D)$, because $A_{new}(D)$ is able to assign lower privacy loss to those who have higher privacy valuation. That is, $\sum_{i=1}^n c(v_i, h_i(\boldsymbol{v})) \leq \sum_{i=1}^n c(v_i, \sqrt{\frac{2}{K}})$.

As we mentioned in this section, the (IR) constraint is always binding under the full information assumption, and

Algorithm 1: Solution to optimization problem (10)

1: input: $\boldsymbol{v}, r, K$;
2: initialization: $a_i^* = 1, \forall i \in \mathcal{N}$, $b^* = \sqrt{K/2}$;
3: $Cost_{\min} = \sum_{k=1}^n v_i \cdot (\sqrt{2/K})^r$;
4: Sort $\boldsymbol{v}$ such that $v_1 \leq v_2 \leq \ldots \leq v_n$;
5: **for** $i = \left\lceil n - 2\sqrt{K} \right\rceil + 1, \ldots, n$ **do**
6: **for** $j = 0, \ldots, i$ **do**
7: **if** $j > 0$ **then**
8: $a_k = 1, \forall k = 1, \ldots, j.$
9: **end if**
10: **if** $i < n$ **then**
11: $a_k = 0, \forall k = i+1, \ldots, n.$
12: **end if**
13: **if** $j < i$ **then**
14: $A = \sum_{k=j+1}^{i} \sqrt[(r-1)]{v_{j+1}/v_k}$;
15: $a = \frac{(n-j)^2 - 4K}{A\cdot(n-j)}$;
16: **if** $0 \leq a \leq 1$ **then**
17: $a_k = \sqrt[r-1]{(\frac{v_{j+1}}{v_k})} \cdot a,\ k = j+1, \ldots, i$
18: **else**
19: go to line 5;
20: **end if**
21: **end if**
22: $b = \sqrt{0.5 \times (K - (\frac{\sum_{k=1}^n 1 - a_k}{2})^2)}$;
23: $C = \sum_{k=1}^n v_k \cdot (\frac{a_k}{b})^r$;
24: **if** $C < Cost_{\min}$ **then**
25: $a_k^* = a_k, \forall k \in \mathcal{N}$;
26: $b^* = b$;
27: $Cost_{\min} = C$;
28: **end if**
29: **end for**
30: **end for**
31: $\epsilon_i^* = \frac{a_i^*}{b^*}, \forall i \in \mathcal{N}$;
32: Output: $\{\epsilon_i^*\}_{i\in\mathcal{N}}$

the payment to each individual is equal to its privacy cost. In the next section, we study the contract design problem under information asymmetry where privacy valuation v_i is only known to individual i.

VI. CONTRACT DESIGN UNDER INFORMATION ASYMMETRY

We will now turn to scenarios where the privacy attitude of each seller is its own private information and remains unknown to the buyer, the broker, and the other sellers. The goal of this section is to design a mechanism to incentivize the sellers to report their actual privacy valuations as well as their data to the broker. In order to make the mechanism design problem tractable, we make the following assumption.

Assumption 1: $c(v, \epsilon) = v \cdot l(\epsilon)$, where $l(.)$ is an increasing function. Moreover, $v_i \in [0, \overline{v}]$, $\forall i \in \mathcal{N}$, where $\overline{v}$ is a positive constant.

Next we design two mechanisms that comply with Principle 1 and Principle 2, respectively, under incomplete information.

A. MECHANISM UNDER PRINCIPLE 1

Under Principle 1, the broker would like to assign privacy loss $\hat{\epsilon}$ to individual i ($\hat{\epsilon}$ has been defined in (6)). However, v_i is not known to the broker, and he cannot determine the sufficient amount of compensation for the privacy cost incurred by individual i. In order to overcome the issue of information asymmetry, the broker ask individuals to report their privacy attitudes and induces a one-shot game among the sellers by announcing mechanism $M_1 = \{t(\hat{\boldsymbol{v}}) = [t_1(\hat{\boldsymbol{v}}), \ldots, t_n(\hat{\boldsymbol{v}})],\ g(\hat{\boldsymbol{v}}) = [g_1(\hat{\boldsymbol{v}}), \ldots, g_n(\hat{\boldsymbol{v}})]\}$, where $\hat{v}_i$ is the reported privacy attitude by individual i, $\hat{\boldsymbol{v}} = [\hat{v}_1, \ldots, \hat{v}_n]$ is a vector of reported privacy attitudes, $g_i(\hat{\boldsymbol{v}}) = \hat{\epsilon}$ is the privacy loss of individuals which complies with Principle 1, and $t_i(\hat{\boldsymbol{v}})$ is the payment to individual i as a function of $\hat{\boldsymbol{v}}$.[2] After announcing function $g(.)$ and $t(.)$, the individuals report their privacy attitudes and receive the payment based on $t(\hat{\boldsymbol{v}})$. Lastly, individual i experiences privacy loss $\hat{\epsilon}_i$.

Let $G_1 = \{\mathcal{N}, \{u_i(\hat{\boldsymbol{v}}|v_i)\}_{i\in\mathcal{N}}, A = [0, \overline{v}]^n\}$ be the game induced by mechanism M_1 where $u_i(\hat{\boldsymbol{v}}|v_i) = +t_i(\hat{\boldsymbol{v}}) - c(v_i, \hat{\epsilon})$ is the utility of individual i, $v_i \in [0, \overline{v}]$ and $\hat{v}_i \in [0, \overline{v}]$ are his true privacy valuation and his action/reported privacy attitude, respectively, and $A = [0, \overline{v}]^n$ is the action space.

We use Nash Equilibrium (NE) as the solution concept for game G_1. We say the strategy profile $\boldsymbol{v}^*$ is the NE of game G_1 if we have,

$$u_i(v_i^*, v_{-i}^*|v_i) \geq u_i(\hat{v}_i, v_{-i}^*|v_i),\ \forall \hat{v}_i \in [0, +\infty),\ \forall i \in \mathcal{N}$$

where v_{-i}^* denotes the strategy profile of the sellers excluding individual i at the NE.

In order to comply with Principle 1, the NE of game G_1 must satisfy the two following conditions,

$$\begin{aligned} (IC) \qquad & \boldsymbol{v}^* = \boldsymbol{v}, \\ (IR) \qquad & u_i(\boldsymbol{v}|v_i) \geq 0, \end{aligned} \tag{14}$$

where the Incentive Compatibility (IC) condition implies that the individuals report their privacy valuations truthfully at the NE, and the Individual Rationality (IR) ensures that the individuals obtain higher utility as compared to their outside option (i.e., not selling the data).

The final goal of the broker is to find a mechanism to implement Principle 1 (i.e., $g(\boldsymbol{v})$) with a minimum payment subject to the IR and IC constraints. The next theorem introduces such a mechanism.

Theorem 7: Under Assumption 1, mechanism M_1 satisfies (IR) and (IC) constraints and minimizes the payment if and only if,

$$t_i(\hat{\boldsymbol{v}}) \quad = \quad \overline{v} \cdot l(\hat{\epsilon}) = \overline{v} \cdot l(\hat{\epsilon}) = \frac{\overline{v}}{n}\sqrt{(2n^2 - 8K)/K}, \forall i \in \mathcal{N}.$$

[2]Under Principle 1, the individuals' privacy loss does not depend on $\hat{\boldsymbol{v}}$. However, we will see individuals' privacy loss should be a function of reported privacy valuations $\hat{\boldsymbol{v}}$ under Principle 2.

Theorem 7 implies that the payment to each individual does not depend on reported privacy attitudes. This is because individuals' privacy loss under Principle 1 (i.e., $g(\hat{\boldsymbol{v}})$) does not depend on $\hat{\boldsymbol{v}}$. It is worth mentioning that the payment to the individual i under information asymmetry at NE of game G_1 (i.e., $t_i(\boldsymbol{v})$) is always larger than the payment under full information (i.e., $\hat{p}_i$) because under information asymmetry the broker has to be conservative and offers a higher payment to guarantee the sellers' participation.

B. MECHANISM UNDER PRINCIPLE 2

Under Principle 2, the broker designs a mechanism to incentivize seller i to share his data with the while he experiences privacy loss $h_i(\boldsymbol{v})$. Let $M_2 = \{\tau(\hat{\boldsymbol{v}}) = [\tau_1(\hat{\boldsymbol{v}}), \ldots, \tau_n(\hat{\boldsymbol{v}})],\ h(\hat{\boldsymbol{v}}) = [h_1(\hat{\boldsymbol{v}}), \ldots, h_n(\hat{\boldsymbol{v}})]\}$ be such a mechanism implementing Principle 2 with a minimum payment, where $\tau_i(\hat{\boldsymbol{v}})$ is the payment to individual i and $h_i(\hat{\boldsymbol{v}})$ is the privacy loss experienced by individual i as a function of reported privacy valuations $\hat{\boldsymbol{v}}$ ($h_i(\hat{\boldsymbol{v}})$ is calculated by solving (10)). Similar to M_1, M_2 induces game $G_2 = \{\mathcal{N}, \{w_i(\hat{\boldsymbol{v}}|v_i)\}_{i\in\mathcal{N}}, A = [0, \overline{v}]^n\}$ among the sellers, where $w_i(\hat{\boldsymbol{v}}|v_i) = +\tau_i(\hat{\boldsymbol{v}}) - c(v_i, h_i(\hat{\boldsymbol{v}}))$ is the utility of seller i inside mechanism M_2. The next theorem identifies payment function $\tau(.)$ such that the NE of game G_2 satisfies IC and IR constraints.

Theorem 8: Under Assumption 1, mechanism M_2 satisfies the (IR) and (IC) constraints and implement Principle 2 with a minimum payment if and only if,

$$\tau_i(\hat{\boldsymbol{v}}) \quad = \quad \int_{\hat{v}_i}^{\overline{v}} l(h_i(s_i, \hat{v}_{-i}))ds_i + \hat{v}_i \cdot l(h_i(\hat{\boldsymbol{v}})) \tag{15}$$

Note that both privacy profiles $g(\boldsymbol{v})$ and $h(\boldsymbol{v})$ used in mechanisms M_1 and M_2 are calculated using algorithm $A_{new}(D)$. In the next part, we study the mechanism design problem under algorithm $A_u(D)$.

C. MECHANISM USING ALGORITHM $A_U(D)$

Algorithm $A_u(D) = Q(D) + N(\sqrt{K/2})$ is K-accurate, and all the individuals experience privacy loss $\sqrt{\frac{2}{K}}$. Let $\boldsymbol{e} = [\sqrt{\frac{2}{K}}, \ldots, \sqrt{\frac{2}{K}}]$ be a vector with length n which denotes the sellers' privacy loss under $A_u(D)$. Let $M_3 = \{\rho(\hat{\boldsymbol{v}}) = [\rho_1(\hat{\boldsymbol{v}}), \ldots, \rho_n(\hat{\boldsymbol{v}})],\ \boldsymbol{e}\}$, where $\rho_i(\hat{\boldsymbol{v}})$ is the payment to individual i. The goal of mechanim M_3 is to incentive the seller to report their privacy attitude truthfully. Next theorem shows that payment profile $\rho(\hat{\boldsymbol{v}})$ would be a constant function.

Theorem 9: Mechanism M_3 satisfies the IR and IC constraint with a minimum payment if and only if,

$$\rho_i(\hat{\boldsymbol{v}}) = \overline{v} \cdot l(\sqrt{\frac{2}{K}}), \forall i \in \mathcal{N} \tag{16}$$

Theorem 9 implies that under algorithm $A_u(D)$, the payment to the individual does not depend on the reported privacy valuation and is a constant. Note that $t_i(\hat{\boldsymbol{v}}) \leq$

$\rho_i(\hat{\boldsymbol{v}}), \forall i \in \mathcal{N}$. Therefore, the total payment under mechanism M_1 is less than that under mechanism M_3.

In the next section, we perform a numerical experiment to compare the proposed mechanisms M_1, M_2, M_3.

VII. NUMERICAL EXAMPLE

A. CONTRACT DESIGN UNDER FULL INFORMATION

Consider a case of two sellers and linear cost under. Let $v_1 = 1,\ v_2 = 2$ and $K = \frac{1}{4}$. By Theorem 4, the optimal contract under Principle 1 and full information is given by,

$$\begin{aligned}
\hat{a} &= \frac{3}{4},\ \hat{b} = \frac{1}{4}\sqrt{1.5},\ \hat{\epsilon} = \sqrt{6} \\
\hat{\epsilon}_1 &= \hat{\epsilon}_2 = \hat{\epsilon} = \sqrt{6} \\
\hat{p}_1 &= v_1 \cdot \hat{\epsilon}_1 = \sqrt{6} \\
\hat{p}_2 &= v_2 \cdot \hat{\epsilon}_2 = 2\sqrt{6}
\end{aligned} \tag{17}$$

By Theorem 5, under Principle 2, the solution to (10) is given by,

$$\begin{aligned}
s_1 &= \frac{3}{2},\ s_2 = \frac{1}{3} \\
a_1^* &= 1,\ a_2^* = \frac{1}{3},\ b^* = \frac{1}{6}\sqrt{2.5} \\
\epsilon_1^* &= \frac{6}{\sqrt{2.5}},\ \epsilon_2^* = \frac{2}{\sqrt{2.5}}
\end{aligned} \tag{18}$$

Therefore, the optimal contract under Principle 2 is given by,

$$\begin{aligned}
p_1^* &= v_1 \cdot \epsilon_1^* = \frac{6}{\sqrt{2.5}} \\
p_2^* &= v_2 \cdot \epsilon_2^* = \frac{4}{\sqrt{2.5}}
\end{aligned} \tag{19}$$

The optimal contract under algorithm $A_u(.)$ is given by,

$$\begin{aligned}
\bar{b} &= \sqrt{\frac{K}{2}} = \frac{\sqrt{2}}{4} \\
\bar{\epsilon}_1 &= \bar{\epsilon}_2 = \frac{1}{\bar{b}} = 2\sqrt{2},\ \bar{p}_1 = 2\sqrt{2},\ \bar{p}_2 = 4\sqrt{2}
\end{aligned} \tag{20}$$

This example helps highlight the two reasons why $A_{new}(.)$ outperforms $A_u(.)$:

- Using $A_{new}(D)$ and under Principle 1, both sellers experience the same privacy loss. We can observe that the broker assigns the same privacy loss to the sellers under $A_u(D)$ as well. However, $A_{new}(D)$ has more degree of freedom than $A_u(D)$ and is able to decrease the privacy loss as compared to $A_u(D)$.
- Under $A_{new}(.)$ and Principle 2, the broker is able to assign different privacy losses to the two individuals. To minimize total cost, an individual with a higher privacy valuation is afforded lower privacy loss in the optimal contract.
- 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, $A_{new}(.)$ improves privacy-accuracy tradeoff.

B. CONTRACT DESIGN UNDER INFORMATION ASYMMETRY

Consider a case of $n = 10$ sellers with cost function $c(v_i, \epsilon_i) = v_i \cdot (\epsilon_i)^2$. Privacy attitude v_i is the individual i's private information. The only information available to the broker is $\overline{v}$. In other words, he knows $v_i \leq \overline{v}, \forall i \in \mathcal{N}$. In this part, we compare the expected total payment under proposed mechanisms.

- Scenario 1: We assume that $\overline{v} = 10$, and $v_1, \ldots, v_n$ are drawn independently and uniformly from interval $[0, 10]$. Under these assumptions, we calculate the expected payment under mechanism M_1, M_2, and M_3. Note that the distribution of an individual's privacy attitude is not available to the broker, and we only use it to calculate the expected payment.
  Figure 3 illustrates the expected total payment as a function of K. First, we observe that the total payment is decreasing as a function K. Second, mechanism M_2 achieves the lowest expected payment as compared to mechanism M_1 and M_3. This observation implies that $A_{new}(D)$ under Principle 1 outperforms $A_u(D)$ in terms of expected total payment.
- Scenario 2: In this scenario, $\overline{v} = 1$, and $v_1, \ldots, v_n$ are i.i.d. random variables and distributed uniformly over interval $[0, 1]$. In this example, mechanism M_2 achieves the lowest expected total payment. Moreover, we observe that the payment under algorithm $A_u(.)$ (mechanism M_3) is lower than that under mechanism M_2. This observation can be justified as follows. Under mechanism M_3, the broker offers the same contract to all the individuals and does not differentiate between them. In particular, $\rho_i(\boldsymbol{v}) = \overline{v} \cdot (\frac{2}{K})^{\frac{r}{2}},\ \forall i \in \mathcal{N}$, and the total payment is independent of the individuals' privacy valuations. In this scenario, since the variance of privacy valuation v_i is much smaller than that in the previous example, it may not be beneficial for the buyer to differentiate between the sellers.

VIII. NON-LINEAR QUERIES

Various machine learning applications such as kernel methods [23] require non-linear queries. In this section, we discuss how the proposed algorithm $A_{new}(.)$ can be generalized for any non-linear queries.

Let $f_i(.) : X^n \rightarrow [0, 1]$ be a non-linear function and $\Delta_i^j = \max_{\{D, D^{(j)}\}} |f_i(D) - f_i(D^{(j)})|$, and $\mathcal{I}_j = \{i | \Delta_i^j \neq 0\}$. Suppose the buyer's goal is to obtain $Q(D) = \sum_{i=1}^{q} f_i(D)$ where q is a constant positive integer. We generalize $A_{new}(.)$ as follows:

$$A_{new}(D) = \sum_{i=1}^{q} a_i f_i(D) + \frac{1 - a_i}{2} + N(b)\ . \tag{21}$$

We then have the following theorem on the accuracy and privacy of $A_{new}(D)$:

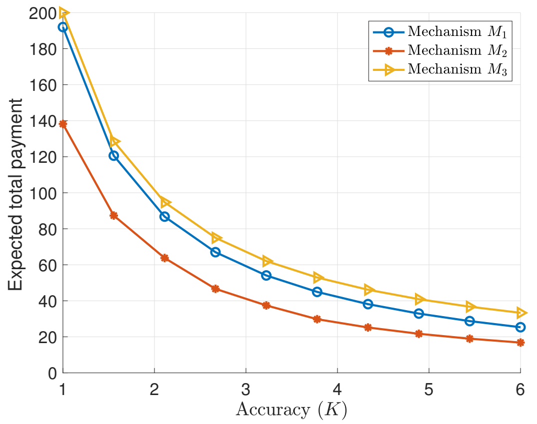


FIGURE 3: Expected payment under different mechanisms when $\overline{v} = 10$. In this scenario, $A_{new}(.)$ under Principle 1 always outperforms $A_u(.)$.

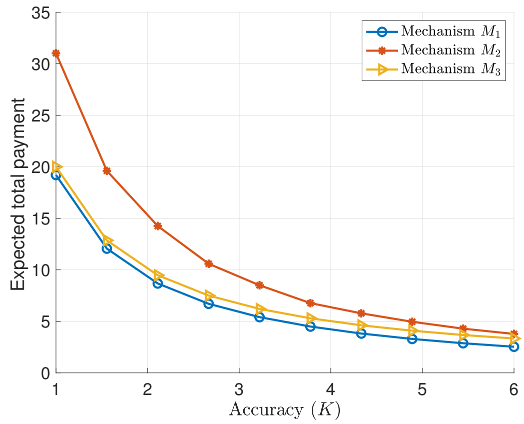


FIGURE 4: Expected payment under different mechanisms when $\overline{v} = 1$. In this scenario, $A_{new}(.)$ under Principle 1 and Principle 2 leads to a higher payment as compared to $A_u(.)$.

Theorem 10: Algorithm $A_{new}(D)$ is $[2b^2 + (\sum_{i=1}^{q} \frac{1-a_i}{2})^2]$-accurate. Moreover, it is $[\sum_{i \in \mathcal{I}_j} \frac{a_i \Delta_i^j}{b}]$-differentially private with respect to individual j.

Proof. See Appendix. ■

Consider Principle 1.[3] The following optimization problem finds the minimum total privacy loss under algorithm $A_{new}(D)$,

$$\begin{aligned} \min_{a_1,\cdots,a_q,b} \quad & \sum_{i=1}^{q} (\sum_{j=1}^{n} \Delta_i^j) \cdot \frac{a_i}{b} \\ \text{s.t.,} \quad & (2b^2 + (\sum_{i=1}^{q} \frac{1-a_i}{2})^2) = K \\ & b > 0, 1 \geq a_i \geq 0,\ i = 1, 2 \cdots, q \end{aligned} \tag{22}$$

It is worth noting that a closed-form solution to optimization problem (22) can be calculated using Theorem 5. Moreover, finding optimal parameters for algorithm $A_{new}(D)$ under Principle 2 can be written as follows,

$$\begin{aligned} \min_{a_1,\cdots,a_q,b} \quad & \sum_{j=1}^{n} c(v_j, \sum_{i=1}^{q} \Delta_i^j \cdot \frac{a_i}{b}) \\ \text{s.t.,} \quad & (2b^2 + (\sum_{i=1}^{q} \frac{1-a_i}{2})^2) = K \\ & b > 0, 1 \geq a_i \geq 0,\ i = 1, 2 \cdots, q \end{aligned} \tag{23}$$

It is worth mentioning that if the cost functions are linear, the optimization problem can simplified as follows and can be solved using Theorem 5.

$$\min_{a_1,\cdots,a_q,b} \quad \sum_{i=1}^{q} (\sum_{j=1}^{n} v_j \Delta_i^j) \cdot \frac{a_i}{b} \tag{24}$$

$$\begin{aligned} \text{s.t.,} \quad & (2b^2 + (\sum_{i=1}^{q} \frac{1-a_i}{2})^2) = K \\ & b > 0, 1 \geq a_i \geq 0,\ i = 1, 2 \cdots, q \end{aligned} \tag{25}$$

After solving optimization problems (22) and (23) and finding optimal privacy loss for each individual, we can use the same contract design approach provided in Sections V and VI to find the optimal contract. Therefore, proposed biased algorithm $A_{new}(D)$ and mechanism design techniques provided for linear queries remain valid for non-linear queries.

Example 1: Consider the following nonlinear query:

$$\begin{aligned} Q(D) &= f_1(D) + f_2(D) = \frac{d_1}{1+d_2^2} + \frac{1}{d_1^2+1}, \\ d_1 &\in [0,1],\ d_2 \in [0,1], \\ \Delta_1^1 &= 1,\ \Delta_1^2 = \frac{1}{2},\ \Delta_2^1 = \frac{1}{2},\ \Delta_2^2 = 0 \end{aligned} \tag{26}$$

By Theorem 10, the privacy loss and accuracy under $A_{new}(D) = a_1 \cdot \frac{d_1}{1+d_2^2} + a_2 \cdot \frac{1}{1+d_1^2} + \frac{1-a_1}{2} + \frac{1-a_2}{2} + N(b)$ are:

$$\begin{aligned} \epsilon_1 &= \frac{a_1 + \frac{1}{2} \cdot a_2}{b},\ \epsilon_2 = \frac{\frac{1}{2} \cdot a_1}{b}, \\ \text{accuracy} &= (\frac{1-a_1+1-a_2}{2})^2 + 2b^2\,. \end{aligned} \tag{27}$$

[3]In this section we do not try to assign the same privacy loss to the individuals because it is not always possible.

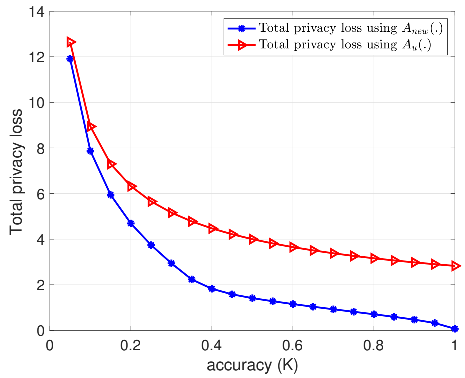


FIGURE 5: Non-linear query: privacy loss under different algorithms.

The optimal values for a_1, a_2, b under Principle 1 are obtained by the following optimization problem:

$$\begin{aligned} \min_{a_1,a_2,b} \quad & \frac{3}{2}\cdot\frac{a_1}{b}+\frac{1}{2}\cdot\frac{a_2}{b} \\ \text{s.t.} \quad & (\frac{1-a_1+1-a_2}{2})^2+2b^2=K, \\ & 0\le a_1\le 1, 0\le a_2\le 1, b>0 \end{aligned} \quad (28)$$

By comparison, using $A_u(D) = Q(D) + N(b)$, the accuracy is $2b^2$ and $\epsilon_1 = \frac{3}{2}\frac{1}{b}$ and $\epsilon_2 = \frac{1}{2}\frac{1}{b}$. In order to achieve accuracy K using $A_u(.)$, $b = \sqrt{\frac{K}{2}}$ and the total privacy loss is $\epsilon_1 + \epsilon_2 = 2\sqrt{\frac{2}{K}}$. Figure 5 shows that the minimum total privacy loss using $A_{new}(.)$ is (much) lower than that under $A_u(.)$ for this nonlinear query.

IX. MULTI-DIMENSIONAL DATA

We now discuss the extension to multi-dimensional databases. Let $D = (\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n)$ where $\boldsymbol{d}_i = [d_i^1, d_i^2, \cdots, d_i^m]^T \in [0,1]^m$. Similarly, we define neighboring database $D^{(i)} = (\boldsymbol{d}_1^{(i)}, \boldsymbol{d}_2^{(i)}, \cdots, \boldsymbol{d}_n^{(i)})$ such that $\boldsymbol{d}_j = \boldsymbol{d}_j^{(i)}$ for all $j \neq i$ and $\boldsymbol{d}_i \neq \boldsymbol{d}_i^{(i)}$. We say a randomized algorithm $\boldsymbol{A}(D)$ is ϵ_i-differentially private with respect to individual i if for any possible output S and for any D and $D^{(i)}$ we have [19]:

$$\frac{Pr(\boldsymbol{A}(D)\in S)}{Pr(\boldsymbol{A}(D^{(i)})\in S)} \le \exp\{\epsilon_i\}\ . \quad (29)$$

Consider linear query $Q(D) = \sum_{i=1}^n \boldsymbol{d}_i$ and noise vector $\boldsymbol{N}(b) = \left[N_1(b), N_2(b), \cdots, N_m(b)\right]^T$ where $N_i(b)$ and $N_j(b)$ are two independent Laplacian noise random variables with parameter b. Similar to Section III, we define randomized algorithms $\boldsymbol{A}_u(D)$ and $\boldsymbol{A}_{new}(D)$ as follows:

$$\begin{aligned} \boldsymbol{A}_u(D) &= Q(D)+\boldsymbol{N}(b) \\ \boldsymbol{A}_{new}(D) &= \sum_{i=1}^n a_i\boldsymbol{d}_i + \frac{(1-a_i)}{2}\cdot\boldsymbol{1}+\boldsymbol{N}(b), \end{aligned} \quad (30)$$

where, $\boldsymbol{1}$ is an all-1 vector. We have the following theorem on the privacy of algorithm $\boldsymbol{A}_u(.)$ and $\boldsymbol{A}_{new}(.)$.

Theorem 11: $\boldsymbol{A}_u(.)$ is $m\cdot\frac{1}{b}$-differentially private with respect to individual i, while $\boldsymbol{A}_{new}(.)$ is $m\cdot\frac{a_i}{b}$-differentially private with respect to individual i.

Proof. See Appendix. ■

Definition 4 (Accuracy): We say algorithm $\boldsymbol{A}(D)$ is K-accurate if $\frac{1}{m}E(||\boldsymbol{A}(D)-Q(D)||_2^2) \le K$ for all possible database D.

Note that this definition reduces to Definition 2 when $m = 1$. Using this definition, we are able to find the accuracy of algorithm $\boldsymbol{A}_u(D)$ and $\boldsymbol{A}_{new}(D)$ as follows.

Theorem 12: Algorithms $\boldsymbol{A}_u(D)$ is $2b^2$-accurate; $\boldsymbol{A}_{new}(D)$ is $\left[(\sum_{i=1}^n \frac{1-a_i}{2})^2+2b^2\right]$-accurate.

Proof. See Appendix. ■

Theorem 11 and 12 together imply that the contract design problem with multi-dimensional database is exactly the same as the problem for single-dimensional database. Therefore, the proposed algorithm $A_{new}(.)$ and results presented earlier are equally applicable to the multi-dimensional case.

X. 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 provides more degree of freedom in contract design problem as compared to the traditional unbiased differentially private algorithm. We showed that the broker can take advantage of our proposed algorithm under both full information and information asymmetric cases, and offer lower privacy loss to individuals and decrease the cost to the buyer as compared to using a common unbiased algorithm.

Lastly, we showed that the proposed differentially private algorithm and contract design techniques are applicable to non-linear queries as well as multi-dimensional databases.

XI. APPENDIX

Proof. [Theorem 1] Let $S_\delta(x) = [x, x+\delta]$ be an arbitrary set, and $f_{A(D)}(s)$ be the pdf of algorithm $A(D)$. Moreover, Let $D = (1, 1, \cdots, 1)$ be a database of all ones, and $D' = (0, 0, \cdots, 0)$ be a database of all zeros. By the definition of differential privacy, we have,

$$\begin{aligned}
\frac{Pr(A(D) \in S_\delta(x))}{Pr(A(D') \in S_\delta(x))} &\leq \exp\{\epsilon_1 + \epsilon_2 + \dots + \epsilon_n\} \\
\lim_{\delta\to 0} \frac{Pr(A(D) \in S_\delta(x))}{Pr(A(D') \in S_\delta(x))} &= \lim_{\delta\to 0} \frac{\delta \cdot f_{A(D)}(x)}{\delta \cdot f_{A(D')}(x)} = \frac{f_{A(D)}(x)}{f_{A(D')}(x)} \\
\frac{f_{A(D)}(x)}{f_{A(D')}(x)} &\leq \exp\{\sum_{i=1}^{n} \epsilon_i\} \forall\, x \in R \implies \\
E(A(D)^2) &= \int s^2 f_{A(D)}(s) ds \\
&\leq \int \exp\{\sum_{i=1}^{n} \epsilon_i\} s^2 f_{A(D')}(s) ds \\
&= \exp\{\sum_{i=1}^{n} \epsilon_i\} E(A(D')^2) \implies \\
\frac{E(A(D)^2)}{E(A(D')^2)} &\leq \exp\{\sum_{i=1}^{n} \epsilon_i\} \qquad (31)
\end{aligned}$$

By the definition of accuracy, and inequality $E(X)^2 \leq E(X^2)$ for random variable X, we have,

$$\begin{aligned}
&Q(D') = 0 \to \\
&\qquad E((A(D') - Q(D'))^2) = E(A(D')^2) \leq K(*) \\
&Q(D) = n \to \\
&\qquad E((A(D) - Q(D))^2) = E((A(D) - n)^2) \leq K \\
&E(n - A(D)) \leq \sqrt{K} \to E(A(D)) \geq n - \sqrt{K} \underbrace{\geq}_{\text{as } K \leq (n/2)^2} 0 \\
&E(A(D)^2) \geq E(A(D))^2 \geq (n - \sqrt{K})^2 (**)
\end{aligned}$$

Using (*) and (**), we have,

$$\begin{aligned}
&\tfrac{(n-\sqrt{K})^2}{K} \leq \tfrac{E(A(D)^2)}{E(A(D')^2)} \leq \exp\{\textstyle\sum_{i=1}^{n} \epsilon_i\} \\
&\ln \tfrac{(n-\sqrt{K})^2}{K} \leq \textstyle\sum_{i=1}^{n} \epsilon_i
\end{aligned} \qquad (32)$$

Now assume that $K \leq (\frac{m}{2})^2$. Let $D'' = (\underbrace{1, 1, \cdots, 1}_{m \text{ ones}}, 0, \cdots, 0)$. Then, similar to (31) , we can show that,

$$\frac{E(A(D'')^2)}{E(A(D')^2)} \leq \exp\{\sum_{i=1}^{m} \epsilon_i\}$$

Moreover, using the definition of differential privacy, we have,

$$\begin{aligned}
&Q(D'') = m \to \\
&E((A(D'') - Q(D''))^2) = E((A(D) - m)^2) \leq K \implies \\
&E(m - A(D'')) \leq \sqrt{K} \implies \\
&\qquad E(A(D'')) \geq m - \sqrt{K} \underbrace{\geq}_{\text{as } K \leq (m/2)^2} 0 \implies \\
&E(A(D'')^2) \geq E(A(D''))^2 \geq (m - \sqrt{K})^2 (***)
\end{aligned}$$

Using $(**)$ and $(***)$, we have,

$$\begin{aligned}
&\tfrac{(m-\sqrt{K})^2}{K} \leq \tfrac{E(A(D'')^2)}{E(A(D')^2)} \leq \exp\{\textstyle\sum_{i=1}^{m} \epsilon_i\} \\
&\ln \tfrac{(m-\sqrt{K})^2}{K} \leq \textstyle\sum_{i=1}^{m} \epsilon_i
\end{aligned} \qquad (33)$$

Because $K \leq (\frac{m}{2})^2$, then $\ln \frac{(m-\sqrt{K})^2}{K} > 0$. This implies that,

$$\sum_{i=1}^{m} \epsilon_i > 0 \qquad (34)$$

This means that at most $m - 1$ individuals can experience zero privacy loss. As a result, at least $n - m + 1$ individuals experiences non-zero privacy loss. ■

Proof. [Theorem 2] Let $D = (d_1, d_2, \cdots, d_n)$ and $D' = (d'_1, d_2, d_3, \cdots, d_n)$ be the two neighboring databases. Moreover, $s = Q(D) = \sum_{i=1}^{n} d_i$ and $s' = Q(D') = d'_1 + \sum_{i=2}^{n} d_i$. Using triangle inequality we have,

$$\begin{aligned}
Pr(A_u(D) \in S) &= \int_{x \in S-s} \frac{1}{2b} \exp\{-\frac{|x|}{b}\} dx \\
&= \int_{x \in S-s'} \frac{1}{2b} \exp\{-\frac{|x + d_1 - d'_1|}{b}\} dx \\
&\leq \exp\{\frac{|d_1 - d'_1|}{b}\} \int_{x \in S-s'} \frac{1}{2b} \exp\{-\frac{|x|}{b}\} dx \\
&\leq \exp\{\frac{1}{b}\} Pr(A_1(D') \in S) , \qquad (35)
\end{aligned}$$

where, $S - t = \{x - t | x \in S\}$. The first inequality holds because $|x + d_1 - d'_1| \leq |x| + |d_1 - d'_1|$. The second inequality holds as $|d_1 - d'_1| \leq 1$.

Moreover, $E((A_u(D) - Q(D))^2) = E(N(b)^2) = 2b^2$. ■

Proof. [Theorem 4] Since $b = \sqrt{\frac{1}{2}(K - \frac{n^2(1-a)^2}{4})}$, optimization problem (5) can be written as follows,

$$\begin{aligned}
\min_{a} \quad & \frac{a}{\sqrt{\frac{1}{2}(K - \frac{n^2(1-a)^2}{4})}} \\
\text{s.t.} \quad & 0 \leq a \leq 1
\end{aligned} \qquad (36)$$

The above optimization problem can be solved using the first order condition. We have,

$$\begin{aligned}
\frac{d \frac{a}{\sqrt{\frac{1}{2}(K - \frac{n^2(1-a)^2}{4})}}}{da} = 0 &\implies \hat{a} = 1 - \frac{4K}{n^2} \\
(\frac{n(1-a)}{2})^2 + 2b^2 = K &\implies \hat{b} = \sqrt{\frac{K(n^2 - 4K)}{2n^2}} \\
\hat{\epsilon} = \frac{\hat{a}}{\hat{b}} &= \frac{1}{n} \sqrt{\frac{2n^2 - 4K}{K}}.
\end{aligned}$$

■

Proof. [Theorem 5] As the cost function is linear, at most one a_i^* can be between zero and one. Otherwise, if $0 < a_i^* < 1$ and $0 < a_j^* < 1, i < j$, then we can decrease a_j^* and increase a_i^* to keep the accuracy equal to K and decrease the total cost/payment.

Now let's assume that $a_1^* = a_2^* = a_3^* = \cdots = a_m^* = 1$ and $a_{m+1}^* < 1$, and $a_{m+2}^* = \cdots , a_n^* = 0$. Notice that $m + 1 > n - 2\sqrt{K}$, otherwise the accuracy constraint cannot be satisfied. To find optimal value a^*_{m+1}, we solve the following optimization problem,

$$\min_{a_{m+1},b} \frac{v_{m+1}a_{m+1}+\sum_{i=1}^{m} v_i}{b}$$
$$s.t., (\frac{n-m-a_{m+1}}{2})^2 + 2b^2 = K \quad (37)$$

We can simplify the above optimization problem as follows,

$$\min_{a_{m+1}} \frac{v_{m+1}a_{m+1}+\sum_{i=1}^{m} v_i}{\sqrt{\frac{1}{2}(K-(\frac{n-m-a_{m+1}}{2})^2)}} \quad (38)$$

Using the first order condition, we can find $a^*_{m+1} = (n-m) - 4\cdot K \cdot \frac{v_{m+1}}{(n-m)v_{m+1}+\sum_{i=1}^{m} v_i}$. Three cases can happen,

- $(n-m) - 4\cdot K \cdot \frac{v_{m+1}}{(n-m)v_{m+1}+\sum_{i=1}^{m} v_i} \leq 0$: We should solve optimization problem (37) for m instead of $m+1$. The optimal value a_{m+1} is equal to zero.
- $(n-m) - 4\cdot K \cdot \frac{v_{m+1}}{(n-m)v_{m+1}+\sum_{i=1}^{m} v_i} > 1$: the optimal value of a_{m+1} is equal to one. We should solve (37) for $m+2$ to find optimal value of a_{m+2}.
- $0 < (n-m) - 4\cdot K \cdot \frac{v_{m+1}}{(n-m)v_{m+1}+\sum_{i=1}^{m} v_i} < 1$: optimal value a_{m+1} is equal to $(n-m) - 4\cdot K \cdot \frac{v_{m+1}}{(n-m)v_{m+1}+\sum_{i=1}^{m} v_i}$.

Given above cases, we can find the optimal solution as follows,

if $m+1$ is the first index where $s_{m+1} \leq 0$ (if s_i is non-negative $\forall i$, then set $m = n$), then the solution to optimization problem (10) is given by,

$$\begin{aligned} a^*_1 &= a^*_2 = \cdots = a^*_{m-1} = 1,\ a^*_m = \min\{s_m, 1\}, \\ a_{m+1} &= \cdots = a_n = 0 \\ b^* &= \sqrt{\frac{1}{2}(K-(\frac{2K\cdot v_m}{(n-m+1)\cdot v_m + \sum_{j=1}^{m-1} v_j})^2)} \end{aligned} \quad (39)$$

■

Proof. [Theorem 6] We assume that $v_1 \leq v_2 \leq \ldots \leq v_n$. Let $a^*_i, i \in \mathcal{N}$ and b^* be the solution to problem (10). Since $c(v,\epsilon)$ is convex, $\frac{dc(v_i,\frac{a}{b^*})}{da}|_{a=a^*_i} = \frac{dc(v_j,\frac{a}{b^*})}{da}|_{a=a^*_j}$ if $0 < a^*_i < 1$ and $0 < a^*_j < 1$. [4] Therefore, we have,

$$\begin{aligned} \frac{dc(v_i,\frac{a}{b^*})}{da}|_{a=a^*_i} &= v_i \cdot r \cdot (a^*_i)^{r-1}/(b^*)^r \\ &= v_j \cdot r \cdot (a^*_j)^{r-1}/(b^*)^r = \frac{dc(v_j,\frac{a}{b^*})}{da}|_{a=a^*_j} \\ v_i \cdot (a^*_i)^{r-1} &= v_j \cdot (a^*_j)^{r-1}, \\ &\forall\, 0 < a^*_i < 1,\ 0 < a^*_j < 1. \end{aligned} \quad (40)$$

[4]Otherwise, if $\frac{dc(v_i,\frac{a}{b^*})}{da}|_{a=a^*_i} < \frac{dc(v_j,\frac{a}{b^*})}{da}|_{a=a^*_j}$ and $0 < a^*_i < 1$ and $0 < a^*_j < 1$, then the broker can improve the objective function by increasing a^*_i and decreasing a^*_j and keeping the accuracy equal to K.

Since $v_1 \leq v_2 \leq \ldots \leq v_n$, it is easy to see that $a^*_i, i \in \mathcal{N}$ can be divided into three different categories:

$$\begin{aligned} &a^*_i = 1,\ \forall i \in \{1,\ldots,m_1\} \\ &v_i \cdot (a^*_i)^{r-1} = v_j \cdot (a^*_j)^{r-1},\ \forall i,j \in \{m_1+1,\ldots,m_2\} \\ &a^*_i = 0,\ \forall i \in \{m_2+1,\ldots,n\} \end{aligned} \quad (41)$$

Note that $m_2 \geq \lceil n - 2\sqrt{K} \rceil + 1$, otherwise accuracy K is not achievable. The main goal of algorithm 1 is to find m_1 and m_2 through exhaustive search. It is worth mentioning that if m_1 and m_2 are known, then $a^*_i, i \in \{m_1+1,\ldots,m_2\}$ can be calculated by following optimization problem,

$$\begin{aligned} &\min_{a_{m_1+1},b} \frac{a_{m_1+1}}{b} \\ &s.t., \\ &(\frac{n-m_1-\sum_{j=m_1+1}^{m_2}(\frac{v_{m_1+1}}{v_j})^{\frac{1}{r-1}} a_{m_1+1}}{2})^2 + 2b^2 = K \\ &b > 0, 0 < a_{m_1+1} < 1, \\ &a_j = (\frac{v_{m_1+1}}{v_j})^{\frac{1}{r-1}} a_{m_1+1} \forall m_1+1 \leq j \leq m_2. \end{aligned} \quad (42)$$

The above optimization problem can be simplified as follows,

$$\begin{aligned} &\min_{a_{m_1+1}} \frac{a_{m_1+1}}{\sqrt{\frac{1}{2}K - \frac{1}{2}(\frac{n-m_1-\sum_{j=m_1+1}^{m_2}(\frac{v_{m_1+1}}{v_j})^{\frac{1}{r-1}}\cdot a_{m_1+1}}{2})^2}} \\ &s.t., 0 < a < 1 \end{aligned} \quad (43)$$

The solution to the above optimization problem can be found by the first order condition and is given by,

$$\begin{aligned} A &= \sum_{k=m_1+1}^{m_2} \sqrt[(r-1)]{v_{m_1+1}/v_k}, \\ a^*_{m_1+1} &= \frac{(n-m_1)^2 - 4K}{A\cdot(n-m_1)} \\ a^*_j &= (\frac{v_{m_1+1}}{v_j})^{\frac{1}{r-1}} \cdot a^*_{m_1+1}, \forall j \in \{m_1+1,\ldots,m_2\} \end{aligned} \quad (44)$$

Because m_1, m_2 are not known beforehand, Algorithm 1 solves optimization problem (43) for all possible values of m_1 and m_2, and finds the optimal values for m_1 and m_2 and $a^*_{m_1+1}$ such that the total privacy cost is minimized. Note that if $a^*_{m_1+1}$ obtained from (43) is larger than 1 or less than zero, the m_1 and m_2 are not chosen correctly, and Algorithm 1 ignores these cases.

■

Next, we introduce the following theorem which will be used in the proof of Theorems 7, 8, and 9.

Theorem 13 (Envelope Theorem [24]): Let $c(v_i,\epsilon_i) = v_i \cdot l(\epsilon_i)$. Then, a mechanism $M =< t(\hat{\boldsymbol{v}}), f(\hat{\boldsymbol{v}}) >$ implements $f(\hat{\boldsymbol{v}})$ and satisfies the IC constraint if and only if,

1) $-l(f(\hat{v}_i,\hat{v}_{-i}))$ is non-decreasing in $\hat{v}_i$ for all $\hat{v}_{-i}$.
2) $U_i(\hat{\boldsymbol{v}}|v_i) = y_i(\hat{v}_{-i}) - \int_0^{\hat{v}_i} l(f(s_i,\hat{v}_{-i}))ds_i$, where $y_i(\hat{v}_{-i})$ is an arbitrary function and $U_i(\hat{\boldsymbol{v}}|v_i) = t_i(\hat{\boldsymbol{v}}) - c(v_i, f_i(\hat{\boldsymbol{v}}))$ is the utility function of individual i after introduction of the truthful mechanism M.

Proof. [Theorem 7] Since $g(\hat{\boldsymbol{v}})$ is constant, $-l(g_i(\hat{\boldsymbol{v}}))$ is non-decreasing in $\hat{v}_i$. In order to satisfy the second condition in the Envelope Theorem, we have,

$$\begin{aligned} t_i(\hat{\boldsymbol{v}}) - c(v_i, g_i(\hat{\boldsymbol{v}})) &= y_i(\hat{v}_{-i}) - \int_0^{\hat{v}_i} l(g_i(s_i, \hat{v}_{-i}))ds_i \\ &= y_i(\hat{v}_{-i}) - \hat{v}_i l(\hat{\epsilon}), \\ \implies t_i(\hat{\boldsymbol{v}}) &= y_i(\hat{v}_{-i}). \end{aligned} \tag{45}$$

By choosing $t_i(\hat{\boldsymbol{v}}) = y_i(\hat{v}_{-i})$, mechanism M_1 would satisfy the IC constraint because it satisfies the second condition of the Envelope Theorem . Since M_1 is incentive compatible, by the IR constraint we have,

$$\begin{aligned} t_i(\boldsymbol{v}) - c(v_i, g_i(\boldsymbol{v})) &= y_i(v_{-i}) - v_i \cdot l(\hat{\epsilon}) \geq 0 \\ \implies y_i(v_{-i}) \geq v_i \cdot l(\hat{\epsilon}) & \end{aligned} \tag{46}$$

The smallest $y_i(\hat{v}_{-i})$ which satisfies the above equation is $\overline{v} \cdot l(\hat{\epsilon})$. Therefore, M_1 satisfy the IC and IR constraints with a minimum payment if and only if $t_i(\boldsymbol{v}) = \overline{v} \cdot l(\hat{\epsilon})$. ■

Proof. [Theorem 8] The proof is similar to the proof of Theorem 7. It is easy to see that $h_i(\hat{v}_i, \hat{v}_{-i})$ is non-increasing in $\hat{v}_i$. Therefore, $-l(h_i(\hat{v}_i, \hat{v}_{-i}))$ is not-decreasing in $\hat{v}_i$ for all $\hat{v}_{-i}$. In order to satisfy the second condition of the Envelope Theorem, we have,

$$\begin{aligned} \tau_i(\hat{\boldsymbol{v}}) - c(v_i, h_i(\hat{\boldsymbol{v}})) &= y_i(\hat{v}_{-i}) + \int_0^{\hat{v}_i} l(h(s_i, \hat{v}_{-i}))ds_i, \\ \tau_i(\hat{\boldsymbol{v}}) &= y_i(\hat{v}_{-i}) + v_i h_i(\hat{\boldsymbol{v}}) + \int_0^{\hat{v}_i} l(h_i(s_i, \hat{v}_{-i}))ds_i \end{aligned} \tag{47}$$

The above $\tau_i(\hat{\boldsymbol{v}})$ function satisfies the conditions in Envelope Theorem. Therefore, M_2 is incentive compatible. Next, we find $y_i(\hat{v}_{-i})$ using the IR constraint. Since the sellers report their privacy attitudes truthfully at NE, we have,

$$\begin{aligned} \tau_i(\boldsymbol{v}) - c(v_i, h_i(\boldsymbol{v})) \geq 0, \\ y_i(v_{-i}) - \int_0^{v_i} l(h_i(s_i, v_{-i}))ds_i \geq 0, \forall v_i. \end{aligned} \tag{48}$$

Therefore, if $y_i(v_{-i}) = \max_{v_i} \int_0^{v_i} l(h_i(s_i, v_{-i}))ds_i$, the payment would be minimized and IR constraint would be satisfied. Since $l(.)$ is a non-negative function, $\max_{v_i} \int_0^{v_i} l(h_i(s_i, v_{-i})) = \int_0^{\overline{v}} l(h_i(s_i, v_{-i}))$. As a result, M_2 satisfies both IR and IC constraints with the minimum payment if and only if,

$$\tau_i(\hat{\boldsymbol{v}}) = \int_{\hat{v}_i}^{\overline{v}} l(h_i(s_i, \hat{v}_{-i}))ds_i + \hat{v}_i \cdot l(h_i(\hat{\boldsymbol{v}})) \tag{49}$$

■

Proof. [Theorem 9] The proof is similar to the proof of Theorem 7. ■

Proof. [Theorem 10]

$$\begin{aligned} & E([A_{new}(D) - Q(D)]^2) \\ &= [\sum_{i=1}^{q}(a_i - 1)f_i(D) + \frac{1-a_i}{2}]^2 + E(N(b)^2) \\ &\leq (\sum_{i=1}^{q} \frac{1-a_i}{2})^2 + 2b^2, \end{aligned} \tag{50}$$

where the inequality holds because $0 \leq f_i(D) \leq 1$. Let $D = (d_1, d_2, \cdots, d_n)$ and $D' = (d_1', d_2, d_3, \cdots, d_n)$ be the two neighboring databases. Moreover, let $s = \sum_{i=1}^{q}\left[a_i \cdot f_i(D) + \frac{1-a_i}{2}\right]$ and $s' = \sum_{i=1}^{n}\left[a_i \cdot f_i(D') + \frac{1-a_i}{2}\right]$. We then have,

$$\begin{aligned} & Pr\{A_{new}(D) \in S\} = \int_{x \in S-s} \frac{1}{2b} \exp\{-\frac{|x|}{b}\}dx \\ &= \int_{x \in S-s'} \frac{1}{2b} \exp\{-\frac{|x + \sum_{i=1}^{q} a_i[\cdot f_i(D) - g_i(D')]|}{b}\}dx \\ &\leq \exp\{\frac{\sum_{i=1}^{q} a_i|f_i(D) - f_i(D')|}{b}\} \int_{x \in S-s'} \frac{\exp\{-\frac{|x|}{b}\}}{2b}dx \\ &\leq \exp\{\frac{\sum_{i \in \mathcal{I}_1} a_i \Delta_i^1}{b}\} Pr(A_{new}(D') \in S) , \end{aligned} \tag{51}$$

where $S - t = \{x - t | x \in S\}$. Therefore, $A_{new}(D)$ is $\frac{\sum_{i \in \mathcal{I}_1} a_i \Delta_i^1}{b}$-differentially private with respect to individual 1. Similarly, we can show that $A_{new}(D)$ is $\frac{\sum_{i \in \mathcal{I}_j} a_i \Delta_i^j}{b}$-differentially private with respect to individual j.

■

Proof. [Theorem 11]

Let $D = (\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_n)$ and $D' = (\hat{\boldsymbol{d}}_1, \boldsymbol{d}_2, \boldsymbol{d}_3, \cdots, \boldsymbol{d}_n)$ be the two neighboring databases where $||\boldsymbol{d}_1 - \hat{\boldsymbol{d}}_1||_1 \leq 1$. Moreover, let $s = \sum_{i=1}^{n} a_i \cdot \boldsymbol{d}_i + \frac{1-a_i}{2}\mathbf{1}$ and $s' = a_1\hat{\boldsymbol{d}}_1 + \frac{1-a_1}{2}\mathbf{1} + \sum_{i=2}^{n} a_i \cdot \boldsymbol{d}_i + \frac{1-a_i}{2}\mathbf{1}$. We then have

$$\begin{aligned} & Pr\{\boldsymbol{A}_{new}(D) \in S\} \\ &= \int_{\boldsymbol{x} \in S-s} \prod_{i=1}^{m} \left(\frac{1}{2b} \exp\{-\frac{|x_i|}{b}\}\right) dx_1 \cdots dx_m \\ &= \int_{\boldsymbol{x} \in S-s'} \prod_{i=1}^{m} \left(\frac{1}{2b} \exp\{-\frac{|x_i + a_1 \cdot d_1^i - a_1 \cdot \hat{d}_1^i|}{b}\}\right) dx_1 \cdots dx_m \\ &\leq \exp\{\frac{a_1 \cdot \sum_{i=1}^{m} |d_1^i - \hat{d}_1^i|}{b}\} \int_{\boldsymbol{x} \in S-s'} \prod_{i=1}^{m} \frac{1}{2b} \exp\{-\frac{|x_i|}{b}\}dx_1 \cdots dx_m \\ &\leq \exp\{m \cdot \frac{a_1}{b}\} Pr(\boldsymbol{A}_{new}(D') \in S) , \end{aligned} \tag{52}$$

Therefore, $\boldsymbol{A}_{new}(D)$ is $m \cdot \frac{a_1}{b}$-differentially private with respect to individual 1. Similarly, we can show that $\boldsymbol{A}_{new}(D)$ is $m \cdot \frac{a_i}{b}$-differentially private with respect to individual i. Similarly, we can show that $\boldsymbol{A}_u(.)$ is $\frac{m}{b}$-differentially private with respect to each agent.

■

Proof. [Theorem 12]

$$\frac{1}{m}E\{||\boldsymbol{A}_{new}(D) - Q(D)||_2^2\}$$
$$= \frac{1}{m}E\left\{\sum_{j=1}^{m}\left(N_j(b) + \sum_{i=1}^{n} a_i d_i^j + \frac{1-a_i}{2}\right)^2\right\} =$$
$$\frac{1}{m}\sum_{j=1}^{m} E\left\{N_j(b)^2 + (\sum_{i=1}^{n} a_i d_i^j + \frac{1-a_i}{2})^2 + N_j(b)\sum_{i=1}^{n} a_i d_i^j + \frac{1-a_i}{2}\right\}$$
$$\le \frac{1}{m}\sum_{j=1}^{m}[2b^2 + (\sum_{i=1}^{n}\frac{1-a_i}{2})^2] = 2b^2 + (\sum_{i=1}^{n}\frac{1-a_i}{2})^2$$

Similarly, we can show that $\boldsymbol{A}_u(D)$ is $2b^2$-accurate. ■

REFERENCES

[1] M. M. Khalili, X. Zhang, and M. Liu, "Contract design for purchasing private data using a biased differentially private algorithm," in *14th Workshop on the Economics of Networks, Systems and Computation (NetEcon)*, 2019, available at http://bit.ly/2VwfLo9.

[2] *Datacoup*, http://datacoup.com/.

[3] C. Li, D. Y. Li, G. Miklau, and D. Suciu, "A theory of pricing private data," *ACM Trans. Database Syst.*, vol. 39, no. 4, pp. 34:1–34:28, Dec. 2014. [Online]. Available: http://doi.acm.org/10.1145/2691190.2691191

[4] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. R. Liu, "Privacy or utility in data collection? a contract theoretic approach," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 7, pp. 1256–1269, Oct 2015.

[5] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proceedings of the 12th ACM Conference on Electronic Commerce*, ser. EC '11. New York, NY, USA: ACM, 2011, pp. 199–208. [Online]. Available: http://doi.acm.org/10.1145/1993574.1993605

[6] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, ser. EC '12. New York, NY, USA: ACM, 2012, pp. 568–585. [Online]. Available: http://doi.acm.org/10.1145/2229012.2229054

[7] L. Xu, C. Jiang, Y. Qian, Y. Zhao, J. Li, and Y. Ren, "Dynamic privacy pricing: A multi-armed bandit approach with time-variant rewards," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 2, pp. 271–285, Feb 2017.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[9] C. Xia and S. Muthukrishnan, "Arbitrage-free pricing in user-based markets," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '18. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 327–335. [Online]. Available: http://dl.acm.org/citation.cfm?id=3237383.3237436

[10] F. E. Benth, L. Ekeland, R. Hauge, and B. R. F. Nielsen, "A note on arbitrage-free pricing of forward contracts in energy markets," *Applied Mathematical Finance*, vol. 10, no. 4, pp. 325–336, 2003.

[11] J. K. Fung and K. C. Chan, "On the arbitrage-free pricing relationship between index futures and index options: A note," *Journal of Futures Markets*, vol. 14, no. 8, pp. 957–962, 1994.

[12] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *Journal of the ACM (JACM)*, vol. 62, no. 5, p. 43, 2015.

[13] S. Deep and P. Koutris, "The design of arbitrage-free data pricing schemes," *CoRR*, vol. abs/1606.09376, 2016. [Online]. Available: http://arxiv.org/abs/1606.09376

[14] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *Proc. VLDB Endow.*, vol. 7, no. 9, pp. 757–768, May 2014. [Online]. Available: http://dx.doi.org/10.14778/2732939.2732948

[15] R. Cummings, K. Ligett, A. Roth, Z. S. Wu, and J. Ziani, "Accuracy for sale: Aggregating data with a variance constraint," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ser. ITCS '15. New York, NY, USA: ACM, 2015, pp. 317–324. [Online]. Available: http://doi.acm.org/10.1145/2688073.2688106

[16] I. Vakilinia, J. Xin, M. Li, and L. Guo, "Privacy-preserving data aggregation over incomplete data for crowdsensing," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[17] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in *2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2016, pp. 344–353.

[18] I. Vakilinia, D. K. Tosh, and S. Sengupta, "Privacy-preserving cybersecurity information exchange mechanism," in *2017 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, July 2017, pp. 1–7.

[19] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[20] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the sulq framework," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2005, pp. 128–138.

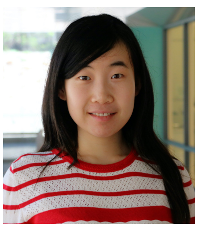
[21] *PRIVACY ACT OF 1974*, https://www.justice.gov/opcl/privacy-act-1974.

[22] *What is the GDPR Lawfulness, Fairness, and Transparency Principle?*, https://www.dataguise.com/gdpr-knowledge-center/lawfulness-fairness-transparency-principle/.

[23] A. J. Smola and B. Schölkopf, *Learning with kernels*. Citeseer, 1998, vol. 4.

[24] S. Tadelis and I. Segal, "Lectures in contract theory," 2005.

MOHAMMAD MAHDI KHALILI is an assistant professor in the Computer and Information Sciences Department at the University of Delaware. Before joining the University of Delaware, he was a postdoctoral researcher at the University of California, Berkeley. He received his Ph.D. degree in Electrical and Computer Engineering from the University of Michigan, Ann Arbor, in 2019. His research interest is in the societal aspect of machine learning algorithms, specifically in the areas of data privacy, fairness, game theory and mechanism design, and security economics.



XUERU ZHANG is a Ph.D. candidate in the Department of Electrical and Computer Engineering at the University of Michigan. She received her M.Sc. degree in Electrical and Computer Engineering from the University of Michigan in 2016 and B.Eng. degree in Electronic and Information Engineering from Beihang University (BUAA), Beijing, China, in 2015. Her research lies at the intersection of machine learning, optimization, statistics and economics, including topics such as data privacy, algorithmic fairness and security economics.

MINGYAN LIU received her Ph.D in electrical engineering from the University of Maryland, College Park, in 2000. She is a professor and the Peter and Evelyn Fuss Chair of Electrical and Computer Engineering at the University of Michigan, Ann Arbor. Her interests are in sequential decision and learning theory, game theory and incentive mechanisms, with applications to large-scale networked systems. She is a Fellow of the IEEE and a member of the ACM.