Selling Data at an Auction under Privacy Constraints

Mengxiao Zhang
Business School
The University of Auckland
Auckland, NZ
mengxiao.zhang@auckland.ac.nz

Fernando Beltran
Business School
The University of Auckland
Auckland, NZ
f.beltran@auckland.ac.nz

Jiamou Liu
School of Computer Science
The University of Auckland
Auckland, NZ
jiamou.liu@auckland.ac.nz

Abstract

Private data query combines mechanism design with privacy protection to produce aggregated statistics from privately-owned data records. The problem arises in a data marketplace where data owners have personalised privacy requirements and private data valuations. We focus on the case when the data owners are *single-minded*, i.e., they are willing to release their data only if the data broker guarantees to meet their announced privacy requirements. For a data broker who wants to purchase data from such data owners, we propose the Single-MindedQuery (SMQ) mechanism, which uses a reverse auction to select data owners and determine compensations. SMQ satisfies interim incentive compatibility, individual rationality, and budget feasibility. Moreover, it uses *purchased privacy expectation maximisation* as a principle to produce accurate outputs for commonly-used queries such as counting, median and linear predictor. The effectiveness of our method is empirically validated by a series of experiments.

1 INTRODUCTION

The increasing reliance on data-driven technologies has led to the formation of an economy that is built on data trading. Many *data marketplaces* emerged that bring *data buyers*, i.e., those who are seeking data to purchase, together with *data owners*, i.e., those who are willing to release data for a compensation. Examples of data marketplaces include Datacoup, Datum, CitizenMe and DataWallet, many of which enable data buyers to purchase personal data from individual data owners (Laudon, 1996; Nget, Cao, & Yoshikawa, 2017).

Imagine a *data broker* who would like to query a set of privately held data records – such as income record, energy consumption data, or online service rating – to produce aggregated statistics. This may mean that consents must be purchased from individual data owners to access their data. In designing a query mechanism for this task, the data broker faces a number of challenges: The first is *privacy*. By giving out their data, the data owners give up a certain privacy. An attacker may use, say, income data to infer confidential information such as personal identities (Brankovic & Estivill-Castro, 1999). It thus makes sense for data owners to demand privacy protection when their data are released. Secondly, every data owner associates a value to their data. The value denotes the minimum compensation for the individual to release the data. Naturally, this value is hidden from the data broker. The imbalance in the access to data valuation between data owners and the data broker, i.e., *information asymmetry*, is the second challenge faced by the data broker. Thirdly, when using compensations to incentivise the data owners to release their data, decisions need to be made regarding which (and how many) data owners to procure data from, subject to budget constraints. The utility to the data broker is determined by the accuracy of queries made on the purchased data records. The third challenge thus lies in *optimising accuracy*. These challenges make private data query an attractive topic that sits between data analysis, algorithmic game theory, and data security.

A paradigm of *privacy trading* can be found in Ghosh and Roth (2011). The premises involve a data broker who aims to query private data held by individuals who have varying attitudes towards privacy. To reveal the hidden privacy attitudes of data owners, the data broker uses an auction where each data owner submits a bid reflecting their privacy valuation. Based on all bids received, the data broker decides on a level of privacy to be purchased from the data owners. A noisy query output is then produced which preserves the purchased level of privacy. The paradigm laid out in (Ghosh & Roth, 2011) implic-

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

itly makes two assumptions: (1) privacy trading may be done in ways that resemble the trading of other commodities such as stocks and bonds; and (2) the data broker, by giving sufficient incentive, could purchase an arbitrary amount of privacy from every data owner. Hence it is up to the data broker to determine the level of privacy to be purchased from the data owners. A number of work have subsequently adopted these views in other settings (Dandekar, Fawaz, & Ioannidis, 2012; Fleischer & Lyu, 2012; Ligett & Roth, 2012; Cummings, Ligett, Roth, Wu, & Ziani, 2015).

Contribution. In this paper, we adopt a different view towards private data query. Under the assumptions of privacy trading, as the data broker applies more stringent methods to protect data privacy, the cost (and therefore payment) to the data owner would tend towards 0. This is often not the case in reality. In many situations, the data owners are "single-minded", i.e., they would demand a level of privacy protection and would not release their data *unless their requested level of privacy protection is met*. Moreover, once the level of privacy protection is guaranteed, privacy is decoupled from the cost of releasing the data. In other words, the payment demanded by a data owner to release their data would not decrease as the data broker provides more privacy protection to the data. This paradigm can be considered as *data trading* as the purchased commodity is the access to data rather than privacy. To our knowledge, despite its simple and relatively straightforward setup, no work has addressed private data query under this assumption. This is the major conceptual contribution of our work.

Our technical contribution lies in *SingleMindedQuery (SMQ)*, a new private data query mechanism over single-minded data owners as described above. We adopt the personalised variant of the classical *differential privacy* (DP) to quantify data privacy. Our goal is to perform queries to a high accuracy while preserving every data owner's declared privacy requirement. The main technical novelty is a reverse auction mechanism that determines whose data to purchase and how much compensation should be paid. Note that unlike mechanisms that follow the privacy trading paradigm (Ghosh & Roth, 2011), SMQ does not decide the level of privacy to be purchased from the data owners. SMQ incentivises the data owners so that their data valuation is truthfully revealed, thereby resolving information asymmetry. Furthermore, we use *purchased privacy expectation maximisation (PPEM)* as a principle that guides SMQ to achieve high accuracy for commonly-used queries such as counting, median and linear predictor (See Lemma 2). We validate empirically the effectiveness of our method using a series of experiments.

Related work. We review research on the query of private data. The seminal work of Ghosh and Roth (2011) has laid down some main assumptions. The authors propose the FairQuery (FQ) mechanism to perform count query on binary (0/1-valued) data. FQ uses a reverse auction to obtain data owners' privacy valuation. When combined with a Laplace mechanism that obfuscates query output, FQ guarantees (certain exact formulations of) incentive compatibility, individual rationality, budget feasibility, query accuracy, and privacy protection. These conditions have since been key indicators of the effectiveness of any query mechanism for private data. The notion of *ε-differential privacy (DP)* (Dwork, McSherry, Nissim, & Smith, 2006) has been chosen to quantify data privacy as the parameter ε captures in some precise sense the loss on utility a person experiences if her data is used in an ε-DP manner. This supports an argument that a data owner's privacy can be regarded as a "soft constraint" to be captured by a real-valued cost. The cost increases as a linear function in terms of privacy loss. The FQ mechanism in (Ghosh & Roth, 2011) associates this cost with the compensation paid to the data owner.

It is important to point out that, even though the data owners have different privacy valuations, FQ would compute a single level $\varepsilon > 0$ of DP, and compensate for ε-DP to *all* data owners whose data are used. This means that the mechanism could over-protect some data owners by offering them too stringent privacy protection. Therefore, we can consider this mechanism as "pseudo-personalised" as it fails to account for the differences in the data owners' privacy valuations. Such pseudo-personalised approach has been uptaken by Fleischer and Lyu (2012) and Ligett and Roth (2012) which instead of proposing auction mechanisms, design indirect mechanisms, i.e., take-it-or-leave-it offers, to reveal data owners' privacy valuation.

In contrast, a "truly-personalised" private data query mechanism enables different levels of privacy requirements to be applied to different data owners. Such a mechanism has the potential to avoid over-protecting the data owners' privacy, thus achieving higher accuracy. For example, Dandekar et al. (2012) design the FairInnerProduct (FIP) mechanism for linear predictor queries over real-valued data. FIP also uses a reverse auction mechanism while enforcing different levels of privacy protection for different data owners. Such an approach is also adopted by (Cummings et al., 2015). There, a data broker provides a menu of different variance levels and asks the data owners to report the valuation under each level. However, a crucial limitation exists in (Cummings et al., 2015)'s mechanism as it does not explicitly state its differential privacy level.

As mentioned earlier, all mechanisms above follow the privacy trading paradigm where the cost to a data owner from releasing their data is assumed to only arise from privacy loss. This makes sense assuming (1) the data owners fully trust the data broker to protect their purchased privacy level, and (2) the data owners do not have an intrinsic valuation to their data records. These assumptions may not hold in practice. In the light of this, we will put forward a mechanism that compensates the data owners for their intrinsic data valuation while treating the privacy requirement as a hard constraint.

2 PRELIMINARIES

PDP Queries. We regard a *dataset* as a tuple $\vec{d} = (d_1, \ldots, d_n) \in \mathbb{R}^n$ where $n \geq 1$ and each *data entry* $d_i \in \mathbb{R}$. $\mathcal{D} \subseteq \mathbb{R}^\star$ denotes the collection of all possible datasets. A *query* is a function $\varphi \colon \mathcal{D} \to \mathbb{R}$, such as median and mean. To achieve privacy protection, a randomised function g is applied to the query result to obtain the obfuscated query $\Phi = g \circ \varphi$.

The notion of ε*-personalised differential privacy* (PDP) quantifies the level of privacy achieved by this randomised function: Call two datasets $\vec{d} \in \mathbb{R}^n$ and $\vec{d'} \in \mathbb{R}^n$ i*-neighbouring* if they differ on exactly the ith entry.

Definition 1. *(Jorgensen, Yu, & Cormode, 2015) Given a vector $\vec{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$, a randomised function $\Phi \colon \mathcal{D} \to \mathbb{R}$ is $\vec{\varepsilon}$-*PDP *if for any pair of i-neighbouring datasets $\vec{d}, \vec{d'} \in \mathbb{R}^n$ where $1 \leq i \leq n$:*

$$\frac{\Pr(\Phi(\vec{d}) \in R)}{\Pr(\Phi(\vec{d'}) \in R)} \leq e^{\varepsilon_i}, \forall R \subset \mathrm{Range}(\Phi) \qquad (1)$$

In other words, suppose $\vec{d'}$ is an i-neighbouring dataset from the true dataset $\vec{d}$. As the ratio above moves closer to 1, Φ is more likely to output the same result on $\vec{d}$ and $\vec{d'}$, hiding the true value of the ith data entry. Hence a ε-PDP query mechanism with smaller ε leads to a higher level of privacy protection for data entry d_i[1].

The $P\mathcal{E}$ mechanism generates $\vec{\varepsilon}$-PDP queries (Jorgensen et al., 2015): For $\vec{d'}, \vec{d} \in \mathbb{R}^n$, let $I_{\vec{d} \oplus \vec{d'}} := \{1 \leq i \leq n \mid d'_i \neq d_i\}$. Fix a query φ. Set $\sigma_\varphi(\vec{d}, r) := \max_{\varphi(\vec{d'})=r} \left\{ \sum_{i \in I_{\vec{d} \oplus \vec{d'}}} -\varepsilon_i \right\} \forall r \in \mathbb{R}$. Given a dataset $\vec{d}$, the *$P\mathcal{E}$ mechanism* $\Phi_\varphi(\vec{d})$ generates output $r \in \mathbb{R}$ with probability

$$\Pr\left(\Phi_\varphi(\vec{d}) = r\right) = \frac{\exp(\frac{1}{2}\sigma_\varphi(\vec{d}, r))}{\sum_{r' \in \mathrm{Range}(\Phi)} \exp(\frac{1}{2}\sigma_\varphi(\vec{d}, r'))}$$

We will implement our query using the $P\mathcal{E}$ mechanism, as it can be applied to arbitrary real-valued queries and adds a relatively smaller amount of random noise as compared with other existing methods who claim to achieve $\vec{\varepsilon}$-PDP (Alaggan, Gambs, & Kermarrec, 2016; Li, Xiong, Ji, & Jiang, 2017).

Procurement mechanism. We consider a market that consists of a single buyer and multiple sellers denoted by $s_1, s_2, \ldots, s_n$. The following assumptions are made on every seller s_i, $1 \leq i \leq n$:

(A1) We assume that once an appropriate amount of compensation is given, s_i is willing to sell her good to the buyer. The required level of compensation depends on the inherent valuation θ_i that s_i puts on the good. This is a real value in the range $\Theta := [\underline{\theta}, \overline{\theta}]$ where constants $0 \leq \underline{\theta} \leq \overline{\theta}$ are the lower- and upper-bound, respectively; θ_i represents the loss s_i suffers when she sells the good.

(A2) The valuation θ_i of s_i is a random sample from a distribution with cumulative probability function F_i and density function f_i. The distribution is assumed to be *regular*. In other words, the function $f_i(v)/(1 - F_i(v))$, i.e., the probability that $\theta_i = v$ conditioned on $\theta_i > v$, is monotonically non-decreasing on $v \in \Theta$. This assumption is commonly made in mechanism design literature and is satisfied by most distributions (Börgers, 2015).

We further assume that $F_1 = \cdots = F_n$ and $\theta_1, \ldots, \theta_n$ are i.i.d. random variables. The *valuation vector* $\vec{\theta} := (\theta_1, \ldots, \theta_n) \in \Theta^n$ has joint distribution and density functions F and f, respectively. While the functions F and f are common knowledge among the buyer and sellers, the valuation θ_i is only known by s_i and hidden from anyone else. Therefore, it is crucial for the buyer to incentivise the sellers to reveal their true valuations.

A *procurement mechanism* acts on behalf of the buyer to select a subset of sellers and decides on the amount of compensation for each seller. A direct mechanism, defined below, is a form of procurement mechanism where the buyer makes decisions solely based on the sellers' reported valuations (Börgers, 2015):

Definition 2. *A* direct mechanism Ψ *consists of a pair of functions (q, p) where $q \colon \Theta^n \to \{0, 1\}^n$ is called* allocation function *and $p \colon \Theta^n \to \mathbb{R}^n$ is called* payment function. *For any $\vec{\psi} \in \Theta^n$, the tuples $q(\vec{\psi}) := (q_1(\vec{\psi}), \ldots, q_n(\vec{\psi}))$ and $p(\vec{\psi}) := (p_1(\vec{\psi}), \ldots, p_n(\vec{\psi}))$ are called* allocation vector *and* payment vector, *resp.*

Intuitively, the buyer first receives reported valuation $\psi_i \in \Theta$ from each s_i to form a vector $\vec{\psi} = (\psi_1, \ldots, \psi_n)$. The mechanism computes $q_i(\vec{\psi})$ and $p_i(\vec{\psi})$. When $q_i(\vec{\psi}) = 1$, the buyer "selects" s_i and purchases the good from s_i with a compensation $p_i(\vec{\psi})$.

[1] PDP is generalised from the classical differential privacy (Dwork et al., 2006) to accommodate the diversity in people's privacy attitudes (Acquisti & Grossklags, 2005; Berendt, Günther, & Spiekermann, 2005).

We need a procurement mechanism that leads to certain desirable actions of the sellers. At the time of submitting a valuation ψ_i, the seller s_i makes decision with only the private information θ_i and knowledge regarding the distribution F. In other words, the outcome of the mechanism is made based on *ex-interim* utility expectations (Conitzer, 2009). It therefore makes sense to adopt *Bayesian Nash* utility in the solution concept of mechanism design (Mas-Colell, Whinston, Green, et al., 1995):

Set $\vec{\psi}_{-i} := (\psi_1, \ldots, \psi_{i-1}, \psi_{i+1}, \ldots, \psi_n) \in \Theta^{n-1}$ as the reported valuation vector of the sellers other than s_i. We abuse the notation writing $p_i(\psi_i, \vec{\psi}_{-i})$ for $p_i(\vec{\psi})$ and $q_i(\psi_i, \vec{\psi}_{-i})$ for $q_i(\vec{\psi})$. Set f_{-i} as the density function of the joint probability distribution of $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \theta_n)$. We define

$$Q_i(\psi_i) := \int_{\Theta^{n-1}} q_i(\psi_i, \vec{\psi}_{-i}) f_{-i}(\vec{\psi}_{-i}) \, \mathrm{d}\vec{\psi}_{-i} \tag{2}$$

$$P_i(\psi_i) := \int_{\Theta^{n-1}} p_i(\psi_i, \vec{\psi}_{-i}) f_{-i}(\vec{\psi}_{-i}) \, \mathrm{d}\vec{\psi}_{-i} \tag{3}$$

as the *expected allocation* and the *expected payment* when the reported valuation of s_i is ψ_i, resp. And her *expected utility* is $U_i(\psi_i|\theta_i) := P_i(\psi_i) - \theta_i Q_i(\psi_i)$.

The celebrated *revelation principle* asserts that to find the optimal procurement process, it is sufficient to restrict to direct mechanisms where the data owners truthfully report their valuation in the Bayesian Nash equilibrium (see (Börgers, 2015)). Formally, we would like to design a direct mechanism Ψ with the following properties:

(1) Incentive compatibility (IC): This property ensures that each seller truthfully reports her valuation, as she expects to gain the maximum utility by doing this, i.e.,

$$U_i(\theta_i|\theta_i) \geq U_i(\psi_i|\theta_i), \forall i \in \{1, \ldots, n\}, \forall \theta_i, \psi_i \in \Theta \tag{4}$$

(2) Individual rationality (IR): This property ensures that every seller is willing to participate in the mechanism, as her gain of participating is not less than that of non-participation. Here, we assume that the utility of non-participation is zero, i.e.,

$$U_i(\theta_i|\theta_i) \geq 0, \forall i \in \{1, \ldots, n\}, \forall \theta_i \in \Theta \tag{5}$$

It is also reasonable to assume that the buyer has a limited budget $B \leq \overline{\theta} n$, and thus we need the following.

(3) Budget feasibility (BF): The expected compensation received by all sellers should not exceed budget B, i.e.,

$$\sum_{i=1}^{n} \int_{\Theta^n} p_i(\vec{\psi}) f(\vec{\psi}) \, \mathrm{d}\vec{\psi} \leq B \tag{6}$$

Remark: The definition of BF here is *interim BF*, which merely considers the expected value of total payment. This may seem counter-intuitive – as the buyer would really like to ensure ex post BF, i.e., $\sum_{i=1}^{n} p_i(\vec{\psi}) \leq B$. It has turned out that these two notions are equivalent, as shown by the following lemma.

Lemma 1. *(Börgers, 2015) For any direct mechanism that is individually rational, incentive compatible and interim budget feasible, there is a direct mechanism with the same allocation rule that is individually rational, incentive compatible and ex post budget feasible.*

3 PRIVACY-AWARE DATA OWNERS

We consider a data marketplace that involves a group of *data owners*, each of whom holds a private (real-valued) data entry, and a *data broker* who would like to collect these data entries. Denote the data owners by $s_1, \ldots, s_n$ and s_i's data entry by d_i. Every s_i has a *data valuation* $\theta_i \in \mathbb{R}$ for her data entry. In this way, the data owners and broker are respectively the sellers and buyer in the market with data entries as goods. We make some further assumptions regarding every data owner s_i:

(A3) s_i is a *single-minded* data owner, i.e., the data owner s_i has a *privacy requirement* $\varepsilon_i \in \mathbb{R}$. To release the data, s_i requires the data broker to meet ε_i-PDP for any query made on the collected dataset.

(A4) The data entry d_i, once released, is verifiable and thus data owners cannot misreport their data. This assumption is reasonable because in some data marketplaces such as Datacoup, the data brokers do not directly collect data from data owners, but rather, they seek for data access permission from data owners and the data are provided by certain intermediary service.

(A5) No correlation exists between θ_i and the data value d_i. This means that the output of the allocation function does not reveal any information about the private data.

(A1)–(A5) naturally infer the following definition:

Definition 3. *A* privacy-aware data owner s_i *is formally a tuple* $s_i := (d_i, \theta_i, \varepsilon_i)$*, where* $d_i \in \mathbb{R}$ *is* s_i*'s data entry,* $\theta_i \in \Theta$ *is* s_i*'s* data valuation*, and* $\varepsilon_i \geq 0$ *is her* privacy requirement*.*

Remark: The definition above is different from the one in (Ghosh & Roth, 2011) where instead of the value ε_i, a data owner s_i is associated a *cost function* $c_i(\varepsilon) \in \mathbb{R}$ that captures the amount of loss s_i experiences when the data broker "purchases" ε amount of privacy. In particular, the cost function is defined as $c_i(\varepsilon) = \varepsilon \cdot v_i$ where $v_i \in \mathbb{R}$ is a *privacy valuation*. When using c_i to denote the minimum compensation s_i requires to release the data entry

d_i, in our setting, a privacy-aware data owner would have a stepwise cost function defined as

$$c_i(\varepsilon) = \begin{cases} \theta_i & \text{if } 0 \leq \varepsilon \leq \varepsilon_i \\ \infty & \text{otherwise} \end{cases} \tag{7}$$

In particular, $c_i(0)$ may be non-zero as it represents s_i's valuation to the data entry rather than privacy. This crucial difference makes the analysis in (Ghosh & Roth, 2011) not applicable to our setting.

4 DATA QUERY MECHANISM

A *data query mechanism* A combines a procurement mechanism Ψ with a PDP query Φ (See Fig. 1). Suppose $\mathcal{S} = \{s_1, \ldots, s_n\}$ is a set of privacy-aware data owners. Given a query $\varphi\colon \mathcal{D} \to \mathbb{R}$ and a budget $B > 0$, the mechanism A first applies Ψ which purchases data entries from a subset of data owners and constructs a sampled dataset; the PDP query Φ is then applied to return a final query result $A(\mathcal{S})$. As argued in Section 2, we will apply the $P\mathcal{E}$ mechanism as Φ.

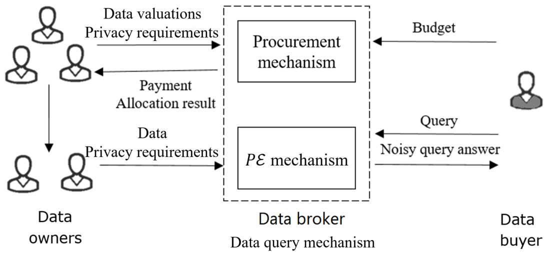

Figure 1: An illustration of a data query mechanism

We denote the *ground-truth dataset* $(d_1, \ldots, d_n)$ by $\vec{d}_{\text{gt}}$ and the sampled dataset by $\vec{d}_{\text{c}}$. The next definition captures the query accuracy of A.

Definition 4. *A query mechanism A is* (α, δ)-probably approximately correct *(PAC) if for any $\vec{d}_{\text{gt}}$, $\Pr(|A(\vec{d}_{\text{c}}) - \varphi(\vec{d}_{\text{gt}})| \geq \alpha) \leq 1 - \delta$.*

For brevity we write q_i for the allocation result $q_i(\psi_i, \vec{\psi}_{-i}) \in \{0, 1\}$ of s_i given the reported valuation $\vec{\psi}$. As each data owner is single-minded, s_i's privacy requirement ε_i must be met in case $q_i = 1$. One can thus view $\varepsilon_i q_i$ as the amount of privacy "purchased" by the procurement mechanism. We next establishes a connection between the query accuracy and the total amount $\sum_{i=1}^{n} \varepsilon_i q_i$ of purchased privacy.

We consider the following commonly-used query functions: *Linear predictor* captures a wide range of potential queries over real numbers that include k-nearest neighbours, Nadaranya-Watson weighted average, ridge regression and support vector machines (Dandekar et al., 2012). It is defined as $\varphi(\vec{d}_{\text{gt}}) := \sum_{i=1}^{n} w_i d_i$, where $w_i \neq 0$ is the weight of the data owner s_i. If $d_i \in \{0, 1\}$ and $w_i = 1$ for all $i \in \{1, \ldots, n\}$, the query is called a *count query*. More generally, a linear predictor can be used in a recommender system where w_i represents the similarity between a user s_i and a new user; $\varphi(\vec{d}_{\text{gt}})$ is then the prediction about the new user's data value. We also consider *median query* which aims to find the median value among *mutually distinct positive integers*. In the next lemma, the query φ belongs to one of the three types defined above.

Lemma 2. *For any integer $1 \leq \alpha \leq n/4$ and $\delta \in (0, 1)$, if the query mechanism A is (α, δ)-PAC, then $\alpha \geq \frac{n}{4 \sum_{i=1}^{n} \varepsilon_i q_i} \cdot (\ln \delta - \ln(1 - \delta))$.*

Proof. We prove the equivalent form, if A is (α, δ)-PAC, then $\sum_{i=1}^{n} \varepsilon_i q_i \geq \frac{n(\ln \delta - \ln(1-\delta))}{4\alpha}$. We first consider count query. Recall that this case assumes each data entry d_i is a 0/1-value. We assume for a contradiction that $\sum_{i=1}^{n} \varepsilon_i q_i < \frac{n(\ln \delta - \ln(1-\delta))}{4\alpha}$ and the query mechanism is (α, δ)-PAC. Let $R = \{r \in \mathbb{R} \mid |r - \varphi(\vec{d}_{\text{gt}})| < \alpha\}$. By the definition of (α, δ)-PAC, $\Pr\left(\Phi\left(\vec{d}_{\text{gt}}\right) \in R\right) \geq \delta$.

Sort the data owners so that $\varepsilon_i q_i$ are in ascending order, i.e., $\varepsilon_1 q_1 \leq \varepsilon_2 q_2 \leq \ldots \leq \varepsilon_n q_n$. Consider the first 4α data owners (Note that $4\alpha \leq n$). Clearly,

$$\sum_{i=1}^{4\alpha} \varepsilon_i q_i < \frac{n(\ln \delta - \ln(1 - \delta))}{4\alpha} \frac{4\alpha}{n} = \ln \delta - \ln(1 - \delta).$$

Let $\vec{d}^0 := (d_i)_{i \in I_0}$ and $\vec{d}^1 := (d_i)_{i \in I_1}$ where $I_j = \{1 \leq i \leq 4\alpha \mid d_i = j\}$ for $j \in \{0, 1\}$. Without loss of generality, assume that $|\vec{d}^0| > 2\alpha$. Let $I' \subseteq I_0$ that contains exactly 2α elements, and define a dataset $\vec{d'} := (b_1, \ldots, b_n)$ where $b_i = 1$ if $i \in I'$, and $b_i = d_i$ otherwise. It follows that $\varphi(\vec{d'}) = \varphi(\vec{d}_{\text{gt}}) + 2\alpha$.

It is straightforward to verify by definition of PDP that

$$\Pr\left(\Phi(\vec{d'}) \in R\right) \geq \exp\left(-\sum_{i \in I'} \varepsilon_i q_i\right) \Pr\left(\Phi(\vec{d}_{\text{gt}}) \in R\right)$$

$$> \exp\left(-(\ln \delta - \ln(1 - \delta))\right) \times \delta = \frac{1 - \delta}{\delta} \cdot \delta = 1 - \delta$$

Since $\varphi(\vec{d'}) = \varphi(\vec{d}_{\text{gt}}) + 2\alpha$, by the triangle inequality, we have $\Pr\left(|\Phi(\vec{d'}) - \varphi(\vec{d'})| > \alpha\right) \geq \Pr\left(|\Phi(\vec{d'}) - \varphi(\vec{d}_{\text{gt}})| < \alpha\right) > 1 - \delta$, which contradicts the (α, δ)-PAC assumption.

The proof is similar for the case when φ is the general linear predictor where the data entries are real values. The only difference is that we define the set I' as

$\{1, \dots, 2\alpha\}$ and the dataset $\vec{d'}$ by $b_i = d_i + \frac{1}{w_i}$ for all $i \in I'$ and $b_i = d_i$ otherwise. The case when φ is a median query (over mutually distinct integers) can be found in Appendix A. □

5 SINGLEMINDEDQUERY (SMQ)

Fixing the confidence level δ ($> 1 - \delta$), Lemma 2 asserts that $\sum_{i=1}^{n} \varepsilon_i q_i \in \Omega(n/\alpha)$ is necessary for any query mechanism to achieve (α, δ)-PAC. This suggests the total amount of purchased privacy $\sum_{i=1}^{n} \varepsilon_i q_i$ plays a significant role in determining query accuracy. Moreover, since we concern with ex-interim utility, the term q_i here should be considered in expectation, i.e., $\int_\Theta \sum_{i=1}^{n} \varepsilon_i q_i(\vec{\psi}) f(\vec{\psi}) \, \mathrm{d}\vec{\psi}$. In our query mechanism, we thus aim to *maximise purchased privacy expectation* (PPEM) in order to obtain accurate query results, i.e., we aim to solve the following optimisation problem:

$$\begin{aligned} &\text{maximise} \quad \int_\Theta \sum_{i=1}^{n} \varepsilon_i q_i(\vec{\psi}) f(\vec{\psi}) \, \mathrm{d}\vec{\psi} \\ &\text{such that} \quad (4), (5) \text{ and } (6) \text{ are satisfied} \end{aligned} \tag{8}$$

In the rest of the paper, we describe our data query mechanism, namely *SingleMindedQuery* (SMQ), to solve Problem (8). Note that the problem can be regarded as a *knapsack auction problem* (see (Ensthaler & Giebe, 2014)): Treat B as the *capacity* of the knapsack, ε_i as the *value* and θ_i as the *weight* of the ith item ($1 \le i \le n$).

Definition 5. *A* simple direct mechanism Ψ *consists of allocation rule $q_i(\vec{\psi})$ and payment rule $p_i(\vec{\psi})$ as follows,*

$$q_i(\vec{\psi}) = Q_i(\psi_i) \coloneqq \begin{cases} 1 & \text{if } \psi_i \le \theta_i^* \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

$$p_i(\vec{\psi}) = P_i(\psi_i) \coloneqq \psi_i Q_i(\psi_i) + \int_{\psi_i}^{\overline{\theta_i}} Q_i(s) \, \mathrm{d}s \tag{10}$$

In a simple direct mechanism, each data owner s_i has a take-it-or-leave-it offer with s_i's valuation bounded by a *threshold* θ_i^*. If the reported ψ_i is smaller than the threshold, the data owner will be selected and get a compensation $p_i(\vec{\psi})$, which is higher than ψ_i.

Lemma 3. *Assuming that θ_i^* is independent from the reported valuation ψ_i for all $1 \le i \le n$, a simple direct mechanism Ψ is incentive compatible and individually rational.*

Proof. For IR, suppose $\theta_i \le \theta_i^*$. Then $Q_i(\theta_i) = 1$. By (10), $P_i(\theta_i)$ equals

$$\theta_i Q_i(\theta_i) + \int_{\theta_i}^{\overline{\theta_i}} Q_i(s) \, \mathrm{d}s = \theta_i + \int_{\theta_i}^{\theta_i^*} 1 \, \mathrm{d}s = \theta_i^*, \tag{11}$$

and $U_i(\theta_i|\theta_i) = P_i(\theta_i) - \theta_i Q_i(\theta_i) = \theta_i^* - \theta_i \ge 0$. If $\theta_i > \theta_i^*$, $Q_i(\psi_i) = 0$ which implies $P_i(\theta_i) = 0$ and $U_i(\theta_i|\theta_i) = 0$. In either case, the expected utility of reporting the valuation truthfully is non-negative.

The IC condition can be proved easily by considering the two possible cases of $\psi_i > \theta_i$ and $\psi_i < \theta_i$. The full proof is in Appendix B. □

We would like to define a simple direct mechanism to solve (8). It remains to find the appropriate threshold $\vec{\theta^*}$. For the following lemma, by an *optimal threshold*, we mean a vector $\vec{\theta^*} = (\theta_1^*, \dots, \theta_n^*)$ whose corresponding allocation and payment rules as defined in (9) and (10), respectively, is an optimal solution for (8).

Lemma 4. *The optimal solution to the following optimisation problem* (12) *is an optimal threshold.*

$$\begin{aligned} \max_{\Theta} \quad & \sum_{i=1}^{n} \varepsilon_i F_i(\theta_i^*) \\ \text{such that} \quad & \sum_{i=1}^{n} \theta_i^* F_i(\theta_i^*) = B \\ & \underline{\theta} \le \theta_i^* \le \overline{\theta} \qquad \forall i \in \{1, \dots, n\} \end{aligned} \tag{12}$$

Proof. By substituting (2) the objective function of (8) becomes $\sum_{i=1}^{n} \int_{\underline{\theta}}^{\overline{\theta}} \varepsilon_i Q_i(\psi_i) f_i(\psi_i) \, \mathrm{d}\psi_i$, which, by (9), is

$$\sum_{i=1}^{n} \int_{\underline{\theta}}^{\theta_i^*} \varepsilon_i f_i(\psi_i) \, \mathrm{d}\psi_i = \sum_{i=1}^{n} \varepsilon_i F_i(\theta_i^*).$$

IC and IR are satisfied due to Lemma 3 and the fact that θ_i^* is chosen by solving (12), which is independent from ψ_i. BF is equivalent to $\sum_{i=1}^{n} \theta_i^* F_i(\theta_i^*) \le B$ which can be derived using (3) and (10). Moreover, it is easy to see that (6) is binding, i.e., $\sum_{i=1}^{n} \theta_i^* F_i(\theta_i^*) = B$. Otherwise, we can always increase the value of θ_i^* and select more data owners. See full proof at Appendix C. □

To solve problem (12), take the Lagrange function

$$\begin{aligned} L(\vec{\theta^*}, \lambda, \mu_1, \dots, \mu_n, \gamma_1, \dots, \gamma_n) \coloneqq \sum_{i=1}^{n} \varepsilon_i F_i(\theta_i^*) \\ - \lambda \sum_{i=1}^{n} (\theta_i^* F_i(\theta_i^*) - B) - \mu_i(\theta_i^* - \overline{\theta}) - \gamma_i(\underline{\theta} - \theta_i^*) \\ \forall i \in \{1, \dots, n\}, \end{aligned} \tag{13}$$

where λ, μ_i and γ_i are Lagrange multipliers. Setting the first order derivative to 0, we get:

$$\varepsilon_i f_i(\theta_i^*) - \lambda(F_i(\theta_i^*) + \theta_i^* f_i(\theta_i^*)) + \mu_i - \gamma_i = 0 \; \forall 1 \le i \le n \tag{14}$$

The desirable threshold vector $\vec{\theta^*}$ is the solution to the system that contains (14) and the following conditions:

$$\begin{aligned}
&\mu_i(\underline{\theta} - \theta_i^*) = 0 && \forall i = 1, \ldots, n \\
&\gamma_i(\theta_i^* - \overline{\theta}) = 0 && \forall i = 1, \ldots, n \\
&\sum_{i=1}^{n} \theta_i^* F_i(\theta_i^*) - B = 0 && \\
&\lambda, \mu_1, \ldots, \mu_n, \gamma_1, \ldots, \gamma_n \geq 0 &&
\end{aligned} \tag{15}$$

Our procurement mechanism Ψ takes $\vec{\psi}$ and $\vec{\varepsilon}$ as inputs. It first solves the system above and obtains a threshold vector $\vec{\theta^*}$. Ψ then selects data owners based on this vector: For $1 \leq i \leq n$, select the data owner s_i if s_i's reported valuation ψ_i is lower than θ_i^*. In this case, make a payment of θ_i^* to s_i. Otherwise, s_i is not chosen and the payment is 0. We propose an algorithm to implement the procurement mechanism as shown in Alg. 1.

Algorithm 1 Procurement mechanism Ψ

```
Solve the system (14),(15) to obtain $\theta_i^*$ for $1 \leq i \leq n$.
for $i \in \{1, \ldots, n\}$ do
    if $\psi_i \leq \theta_i^*$ then
        set $q_i := 1$ and pay $p_i := \theta_i^*$;
    else
        set $q_i := 0$ and pay $p_i := 0$.
    end if
end for
```

The next theorem follows from Lemma 4, Karush-Kuhn-Tucker theorem (see (Luenberger, 1997)), and the convexity of Problem (12). See full proof at Appendix D.

Theorem 1. *The procurement mechanism Ψ guarantees to find the optimal solution of Problem* (8).

After data procurement, the data entries of selected data owners form a dataset $\vec{d_{\mathrm{c}}}$, and $P\mathcal{E}$ mechanism is applied to it. SMQ can meet the hard privacy constraints of all data owners. For each s_i, the achieved privacy is denoted as ε_i'. For those who are not selected, $\varepsilon_i' = 0 < \varepsilon_i$; for the selected ones, $\varepsilon_i' \leq \varepsilon_i$ is guaranteed by $P\mathcal{E}$ mechanism.

6 EXPERIMENT SETUP

Through the experiments, we aim to evaluate the performance of SMQ in terms of its query accuracy under different query types, budgets, and dependence relationships between $\vec{\theta}$ and $\vec{\varepsilon}$. We consider three query types, count, median and linear predictor. As performance metric, we compare the mean and 95% confidential interval (CI) of the returned query results against the true query answer $\varphi(\vec{d_{\mathrm{gt}}})$, and use root mean squared error (RMSE) to measure the error.

Datasets $\vec{d_{\mathrm{gt}}}$. We use three real-world datasets, including Adults dataset[2], MovieLense 1M dataset (Harper & Konstan, 2016), and Residential energy consumption survey (RECS) dataset (EIA, 2009). Adults dataset consists of $32,561$ entries, each representing an adult living in the US (Dua & Graff, 2017). It has 15 attributes, including age, income, education, marital status, etc. The RECS dataset has $12,084$ records, each record with 940 attributes, including identifier, region, division, etc. The MovieLense 1M dataset contains the information of $6,040$ audience, $3,952$ movies and $1,000,209$ ratings.

Privacy parameters $\vec{\varepsilon}$ and data valuations $\vec{\theta}$. The three datasets contain no information about privacy attitudes of the data owners, so we generate two sets of random numbers, representing data valuations θ_i and privacy requirements ε_i, respectively. ε_i is a small non-negative number, hence, we restrict the privacy parameter to be bounded by 1. The $\vec{\varepsilon}$ and $\vec{\theta}$ are correlated uniformly distributed random variables in the range $(0, 1)$. They are generated based on a correlation coefficient ρ. As a smaller ε_i denotes a more stringent privacy requirement, negative values of ρ expresses positive correlation. We consider three different dependence relationships between $\vec{\theta}$ and $\vec{\varepsilon}$: (1) *independence*, where $\rho := 0$, (2) *partial positive correlation*, where $\rho := -0.5$, and (3) *perfect positive correlation*, where $\rho := -1$. In scenario (1), $\vec{\theta}$ and $\vec{\varepsilon}$ are irrelevant. In other words, having a high privacy requirement does not necessarily mean that this data owner attaches high valuation on her private data. In contrast, in scenarios (2) and (3), a data owner with high privacy requirement tends to have high data valuation.

Budget B. The broker has a budget $B \leq \overline{\theta}n$ for data procurement. We investigate the performance of the SMQ under different budgets, $B = \{0.1\overline{\theta}n, 0.2\overline{\theta}n, \ldots, 0.9\overline{\theta}n\}$.

Query types φ. For count query, we use the income attribute of the Adults dataset, gender attribute of the MovieLense 1M dataset and the total site electricity usage of the RECS dataset. The count queries ask: *How many adults have income higher than* 50 *k? How many female audiences? And how many households consume more than 10 thousand kwh?*

For median query, we use the age attribute of the Adults dataset and the MovieLense 1M dataset and the total site electricity usage (integer-valued) of the RECS dataset.

For the linear predictor, we use the data in the last row to represent s_{n+1} and the data in the other rows to represent existing data owners. We choose the data of one attribute as $\vec{d_{\mathrm{gt}}} = (d_1, \ldots, d_n)$ and the data

[2] https://archive.ics.uci.edu/ml/datasets/Adult

of the other attributes as the profile, denoted as $Y = (y_1, \dots, y_n, y_{n+1})$, where y_{n+1} is the profile of s_{n+1}, and used to calculate the similarities. Here, we use a common measure, *cosine similarity*, to quantify the similarity between s_i and s_{n+1}, i.e., for each $i \in \{1, \dots, n\}$, $w_i := \cos\_sim(s_i, s_{n+1}) = \frac{y_i \cdot y_{n+1}}{\|y_i\|\|y_{n+1}\|}$. We use the Adult dataset, the MovieLense 1M dataset and the RECS dataset, to predict whether a new individual's income is higher than 50 k, whether a new individual likes the movie and whether a new household consumes more than 10 thousand kwh, respectively.

In SMQ, the set $\mathrm{Range}_\varphi(\vec{d_c})$ is constructed differently for different query types. For instance, $\mathrm{Range}_\varphi(\vec{d_c})$ can be enumerated as $\frac{\sum_{i:s_i \in \mathcal{S}} q_i}{\sum_{i:s_i \in \mathcal{S}_c} q_i} t$ in a count query; as t in a median query; and as $\frac{\sum_{i:s_i \in \mathcal{S}} w_i}{\sum_{i:s_i \in \mathcal{S}_c} w_i} t$ in a linear predictor, where $t \in \mathrm{Range}(\varphi)$ and $\mathcal{S}_c$ is the set of the selected data owners.

Baselines. For count and median queries, we compare SMQ with FQ (Ghosh & Roth, 2011). For linear predictor, we compare with FIP (Dandekar et al., 2012).

FQ. FQ achieves $\frac{1}{n-k}$-DP and pays equally for all k selected data owners. FQ first computes privacy valuation $v_i := \theta_i / \varepsilon_i$ for each s_i. It then selects the k data owners with the least v_i where k is the largest integer satisfying $kv_k \leq B$. FQ then pays each selected data owner $\min\{B/k, v_{k+1}/(n-k)\}$ as compensation. The query answer is $r := \sum_{i=1}^{k} d_i + (n-k)/2 + Lap(n-k)$, where $Lap(n-k)$ is Laplace noise with variance $(n-k)$.

We also adjust the original FQ formulation for median queries. The mechanism follows the same allocation rule and pricing rule as those for count queries. The query answer is $r := \varphi(\vec{d_c}) + Lap(\Delta\varphi(n-k))$, where $\Delta\varphi$ is the sensitivity of φ on $\vec{d_c}$ (Dwork et al., 2006).

FIP. FIP firstly sorts v_i in ascending order. If there exists an s_{i^*} whose weight w_{i^*} satisfies $w_{i^*} > \sum_{i:\, \mathcal{S} \setminus \{s_{i^*}\}} w_i$, FIP only selects s_{i^*}. Otherwise, FIP selects the first k data owners subject to B, i.e., $\frac{B}{\sum_{i=1}^{k} w_i} \geq \frac{v_k}{\sum_{i=k+1}^{n} w_i}$ and pays $p_i = w_i \min\left\{\frac{B}{\sum_{i=1}^{k} w_i}, \frac{v_{k+1}}{\sum_{i=k+1}^{n} w_i}\right\}$ for each selected data owner. Assuming the range of the dataset is known as $[\underline{d}, \overline{d}]$, FIP returns $r := \sum_{i=1}^{k} w_i d_i + \frac{1}{2}(\underline{d} + \overline{d}) \sum_{i=k+1}^{n} w_i + Lap\left((\overline{d} - \underline{d}) \sum_{i=k+1}^{n} w_i\right)$.

As neither FQ nor FIP collects data owners' privacy requirements, to meet the privacy requirements in median and count queries, we need to ensure that FQ does not select data owners who have $\varepsilon_i \leq 1/(n-k)$. Also, FIP sets the privacy level $\varepsilon_i := w_i / \sum_{i: \in \mathcal{S} \setminus \mathcal{S}_c} w_i$, which is used for both SMQ and FIP, in order to make a fair comparison between these two mechanisms.

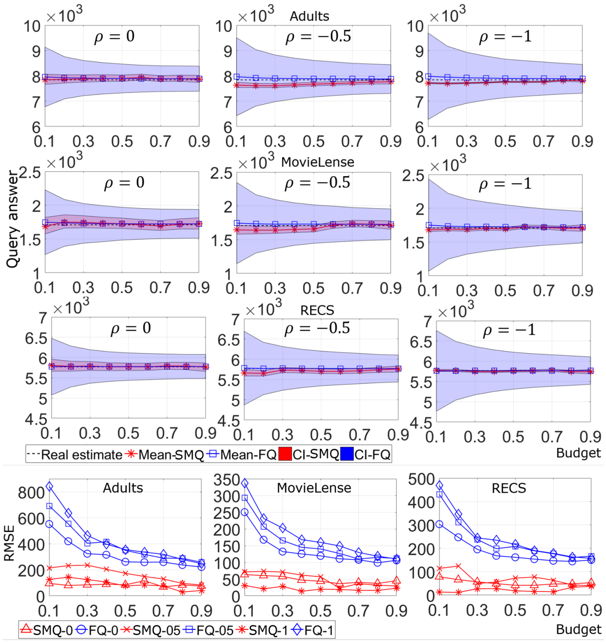

Figure 2: CI and RMSE of SMQ and FQ for count query. The top three rows show the mean and 95% CI of the query answers of SMQ and FQ, where each row denotes a different dataset, and each column denotes a dependence relationship between $\vec{\theta}$ and $\vec{\varepsilon}$. The last row shows RMSE for different datasets. The horizontal axis indicates budget between $0.1\overline{\theta}n$ and $0.9\overline{\theta}n$.

For each query type, we test two mechanisms under different budgets and different dependence relationships between $\vec{\theta}$ and $\vec{\varepsilon}$. Under each experiment setup, 500 trials are carried out, and the average, the 95% CI and the RMSE for each mechanism are reported. The experiment setups are summarised in Table 1.

Table 1: Experiment Setups

Query φ	Count, median, linear predictor
Dataset $\vec{d}_{\mathrm{gt}}$	Adults, RECS, MovieLense 1M
Data valuations $\vec{\theta}$	$\vec{\theta} \sim U(0, 1)$
Privacy parameters $\vec{\varepsilon}$	$\vec{\varepsilon} \sim U(0, 1)$
Correlation ρ	$\{0, -0.5, -1\}$
Budget B	$\{0.1\overline{\theta}n, 0.2\overline{\theta}n, \dots, \overline{\theta}n\}$
Mechanism	SMQ, FQ, FIP mechanisms

7 RESULTS

Experiment 1: Count query. We apply SMQ and FQ to count queries on the three datasets. The two mechanisms demonstrate considerably different results on their allocation and payments. While FQ tends to select data owners with the lowest privacy valuations (regardless of their privacy requirements), SMQ tends to choose those data owners who have larger ε_i. This is because θ_i^* is determined by ε_i and are different across data owners. Furthermore, FQ compensates the selected data owners uniformly while SMQ sets a price of $p_i = \theta_i^*$ which varies among data owners.

We compare the mechanisms in terms of accuracy under different budgets and different dependence relationships between $\vec{\theta}$ and $\vec{\varepsilon}$; See Fig. 2. As B increases, CI narrows down and RMSE displays a descending trend for both SMQ and FQ. SMQ outperforms FQ in terms of accuracy for all datasets by a large margin. The CI for SMQ is significantly narrower and the RMSE for the SMQ is significantly lower than those for FQ across all cases.

The results also show that SMQ's performance improves as a stronger dependence relationship exists, whilst FQ's performance worsens, as CI becomes narrower and the RMSE becomes smaller for SMQ when ρ decreases from 0 to -0.5 and -1. Those for FQ show an opposite trend. When ρ becomes smaller, the negative correlation between θ_i and ε_i is higher. In other words, when θ_i is large, ε_i is small, which makes v_i large. As a result, under the same budget, the number k is smaller and the variance $n - k$ of Laplace noise is larger, which worsens the performance of FQ.

Experiment 2: Median query. We implement SMQ and FQ on median queries. The allocation and payment results are similar to those for count queries. In terms of accuracy, SMQ significantly outperforms FQ for median queries. As shown in Figure 3, the error for SMQ is negligible comparing to FQ. For the Adults and the MovieLense 1M datasets, the RMSE for the SMQ is zero. As for the RECS dataset, since the range is comparatively larger, the results are less accurate, but much better than the results for FQ. Also, under different value of ρ, SMQ returns reliable results while FQ performs even worse when ρ becomes smaller.

Experiment 3: Linear predictor. We implement SMQ and FIP for linear predictors. The allocation results of FIP is similar to FQ, where the data owners with low privacy valuations are chosen. The results show that for the Adult dataset, the number of data owners chosen by SMQ is larger than that by FIP, while this number is similar for the other two datasets. Consistent with the previous experiments, SMQ outperforms FIP in most cases with respect to accuracy, with the exception of when $B > 0.5\overline{\theta}n$. As shown in Figure 4, for the Adult dataset, the CI is narrower and the RMSE is smaller for SMQ . For the MovieLense 1M dataset and RECS dataset, when the budget is low, the performance of SMQ is better than FIP. When the the budget is high, SMQ outputs slightly less accurate results.

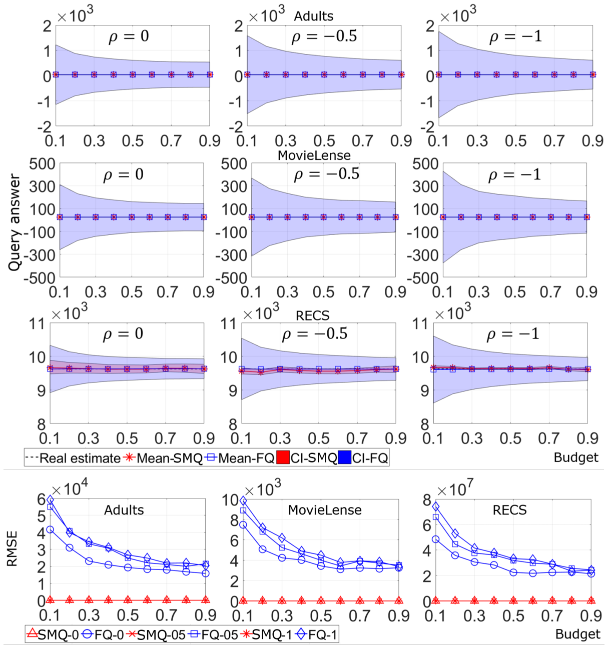

Figure 3: CI and RMSE of SMQ and FQ for median query. All setting are the same as in Fig. 2.

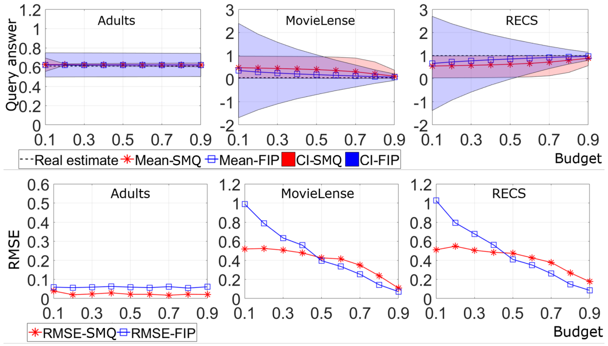

Figure 4: CI and RMSE of SMQ and FIP for linear predictor. The top row shows the mean and 95% CI of the query answers of SMQ and FIP. The second row shows RMSE

8 CONCLUSION

We consider private data query problem where the data are held by single-minded data owners. We propose the data query mechanism SMQ that satisfies IC, IR, BF and $\vec{\varepsilon}$-PDP for every $i \in \{1, \dots, n\}$. The empirical results show that SMQ effectively improves the query accuracy with the same budget than existing mechanisms. An assumption in the work assumes that all data owners truthfully announce their privacy protection requirements. As future work, we will explore the data broker problem where hidden information comes from data valuation and privacy protection requirements. We also could study the problem with more sophisticated query types, e.g., queries whose output are beyond real values.

References

Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE security & privacy*, *3*(1), 26–33.

Alaggan, M., Gambs, S., & Kermarrec, A.-M. (2016). Heterogeneous differential privacy. *Journal of Privacy and Confidentiality*, *7*(2).

Berendt, B., Günther, O., & Spiekermann, S. (2005). Privacy in e-commerce: stated preferences vs. actual behavior. *Communications of the ACM*, *48*(4), 101–106.

Börgers, T. (2015). *An introduction to the theory of mechanism design*. Oxford University Press, USA.

Brankovic, L., & Estivill-Castro, V. (1999). Privacy issues in knowledge discovery and data mining. In *Australian institute of computer ethics conference* (pp. 89–99).

Conitzer, V. (2009). Prediction markets, mechanism design, and cooperative game theory. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence (uai-09)* (pp. 101–108).

Cummings, R., Ligett, K., Roth, A., Wu, Z. S., & Ziani, J. (2015). Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 conference on innovations in theoretical computer science* (pp. 317–324).

Dandekar, P., Fawaz, N., & Ioannidis, S. (2012). Privacy auctions for recommender systems. In *Proceedings of the 8th international conference on internet and network economics* (pp. 309–322).

Dua, D., & Graff, C. (2017). Uci machine learning repository (2017). *URL http://archive. ics. uci. edu/ml*.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference* (pp. 265–284).

EIA, U. (2009). *2009 residential energy consumption survey dataset.* Retrieved Oct 22, 2019, from `https://www.eia.gov/consumption/residential/data/2009/index.php?view=microdata`

Ensthaler, L., & Giebe, T. (2014). Bayesian optimal knapsack procurement. *European Journal of Operational Research*, *234*(3), 774–779.

Fleischer, L. K., & Lyu, Y.-H. (2012). Approximately optimal auctions for selling privacy when costs are correlated with data. In *Proceedings of the 13th ACM conference on electronic commerce* (pp. 568–585).

Ghosh, A., & Roth, A. (2011). Selling privacy at auction. In *Proceedings of the 12th ACM conference on electronic commerce* (pp. 199–208).

Harper, F. M., & Konstan, J. A. (2016). The movielens datasets: History and context. *ACM transactions on interactive intelligent systems (tiis)*, *5*(4), 19.

Jorgensen, Z., Yu, T., & Cormode, G. (2015). Conservative or liberal? personalized differential privacy. In *2015 IEEE 31st international conference on data engineering* (pp. 1023–1034).

Laudon, K. C. (1996). Markets and privacy. *Communications of the ACM*, *39*(9), 92–104.

Li, H., Xiong, L., Ji, Z., & Jiang, X. (2017). Partitioning-based mechanisms under personalized differential privacy. In *Pacific-asia conference on knowledge discovery and data mining* (pp. 615–627).

Ligett, K., & Roth, A. (2012). Take it or leave it: Running a survey when privacy comes at a cost. In *International workshop on internet and network economics* (pp. 378–391).

Luenberger, D. G. (1997). *Optimization by vector space methods*. John Wiley & Sons.

Mas-Colell, A., Whinston, M. D., Green, J. R., et al. (1995). *Microeconomic theory* (Vol. 1). Oxford university press New York.

Nget, R., Cao, Y., & Yoshikawa, M. (2017). How to balance privacy and money through pricing mechanism in personal data market. *arXiv preprint arXiv:1705.02982*.