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

Appendix A.

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 that 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, $\Pr\left(\Phi\left(\vec{d}_{\text{gt}}\right) \in R\right) \geq \delta$.

Assume, w.l.o.g., that $\varepsilon_i q_i$ are sorted 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

$$\begin{aligned}\Pr\left(\Phi(\vec{d'}) \in R\right) &\geq \exp\left(-\sum_{i\in I'}\varepsilon_i q_i\right)\Pr\left(\Phi(\vec{d}_{\text{gt}}) \in R\right)\\ &> \exp\left(-(\ln\delta - \ln(1-\delta))\right) \times \delta\\ &= \frac{1-\delta}{\delta}\cdot\delta = 1-\delta\end{aligned}$$

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 $\{1,\ldots,2\alpha\}$ 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.

For the case when φ is a median query. Assume $d_1, d_2, \ldots, d_n$ are distinct positive integers. We only deal with the case when n is odd (the case when n is even can be proven in a similar way). Let m denote the median among $d_1,\ldots,d_n$. Let $I_0 := \{i \mid d_i < m\}$ and $I_1 := \{i \mid d_i > m\}$. Suppose, w.l.o.g., that $\sum_{i\in I_0}\varepsilon_i q_i < \frac{n(\ln\delta-\ln(1-\delta))}{8\alpha}$. Let $k := |\{i \mid m \leq d_i < m+2\alpha\}|$. Note that by mutual distinction of data values, $k \leq 2\alpha$. For every $i \in I_0$, put i into H if the data owner s_i's privacy requirement ε_i is among the smallest k among data owners in I_0. Clearly, $\sum_{i\in H}\varepsilon_i q_i \leq \frac{n(\ln\delta-\ln(1-\delta))}{4\alpha}\frac{2\alpha}{n} < \ln\delta - \ln(1-\delta)$. Let $d_{\max} := \max\{d_1,\ldots,d_n\}$. Define a new dataset $\vec{d'} := (b_1,\ldots,b_n)$ by $b_i = d_i + d_{\max}$ if $i \in H$; and $b_i = d_i$ otherwise. It then follows that the median of $\vec{d'}$ is at least $m+2\alpha$ and thus $\varphi(\vec{d'}) \geq \varphi(\vec{d}_{\text{gt}})+2\alpha$. By PDP of Φ, we have $\Pr(|\Phi(\vec{d'})-\varphi(\vec{d}_{\text{gt}})| < \alpha) > 1-\delta$. 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 accuracy assumption. □

Appendix B.

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

Proof. For IR, suppose $\theta_i \leq \theta_i^*$. Then $Q_i(\theta_i) = 1$. By

(10), $P_i(\theta_i)$ equals

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

and $U_i(\theta_i|\theta_i) = P_i(\theta_i) - \theta_i Q_i(\theta_i) = \theta_i^* - \theta_i \geq 0$. 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.

For IC, note that θ_i^* for all $i \in \{1, \ldots, n\}$ is independent from the reported valuation. When data owners report their valuations untruthfully, there are two cases:

Case (1) Suppose s_i reports a valuation $\psi_i > \theta_i$.

a. if $\theta_i < \psi_i \leq \theta_i^*$, $U_i(\psi_i|\theta_i) = U_i(\theta_i|\theta_i) = \theta_i^* - \theta_i$.

b. if $\theta_i \leq \theta_i^* < \psi_i$, $U_i(\theta_i|\theta_i) = \theta_i^* - \theta_i \geq 0 = U_i(\psi_i|\theta_i)$.

c. if $\theta_i^* < \theta_i < \psi_i$, $U_i(\psi_i|\theta_i) = U_i(\theta_i|\theta_i) = 0$.

Case (2) Suppose s_i reports a valuation $\psi_i < \theta_i$.

a. if $\psi_i < \theta_i \leq \theta_i^*$, $U_i(\psi_i|\theta_i) = U_i(\theta_i|\theta_i) = \theta_i^* - \theta_i$.

b. if $\psi_i \leq \theta_i^* < \theta_i$, $U_i(\psi_i|\theta_i) = \theta_i^* - \theta_i < 0 = U_i(\theta_i|\theta_i)$.

c. if $\theta_i^* < \psi_i < \theta_i$, $U_i(\psi_i|\theta_i) = U_i(\theta_i|\theta_i) = 0$.

The above argument shows that each data owner can maximise her expected utility by truthfully reporting the valuation. □

Appendix C.

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

Proof. Firstly, since the threshold θ_i^* is determined by solving (12), it is independent from ψ_i. By Lemma 3, IC and IR constraints are satisfied by allocation rule (9) and payment rule (10).

For the objective function, by substituting (2) the objective function 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

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

For BF, by (3) the left hand side of the constraint (6) is

$$\begin{aligned}
&\sum_{i=1}^n \int_{\underline{\theta}}^{\overline{\theta}} P_i(\psi_i) f_i(\psi_i)\,\mathrm{d}\psi_i \\
=&\sum_{i=1}^n \int_{\underline{\theta}}^{\overline{\theta}} \left(\psi_i Q_i(\psi_i) + \int_{\psi_i}^{\overline{\theta}} Q_i(s)\,\mathrm{d}s\right) f_i(\psi_i)\,\mathrm{d}\psi_i \quad \text{by (10)} \\
=&\sum_{i=1}^n \int_{\underline{\theta}}^{\theta_i^*} \theta_i^* f_i(\psi_i)\,\mathrm{d}\psi_i = \sum_{i=1}^n \theta_i^* F_i(\theta_i^*)
\end{aligned}$$

Thus (6) is equivalent to $\sum_{i=1}^n \theta_i^* F_i(\theta_i^*) \leq B$. 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. □

Appendix D.

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

Proof. By Lemma 4, we only need to show that the procurement mechanism Ψ solves Problem (12). Define B_i as $\theta_i^* F_i(\theta_i^*)$. The first constraint in (12) then becomes $\sum_{i=1}^n B_i = B$, which is affine in terms of B_i.

Also, since any B_i corresponds to a θ_i^*, we can view θ_i^* as a function of B_i and thus write $B_i = \theta_i^*(B_i) F_i(\theta_i^*(B_i))$. The derivative in terms of B_i is

$$1 = \theta_i^{*'}(B_i) F_i(\theta_i^*(B_i)) + \theta_i(B_i)^* f_i(\theta_i^*(B_i)) \theta_i^{*'}(B_i)$$

Reorganise the equation, we can get

$$f_i(\theta_i^*)\theta_i^{*'} = \frac{1}{\frac{F_i(\theta_i^*)}{f_i(\theta_i^*)} + \theta_i^*}.$$

Because of the regularity assumption, the denominator is strictly increasing. Thus, $f_i(\theta_i^*)\theta_i^{*'}$ is strictly decreasing. Furthermore, the derivative of the objective function in terms of B_i is

$$\sum_{i=1}^n \varepsilon_i f_i(\theta_i^*(B_i)) \theta_i^{*'}(B_i).$$

It is strictly decreasing as well. Therefore, the objective is to maximise a concave function. The above arguments asserts the convexity of Problem (12).

Since Problem (12) is convex and the vector $\vec{\theta^*}$ satisfies conditions (14) and (15), Karush-Kuhn-Tucker theorem (see (Luenberger, 1997)) implies that $\vec{\theta^*}$ is the optimal solution to (12). □