Pricing for Revenue Maximization in IoT Data Markets: An Information Design Perspective

Weichao Mao, Zhenzhe Zheng, Fan Wu
Shanghai Key Laboratory of Scalable Computing and Systems, Shanghai Jiao Tong University, China
{maoweichao,zhengzhenzhe}@sjtu.edu.cn, {fwu}@cs.sjtu.edu.cn

***Abstract*—Data is becoming an important kind of commercial good, and many online data marketplaces are set up to facilitate the trading of data. However, most existing data market models and the corresponding pricing mechanisms are simple, and fail to capture the unique economic properties of data. In this paper, we first characterize the distinctive features of IoT data as a commodity, and then present a new IoT data market model from an information design perspective. We further propose a family of data pricing mechanisms for revenue maximization under different market settings. Our *MSimple* mechanism extracts full surplus from the market for the model with one type of buyer. When multiple types of buyers coexist, our *MGeneral* mechanism optimally solves the problem of revenue maximization by formulating it as a polynomial size convex program. For a more practical setting where buyers have bounded rationality, we design *MPractical* mechanism with a tight logarithmic approximation ratio. We evaluate our pricing mechanisms on a real-world ambient sound dataset. Evaluation results show our pricing mechanisms achieve good performance and approach the revenue upper bound.**

I. INTRODUCTION

Data is becoming a commodity. It has tremendous value to both its owner and other parties who want to integrate it into their services. A number of online data marketplaces are emerging to enable data sharing and trading over the Internet, facilitating various data-based services, such as targeted advertising and business decision making. For example, Gnip [1] aggregates and sells social media data from Twitter before the General Data Protection Regulation(GDPR); Xignite [2] vends real-time financial data; and Here [3] trades tracking and positioning data for location-based advertising.

Among various online data marketplaces, several companies [4]–[6] focus on data from Internet of Things (IoT). IoT data marketplaces allow different stakeholders to share the sensor network infrastructure, and facilitate many city services such as waste management and environment monitoring [7], traffic jam avoidance [8], smart agriculture and weather forecasting [9], and etc. Data from widely deployed sensors are more accurate and granular than the coarse-grained data from the national weather or traffic services. For example, IOTA [4] is a blockchain-based data marketplace for aggregating and selling IoT data, and DataBroker DAO [6] is a peer to peer marketplace for IoT sensor data.

One major drawback of existing data marketplaces is that the pricing mechanism for data trading is still very primary: the data brokers either adopt a fixed price mechanism [4], or choose to negotiate with the buyer offline [2]. Although there are some existing work dedicated to designing flexible data pricing mechanisms, most of them only support structured and relational data [10], [11], fail to capture the unique features of IoT data as a commodity, and ignore the economic objective of the data seller. In this work, we aim to analyze the unique economic properties of IoT data, and then design proper pricing mechanisms to achieve revenue maximization.

In the following, we list four properties of IoT data as a commodity that could heavily influence the trading model and pricing mechanism design.

• First, IoT data generally falls into the category of digital goods, and can be reproduced with a neglectable marginal cost. Due to such cost feature, a buyer can easily generate a new copy of the raw data, and resell it at a lower price, making the data easy to pirate. Traditional copyright techniques can hardly resist such piratical behavior. To resolve this problem, we propose that the data seller should sell data services instead of raw data. This sale strategy can also preserve the privacy of data owner to some extent. The data services could be the mean, median, and maximum values of aggregated data, or the results of performing data mining techniques on the data.

• Second, the valuation of IoT data does not necessarily depend on data volume, but depend on the amount of information it provides. This property differentiates IoT data from traditional commodities including digital goods. A large volume of noisy data from low standard sensors could have less valuation than a small set of precise data from a professional sensor. One fundamental question in data trading is: how to quantify the valuation of data in the market? In the context of IoT applications, based on the information extracted from sensed data, buyers always take actions to earn certain utility. Therefore, we measure buyer's valuation towards a set of data by how the data guides the buyer's action. For example, suppose you are going out on a sunny day and you do not consider it necessary to take an umbrella with you. In this case, a large set of humidity sensor data confirming a long sunny day does not generate much valuation to you. On the contrary, a set of sensor data forecasting a heavy rain one hour later generates high valuation to you, as it changes your action

This work was supported in part by the National Key R&D Program of China 2018YFB1004703, in part by China NSF grant 61672348, 61672353, and 61472252. The opinions, findings, conclusions, and recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies or the government.

F. Wu is the corresponding author.

by guiding you to take an umbrella with you.

• Third, the price of data might have correlation with the information behind the data, and thus releasing the price of data could leak the information of data. Suppose the data seller in the previous "umbrella" example sets $1 and $2 for the "sunny" data set and the "rainy" data set, respectively. A buyer could distinguish these two data sets through only observing the corresponding prices, as the rainy data set contains more information and has a higher price. Since buyers are willing to pay a higher price for the rainy data set, the seller would lose revenue if he reduces the price of this data set to $1. We therefore argue that, in order to avoid information leakage before data trading, the seller should not set explicit prices related to data content, but instead should declare prices independent of the specific values of data. For example, one possible qualified pricing scheme is: charge $1 for a weather data set with 75% accuracy, and charge $2 for 95% accuracy. Such content-independent pricing schemes do not leak information about the actual data values.

• Fourth, time-sensitive IoT applications require the price of IoT data should be determined before data is actually generated. In many real world use cases of IoT data, the data buyer requires the data stream to be fed in real-time [12], and thus it is impractical to calculate the price in an online manner. Most of existing work cannot handle this unique feature of IoT data, as they always assume the data is sold after it is collected, structured or modeled [10], [11]. Considering the valuation of data is highly sensitive to the timing, the commodity being sold in IoT data marketplaces should not be the data itself, but rather the permission of data access in a future period of time. This feature of IoT data raises a challenging problem: how do sellers persuade buyers to purchase the data when they still have not collected data?

Besides the four features mentioned above, the revenue maximization problem has many additional challenges. As the seller does not know the valuation of each individual buyer, he has to determine the price of data under incomplete information. Moreover, with various types of buyers in the market, an optimal pricing mechanism should perform market segmentation or price discrimination among different buyers. Under such flexible pricing mechanisms, the data seller should manage to preclude the potential strategic behavior of buyers.

Jointly considering the previous challenges, in this paper, we present a market model of trading IoT data from an information design perspective that captures the aforementioned unique features of IoT data. First, the seller in our model provides data services to buyers by sending various signals, rather than feeding raw data. Second, we define the valuation of data as buyer's utility increment due to the action change after buying data. Finally, the seller designs and publishes the pricing schemes before actually collecting the data, and incentivizes buyers to purchase data by giving them high expected utility increments. This timing enables the trading data to be collected in the future, and ensures that prices never leak information about the actual data values.

We summarize our contributions in this paper as follows:

• First, we characterize four unique features of IoT data as a commodity that differentiate IoT data from traditional goods. We present a market model from an information design perspective that fully captures these features.

• Second, we design revenue-maximizing pricing mechanisms under different market settings. We first consider a simple setting where only one type of buyer exists in the market, and propose *MSimple* mechanism that extracts full surplus from the market. We then consider the general setting where different types of buyers coexist in the market. We present *MGeneral* mechanism to this setting, and prove there exists a polynomial time solution by formulating the problem of revenue maximization as a convex program. We further present *MPractical* mechanism to a more practical setting where buyers have bounded rationality. We prove *MPractical* achieves logarithmic approximation ratio towards the optimal revenue, which is the upper bound of any mechanism of constant size.

• Finally, we evaluate our pricing mechanisms on a real-world ambient sound dataset. We test the influence of different parameters in the market model, and show that our mechanisms achieve good performance.

The rest of the paper is organized as follows. In Section II, we introduce our market model and necessary notations. In Section III, we present our pricing mechanisms to different market settings. We evaluate our pricing mechanisms in Section IV. In Section V, we briefly review related work in the literature. Finally, we conclude the paper in Section VI.

II. Preliminaries

We consider the intersection between a data seller and multiple data buyers. The data commodity in the IoT data market is the *state* of nature, denoted by a random variable ω drawn from a sample space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. The random variable ω could denote a particular numerical value. For example, ω can be the mean value of a set of noise sensors near street, and correspondingly Ω is a discrete set of numerical values for possible noise levels. The random variable ω could also denote the data service extracted from raw data. For example, the seller can aggregate data from various sources–street noise sensors, traffic camera videos, crowdsourced pedestrian traces–to analyze the traffic condition of a certain street. The analytical result is sold to buyers as a data service. In this case the nature state ω is chosen from a binary set $\Omega = \{Crowded, NotCrowded\}$.

The seller trades the data through publishing a *menu* $\mathcal{M} = \{(I, t_I)\}$, which contains multiple pricing schemes. Each buyer chooses a pricing scheme (I, t_I) that maximizes her expected utility, which will be defined later. Each pricing scheme contains an *experiment*[1] I and a corresponding *price* t_I. An experiment $I = \{S, P\}$ contains a set S of possible *signals*[2], and an $n \times |S|$ right probability matrix $P = [\ p_{ij}\ ]$,

[1] An experiment is also called an information structure in the literature.

[2] We abstract different responses from the seller as different signals. In the context of IoT data market, reporting different probabilities of precipitation to the buyer can be regarded as sending different signals.

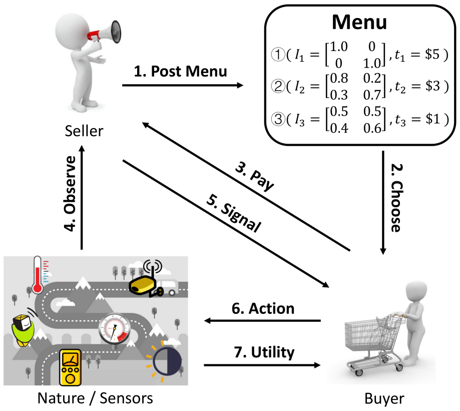

Menu
①($I_1 = \begin{bmatrix}1.0 & 0\\ 0 & 1.0\end{bmatrix}, t_1 = \$5$)
②($I_2 = \begin{bmatrix}0.8 & 0.2\\ 0.3 & 0.7\end{bmatrix}, t_2 = \$3$)
③($I_3 = \begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\end{bmatrix}, t_3 = \$1$)
1. Post Menu
Seller
2. Choose
3. Pay
4. Observe
5. Signal
6. Action
7. Utility
Nature / Sensors
Buyer

Fig. 1. Data trading process.

$1 \le i \le n$, $1 \le j \le |S|$, where $0 \le p_{ij} \le 1$ and $\sum_{j=1}^{|S|} p_{ij} = 1$. The interpretation of p_{ij} is the probability that the seller sends signal $s_j \in S$ to the buyer when the true nature state is ω_i, i.e., $p_{ij} = \Pr(s_j \mid \omega_i)$.

Consider two special types of experiment here: full-information experiment $\bar{I}$ and no-information experiment $\underline{I}$. In the full information case, we assume that $|S| = n$ and P is a diagonal matrix of size $n \times n$. In such case, the seller directly tells the buyer his entire knowledge about the nature state. Specifically, once the seller observes the nature state as ω_i, he will always send signal s_i to the buyer. From the buyer's perspective, upon receiving signal s_i, she is fully confident that the true nature state is ω_i. In the no-information experiment $\underline{I}$, the seller uniformly selects a signal from S and sends it to the buyer, regardless of the true nature state, i.e., $p_{ij} = 1/|S|$ for all $1 \le i \le n, 1 \le j \le |S|$. The buyer gains no information from this experiment, and thus the no-information experiment can be used to fully obfuscate the nature state.

The buyer is uncertain about the true state of nature, and seeks to buy data (or data services) from the seller to supplement her knowledge. We assume the buyer has a prior estimation of the nature state before buying data. The buyer may have previously bought data from the same sensors, and this relatively out-of-date data can help her form a good estimation, due to data correlation in time dimension. We denote the prior distribution for the random variable ω as $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \Delta\Omega$,[3] where $0 \le \theta_i \le 1$ and $\sum_{i=1}^{n} \theta_i = 1$. The parameter θ_i denotes the probability that the nature state is ω_i, i.e., $\theta_i = \Pr(\omega_i)$. We also call the prior distribution θ as the private *type* of buyer. We assume that the type θ is drawn from a finite set Θ with an independent and identical distribution $F(\theta) \in \Delta\Theta$. We further assume the cumulative distribution function $F(\theta)$ is public information.

In many IoT applications, the buyer usually faces a decision problem, and has to choose an action a from a finite set A, based on her perception over the nature state. Let $u_\theta(\omega, a)$ denote the *utility* of the buyer with type θ when the nature state is ω and action a is taken. Without purchasing data from the seller, the buyer has to base her decision only on her prior estimation θ, and the expected utility is

$$u(\theta) = \max_a \mathbb{E}_\omega \left[u_\theta(\omega, a)\right] = \max_a \sum_{i=1}^{n} \theta_i u_\theta(\omega_i, a). \quad (1)$$

After receiving signal s_j from the seller, the buyer θ updates her estimation of the nature state using Bayes' rule:

$$\Pr(\omega_i \mid s_j) = \frac{\Pr(s_j \mid \omega_i)\Pr(\omega_i)}{\Pr(s_j)} = \frac{p_{ij} \cdot \theta_i}{\sum_{k=1}^{n} p_{kj} \cdot \theta_k}, \quad (2)$$

and her expected utility turns to

$$\begin{aligned} u(\theta, s_j) &= \max_a \mathbb{E}_\omega \left[u_\theta(w, a) \mid s_j\right] \\ &= \max_a \sum_{i=1}^{n} \Pr(\omega_i \mid s_j)\, u_\theta(\omega_i, a). \end{aligned} \quad (3)$$

Given an experiment I, from buyer θ's point of view, the probability of receiving signal s_j is

$$\Pr(s_j) = \sum_{i=1}^{n} \Pr(\omega_i) \Pr(s_j \mid \omega_i) = \sum_{i=1}^{n} \theta_i p_{ij}.$$

The buyer's expected utility after buying experiment I is

$$u(\theta, I) = \sum_{j=1}^{|S|} \Pr(s_j) u(\theta, s_j). \quad (4)$$

Therefore, combining the four equations above, buyer θ's *utility increment* for buying experiment I is

$$\begin{aligned} v(\theta, I) = u(\theta, I) - u(\theta) &= \sum_{j=1}^{|S|} \Pr(s_j) u(\theta, s_j) - u(\theta) \\ &= \sum_{j=1}^{|S|} \left(\sum_{i=1}^{n} \theta_i p_{ij}\right) \left(\max_a \sum_{i=1}^{n} \frac{p_{ij}\theta_i}{\sum_{k=1}^{n} p_{kj}\theta_k} u_\theta(\omega_i, a)\right) \\ &\quad - \max_a \sum_{i=1}^{n} \theta_i u_\theta(\omega_i, a) \\ &= \sum_{j=1}^{|S|} \max_a \sum_{i=1}^{n} \theta_i p_{ij} u_\theta(\omega_i, a) - \max_a \sum_{i=1}^{n} \theta_i u_\theta(\omega_i, a). \end{aligned}$$

We assume buyer θ is willing to buy experiment I if and only if the price t_I is no larger than her utility increment. More specifically, we have the following **Individual Rationality** (I.R.) constraint:

$$v(\theta, I) - t_I \ge 0, \quad \forall \theta \in \Theta. \quad \text{(I.R.)}$$

We use Figure 1 to summarize the whole data trading process. The timing of this process is as follows: First, the seller designs a menu $\mathcal{M} = \{(I, t_I)\}$ and posts it to the public. Second, the buyer θ chooses a pricing scheme (I, t_I) with the largest utility increment $v(\theta, I)$ from the menu, and pays the corresponding price t_I. Third, the true nature state ω is realized and revealed to the seller. The seller sends a signal s_j to the buyer following the rule in the selected experiment I. Finally, after receiving the signal s_j, the buyer chooses an action a

[3] $\Delta\Omega$ denotes the probability distributions over Ω.

with the maximum expected utility according to Equation (3). The buyer's utility $u_\theta(\omega, a)$ is then realized.

We remark on four facts about our data market model. First, the seller does not sell raw data to the buyer, but instead extracts different signals from data. Second, we measure the valuation of data as buyer's utility increment due to the action change after buying data, which is independent of the data volume. Third, the seller sets prices to different experiments rather than data content, and the buyer pays the seller before the nature state is realized. In this case, the prices in the menu never leak information about the actual data values. Finally, the seller prices the data before it is actually collected, which satisfies the real-time data requirement of IoT applications. We further assume the seller is committed to the designed pricing schemes. Once the buyer selects a pricing scheme, the seller will strictly follows the rule of the experiment and sends signals to the buyer according to the predefined probability matrix. Such seller commitment can be implemented in practice via a smart contract inside a blockchain [13].

In the full version [14] of this paper, we provide a simple and concrete example to demonstrate our data trading process.

III. Data Pricing Mechanism

In this section, we present our data pricing mechanisms for the problem of revenue maximization. We begin with a special case where there exists only one type of buyer, and design a simple mechanism, namely *MSimple*, that extracts full surplus from buyers. We then step into the general setting with multiple different buyer types. We present the *MGeneral* mechanism to this setting, and prove there exists a polynomial time solution by formulating the problem of revenue maximization as a convex program. Finally, we consider a simple but practical case where buyers have bounded rationality [15], which additionally requires the seller's menu to be constant size. We present the *MPractical* mechanism, and prove the revenue loss with respective to the optimal revenue.

A. A Warm-Up Case

We first consider a simple case, where only one type of buyer θ exists in the market, i.e., $\Theta = \{\theta\}$ and $F(\theta) = 1$. This corresponds to the situation where buyers have no other source of data, and have a common prior estimation of nature. Since buyers are homogeneous, the only constraint in this problem is the I.R. property. In this case, the optimal menu contains only one pricing scheme (I, t_I). The revenue maximization problem of *MSimple* can be formulated by

$$\begin{aligned} \max \quad & t_I, \\ \text{s.t.} \quad & v(\theta, I) - t_I \geq 0, && \text{(I.R.)} \\ & \sum_{j=1}^{|S|} p_{ij} = 1, && \forall i, \\ & p_{ij}, t_I \geq 0, && \forall i, j. \end{aligned}$$

This is equivalent to finding an experiment that maximizes buyer's utility increment $v(\theta, I)$. As we will prove in Theorem 1, the full-information experiment $\bar{I}$ is always a utility-maximizing experiment. Therefore, the optimal pricing scheme is simply a full-information experiment $\bar{I}$, along with a price that is equal to the utility increment of the buyer.

Theorem 1. *For the single buyer type case, the optimal pricing scheme is a full-information experiment $\bar{I}$ with price $t_{\bar{I}} = \sum_{i=1}^n \theta_i \max_a u_\theta(\omega_i, a) - \max_a \sum_{i=1}^n \theta_i u_\theta(\omega_i, a)$.*

Proof. For any experiment I, define a_j as buyer's optimal action when she receives signal s_j, i.e., $a_j = \arg\max_a \mathbb{E}_\omega [u_\theta(\omega, a) \mid s_j]$. We then have

$$t_I \leq v(\theta, I) = \sum_{j=1}^{|S|} \max_a \sum_{i=1}^n \theta_i p_{ij} u_\theta(\omega_i, a) - u(\theta) \tag{5}$$

$$= \sum_{j=1}^{|S|} \sum_{i=1}^n \theta_i p_{ij} u_\theta(\omega_i, a_j) - u(\theta) \tag{6}$$

$$= \sum_{i=1}^n \theta_i \left(\sum_{j=1}^{|S|} p_{ij} u_\theta(\omega_i, a_j) \right) - u(\theta) \tag{7}$$

$$\leq \sum_{i=1}^n \theta_i \max_a u_\theta(\omega_i, a) - u(\theta) \tag{8}$$

$$= u(\theta, \bar{I}) - u(\theta) \tag{9}$$

The first inequality comes from the I.R. constraint. The first equality is by the definition of utility increment. The equality (6) holds due to the definition of a_j. The equality (7) is derived from switching the order of summation. The inequality (8) is by setting the p_{ij} with largest $u_\theta(\omega_i, a_j)$ value to be 1 and others to be 0. The equality (9) follows from the definition of $u(\theta, \bar{I})$ — the buyer is fully informed about the true nature state, and she can take exactly the optimal action for any nature state ω_i.

From the preceding derivations, we can easily verify that the full-information experiment generates the largest utility increment among all possible experiments, and maximizes the revenue of the seller. Replacing $u(\theta)$ in equality (8) with its definition in Equation (1), we get the optimal price $t_{\bar{I}}$ for the experiment $\bar{I}$. Since there is only one type of buyer in this simple case, the seller knows the value of every θ_i. Therefore, the optimal price can be exactly calculated by the seller. □

In this simple case, there is only one kind of buyer in the market, and the seller is clear about the type of every buyer he intersects with, and thus can extract full surplus from buyers.

B. The General Case

We further consider the general setting, in which different types of buyers coexist in the market. Considering more fine-grained menu can extract higher revenue from the market, we seek to design a discriminatory pricing scheme (I_θ, t_θ) for each type θ. To avoid the potential strategic behavior of buyers, we need to guarantee that each buyer will indeed choose the pricing scheme we design for her, and has no incentive to choose other pricing schemes. This leads to the following **Incentive Compatible** (I.C.) constraint:

$$v(\theta, I_\theta) - t_\theta \geq v(\theta, I_{\theta'}) - t_{\theta'}, \qquad \forall \theta, \theta' \in \Theta. \qquad \text{(I.C.)}$$

Without loss of generality, we assume whenever the buyer is indifferent between buying (I_θ, t_θ) and not buying, she always chooses to buy. The problem of revenue maximization in this general case can be formulated as follows:

$$\begin{aligned}
\max \quad & \sum_{\theta \in \Theta} F(\theta) t_\theta, & & \\
\text{s.t.} \quad & v(\theta, I_\theta) - t_\theta \geq 0, & \forall \theta, \quad & \text{(I.R.)} \\
& v(\theta, I_\theta) - t_\theta \geq v(\theta, I_{\theta'}) - t_{\theta'}, & \forall \theta, \theta', \quad & \text{(I.C.)} \\
& \sum_{j=1}^{|S|} p_{ij} = 1, & \forall i, I_\theta, & \\
& p_{ij}, t_\theta \geq 0, & \forall i, j, I_\theta, t_\theta. &
\end{aligned}$$

In such formulation, the feasible region is not convex, resulting in high computational complexity of directly solving this problem. To remove this non-convexity, we base our solution on the classical idea of designing posteriors [16], [17]. In the following, we will only sketch the main idea of our solution, and leave the complete proofs to the full version [14] of this paper.

The previous formulation considers an experiment from the "row perspective": In the experiment I_θ, we aim to assign proper row probability p_i over different signals in S to buyer θ when the nature state is ω_i. From the "row perspective", the experiment I_θ can be expressed by the matrix P and prior distribution θ. Now we present a different perspective to express an experiment I_θ. For easy illustration, we define two notations. We use vector $q_j = (q_{1j}, q_{2j}, \cdots, q_{nj})^T \in \Delta(\Omega)$ to denote the posterior distribution $\Pr(\omega \mid s_j)$ after receiving the signal s_j, where q_{ij} is the posterior probability that the nature state is ω_i, i.e., $q_{ij} = \Pr(\omega_i \mid s_j)$. We also denote matrix $Q = (q_1, q_2, \cdots, q_{|S|})$. Let $x^\theta = \{x_j^\theta : s_j \in S\}$. Here, x_j^θ denotes the probability of receiving signal s_j in the experiment I_θ, i.e., $x_j^\theta = \Pr(s_j)$. From this "column perspective", the experiment can be expressed via matrix Q and vector x^θ. The following lemma states that we can express the experiment I_θ equivalently from the "row perspective" and "column perspective" under certain conditions.

Lemma 1. *It is equivalent to define an experiment I_θ from the row perspective with $P = [p_{ij}]$ and from the column perspective with $x^\theta, Q = [q_{ij}]$, if and only if:*

$$\sum_{j=1}^{|S|} x_j^\theta q_{ij} = \theta_i, \quad \forall i \in [n]. \qquad \text{(E.Q.)}$$

Proof. We leave the proof of Lemma 1 to the full version [14] of this paper due to space limitation. □

We propose the following lemma to show that assuming the posterior q_j is chosen from a pre-computed finite subset of $\Delta(\Omega)$ will generate equivalent revenue as assuming it is chosen from an infinite continuous space. The idea of this lemma corresponds to the "interesting posteriors" defined in [17].

Lemma 2. *Given the buyer type space Θ, restricting the candidate values of posterior distribution q_j to a finite set $Q^* \subset \Delta(\Omega)$ that can be pre-computed does not reduce the optimal revenue.*

Proof. We leave the proof of Lemma 2 to the full version [14] of this paper due to space limitation. □

Now we can rewrite the problem of revenue maximization from the column perspective as a linear program:

$$\begin{aligned}
\text{LP} \quad & \max \sum_{\theta \in \Theta} F(\theta) t_\theta, & \\
\text{s.t.} \quad & \sum_{j=1}^{|S|} x_j^\theta u(\theta, s_j) - u(\theta) - t_\theta \geq 0, \quad \forall \theta, & \text{(I.R.)} \\
& \sum_{j=1}^{|S|} x_j^\theta u(\theta, s_j) - t_\theta \geq \sum_{j=1}^{|S|} x_j^{\theta'} u(\theta, s_j) - t_{\theta'}, \forall \theta, \theta', & \\
& & \text{(I.C.)} \\
& \sum_{j=1}^{|S|} x_j^\theta q_{ij} = \theta_i, \quad \forall i, \theta, & \text{(E.Q.)} \\
& x_j^\theta, t_\theta \geq 0, \quad \forall j, \theta. &
\end{aligned}$$

An immediate corollary of Lemma 2 is that LP contains polynomial number of constraints but exponential number of variables. In seeking a solution with polynomial time complexity, we take the dual of LP as follows:

$$\begin{aligned}
\text{DLP} \quad & \min \sum_{i,\theta} y_{\theta,i} \theta_i - \sum_\theta u(\theta) g_\theta, & \\
\text{s.t.} \quad & \sum_{\theta' \neq \theta} (h_{\theta,\theta'} - h_{\theta',\theta}) + g_\theta \geq F(\theta), & \forall \theta, \\
& \sum_{\theta' \neq \theta} h_{\theta',\theta} u(\theta', s_j) - \sum_{\theta' \neq \theta} h_{\theta,\theta'} u(\theta, s_j) & \\
& \qquad \geq g_\theta u(\theta, s_j) - \sum_i y_{\theta,i} q_{ij}, & \forall j, \theta, \\
& h_{\theta,\theta'} \geq 0, \; g_\theta \geq 0, \; y_{\theta,i} \in \mathbb{R}, & \forall i, \theta, \theta'.
\end{aligned}$$

This dual linear program contains $O(|\Theta|^2 + |\Theta| \cdot |\Omega|)$ variables and finitely many constraints. For a polynomial time solution, we need to find a separation oracle for the second family of constraints. Based on the solution in [17], since $u(\theta, s_j)$ takes the maximum over $|A|$ linear functions, we can substitute each constraint in the second family with $|A|$ equivalent constraints. Checking if all the $|A|$ constraints are satisfied by all q_j is equivalent to solving the following problem:

$$\begin{aligned}
\min \sum_i y_{\theta,i} q_{ij} - \left(g_\theta + \sum_{\theta' \neq \theta} h_{\theta,\theta'} \right) \sum_{i=1}^n q_{ij} u_\theta(\omega_i, a) \\
+ \sum_{\theta' \neq \theta} h_{\theta',\theta} u(\theta', s_j), \quad \forall \theta \in \Theta, a \in A, q_j \in \Delta(\Omega).
\end{aligned}$$

As this is a convex program that can be solved exactly in polynomial time with the standard technique from optimization theory, we can conclude the main result in the general case in the following theorem.

Theorem 2. *MGeneral finds the revenue-maximizing menu in polynomial time of $|\Omega|$ and $|\Theta|$, by solving the dual linear programming problem* DLP.

C. A Practical Case

A potential problem with *MGeneral* mechanism is that its menu size can be as large as the number of buyers, making it hard to be implemented in some practical context. For a large data marketplace with thousands of buyer types, each buyer has to look through all the pricing schemes to find the one that optimizes her utility. Although automated trading agents or bots are commonly utilized in modern online markets, performing Bayesian belief updates for each pricing scheme can still be computationally burdensome even for a computer agent. On the seller's point of view, calculating the optimal menu requires solving a convex program that involves perhaps thousands of variables, which is also time-consuming in high-frequency online markets.

In this part, we present our solution to a more practical scenario, where agents are computationally bounded. We prefer a menu with explicit and closed-form representation, instead of referring to solving a convex programming problem. We further require our menu to have a constant size, listing only a constant number of pricing schemes for human buyers to choose from, as the existing data marketplaces do [1], [18].

Our simple mechanism *MPractical* satisfies the preceding requirements. *MPractical* either offers a buyer the most accurate data with a fixed price, or sells nothing to the buyer. More specifically, this menu contains two pricing schemes: a full-information experiment $\bar{I}$ with a fixed price $\bar{t}$ for all buyers, and a no-information experiment $\underline{I}$ with zero price. The no-information experiment gives the buyer a chance to safely opt out when she cannot extract non-negative utility from the purchase, and thus the I.R. property is always guaranteed. Suppose there are in total N buyers in the market. The price $\bar{p}$ for the full-information experiment is simply the price that maximizes seller's expected revenue:

$$\bar{t} = \arg\max_{t} \sum_{\theta} N \cdot F(\theta) \cdot t \cdot \mathbb{1}\left[v(\theta, \bar{I}) \geq \bar{t}\right],$$

where the indicator function $\mathbb{1}\left[v(\theta, \bar{I}) \geq \bar{t}\right] = \mathbb{1}\left[\sum_{i=1}^{n} \theta_i \max_a u_\theta(\omega_i, a) - u(\theta) \geq t\right]$ denotes whether the buyer can extract non-negative utility.

A natural question is, how much revenue will the seller lose if he employs *MPractical* instead of the optimal *MGeneral*? In the following, we show that *MPractical* can achieve $\Omega(\frac{1}{\log|\Theta|})$ revenue of *MGeneral* even in the worst case.

For easier illustration, we first introduce a few notations. Let $\mathcal{R}$ denote the revenue of *MPractical*, and $\mathcal{S}$ denote the sum of all buyers utility increment towards the full-information experiment, i.e., $\mathcal{S} = \sum_\theta N \cdot F(\theta) \cdot v(\theta, \bar{I})$, which is obviously the revenue upper bound of any pricing mechanism. We assume the number of buyers for each type is upper bounded by a constant c, i.e., $N \leq c\,|\Theta|$. We normalize buyers' utility increment $v(\theta, \bar{I})$ into the range $[1, h]$ by properly scaling the values of the utility function $u_\theta(\omega, a)$. Here, h denotes the largest utility increment of the buyers, i.e., $h = \max_\theta v(\theta, \bar{I})$. We then have the following theorem:

Theorem 3. *Assuming $\mathcal{S} \geq 2h$, the approximation ratio of MPractical is $\mathcal{R}/\mathcal{S} = \Omega(\frac{1}{\log|\Theta|})$.*

Proof. Divide the buyer utility increments into $\log h$ bins by a power of two. For each utility increment $v(\theta, \bar{I})$ in bin B_k ($0 \leq k < \log h$), we have $2^k \leq v(\theta, \bar{I}) < 2^{k+1}$. Since the utility increments sum up to $\mathcal{S}$ and there are $\log h$ bins, there exists a bin B_k such that the sum of all utility increments in B_k is no smaller than $\mathcal{S}/\log h$. If we set the price to be the lowest utility increment in B_k, the generated revenue $\mathcal{R}_k$ will be at least $\mathcal{S}/(2\log h)$, since the lowest utility increment is at least half of any other utility increment in B_k. We now have $\mathcal{R} \geq \mathcal{R}_k \geq \mathcal{S}/(2\log h)$ since $\mathcal{R}$ is the revenue generated by the optimal price $\bar{t}$, which clearly yields revenue no lower than $\mathcal{R}_k$.

Define v^* to be the smallest utility increment such that all the utility increments below v^* sum up to at least $\mathcal{S}/2$. We then have $v^* \geq h/N$, otherwise the sum of utility increments below v^* is smaller than $Nv^* < h \leq \mathcal{S}/2$, which contradicts our definition of v^*. We now ignore all the buyers with utility increment below v^*. Denote the optimal price for the remaining buyers as t^* and the corresponding revenue as $\mathcal{R}^*$. According to the result from last paragraph, we now have

$$\mathcal{R}^* \geq \frac{\mathcal{S}/2}{2\log(h/v^*)} \geq \frac{\mathcal{S}}{4\log N}.$$

Since $\mathcal{R}_k$ is the revenue extracted from a larger set of buyers, we have $\mathcal{R}_k \geq \mathcal{R}^*$. Combining all the results leads to

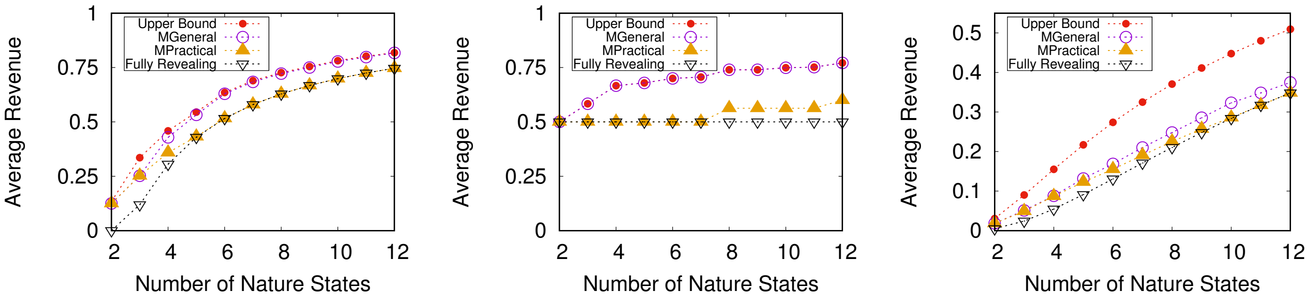
$$\mathcal{R} \geq \mathcal{R}_k \geq \mathcal{R}^* \geq \frac{\mathcal{S}}{4\log N} \geq \frac{\mathcal{S}}{4\log c\,|\Theta|}.$$

This finishes our proof of $\mathcal{R}/\mathcal{S} = \Omega(\frac{1}{\log|\Theta|})$. □

Theorem 3 relies on a simple and reasonable assumption that $\mathcal{S} \geq 2h$. This assumption requires the sum of all buyers' utility increments to be at least twice of any single buyer, which easily holds in practice when the number of buyers is reasonably large. In the following theorem, we will show that the approximation ratio in Theorem 3 is tight in the worst case: this logarithmic lower bound is actually also the upper bound for any menu of constant size.

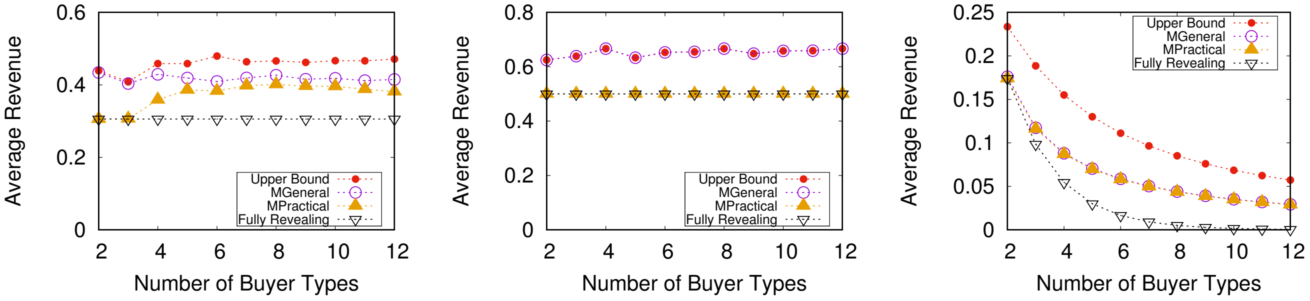
Theorem 4. *There exist cases where no menu of constant size can achieve more than $O(\frac{1}{\log|\Theta|})$ revenue of MGeneral, even when the $\mathcal{S} \geq 2h$ assumption holds true.*

Proof. We explicitly construct the following example. Assume there are N buyers coming from N different types. We number the buyers from 1 to N and set the utility increment of buyer i to be $v_i = \frac{N}{i}$ ($1 \leq i \leq N$). Without loss of generality, we can assume the price for any experiment is chosen from a finite



(a) Gaussian (b) Uniform (c) Pareto

Fig. 2. Average revenue under different number of nature states.



(a) Gaussian (b) Uniform (c) Pareto

Fig. 3. Average revenue under different number of buyer types.

set $\{N, N/2, N/3, \ldots, 1\}$. It is easy to see that when adding a pricing scheme of price $t = N/i$ to the menu, at most i more buyers will have the incentive to buy the data, leading to the additional revenue of no more than N. Since the menu contains constant number of pricing schemes, the revenue of any constant size menu is upper bounded by $O(N)$.

Now we will show the optimal mechanism can indeed extract the full revenue of $\Omega(N \log N)$ in the previous setting. Let the size of nature state set be $|\Omega| = 2N$. In this case, each buyer i can be represented by a type vector $\theta_i = (\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,2N})$. For buyer i, we set $\theta_{i,j}$ to be 0 for all j, except for $\theta_{i,2i-1} = \theta_{i,2i} = \frac{1}{2}$. In this sense, buyer i only cares about the data concerning nature state ω_{2i-1} and ω_{2i}. In our example, all buyers share the same utility function $u(\omega, a)$ defined as: (1) $u(\omega_i, a_j) = 0$ if $i \neq j$. (2) $u(\omega_{2i-1}, a_{2i-1}) = u(\omega_{2i}, a_{2i}) = \frac{2N}{i}, \forall 1 \leq i \leq N$.

We construct the optimal menu as follows. For each buyer, the pricing scheme we design for her gives her full information on the two nature states she cares about, and no information on the other states. Formally, for buyer i, we set $p_{2i-1,2i-1} = p_{2i,2i} = 1$, and all elements in the other rows of the experiment matrix are set to $\frac{1}{2N}$. Since the experiments designed for the others bring no information increment to the buyer but requires a positive price, each buyer is only interested in her own pricing scheme, and hence the I.C. constraint is always satisfied. The readers can verify that the buyers' utility increments are exactly $v_i = N/i$, given the utility function and experiments we designed. Finally, we charge a price of N/i from buyer i $(1 \leq i \leq N)$, and by doing so we extract the full surplus of $\Omega(N \log N)$ from the market.

We conclude that in our example, no constant size menu can extract more than $O(\frac{1}{\log|\Theta|})$ of the optimal revenue, which is achieved by *MGeneral*. Therefore, *MPractical* is indeed one of the optimal mechanisms in the bounded computation case. □

IV. Evaluations

In this section, we evaluate our pricing mechanisms *MGeneral* and *MPractical* on a real-world ambient sound dataset, and compare their performance with our benchmarks. The convex programming parts in our mechanisms are implemented using the Gurobi software [19].

A. Evaluation Setup

We use the Ambient Sound Monitoring Network [20] dataset in our evaluation. The Dublin City Council collected this dataset with a network of sound monitors to measure the ambient sound quality at different sites of Dublin. This dataset contains sound pressure data of every 5 minute from 15 monitoring sites in Dublin on each day from 2012 to 2015. We use the sensory data from the Walkinstown monitoring site on June 1st, 2015 in our evaluation, and we assume the buyer priors are based on the sensory data of the same day in the previous three years, ranging from 44dB to 68dB.

We discretize the interval $[44, 68]$ into n intervals as the sample space of the nature state. We consider three typical families of prior distributions, including Gaussian distribution, uniform distribution and Pareto distribution. Since we consider different types of buyers in the market, we assume buyers of the same distribution family differ from each other by the distribution parameters: Gaussian distributions with different mean values 44, 50, 56 and 62; uniform distribution over the sub-intervals of $[44, 68]$ with different lengths 6, 12, 18 and 24; and Pareto distributions with different values 0.1, 0.5, 0.9 and 1.3 of b for the generating formula $f(x) = \frac{b}{x^{b+1}}$.

We compare the revenue of our mechanisms with two benchmarks, namely the Fully Revealing mechanism and revenue Upper Bound. In the Fully Revealing mechanism, the seller only offers the full-information experiment in his menu, but still guarantees the I.C. and I.R. properties. This mechanism is the optimal solution to a restricted version of *MGeneral* mechanism, by additionally requiring all experiments to be full-information. The revenue Upper Bound is the sum of all buyers' valuations towards the full-information experiment, without guaranteeing the I.C. property. As the Upper Bound extracts full surplus from all buyers, it is obviously the revenue upper bound of any pricing mechanism.

B. Performance of Pricing Mechanisms

We first vary the size of sample space n from 2 to 12, and evaluate its influence on the four pricing mechanisms. In this set of evaluations, we fix the number of buyer types to be $|\Theta| = 4$, and simplify the utility of all buyers to be

$$u_\theta(\omega, a) = \begin{cases} 1, & \text{if } \ \omega = a, \\ 0, & \text{otherwise}, \end{cases}$$

which means that there is only one "correct" action under each possible nature state, and these correct actions generate one unit utility to the buyer. Figure 2 shows the average revenue extracted from each buyer under three different prior distributions. We can observe that for all the cases, *MGeneral* always generates higher revenue than *MPractical* and Fully Revealing, and nearly approaches the revenue Upper Bound. For Gaussian distributions, as the size n of sample space increases, buyer prior distributions are more dispersed over different possible nature states, indicating they are less certain about the true nature state. In this sense, buyers' prior expected utilities $u(\theta)$ are generally low, and data from the seller can bring high valuation to them. *MGeneral* makes use of buyers' uncertainty and almost extracts full surplus when n is relatively large. When $n = 12$, *MGeneral* achieves 99.91% revenue of Upper Bound. For uniform distributions, *MGeneral* extracts full surplus from buyers as Upper Bound does, because uniform estimations indicate that buyers have no prior knowledge of the true nature state. When $n = 12$, *MGeneral* outperforms *MPractical* and Fully Revealing by 28.50% and 54.21%, respectively. For Pareto distributions, the revenue for all mechanisms are lower compared with the other two distributions, because buyers have more confident prior estimations, and their prior expected utilities $u(\theta)$ are already high. In this case, it is hard to extract high revenue by providing data to the buyer, but *MGeneral* still generates 73.56% revenue of the very optimistic Upper Bound when $n = 12$. Under all three distributions, the revenue of our mechanisms increase with n. Since n denotes the discretization level of data, we can conclude that the seller can extract higher revenue by selling fine-grained data.

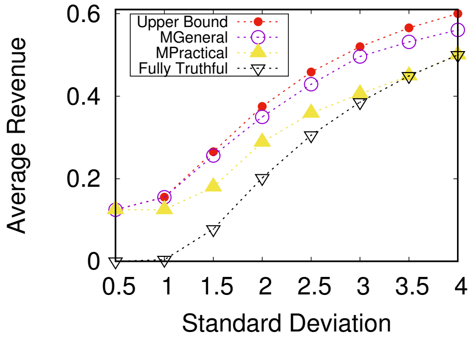


Fig. 4. Revenue under different values of standard deviation for Gaussian distributions.

We then evaluate the impacts of the number of buyer types on the four mechanisms. We report the evaluation results in Figure 3, when the number of types $|\Theta|$ varies from 2 to 12 and the number of possible nature states n is fixed at 4. As $|\Theta|$ increases, more types of buyers with heterogeneous prior estimations appear in the market, and their strategic behaviors raise more challenges to our pricing mechanisms. For Gaussian and uniform distributions, the average revenue of our mechanisms do not decrease as $|\Theta|$ grows. This indicates that our mechanisms are robust against more types of strategic buyers under these two distributions. For Pareto distributions, however, the average revenue of our mechanisms decrease with $|\Theta|$. This is because buyers under Pareto distributions are confident about their prior estimations and have higher prior expected utilities before buying data from the seller. As more confident types of buyers join the market, seller's average revenue from each buyer certainly decreases.

We finally test the influence of the standard deviation σ in Gaussian distributions. We are interested in this parameter because it denotes how confident the buyers are about their prior estimations. We vary σ from 0.5 to 4.0, while fixing both n and $|\Theta|$ to be 4. As we can see in Figure 4, *MGeneral* still outperforms other mechanisms, and achieves 93.40% revenue of Upper Bound when $\sigma = 4.0$. The average revenue of all mechanisms increase with σ, because when buyers are uncertain about the nature state, the data from the seller can bring high utility increments to them. Therefore, we can conclude that when buyers are not confident about their prior knowledge, the seller can take advantage of buyers' uncertainty and extract higher revenue.

V. Related Work

In recent years, designing data pricing frameworks has attracted increasing interests in the database community. Balazinska *et al.* [21] first envisioned the emergence of cloud-based data markets, and outlined potential challenges and research opportunities. Following them, many query-based frameworks have been proposed to price ad-hoc query data.

These frameworks allow the seller to manually assign prices to a few views, and automatically extrapolate the prices to other ad-hoc queries from the buyer. In [22], Koutris *et al.* first identified two key properties that a pricing function must satisfy, namely arbitrage-free and discount-free, and proposed a polynomial time algorithm that derives the price for common types of queries. Similar work include arbitrage-free pricing functions for arbitrary queries [11], and a scalable framework for pricing relational queries [23]. A set of accountable protocols named AccountTrade was proposed in [24] for big data trading among dishonest customers. These work assume that data has already been collected and structured before being priced, and their objective is not to maximize the revenue of the seller.

Data marketplace has also been an active research topic in the community of Internet of Things. Perera *et al.* [7] surveyed smart city applications that can benefit from data markets. An IoT data transfer framework for cloud-based applications was proposed in [25]. The authors in [26] designed a decentralized infrastructure for IoT data trading based on blockchain technologies, but they did not elaborate on the pricing mechanisms. A two-sided market for crowdsensed data was proposed in [27], and secondary market models for mobile data were studied in [28]. In a recent paper, Zheng *et al.* [29] took advantage of the geographical locality of sensor data, and employed a versioning technique based on the accuracy of data. Our work differs from previous work by further revealing and utilizing the unique features of IoT data as a commodity.

Information design is a rapidly growing research area in both computer science and economics literature. Different from providing incentives to participators in mechanism design problems, information design studies how to influence the belief of participators by providing payoff-relevant information to them through strategic interactions. A special yet influential case called Bayesian persuasion, concerning one information sender and one receiver, was studied in [30]. In a model similar to ours [31], Bergemann *et al.* investigated the problem where a buyer seeks supplemental information from the seller to facilitate her decision making. As they sought optimal solutions in the continuous space, they had to put strict restrictions on the model to maintain tractability. In another related work [17], Babaioff *et al.* considered the optimal mechanism for selling information sequentially. [32] and [33] provide excellent surveys of the information design literature.

VI. Conclusions

In this paper, we have studied the problem of revenue maximization in IoT data markets. We have characterized the unique economic properties of IoT data, and proposed a market model accordingly from an information design perspective. We have presented our pricing mechanisms that achieve optimal revenue in different market settings. Evaluation results have shown that our mechanisms achieve good performance and approach the revenue upper bound.

References

[1] "Gnip apis." [Online]. Available: http://support.gnip.com/apis/

[2] "Xignite." [Online]. Available: https://www.xignite.com/

[3] "Here." [Online]. Available: https://www.here.com/en

[4] "Iota." [Online]. Available: https://www.iota.org/

[5] "Ambient maps." [Online]. Available: https://ambientmaps.co/

[6] "Databroker dao." [Online]. Available: https://databrokerdao.com/

[7] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Sensing as a service model for smart cities supported by internet of things," *Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 1, pp. 81–93, 2014.

[8] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (iot): A vision, architectural elements, and future directions," *Future generation computer systems*, vol. 29, no. 7, pp. 1645–1660, 2013.

[9] Z. Liqiang, Y. Shouyi, L. Leibo, Z. Zhen, and W. Shaojun, "A crop monitoring system based on wireless sensor network," *Procedia Environmental Sciences*, vol. 11, pp. 558–565, 2011.

[10] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with querymarket," in *SIGMOD*, 2013.

[11] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," in *VLDB*, 2014.

[12] L. Toka, B. Lajtha, É. Hosszu, B. Formanek, D. Géhberger, and J. Tapolcai, "A resource-aware and time-critical iot framework," in *INFOCOM*, 2017.

[13] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in *SP*, 2016.

[14] W. Mao, Z. Zheng, and F. Wu, "Pricing for revenue maximization in iot data markets: An information design perspective," 2019. [Online]. Available: https://drive.google.com/open?id=1ifmw8tIeZdXFXK0EwAHv5UGybNlDkc-W

[15] H. A. Simon, *Models of bounded rationality: Empirically grounded economic reason.* MIT press, 1997, vol. 3.

[16] D. Bergemann, B. Brooks, and S. Morris, "The limits of price discrimination," *American Economic Review*, vol. 105, no. 3, pp. 921–57, 2015.

[17] M. Babaioff, R. Kleinberg, and R. Paes Leme, "Optimal mechanisms for selling information," in *EC*, 2012.

[18] "Dialogfeed." [Online]. Available: https://www.dialogfeed.com/pricing/

[19] "Gurobi." [Online]. Available: http://www.gurobi.com/

[20] "Ambient sound monitoring network." [Online]. Available: https://data.smartdublin.ie/dataset/ambient-sound-monitoring-network

[21] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," *VLDB*, 2011.

[22] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *PODS*, 2012.

[23] S. Deep and P. Koutris, "Qirana: A framework for scalable query pricing," in *SIGMOD*, 2017.

[24] T. Jung, X.-Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su, "Accounttrade: Accountable protocols for big data trading against dishonest consumers," in *INFOCOM*, 2017.

[25] R. Montella, M. Ruggieri, and S. Kosta, "A fast, secure, reliable, and resilient data transfer framework for pervasive iot applications," in *INFOCOM*, 2018.

[26] P. Missier, S. Bajoudah, A. Capossele, A. Gaglione, and M. Nati, "Mind my value: a decentralized infrastructure for fair and trusted iot data trading," in *IoT*, 2017.

[27] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "Trading data in the crowd: Profit-driven data acquisition for mobile crowdsensing," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 2, pp. 486–501, 2017.

[28] L. Zheng, C. Joe-Wong, C. W. Tan, S. Ha, and M. Chiang, "Secondary markets for mobile data: Feasibility and benefits of traded data plans," in *INFOCOM*, 2015.

[29] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *MobiHoc*, 2017.

[30] E. Kamenica and M. Gentzkow, "Bayesian persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.

[31] D. Bergemann, A. Bonatti, and A. Smolin, "The design and price of information," *American Economic Review*, vol. 108, no. 1, pp. 1–48, 2018.

[32] D. Bergemann and S. Morris, "Information design: A unified perspective," 2017.

[33] S. Dughmi, "Algorithmic information structure design: a survey," *SIGecom Exchanges*, vol. 15, no. 2, pp. 2–24, 2017.