On Designing Market Model and Pricing Mechanisms for IoT Data Exchange

Zhenzhe Zheng, *Member, IEEE*, Weichao Mao, *Student Member, IEEE*, Yidan Xing, *Student Member, IEEE*, and Fan Wu, *Member, IEEE*

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

***Index Terms*—Internet of Things data, data pricing, revenue maximization, information design.**

I. Introduction

DATA is becoming an important commodity in the era of artificial intelligence. Data has tremendous value to both its owner and other parties who want to integrate it into their services. A number of online data exchange platforms are emerging to enable data sharing and trading over the Internet, facilitating various kinds of data-based services, such as personalized advertising and business decision making. For example, Gnip [1] aggregates and sells social media data from Twitter, Xignite [2] vends real-time financial data, and Here [3] trades tracking and positioning data for location-based services.

Manuscript received 7 May 2023; revised 23 January 2024; accepted 27 January 2024. Date of publication 6 March 2024; date of current version 3 October 2024. This work was supported in part by National Key R&D Program of China under Grant 2022ZD0119100, in part by the China NSF under Grant 62322206, Grant 62132018, Grant U2268204, Grant 62025204, Grant 62272307, and Grant 62372296. Recommended for acceptance by D. Niyato. *(Corresponding author: Zhenzhe Zheng.)*

Zhenzhe Zheng, Yidan Xing, and Fan Wu are with the Department of Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhengzhenzhe@sjtu.edu.cn; katexing@sjtu.edu.cn; fwu@cs.sjtu.edu.cn).

Weichao Mao is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA (e-mail: weichao2@illinois.edu).

Digital Object Identifier 10.1109/TMC.2024.3373811

Among various online data exchange platforms, several companies [4], [5], [6] focus on data from Internet of Things (IoT). IoT data marketplaces allow different stakeholders to share their sensor network infrastructures through the exchange of IoT data among different organizations. 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 platform for IoT data exchange. Data from widely deployed IoT devices is accurate and fined-grained, which enables many intelligent city services, such as waste management and environment monitoring [7], traffic jam avoidance [8], smart agriculture and weather forecasting [9], and etc. The massive value of IoT data in diverse applications significantly increases its market demand.

One major drawback of existing data marketplaces is the inefficient data pricing mechanisms: the data sellers either adopt a fixed price mechanism [4], or choose to negotiate with buyers offline [2], introducing obstacles for data exchange. Although there are some existing work dedicated to designing flexible data pricing mechanisms, most of them only considered structured and relational data [10], [11]. These work fail to capture the features of IoT data, and ignore the economic objective of the data seller. In this work, we aim to analyze the new economic properties of IoT data compared to the traditional digital goods, and design efficient pricing mechanisms to achieve the objective of revenue maximization.

In the following, we first present four properties of IoT data as a new kind of digital 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 negligible marginal cost. Due to such a cost structure, a buyer can easily generate a new copy of the raw data, and resell it at a lower price, introducing the problem of data piracy. Traditional copyright techniques for digital goods, such as software and movie, can hardly resist such piratical behaviors over data. To resolve this issue, we argue that the data seller should share data services instead of raw data in data marketplaces. The data services could be the mean, median, maximum values and even statistical models of aggregated data, or the results of performing data mining techniques on the data. We note this trading strategy can also preserve the privacy of data owners to some extent.

• Second, the valuation over IoT data does not necessarily depend on data volume, but should be highly related to the amount of information it provides. This property differentiates IoT data from traditional (digital) commodities. 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 exchange is how to quantify the valuation of data? In the context of IoT applications, buyers make some decision to earn certain utilities based on the information extracted from sensory data. With this observation, we can measure buyer's valuation towards a set of data by the utility increment (information gain) after purchasing this data set. For example, suppose you are going out on a sunny day and consider it is not necessary to take an umbrella. In this case, a large set of humidity sensory data confirming a long sunny day does not generate much valuation, as the provided information is consistent with your prior belief. By contrast, a set of sensory data forecasting a heavy rain one hour later generates high valuation to you, as it changes your belief and decision, guiding you to take an umbrella.

• Third, the price of data might have a certain correlation with the information behind the data, and thus directly releasing the data price may leak the content of data. Suppose the data seller in the previous "umbrella" example sets prices $\$1$ and $\$2$ for the "sunny" data set and the "rainy" data set, respectively. A buyer could distinguish these two data sets through observing the corresponding prices, as the rainy data set is more valuable and deserves a higher price. Since buyers are willing to pay a high price for valuable information, the seller would lose revenue if she simply reduces the price of the rainy data set to $\$1$. We therefore argue that, in order to avoid information leakage before data trading, the seller should decouple the data price and data content. For example, one possible qualified pricing scheme is: charge $\$1$ for a weather data set with 25% uncertainty, and charge $\$2$ for 5% uncertainty, where a $x\%$ uncertainty indicates the contained data is disturbed or messed up by a probability of $x\%$. Such a content-independent pricing scheme do not leak information about the actual data.

• Fourth, IoT applications require the price of data to be determined before data is actually generated. The data buyer would like IoT data streams to be fed in real-time for time-sensitive decision making [12], and thus it is impractical to calculate the price in an online manner. Most of existing work cannot handle the real-time feature of IoT data, as they always consider the static data, which is sold after it is collected, structured or modeled [10], [11]. This feature of IoT data raises a challenging problem: how do sellers persuade buyers to purchase the data with appropriate prices when they still have not collected data?

Besides the four features mentioned above, the objective of revenue maximization introduces additional challenges. As the seller does not know the valuation of each individual buyer, she has to determine the price of data under a Bayesian valuation setting, where only the distribution of valuations is available. Moreover, with various types of buyers in the market, an optimal pricing mechanism should perform market segmentation or price discrimination among buyers, which would introduce the potential strategic behaviors of buyers.

Jointly considering the discussed challenges, in this paper, we present a new market model for IoT data exchange from an information design perspective, which captures the aforementioned unique properties of IoT data. First, the seller in our model provides data services to buyers by sending signals with various amount of information, 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 data collection, and incentivizes buyers to purchase the desired data sets by giving them the highest expected utility increments. The timing of the pricing schemes enables the sharing of future data, and ensures that prices would not leak information about the actual data content.

Based on the new market model for IoT data exchange, we then design a series of pricing mechanisms to maximize revenue from data trading, inspired by the ideas from information design. As a classical result from the information design literature, the information designer is often better off by sometimes obfuscating the receiver, rather than offer completely accurate information [13]. We implement this rule by posting a spectrum of data purchasing options with different accuracy, each catering to a specific type of buyer. We set higher prices to more informative data and lower prices to less informative data. Our pricing mechanisms automatically perform market segmentation among buyers, which resonates the accuracy-based versioning mechanisms [14], to maximize revenue.

We summarize our contributions in this paper as follows:

- First, we characterize four unique properties 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 new properties.
- Second, we design revenue-maximizing pricing mechanisms under different market settings. We first consider a simple setting where only one type of homogeneous buyer exists in the market, and propose the *MSimple* mechanism that extracts full surplus from the market. We then consider the general setting where different types of buyers coexist. We present the *MGeneral* mechanism to this setting, and show there exists a polynomial time solution by formulating the problem of revenue maximization as a convex program. We further present the *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.
- Third, we observe the seller could further increase revenue by offering free data trials to buyers before data trading. We present algorithms for the seller to design profitable free trials both in the single buyer setting and the multiple buyers setting.
- Finally, we implement our pricing mechanisms on a real-world ambient sound dataset. We evaluate the influence of different parameters in the market model, and show that our mechanisms achieve good performance in terms of revenue maximization.

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 revenue-maximizing pricing mechanisms under different market settings. In Section IV, we study how free trials increase the seller's revenue. We evaluate our

pricing mechanisms in Section V. In Section VI, we briefly review related work in the literature. Finally, we conclude the paper in Section VII.

II. Preliminaries

We consider the intersection between a data seller and multiple data buyers. The data commodity in IoT data marketplaces 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 represent a particular numerical value of the sensing result for an environmental phenomenon. For example, ω can be the mean of the readings from a set of noise sensors near the street, and correspondingly Ω is a discrete set of possible numerical values for noise levels. The random variable ω could also denote the data services extracted from multimodal 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 route. The analytical result is sold to buyers as a data service. In this case, the nature state ω is chosen from a binary set $\Omega = \{Crowded, NotCrowded\}$.[1]

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*[2] I and a corresponding *price* t_I. An experiment $I = \{S, P\}$ contains a set S of possible *signals*,[3] and an $n \times |S|$ right probability matrix $P = [\ p_{ij}\ ]$, $1 \le i \le n,\ 1 \le j \le |S|$, 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., $p_{ij} = \Pr(s_j \mid \omega_i)$.

We consider two special types of experiment: *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 her entire knowledge about the nature state. Once the seller observes the nature state as ω_i, she always sends signal s_i to the buyer. From the perspective of buyer, 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. The buyer has a prior estimation of the nature state before purchasing 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. It is also possible that the buyers do not have any prior knowledge about the nature state. For these specific scenarios that buyers do not have any prior knowledge, we assume the prior estimation to be an uniform distribution. Under this situation, there is no evidence or knowledge that can help the buyer gain an informative estimation for the distribution of nature states, and thus it is reasonable to assume the buyer to believe that each nature state has a same and uniform probability to happen. This assumption is also without loss of generality as our solutions would work for arbitrary prior distribution. We denote the prior distribution for the random variable ω as $\theta = (\theta_1, \theta_2, \ldots, \theta_n) \in \Delta\Omega$,[4] 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, and 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, which can be estimated from the historical interaction with buyers or through some survey deployed by sellers. These assumptions about type θ follow from Bayesian mechanism design literature [16], which are necessary for revenue computation. This is because the same mechanism could have highly fluctuating revenue performances for buyers with different prior beliefs, and we need $F(\theta)$, the distribution of these prior beliefs, to estimate the expected revenue of a mechansim over all the buyer population. As we would observe in Section III, an inaccurate $F(\theta)$ would only affect the revenue performances of our proposed pricing mechanisms, while the I.C. conditions are guaranteed to hold regardless of the accuracy of $F(\theta)$ by our design. Therefore, in practice, the seller could first use a $F(\theta)$ estimated from historical interaction data or pre-survey among buyers, then further collect and update the type distribution through interactions with the buyers.

We next define the utility of a buyer in IoT data markets. In IoT applications, the buyer usually faces a decision problem, where she has to choose an action a from a finite set A, based on her perception over the nature state. Let $u(\omega, a)$ denote the *utility* of the buyer when the nature state is ω and an action a is taken. Without loss of generality, we normalize $u(\omega, a)$ to [0,1]. We also assume for each nature state ω, there exists one action to achieve the highest utility 1, i.e., $\max_a u(\omega, a) = 1, \forall \omega \in \Omega$. Without purchasing data from the seller, the buyer has to rely her decision only on her prior estimation θ, and the corresponding expected utility is

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

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

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

[1] Our results can be easily extended to a vector of random variables instead of a scalar random variable. We would keep the scalar random variable for the simplicity of discussion in this work.

[2] An experiment is also called an information structure in the literature [15].

[3] We abstract different responses from the seller as different signals. For example, in the context of IoT data exchange, reporting different probabilities of precipitation to the buyer can be regarded as sending different signals.

[4] $\Delta\Omega$ denotes all the possible probability distributions over Ω.

and her expected utility becomes

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

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

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

The buyer's expected utility after buying experiment I is

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

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

$$\begin{aligned} v(\theta, I) &= u(\theta, I) - u(\theta) = \sum_{j=1}^{|S|} \Pr(s_j) u(\theta, s_j) - u(\theta) \\ &= \sum_{j=1}^{|S|} \left(\sum_{i=1}^{n} \theta_i p_{ij}\right) \left(\max_a \sum_{i=1}^{n} \frac{p_{ij}\theta_i}{\sum_{k=1}^{n} p_{kj}\theta_k} u(\omega_i, a)\right) \\ &\quad - \max_a \sum_{i=1}^{n} \theta_i u(\omega_i, a) \\ &= \sum_{j=1}^{|S|} \max_a \sum_{i=1}^{n} \theta_i p_{ij} u(\omega_i, a) - \max_a \sum_{i=1}^{n} \theta_i u(\omega_i, a). \end{aligned} \tag{5}$$

We assume buyer θ is willing to buy the 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:

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

We remark on four advantages about our new market model for IoT data exchange, which overcomes the challenges we discussed in Introduction. First, the seller does not directly sell raw data to the buyer, but instead trades information (signals) extracted from raw data. This information-based trading model can resist data piracy and preserve data privacy to some extent. Second, we propose a new metric to measure the valuation of data, which depends on the buyer's utility increment due to the action change after buying data. While previous metric for data valuation is simply related to data volume [17], our metric considers the data's effect on decision making and reflects the information contained in the data. Third, the seller sets prices to different experiments independent of the data content, and the buyer pays the seller before the nature state is realized. In this case, the prices in the menu would not leak information about the actual data content. Finally, the seller sets the prices of data before it is actually collected, which is suitable for the exchange of real-time IoT data.

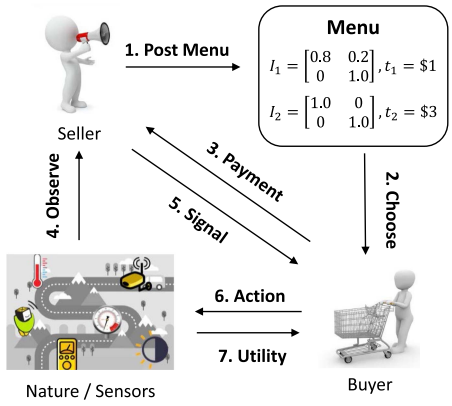


Fig. 1. Data trading process in IoT data marketplaces.

The data trading process in IoT data marketplaces is illustrated in Fig. 1, which could be described in sequence as follows: 1) First, the seller designs a menu $\mathcal{M} = \{(I, t_I)\}$ and posts it to the market; 2) Given the menu, the buyer with prior belief θ chooses a pricing scheme (I, t_I) from the menu to maximize her total gain, i.e., the difference between utility increment and the payment $v(\theta, I) - t_I$, and 3) pays the corresponding price t_I to the seller. 4) Then, the true nature state ω is realized and revealed to the seller. 5) After observing the true nature state, the seller sends a signal s_j to the buyer following the rule in the selected experiment I. 6) After receiving the signal s_j, the buyer chooses an action a with the maximum expected utility according to (3). 7) Finally, The buyer's utility $u(\omega, a)$ is realized based on the chosen the action chosen by the buyer. We note that the seller is committed to the designed pricing schemes, meaning that once the buyer selects a pricing scheme, the seller would strictly follow the rule of the experiment, and send signals to the buyer according to the predefined probability matrix. Such a seller commitment can be implemented in practice via a smart contract inside a blockchain [18].

We provide a simple and concrete example to demonstrate how each step in the above data trading process is proceeded. We consider a data seller to collect and analyse ambient noise data from widely deployed sensors, and sells the prediction results of traffic conditions to data buyers [5], [19]. The basic information of nature states, the buyer's type, action set, and her utility functions are defined as follows. We simplify the nature state set in this example to be binary $\Omega = \{\omega_1 = Crowded,\ \omega_2 = NotCrowded\}$. A buyer with type $\theta = (\theta_1 = 0.3,\ \theta_2 = 0.7)$ needs to make a decision selected from her action set $A = \{a_1 = NotGoingOut,\ a_2 = GoingOut\}$. The buyer would get unit utility if she takes the "right" action, and zero utility otherwise, i.e., $u(\omega_1, a_1) = u(\omega_2, a_2) = 1$ and $u(\omega_1, a_2) = u(\omega_2, a_1) = 0$. Before buying any data, the buyer considers $GoingOut$ would generate higher expected utility based on her prior estimation θ, and by (1), her utility before buying the data could be calculated as $u(\theta) = 0.7$. We are now prepared to go through the details for the trading of traffic condition data. Suppose the seller posts a menu with two pricing schemes as defined in Fig. 1, and the buyer chooses the first pricing scheme containing the following

experiment

$$I = \begin{bmatrix} 0.8 & 0.2 \\ 0 & 1 \end{bmatrix}$$

and pays a price $t_I = 0.1$. This experiment I implies: when the nature state is $NotCrowded$, i.e., $\omega = \omega_2$, the seller would reveal the true state by sending signal s_2 with probability 1, since $p_{22} = 1$; when $\omega = \omega_1$, the seller would obfuscate the buyer a little bit, by telling her the truth (sending signal s_1) with probability 0.8, but lying to her (sending signal s_2) with probability 0.2. On the buyer's side, upon receiving signal s_1, she updates her estimation using Bayes' rule in (2), and gets posterior estimation $\Pr(\omega_1 \mid s_1) = 1$ and $\Pr(\omega_2 \mid s_1) = 0$. From (3), her expected utility after receiving signal s_1 is $u(\theta, s_1) = 1$. Similarly, we have $u(\theta, s_2) = 0.92$. By (4), her expected utility after buying experiment I is $u(\theta, I) = 0.94$. Since the buyer's utility increment $v(\theta, I) = u(\theta, I) - u(\theta) = 0.24$ is higher than the price $t_I = 0.1$, the buyer would like to buy such an experiment, giving the seller a revenue of 0.1. We can verify that the revenue-maximizing pricing scheme in this example is actually the other pricing scheme provided by the seller, i.e., to completely reveal the nature state to the buyer, by making I a diagonal matrix with a price $t_I = 0.3$. However, as various types of buyers co-exist in more complicated market settings, we will see that deliberately obfuscating the buyer might sometimes generate higher revenue to the seller. We summarize the frequently used notations in Table I.

TABLE I
MAJOR NOTATIONS

ω	Nature state
Ω	Space of nature states, with $\lvert\Omega\rvert = n$
$\Delta\Omega$	Space of probability distributions over Ω
s	Signal sent from seller to buyer
S	Space of signals
p_{ij}	Probability that the seller sends signal s_j when the nature state is ω_i
P	Probability matrix consists of p_{ij}, with $i \in [1, n]$ and $j \in [1, \lvert S\rvert]$ (subscripts in subsequent notations have the same ranges)
I	Experiment consists of a signal space S and a probability matrix P
t_I	Price of experiment I
θ	Buyer's prior distribution on nature states with θ_i denotes the probability that ω_i is the nature state
Θ	A finite set of buyer's possible prior distribution θ
$F(\theta)$	Cumulative distribution function of θ
$\Delta\Theta$	Space of probability distributions over Δ
a	Action the buyer chooses in decision problem
A	Space of actions
$u(\omega, a)$	Utility of the buyer when the nature state is ω and an action a is taken
$u(\theta)$	Buyer's expected utility given prior θ
$u(\theta, s)$	Buyer's expected utility given prior θ and signal s
$u(\theta, I)$	Buyer's expected utility after buying experiment I given prior θ
$v(\theta, I)$	Buyer's expected utility increment given prior θ and experiment I
q_j	Posterior distribution after receiving the signal s_j, where q_{ij} is the posterior probability that the nature state is ω_i
Q	Probability matrix consists of p_j
x^θ	Probabilities of receiving different signals in the experiment I_θ, where x_j^θ denotes the probability for signal s_j

III. Data Pricing Mechanisms

In this section, we present a family of data pricing mechanisms for the problem of revenue maximization in IoT data marketplaces. We begin with a special case where there exists only one type of homogeneous buyers, and design an optimal mechanism, namely *MSimple*, which is able to extract full surplus from the buyers. We then step into the general setting with multiple types of buyers co-existing in the market. We formulate the problem of revenue maximization in this setting as a convex program, and propose *MGeneral*, an optimal mechanism with polynomial time complexity. Finally, we consider a more practical case with bounded rational buyers [20], which additionally requires the seller's menu to have a constant size. For this case, we present the *MPractical* mechanism, and bound the revenue loss with respective to the optimal revenue. The characteristics of data pricing mechanisms proposed in this work are summarized in Table II.

TABLE II
CHARACTERISTICS FOR PROPOSED MECHANISMS

Name	Advantage(s)	Disadvantage(s)
MGeneral	Optimality in revenue	High computational complexity
MPractical	Revenue approximation guarantee and low computation complexity	Lack of optimality guarantee in revenue
Free data trial	Additional chances to increase the revenue when there are multiple rounds of interactions	High computational complexity and lack of revenue guarantee

A. A Warm-Up Case

We first consider a simple case, where only one type of buyers θ exists in the market, i.e., $\Theta = \{\theta\}$ and $F(\theta) = 1$. This corresponds to the situation where all buyers have no other private source of data, and have a common prior estimation over the nature state. 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 problem of revenue maximization can be formulated by

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

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 $t_{\bar{I}} = \sum_{i=1}^{n} \theta_i \max_a u(\omega_i, a) - \max_a \sum_{i=1}^{n} \theta_i u(\omega_i, a)$.

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

$$\begin{aligned} t_I \le\ & v(\theta, I) = \sum_{j=1}^{|S|} \max_a \sum_{i=1}^{n} \theta_i p_{ij} u(\omega_i, a) - u(\theta) \\ =\ & \sum_{j=1}^{|S|} \sum_{i=1}^{n} \theta_i p_{ij} u(\omega_i, a_j) - u(\theta) && (6) \\ =\ & \sum_{i=1}^{n} \theta_i \left(\sum_{j=1}^{|S|} p_{ij} u(\omega_i, a_j) \right) - u(\theta) && (7) \\ \le\ & \sum_{i=1}^{n} \theta_i \max_a u(\omega_i, a) - u(\theta) && (8) \\ =\ & u(\theta, \bar{I}) - u(\theta) && (9) \end{aligned}$$

The first inequality comes from the I.R. constraint. The first equation is by the definition of utility increment. Equation (6) uses the definition of a_j. Equation (7) is derived from switching the order of summation. Equality (8) is obtained by setting the probability p_{ij} with the largest $u(\omega_i, a_j)$ to be 1 and others to be 0. Equation (9) follows from the definition of $u(\theta, \bar{I})$, where the buyer is fully informed about the true nature state, and she can take exactly the optimal action for any nature state ω_i. Since we assume the highest achievable utility in every nature state is normalized to 1, (8) also suggests $u(\theta, \bar{I}) = 1, \forall \theta$.

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 (8) with its definition in (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 kinds of buyers in the market, and the seller knows about the type of every buyer she intersects with, and thus can extract full surplus from buyers. Although this assumption of homogeneous buyers might not hold in many situations, the useful idea of formulating a mathematical program as the mechanism in this simple case could be inherited and applied to more complex cases. It is also worth to note that by this simple case with homogeneous buyers, we can clearly illustrate the intuition behind the solution of leveraging the tool of information design to design a data pricing mechanism.

B. The General Case

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

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

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:

$$\begin{aligned} \max \quad & \sum_{\theta \in \Theta} F(\theta) t_\theta, \\ \text{s.t.} \quad & v(\theta, I_\theta) - t_\theta \ge 0, && \forall \theta, && \text{(I.R.)} \\ & v(\theta, I_\theta) - t_\theta \ge v(\theta, I_{\theta'}) - t_{\theta'}, && \forall \theta, \theta', && \text{(I.C.)} \\ & \sum_{j=1}^{|S|} p_{ij} = 1, && \forall i, I_\theta, \\ & p_{ij}, t_\theta \ge 0, && \forall i, j, I_\theta, t_\theta. && (10) \end{aligned}$$

The feasible region in such a formulation is not convex, leading to high computational complexity of directly solving this problem. Specifically, the $v(\theta, I)$ in I.R. and I.C. constraints is a piecewise linear combination of the decision variables p_{ij} due to the "maximum" operation in its definition (5), and this piecewise linearity causes non-convexity. In order to overcome the challenges caused by non-convexity, we need to formulate our problem from a new perspective. The main idea of our reformulation is to replace the decision variables p_{ij} with some constant parameters q_{ij} that can be pre-computed efficiently. By doing this, the "maximum" operations are exerted on the constant parameters, rather than the decision variables, and thus we can detour the difficulty from the non-convexity of feasible region.

The previous formulation considers an experiment from the "row perspective": in the experiment I_θ, we aim to assign a proper row probability p_i over different signals in S when the nature state is ω_i. Now we present a different perspective, namely "column perspective", to express the experiment I_θ. For easy illustration, we define two notations. We use vector $q_j = (q_{1j}, q_{2j}, \ldots, q_{nj})^T \in \Delta(\Omega)$ 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 denote all the posterior distributions after receiving different signals to matrix $Q = (q_1, q_2, \ldots, q_{|S|})$. Let $x^\theta = \{x_j^\theta : s_j \in S\}$, where 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 the matrix Q and the 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^θ and $Q = [q_{ij}]$, if and only if the Bayes plausibility restriction is satisfied:

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

Proof: The "if" side: Suppose (E.Q.) holds true, and we recall that $x_j^\theta = \Pr(s_j)$, $q_{ij} = \Pr(\omega_i \mid s_j)$. Then, for any experiment I_1 defined by p_{ij}, we can construct an identical experiment I_2 from column perspective by setting

$$x_j^\theta = \Pr(s_j) = \sum_{i=1}^{n} \Pr(\omega_i)\Pr(s_j \mid \omega_i) = \sum_{i=1}^{n} \theta_i p_{ij}, \forall j,$$

$$q_{ij} = \Pr(\omega_i \mid s_j) = \frac{\Pr(\omega_i)\Pr(s_j \mid \omega_i)}{\Pr(s_j)} = \frac{\theta_i p_{ij}}{\sum_{i=1}^{n} \theta_i p_{ij}}, \forall i, j. \tag{11}$$

Similarly, for any column experiment, we can construct an identical experiment using only p_{ij}, by setting

$$p_{ij} = \Pr(s_j \mid \omega_i) = \frac{\Pr(s_j)\Pr(\omega_i \mid s_j)}{\Pr(\omega_i)} = \frac{x_j^\theta q_{ij}}{\theta_i}, \forall i, j. \tag{12}$$

The "only if" side: Suppose the row perspective is equivalent to the column perspective in defining an experiment, we have $x_j^\theta q_{ij} = \Pr(\omega_i, s_j) = \theta_i p_{ij}$. Summing over j leads to

$$\sum_{j=1}^{|S|} x_j^\theta q_{ij} = \sum_{j=1}^{|S|} \theta_i p_{ij} = \theta_i.$$

From the above two parts, we have proved this lemma. □

Simply defining an experiment from the column perspective does not make our problem easier, because the posterior probability q_{ij} is a continuous value. In order to achieve the optimal solution, we still need to enumerate all possible posterior distribution q_j (representing all possible experiments) in an infinite continuous space. To overcome this difficulty, we propose the following lemma to show that assuming the posterior q_j is chosen from a pre-computed finite subset of $\Delta(\Omega)$ would generate the equivalent revenue as it is chosen from the original infinite continuous space. The idea of this lemma corresponds to the "interesting posteriors" in [21].

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 in advance does not reduce the optimal revenue.

Proof: According to the column representation, utility $u(\theta, s_j) = \max_a \sum_{i=1}^{n} q_{ij} u(\omega_i, a)$ can be considered as a piece-wise linear function $f(q_j) : \Delta(\Omega) \to \mathbb{R}$ on q_j. For each $\theta \in \Theta$, function $u(\theta, s_j)$ partitions the continuous space $\Delta(\Omega) \subset \mathbb{R}^n$ into $|A|$ polytopes, and within each polytope $u(\theta, s_j)$ is linear. We now combine the $|\Theta|$ sets of these partitions by letting their boundaries cut each other, and we finally get a finite set of newly partitioned polytopes. All $u(\theta, s_j)$'s are linear within each such polytope.

Let Q^* be the set of all the polytope vertex from the above new generalized polytopes. For any posterior $q_j \in \Delta(\Omega)$, we can rewrite q_j with linear combinations of posteriors from Q^*. Specifically, we can express $q_j = \sum_k \gamma_k q_k$, where q_k's are the vertexes of the polytope that contains q_j, and $\gamma_k \geq 0$, $\sum_k \gamma_k = 1$. For each solution to the revenue maximization problem expressed from the column perspective, if it assigns a non-zero probability x_j^θ to any posterior $q_j \notin Q^*$, we can convert this solution to another one with posteriors only in Q^*, by repeatedly increasing x_k^θ for each q_k by $\gamma_k x_j^\theta$, and decreasing x_j^θ to 0. As the number of partitioned polytopes is finite, we can pre-compute Q^* in advance. □

Since the posterior distribution q_j is drawn from a finite set Q^* that could be pre-computed in advance as shown in Lemma 2, in the column perspective, we can regard q_{ij} as constant parameters rather than decision variables. With this result, we can rewrite the problem of revenue maximization from the column perspective as a linear programming:

$$\begin{aligned}
\text{LP} \quad & \max \sum_{\theta \in \Theta} F(\theta) t_\theta, \\
\text{s.t.} \quad & \sum_{j=1}^{|S|} x_j^\theta u(\theta, s_j) - u(\theta) - t_\theta \geq 0, \quad \forall \theta, && \text{(I.R.)} \\
& \sum_{j=1}^{|S|} x_j^\theta u(\theta, s_j) - t_\theta \geq \sum_{j=1}^{|S|} x_j^{\theta'} u(\theta, \theta', s_j) \\
& \quad - t_{\theta'}, \forall \theta, \theta', && \text{(I.C.)} \\
& \sum_{j=1}^{|S|} x_j^\theta q_{ij} = \theta_i, \quad \forall i, \theta, && \text{(E.Q.)} \\
& x_j^\theta, t_\theta \geq 0, \quad \forall j, \theta. && (13)
\end{aligned}$$

Since q_{ij}'s are constant parameters, in the above linear programming, utility $u(\theta, s_j)$ can also be pre-computed and considered as constants, by $u(\theta, s_j) = \max_a \sum_{i=1}^{n} q_{ij} u(\omega_i, a)$. Within this formulation, we only need to determine the values of variables x_j^θ and t_θ. The (I.C.) constraints state that buyer θ is better off choosing the experiment I_θ we design for her rather than the experiment $I_{\theta'}$ for another buyer θ'. In the formulation of column perspective, we need additional computations to express the (I.C.) constraints. Specifically, we need to figure out the posterior distribution of the buyer θ when choosing the experiment $I_{\theta'}$. We recall that the experiment $I_{\theta'}$ maps the prior belief θ' to a distribution $x^{\theta'}$ over the posterior beliefs $q^{\theta'}$. However, $I_{\theta'}$ does not map θ to the same distribution of posterior beliefs. We would first calculate the $I_{\theta'}$ matrix, i.e., the matrix $\{p_{ij}\}$, from θ''s posterior beliefs $x^{\theta'}$ and $q_j^{\theta'}$ according to (12). Then by (11) we could then calculate the distribution over posterior beliefs that $I_{\theta'}$ maps θ to. Specifically, $I_{\theta'}$ maps θ to a distribution $\{x_j\}$ over posterior beliefs $\{q_j\}$, where

$$x_j = \sum_{i=1}^{n} \theta_i \frac{x_j^{\theta'} q_{ij}^\theta}{\theta_i'} \text{ and } q_{ij} = \frac{\theta_i x_j^{\theta'} q_{ij}^\theta}{\theta_i' x_j}.$$

Therefore, in the (I.C.) constraints, $u(\theta, \theta', s_j)$, the expected utility of buyer θ when choosing the experiment $I_{\theta'}$, can be calculated as $u(\theta, \theta', s_j) = \sum_{j=1}^{n} x_j u(q_j)$, where $u(q_j)$ is the expected utility with the belief q_j, similar to that in (1).

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

$$\begin{aligned}
\text{DLP} \quad & \min \sum_{i,\theta} y_{\theta,i}\theta_i - \sum_{\theta} u(\theta) g_\theta, \\
\text{s.t.} \quad & \sum_{\theta' \neq \theta} (h_{\theta,\theta'} - h_{\theta',\theta}) + g_\theta \geq F(\theta), && \forall \theta, \\
& \sum_{\theta' \neq \theta} h_{\theta',\theta} u(\theta', s_j) - \sum_{\theta' \neq \theta} h_{\theta,\theta'} u(\theta, s_j) \\
& \qquad \geq g_\theta u(\theta, s_j) - \sum_i y_{\theta,i} q_{ij}, && \forall j, \theta, \\
& h_{\theta,\theta'} \geq 0,\ g_\theta \geq 0,\ y_{\theta,i} \in \mathbb{R}, && \forall i, \theta, \theta'.
\end{aligned} \tag{14}$$

This dual linear programming contains $O(|\Theta|^2 + |\Theta| \cdot |\Omega|)$ variables and finite constraints. For a polynomial time solution, we need to find a separation oracle for the family of constraints in (14). 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:

$$\begin{aligned}
\min & \sum_i y_{\theta,i} q_{ij} - \left(g_\theta + \sum_{\theta' \neq \theta} h_{\theta,\theta'} \right) \sum_{i=1}^{n} q_{ij} u(\omega_i, a) \\
& + \sum_{\theta' \neq \theta} h_{\theta',\theta} u(\theta', s_j), \quad \forall \theta \in \Theta, a \in A, q_j \in \Delta(\Omega).
\end{aligned} \tag{15}$$

As this is a convex program that can be solved exactly in polynomial time with the standard technique from optimization theory [22], we can conclude the main result for the general case in the following theorem.

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

Therefore, given the set of potential actions and priors of buyers as well as the prior distribution, the procedure of running MGeneral mechanism would be: 1) compute the finite set of posteriors that would not harm the optimality as we introduced in Lemma 2; 2) use these posteriors to formulate the dual problem according to program (14), and solve it through the provided separation oracle. As both the steps could be finished within polynomial time, the total time complexity of MGeneral mechanism would be also polynomial. Once the dual program is solved, the resulted MGeneral mechanism could be applied repeatedly as long as the prior distribution of buyers keeps the same.

C. A Practical Case

While MGeneral mechanism achieves the optimal revenue, its optimality would also bring some practical problems. A potential problem with the *MGeneral* mechanism is that its menu size can be as large as the number of buyers, making it hard to be implemented in practical markets. For a large data marketplace with thousands of buyers, each buyer has to look through all the pricing schemes to find the one that optimizes her utility. Although automated trading agents are commonly applied in modern online marketplaces, performing Bayesian belief updates for large number of pricing schemes 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 online marketplaces. These problems motivate us to find some mechanisms that are more practical in the online marketplace. As these problems are brought by the optimality nature of MGeneral mechanism, to reduce the computational burden and the menu size of provided experiments, we have to allow certain tradeoff from the mechanism's revenue performances. Our aim is to design a mechanism with low computational burden for both the seller and the buyers while guaranteeing the revenue performances of the mechanisms.

In this section, we present our design of a mechanism with low computational burden for both the seller and the buyers, while guaranteeing the revenue performances. We prefer a menu with an 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 in the existing data marketplaces [1], [23].

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 just 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:

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

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

A fundamental question for this simple mechanism is, how much revenue will the seller lose if she 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., $\mathcal{S} \triangleq \sum_\theta N \times F(\theta) \times v(\theta, \bar{I})$, which is obviously the revenue upper bound of any pricing mechanisms. 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 $u(\omega, a)$. We then have the following theorem for the performance guarantee of *MPractical* mechanism.

Theorem 3: Assuming $\mathcal{S} \geq 2\tilde{}h$, 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 increment is at least half of any other 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 reasonable assumption of $\mathcal{S} \geq 2\tilde{}h$. This assumption requires the sum of all buyers' utility increments to be at least twice of the increment from any single buyer, which easily holds in practice when the number of buyers is reasonably large. In the following theorem, we 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 with a constant size.

Theorem 4: There exist cases where no menu with a constant size can achieve more than $O(\frac{1}{\log|\Theta|})$ revenue of *MGeneral*, even when the assumption of $\mathcal{S} \geq 2\tilde{}h$ 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 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 $|\Omega| = 2\tilde{}N$. In this case, each buyer i can be represented by a type vector $\theta_i = (\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,2\tilde{}N})$. 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) $u(\omega_{2i-1}, a_{2i-1}) = u(\omega_{2i}, a_{2i}) = \frac{2\tilde{}N}{i}, \forall 1 \leq i \leq N$.

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 $\frac{1}{2\tilde{}N}$. 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. □

Compared to the optimal but relatively time-consuming MGeneral mechanism, the MPractical mechanism could significantly reduce the computation time while achieving a comparable amount of revenue with a good approximation guarantee. As a result, when applying these IoT data pricing mechanisms in practice, MPractical is more suitable for applications with relatively tight time delay requirement to preserve the freshness of IoT data. In contrast, we need to adopt MGeneral mechanism for those IoT applications with relatively loose time delay requirement, where we could spend sufficient time to compute a mechanism with optimal revenue.

Besides pricing for IoT sensing data that directly indicates the nature state, e.g., data that records whether the road is crowded, our pricing mechanisms could also be applied to more generalized forms of data, such as text or image dataset, so long as the data could provide key information to infer the nature state faced by the buyer. The seller only needs to design one additional function to extract the valuable information about the nature states from the dataset, then the remaining pricing procedure would be the same. As an instance, for the traffic condition example in Section II, instead of selling ambient noise data, if the seller could access the real-time photos taken along a mainstream road from cameras, then the seller could also use these images to infer whether the road is crowded or not, and selling those data with our MGeneral or MPractical mechanism. When applying our pricing mechanisms in reality, another potential problem is how to convince the buyer that the seller would obey the

experiment when sending the signal. To deal with this problem, the seller could either introduce a trusted third-party to build some commitment protocols, or provide the history trading data to enable the verification from a statistical perspective.

IV. Free Data Trials

In previous sections, we consider the case where the seller interacts with the buyer only once during data trading. In practice, the data seller may communicate with the data buyers, such as deploying a data demonstration, before initiating the data trading. Such kinds of additional interactions provide the seller with more chances to affect the buyer's prior beliefs and achieve the seller's own objectives. However, the interactions before the formal trading are more complex to design, since they would significantly change the buyer's behaviors and the resulted revenue in the formal trading when the buyer uses Bayesian update to estimate the distribution of nature state. In this section, we study how the seller could extract higher revenue from buyers if the seller is able to conduct an extra round of interaction with the buyer. Compared with always providing the same data pricing menu, the data seller can achieve such revenue improvement by simply offering a free data trial to buyers before she reveals the menu, which is ubiquitous in practice. We also design an algorithm for the seller to find such a profitable free trial.

To improve the revenue, the data seller should have an information advantage over the buyer *a priori*. In this section, we consider the scenario that the seller knows exactly the nature state, and the nature state remains the same during the data trading process.[5] This happens when the seller has external sources of information or when the true nature state is consistent over a short period of time and the seller learns this information from an early transaction. We also assume the buyer is myopic and only seeks to maximize her utility in each of the separate phases: the free trial phase and the transaction phase. Otherwise, the buyer's optimal dynamic behavior in this two-phase game would rely on the generating probability distribution of the nature states, which heavily complicates the analysis for this problem. We would further relax this assumption in our future work.

We describe the timeline of the new data trading process with an additional free trail phase as follows:

- The buyer enters the marketplace and requests a data service from the seller.
- The free trial phase: The seller offers a free data trial to the buyer, aiming to manipulate the buyer's prior belief into an interim belief. The expected utility of the interim belief would be lower than that of the prior belief, introducing the opportunity of the seller to charge a higher price in the transaction phase.
- The transaction phase: The interaction between the seller and the buyer in this phase directly follows that in Fig. 1. The buyer chooses a pricing scheme according to her interim belief (instead of her prior belief), pays the price, obtains a posterior belief based on the signal she receives from the seller, takes a best action according to her posterior belief, and finally leaves the market.

We use a simple example to illustrate this new data trading process, and show the seller can extract higher revenue through deploying a free data trial. This example follows the same setting as the traffic condition example in Section II. Without the free data trial, the highest possible revenue that the seller can get is 0.3. We now consider how the seller designs a free trial to increase her revenue. In the free trial phase, the seller offers the following experiment to the buyer at a price of 0:

$$I_{trial} = \begin{bmatrix} 0 & 1 \\ 4/7 & 3/7 \end{bmatrix}.$$

After receiving the free trial experiment, the buyer with a prior belief $\theta = (\theta_1 = 0.3,\ \theta_2 = 0.7)$ would estimate the probability of receiving signal s_1 as 0.4, resulting in an interim belief of $\hat{\theta} = (\hat{\theta}_1 = 0,\ \hat{\theta}_2 = 1)$, and the probability of receiving signal s_2 as 0.6, resulting in an interim belief of $\hat{\theta} = (\hat{\theta}_1 = 0.5,\ \hat{\theta}_2 = 0.5)$. Thus, we can calculate that buyer θ's expected posterior utility after employing the free trial experiment I_{trial} would be $u(\theta,\ I_{trial}) = 0.7$, equal to her prior utility and obtain a zero utility increment $v(\theta,\ I_{trial}) = 0$. Since the buyer is myopic and her utility increment is no smaller than the price 0, the buyer would take the free trial. As the actual nature state is ω_1, the seller would deterministically send signal s_2 to the buyer according to the designed free trial experiment I_{trial}, resulting in an interim belief $\hat{\theta} = (\hat{\theta}_1 = 0.5,\ \hat{\theta}_2 = 0.5)$. Based on this interim belief, the seller would offer a full-information experiment to the buyer at a price of 0.5 in the transaction phase. We can calculate the revenue after the transaction phase and extract a higher revenue 0.5 compared with the revenue without deploying the free trial experiment. Note that the interim belief of a buyer is calculated by the Bayesian update rule given her prior type and the signaling schemes. Since the seller knows those required information, the calculation of buyer's interim belief could be conducted by the seller individually without the need to explicitly acquire from the buyer. Therefore, to evaluate the effects of adopting a specific experiment as the free data trial, the buyer could use the prior distribution of buyers to estimate the corresponding interim belief distribution, then use this interim distribution of buyer types to calculate the final revenue generated in the formal trading phase.

We would like to point out that the free trial is not always profitable for the seller. For example, for another buyer with a prior belief $\theta = (\theta_1 = 0.7,\ \theta_2 = 0.3)$, there exists no free trial that could generate higher revenue to the seller. In this case, the seller's revenue-maximizing strategy would be directly selling a full-information experiment at a price of 0.3, without using the free trial experiment. In the following discussion, we would identify the condition, under which the free data trial experiment can increase revenue.

[5]Our solutions can be easily extended to the general case where the seller only knows a probability distribution of the nature states, which still needs to contain more information than the prior beliefs of buyers. In this work, we focus on the case where the seller knows exactly the actual nature state for an easy exposition.

A. The Single Buyer Case

We start with a simple case where there exists only one type of buyer in the market, i.e., $\Theta = \{\theta\}$ and $F(\theta) = 1$, the setting considered in Section III-A. We define $a_\theta = \arg\max_a \mathbb{E}[u(\omega, a)]$ to be the buyer's p optimal *prior action* that generates the highest expected utility according to the prior belief θ. Without loss of generality, we assume the actual nature state is ω_1. A free data trial experiment, similar to other experiments, maps a buyer's prior belief θ to an interim belief, which can be expressed in the column perspective, i.e., a distribution $x = \{x_1, \ldots, x_m\}$ over a set of posterior distributions $Q = \{q_1, \ldots, q_m\} \subset \Delta\Omega$, with the Bayes plausibility restriction of $\theta = \sum_j x_j q_j$ from Lemma 1. Since $u(\theta)$ is a convex function as defined in (1), we have $\sum_j x_j u(q_j) \geq u(\theta)$, and the buyer always believes the free data trial gives her a higher posterior utility in expectation. However, for a broad category of buyers whose prior beliefs are not well aligned with the actual nature state (i.e., θ_1 is relatively small), the seller is able to further mislead the buyer to a certain interim belief θ^l. As the upper bound of price is the utility increment (i.e., the utility difference between the interim belief and the posterior belief), when the utility of the interim belief θ^l is small, the seller could hence charge a potentially high price in the subsequent transaction phase, and extract large revenue. In the following, we focus on the design of free data trial experiments to extract additional revenue from these buyers. In Theorem 5, we first identify a sufficient condition that free data trials can extract increased revenue under certain assumptions.[6]

Theorem 5: The seller is always able to increase revenue using a free trial if the buyer θ is overconfident, i.e., if there exists another belief point $\theta^l \in \Delta\Omega$, such that: (1) $u(\theta^l) < u(\theta)$. (2) $\theta^l_1 > \theta_1$. (3) $a_\theta = a_{\theta^l}$.

Proof: The first condition guarantees that there exists one interim belief θ^l that has a lower expected utility than that of the prior belief, which provides a chance for the seller to mislead the buyer to form an interim belief and to obtain a large utility increment (and then upper bound of price) in the transaction phase. The second condition means that the belief θ is less aligned with the actual nature state ω_1 than θ^l's, which is a technical condition to enable high probability to. The last condition requires the prior action of θ^l is the same as the action of θ, which largely simplifies the searching for the optimal free trial experiments.

Since we have assumed prior belief θ is not on the boundary of $\Delta\Omega$, and is not exactly indifferent between two actions, we could easily find another belief θ^h, such that

- $a_\theta = a_{\theta^l} = a_{\theta^h}$.
- $\exists\, 0 \leq \lambda \leq 1$, such that $\theta = \lambda\theta^l + (1-\lambda)\theta^h$.

The first condition states θ^h is on the same hyperplane as θ and θ^l. The second condition means θ can be expressed as a convex combination of θ^l and θ^h.

With the two beliefs θ^l and θ^h, the seller could design an experiment for θ that leads to an interim belief θ^l with probability λ and interim belief θ^h with probability $1-\lambda$. To make it consistent with the notations in Section III-B, we have $q_1 = \theta^l$, $q_2 = \theta^h$, $x^\theta_1 = \lambda$, and $x^\theta_2 = 1-\lambda$. Since $\theta = \lambda\theta^l + (1-\lambda)\theta^h$, the Bayesian plausibility (E.Q.) restriction is satisfied. According to Lemma 1, we could equivalently define an experiment I_{trial} from the row perspective, where $p_{ij} = \frac{x^\theta_j q_{ij}}{\theta_i}$, $\forall i, j$. Buyer θ's expected utility after buying experiment I_{trial} would be $u(\theta, I_{trial}) = x_1 u(q_1) + x_2 u(q_2)$. Since $\theta = x_1 q_1 + x_2 q_2$ and the belief points θ, θ^l and θ^h are on the same hyperplane ($a_\theta = a_{\theta^l} = a_{\theta^h}$), we have $v(\theta, I_{trial}) = x_1 u(q_1) + x_2 u(q_2) - u(\theta) = 0$. Thus, buyer θ would like to buy I_{trial} at a price of 0.

However, the actual nature state is ω_1, so the seller only sends signal according to the first row of experiment I_{trial}. Specifically, he sends signal s_1 with probability $p_{11} = \frac{x^\theta_1 q_{11}}{\theta_1}$ and sends s_2 with probability $p_{12} = \frac{x^\theta_2 q_{12}}{\theta_1}$. Therefore, buyer θ's actual expected utility after buying I_{trial} should be $\hat{u}(\theta) = p_{11} u(q_1) + p_{12} u(q_2)$ instead of $u(\theta, I_{trial})$ discussed above.

Since I_{trial} is strategically designed by a seller with information advantage, buyer θ's actual expected utility $\hat{u}(\theta)$ for buying I_{trial} is lower than she anticipated, and is hence lower than her prior expected utility $u(\theta)$:

$$\begin{aligned}
\hat{u}(\theta) &= p_{11} u(q_1) + p_{12} u(q_2) \\
&= \frac{1}{\theta_1}\left[x_1 q_{11} u(q_1) + x_2 q_{12} u(q_2)\right] && (16)\\
&= \frac{1}{\theta_1}\left[x_1 q_{11} u(q_1) + x_2 q_{12} u(q_2)\right](x_1 + x_2) && (17)\\
&= \frac{1}{\theta_1}[x_1^2 q_{11} u(q_1) + x_1 x_2 q_{11} u(q_1) \\
&\quad + x_2 x_1 q_{12} u(q_2) + x_2^2 q_{12} u(q_2)] && (18)
\end{aligned}$$

We further have

$$\begin{aligned}
(18) &< \frac{1}{\theta_1}[x_1^2 q_{11} u(q_1) + x_1 x_2 q_{12} u(q_1) \\
&\quad + x_2 x_1 q_{11} u(q_2) + x_2^2 q_{12} u(q_2)] && (19)\\
&= \frac{1}{\theta_1}\left[x_1 u(q_1) + x_2 u(q_2)\right](x_1 q_{11} + x_2 q_{12}) && (20)\\
&= x_1 u(q_1) + x_2 u(q_2) \\
&= u(\theta) && (21)
\end{aligned}$$

The equality (16) is by replacing p_{ij} with $\frac{x_j q_{ij}}{\theta_i}$. The equality (17) comes from $x_1 + x_2 = 1$. The (18) is simply by the expansion of brackets. The inequality (19) is a result of $q_{21} < \theta_1 < q_{11}$, $u(q_2) > u(\theta) > u(q_1)$, and the rearrangement inequality. The (20) is by factorization. The equality (21) follows from $x_1 q_{11} + x_2 q_{12} = \theta_1$.

Since the buyer's actual expected utility is lowered after she takes the free trial, she would have a higher utility increment towards a full-information experiment $\bar{I}$. Therefore, in the transaction phase, the seller is able to increase her revenue by charging the buyer a higher price for $\bar{I}$. □

In Algorithm 1, we present the detailed steps to construct a free trial experiment I_{trial} to improve revenue. We first check

[6] For an easy discussion, we assume θ is not on the boundary of space $\Delta\Omega$ and is exactly different between two actions. Specifically, we assume there exists $\epsilon > 0$, such that: (1) $\theta_i \geq \epsilon$, $\forall i$. (2) $\mathbb{E}[u(\omega, a_\theta)] \geq \mathbb{E}[u(\omega, a')] + \epsilon$, $\forall a' \neq a_\theta$.

Algorithm 1: GenerateFreeTrial.

Input: A vector $\theta = (\theta_1, \ldots, \theta_n)$ indicating the buyer's prior belief.

Output: A profitable data trial experiment I_{trial}.

1 $y^* \leftarrow$ the optimal solution to CVX;
2 $I_{trial} = null$;
3 **if** $u(y^*) < u(\theta)$ **then**
4 $\quad \theta^l \leftarrow y^*$;
5 $\quad$ Binary search for θ^h with the largest $0 \le \lambda \le 1$ such that $\theta = \lambda\theta^l + (1-\lambda)\theta^h$ and $a_{\theta^h} = a_\theta$;
6 $\quad x_1 \leftarrow \lambda,\ x_2 \leftarrow 1-\lambda$;
7 $\quad q_1 \leftarrow \theta^l, q_2 \leftarrow \theta^h$;
8 $\quad$ Construct the experiment $I_{trial} = [p_{ij}]$, where $p_{ij} = \frac{x_j q_{ij}}{\theta_i},\ \forall i, j$;
9 **return** I_{trial};

Algorithm 2: SimulateFreeTrials.

Input: Buyer type space $\Theta \subset \Delta\Omega$ and the distribution function $F(\theta) \in \Delta\Theta$.

1 $R_0 \leftarrow$ the optimal solution to DLP with Θ and $F(\theta)$;
2 $I_{trial} = null$;
3 **foreach** $\theta \in \Theta$ **do**
4 $\quad I_{tmp} \leftarrow GenerateFreeTrial(\theta)$;
5 $\quad$ **if** $I_{tmp} \neq null$ **then**
6 $\quad\quad \Theta', F'(\theta) \leftarrow$ buyers' interim belief distributions after taking I_{tmp};
7 $\quad\quad R \leftarrow$ the optimal solution to DLP with Θ' and $F'(\theta)$;
8 $\quad\quad$ **if** $R > R_0$ **then**
9 $\quad\quad\quad R_0 \leftarrow R$;
10 $\quad\quad\quad I_{trial} \leftarrow I_{tmp}$;
11 **return** I_{trial};

the revenue improvement condition in Theorem 5 by solving the following convex optimization problem CVX.

$$\begin{aligned} \text{CVX} \quad & \min u(y), \\ \text{s.t.} \quad & \sum_{i=1}^{n} y_i = 1, \\ & \sum_{i=1}^{n} y_i u(\omega_i, a_\theta) \ge \sum_{i=1}^{n} y_i u(\omega_i, a), \quad \forall a, \\ & y_1 > \theta_1, \\ & y_i \ge 0, \quad \forall i. \end{aligned}$$

If the optimal solution y^* to CVX satisfies $u(y^*) < u(\theta)$, then the three conditions in Theorem 5 are all satisfied. The optimal solution y^* serves as a legitimate value for θ^l (Line 4). We then use the binary search to find the farthest belief point θ^h such that θ is a convex combination of θ^l and θ^h, and their prior actions are the same (Line 5). In Lines 6 to 8, we construct the free trial experiment I_{trial} based on Lemma 1. The rationale behind such a free trial experiment is described in the proof of Theorem 5.

B. The General Case

We next generalize the above profitable free trials to the general case with multiple buyers. We focus on the solution that offers the same free trial experiment to all the buyers, instead of designing a separate trial experiment for each type of buyer. This is because we cannot use price discrimination to differentiate buyers in the free trial phase, and providing more free experiments only gives buyers more chances to get a utility increment for free.

With multiple types of buyers in the market, whether it is possible to increase revenue through free trials heavily depends on the buyer type distribution $F(\theta)$. It is generally impossible to use a simple mathematical theorem to characterize the condition of revenue improvement, as we have done in Theorem 5. We propose a heuristic method that simulates buyers' interim beliefs after taking the free trial, and compare the resultant revenue with the revenue benchmark to decide whether a trial experiment is indeed profitable. We present a simulation-based algorithm in Algorithm 2 for this purpose.

Algorithm 2 first calculates the seller's revenue without any free trials as the benchmark, by solving the linear programming DLP we proposed in Section III-B. It then enumerates each buyer type θ, and runs Algorithm 1 as a subroutine to calculate the potential experiment I_{tmp} we designed for θ (Line 4). If I_{tmp} is legitimate, we simulate all the buyers' interim belief distribution after taking this free trial experiment (Line 6). We then input the interim belief distributions into DLP and solve DLP for the revenue under the interim distribution (Line 7). If the new revenue is higher than the benchmark, we can conclude the free trial I_{tmp} indeed increases revenue, and hence save the current best revenue and trial experiment.

When applying Algorithm 2 to compute a free data trial, the seller should first ascertain that the potential buyers would have more interactions and purchase data from the seller after they have received some free data from the seller. As the algorithm adopts MGeneral as a subroutine, it is also more suitable for trading the IoT data for applications with relatively loose time delay requirement. We would like to note that Algorithm 2 only simulates the results of $|\Theta|$ trial experiments. Thus, it could not provide the theoretical guarantee to always obtain a profitable data free trial, even if a profitable trial does exist. Finding the optimal free data trials to increase the revenue for the general case is an interesting future work.

V. Evaluation Results

In this section, we evaluate our pricing mechanisms *MGeneral* and *MPractical* on a real-world ambient sound dataset, and compare their performance with the benchmarks. We also evaluate the revenue increase when the seller offers free trials. The convex programming parts in our mechanisms are implemented using the Gurobi software [24].

A. Evaluation Setup

We use the Ambient Sound Monitoring Network [25] 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 minutes from 15 monitoring sites in Dublin on each day from 2012 to 2015. The results of the sound level meters are given in Leq, which denotes the average sound level over each 5-minute period of measurement. 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 44 dB to 68 dB. All the evaluation results are averaged over 200 runs.

We discretize the interval $[44, 68]$ into n intervals as the sample space of the nature state. The default value for n is 4 and the number of buyer types $|\Theta|$ is 4. We consider three typical families of prior distributions, including the uniform distribution, Gaussian distribution and Pareto distribution. For the uniform distribution, we assume buyer types are uniformly sampled from the entire belief simplex $\Delta\Omega$ using the sample algorithm given in [26]. For the other two distribution families, we assume buyers of the same distribution family differ from each other by the distribution parameters: Gaussian distributions with different mean values, and Pareto distributions with different values of b for the generating formula $f(x) = \frac{b}{x^{b+1}}$.

We compare the revenue of our mechanisms with three benchmarks, namely the Fully Revealing mechanism, Grid Search mechanism, and revenue Upper Bound. In the Fully Revealing mechanism, the seller only offers the full-information experiment in her 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. We choose the Grid Search mechanism inspired by the versioning [27] and price discrimination [28] in economics, where we consider the seller to add different extents of noises to the full-information experiment, and set the prices of each experiment to be a constant multiplying the maximum value increment among all the buyer types. The seller would sample the extent of noises added to form different experiments, use grid search to set their prices, and finally choose the mechanism with highest revenue. Since the Grid Search mechanisms are not I.C., we calculate the revenue through simulating buyer's behaviors of choosing the highest-utility experiment. 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.

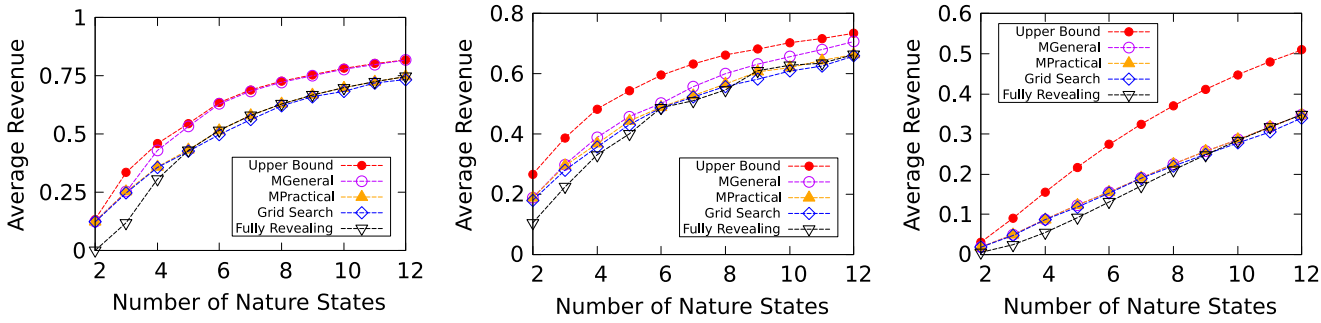
B. Performance of Pricing Mechanisms

We first vary the size of nature state 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

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

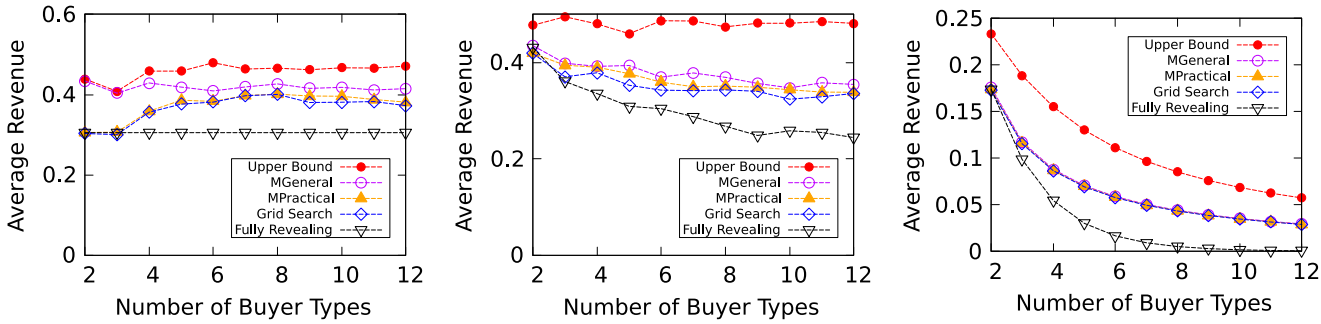
which means that there is only one "correct" action under each possible nature state, and these correct actions generate one unit utility to the buyer. Fig. 2 shows the average revenue extracted from each buyer under three different prior distributions. We can observe that for all the cases, *MGeneral* always generates higher revenue than *MPractical*, Fully Revealing and Grid Search, and nearly approaches the revenue Upper Bound. Due to the sampling and grid search process, the grid search method takes more computation time than all the other methods considered in the experiments. However, as we can observe, the revenue of grid search method does not exceed our simple MPractical mechanism in all the considered settings, and has a distinct revenue gap with our optimal MGeneral mechanism in most of the cases, which demonstrate the effectiveness of our proposed mechanisms in terms of both revenue and computational-efficiency. The Grid Search method achieves a similar revenue to MPractical mechanism in most of the cases since the MPractical mechanism actually belongs to the space of pricing mechanism covered by the Grid Search mechanism, and the Grid Search mechanism may not always achieve the best revenue within its feasible space due to its limited sampling iterations and grid search precision. For Gaussian distributions, as the size of sample space n increases, prior distributions of buyers 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, i.e., utility increase, to them. *MGeneral* makes use of buyers' uncertainty and can extract almost full surplus when n is relatively large. When $n = 12$, *MGeneral* achieves 99.91% revenue of Upper Bound, and outperforms *MPractical* and Fully Revealing by more than 9.5%, in the Gaussian distribution case. For uniform distributions, when $n = 12$, *MGeneral* extracts 96.25% revenue of Upper Bound. For Pareto distributions, the revenue for all mechanisms are lower compared with the other two distributions, because buyers have more confident and accurate prior estimations, and their prior expected utilities $u(\theta)$ are relatively high. In this case, it is hard to extract large additional 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 the parameter n denotes the discretization level of data, we can conclude that the seller can extract higher revenue by selling more fine-grained data.

We then evaluate the impacts of the number of buyer types $|\Theta|$ on the four mechanisms. We report the evaluation results in Fig. 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 distributions appear in the market, and their strategic behaviors raise more challenges to our pricing mechanisms to guarantee the properties of I.C. and I.R.. For Gaussian and uniform distributions, the



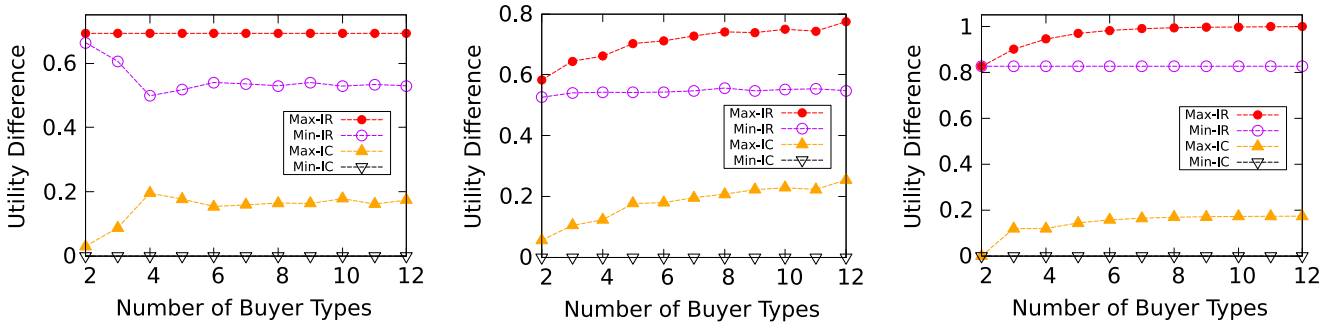
(a) Gaussian (b) Uniform (c) Pareto

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.



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

Fig. 4. Utility difference for I.R. and I.C. conditions under different number of buyer types.

average revenue of our mechanisms do not decrease much as $|\Theta|$ grows. This indicates that our mechanisms are robust against multiple types of strategic buyers under these two distributions. For Pareto distributions, however, the average revenue of our mechanisms decrease with $|\Theta|$ significantly. This is because buyers under Pareto distributions are confident about their prior estimations and thus have higher prior expected utilities before buying data from the seller. As more confident buyers join the market, seller's average revenue from each buyer certainly decreases.

We also validate the I.R. and I.C. properties of the complex MGeneral Mechanism empirically in the same settings of Fig. 3. Given a computed MGeneral mechanism, to verify the I.R. conditions, we compute the utility of each type choosing the corresponding mechanism, and for the I.C. conditions, we compute the utility differences for a specific type of buyer choosing different experiments other than the experiment corresponding to her true type. We record the maximum and minimum values among those I.R. utilities (could be regarded as utility difference between current utility and zero utility) and I.C. utility differences and present them in Fig. 4. As all the values are larger or equal to zero, our proposed mechanisms empirically satisfy the I.R. and I.C. conditions, which aligns with our theoretical results.

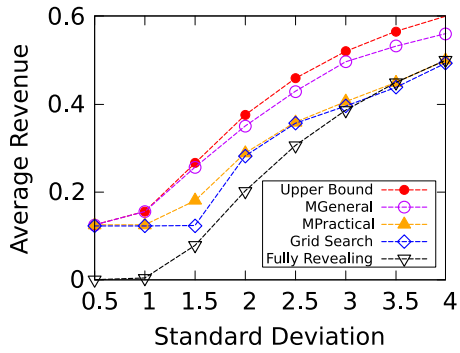


Fig. 5. Revenue under different values of standard deviation for Gaussian distributions.

We finally evaluate 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 observe in Fig. 5, *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.

From the above three sets of experiments, we could observe that the performance of Mpractical mechanism is always quite close to that of MGeneral mechanism. This result demonstrates that MPractical mechanism can still obtain good revenue in practical deployment, even we restrict the size of menu to be a constant.

C. Free Data Trials

We now evaluate the revenue increase when the seller has the option to offer free trials. In Fig. 6, we present the results where buyer types are uniformly sampled from the belief simplex, and the results of using other distributions are similar.

We start with the single overconfident buyer case. We vary the size of nature state space n from 2 to 12, and evaluate the average revenue with and without free trials. As we can see from Fig. 6(a), free trials are more effective when the nature state space is relatively small. When $n = 2$, free trials achieve a revenue increase of 57.44%, and when $n = 12$, the seller increases revenue by only 0.67% with free trials. On average, free trials give a 8.71% revenue increase to the seller.

We then evaluate the performance of Algorithm 2 in the general setting. We vary the number of nature states n from 2 to 12 in Fig. 6(b) and vary the number of buyer types $|\Theta|$ in Fig. 6(c). As we can see, free trials always achieve a stable revenue increase in all parameter settings. On average, free trials increase revenue by 2.86% and 8.20% in Fig. 6(b) and (c), respectively. We can conclude that free trials indeed can guarantee a revenue increase in certain scenarios.

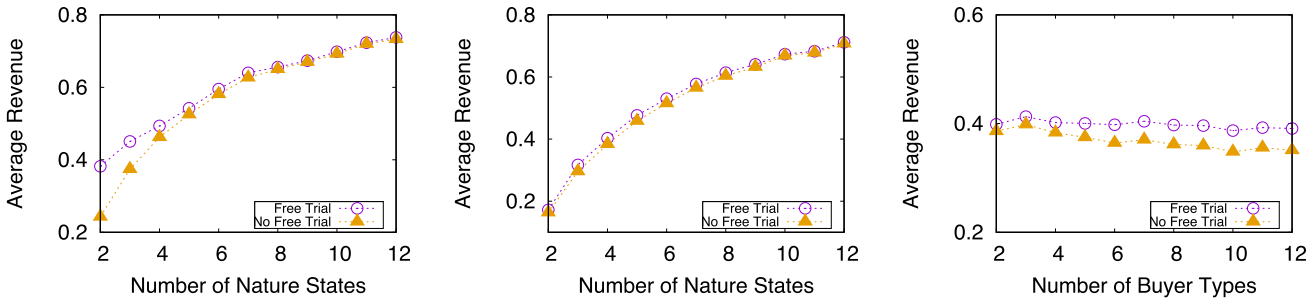
VI. Related Work

In recent years, designing data pricing frameworks has attracted increasing interests. Balazinska et al. [29] first envisioned the emergence of cloud-based data markets, and outlined potential challenges and research opportunities. Following this work, 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 [30], Koutris et al. first identified two key properties that a pricing function must satisfy, namely arbitrage-freeness and discount-freeness, and proposed a polynomial time algorithm that derives the price for common types of queries. Similar work in this direction include arbitrage-free pricing functions for arbitrary queries [11], and a scalable framework for pricing relational queries [31]. A set of accountable protocols named AccountTrade was proposed in [32] for Big Data trading among dishonest customers. These work assume data has been collected and structured before being priced, and their objective is not to maximize the revenue of the seller.

Data trading and exchange 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 [33]. The authors in [34] designed a decentralized infrastructure for IoT data trading based on blockchain technologies, but did not elaborate on the pricing mechanisms. A two-sided market for crowdsensed data was proposed in [35], and secondary market models for mobile data were studied in [36]. In a recent paper, Zheng et al. [14] took advantage of the geographical locality of sensory 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, and propose a new market model for IoT data trading.

Recently, data markets have also drawn increasing attention in the machine learning community. In [37], the authors studied how a machine learning system fairly distributes revenue to its training data contributors. They proposed a family of efficient algorithms to determine the data valuation based on Shapley value. Ghorbani et al. [38] also proposed a data valuation framework for supervised machine learning based on Shapley value. Agarwal et al. revised Shapley value to be robust to freely replicable goods in the context of data marketplaces [39]. The authors in [40] considered a transfer learning setting, and implemented a blockchain-based data marketplace that guarantees privacy and consumer's benefit. In a recent paper [41], the authors envisioned the challenges and opportunities in the valuation, pricing, and governance of AI data.

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



(a) Single Buyer - Nature States (b) Multiple Buyers - Nature States (c) Multiple Buyers - Buyer Types

Fig. 6. Average revenue with and without free trials.

and one receiver, was studied in [42]. In a model similar to ours [13], 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 [21], Babaioff et al. considered the optimal mechanism for selling information sequentially. Smolin [43] studied revenue-maximizing menus for pricing objects with several attributes. Papers [44] and [45] provide excellent surveys of the information design literature.

The preliminary version of this paper was published as a regular conference paper in [46]. In this full version, we give the complete proofs of Lemmas 1 and 2. We further extend the data exchange to a two-phase process, and show that the seller can increase revenue through deploying free data trials. We add two examples to illustrate the data trading process and how the seller extracts higher revenue through free data trials. We also update the references regarding the recent work on machine learning data markets.

VII. Conclusion

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

Acknowledgment

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.

References

[1] Gnip APIs, 2024. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/enterprise
[2] Xignite, 2024. [Online]. Available: https://www.xignite.com/
[3] Here, 2024. [Online]. Available: https://www.here.com/en
[4] IOTA, 2024. [Online]. Available: https://www.iota.org/
[5] Ambient maps, 2024. [Online]. Available: https://ambientmaps.com.au/
[6] DataBroker dao, 2024. [Online]. Available: https://www.databroker.global/
[7] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Sensing as a service model for smart cities supported by Internet of Things," *Trans. Emerg. Telecommun. Technol.*, 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 Gener. Comput. Syst.*, 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 Environ. Sci.*, 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 *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2013, pp. 613–624.
[11] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," in *Proc. VLDB Endowment*, vol. 7, pp. 757–768, 2014.
[12] L. Toka, B. Lajtha, É. Hosszu, B. Formanek, D. Géhberger, and J. Tapolcai, "A resource-aware and time-critical IoT framework," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.
[13] D. Bergemann, A. Bonatti, and A. Smolin, "The design and price of information," *Amer. Econ. Rev.*, vol. 108, no. 1, pp. 1–48, 2018.
[14] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2017, Art. no. 26.
[15] D. Bergemann and S. Morris, "Bayes correlated equilibrium and the comparison of information structures in games," *Theor. Econ.*, vol. 11, no. 2, pp. 487–522, 2016.
[16] J. D. Hartline and B. Lucier, "Bayesian algorithmic mechanism design," in *Proc. 42nd ACM Symp. Theory Comput.*, New York, NY, USA, 2010, pp. 301–310. [Online]. Available: https://doi.org/10.1145/1806689.1806732
[17] D. Niyato, D. T. Hoang, N. C. Luong, P. Wang, D. I. Kim, and Z. Han, "Smart data pricing models for the Internet of Things: A bundling strategy approach," *IEEE Netw.*, vol. 30, no. 2, pp. 18–25, Mar./Apr. 2016.
[18] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, "Hawk: The blockchain model of cryptography and privacy-preserving smart contracts," in *Proc. IEEE Symp. Secur. Privacy*, 2016, pp. 839–858.
[19] J. Treboux, A. J. Jara, L. Dufour, and D. Genoud, "A predictive data-driven model for traffic-jams forecasting in Smart Santader City-scale testbed," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops*, 2015, pp. 64–68.
[20] H. A. Simon, *Models of Bounded Rationality: Empirically Grounded Economic Reason*, vol. 3. Cambridge, MA, USA: MIT Press, 1997.
[21] M. Babaioff, R. Kleinberg, and R. Paes Leme, "Optimal mechanisms for selling information," in *Proc. 13th ACM Conf. Electron. Commerce*, 2012, pp. 92–109.
[22] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[23] Dialogfeed, 2024. [Online]. Available: https://www.dialogfeed.com/pricing/
[24] Gurobi, 2024. [Online]. Available: http://www.gurobi.com/
[25] Ambient sound monitoring network, 2024. [Online]. Available: https://data.smartdublin.ie/dataset/ambient-sound-monitoring-network
[26] D. B. Rubin, "The Bayesian bootstrap," *Ann. Statist.*, vol. 9, pp. 130–134, 1981.

[27] S. Bhattacherjee, A. Chavan, S. Huang, A. Deshpande, and A. G. Parameswaran, "Principles of dataset versioning: Exploring the recreation/storage tradeoff," in *Proc. VLDB Endowment*, vol. 8, no. 12, pp. 1346–1357, 2015.
[28] L. Phlips, *The Economics of Price Discrimination*. Cambridge, U.K.: Cambridge Univ. Press, 1983.
[29] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," in *Proc. VLDB Endowment*, vol. 4, pp. 1482–1485, 2011.
[30] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *Proc. 31st ACM SIGMOD-SIGACT-SIGAI Symp. Princ. Database Syst.*, 2012, pp. 167–178.
[31] S. Deep and P. Koutris, "QIRANA: A framework for scalable query pricing," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2017, pp. 699–713.
[32] T. Jung et al., "AccountTrade: Accountable protocols for Big Data trading against dishonest consumers," in *Proc. IEEE Conf. Comput. Commun.*, 2017, pp. 1–9.
[33] R. Montella, M. Ruggieri, and S. Kosta, "A fast, secure, reliable, and resilient data transfer framework for pervasive IoT applications," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 710–715.
[34] 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 *Proc. 7th Int. Conf. Internet Things*, 2017, Art. no. 15.
[35] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "Trading data in the crowd: Profit-driven data acquisition for mobile crowdsensing," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 486–501, Feb. 2017.
[36] 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 *Proc. IEEE Conf. Comput. Commun.*, 2015, pp. 1580–1588.
[37] R. Jia et al., "Towards efficient data valuation based on the shapley value," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 1167–1176.
[38] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 2242–2251.
[39] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *Proc. ACM Conf. Econ. Computation*, 2019, pp. 701–726.
[40] K. Sarpatwar, V. S. Ganapavarapu, K. Shanmugam, A. Rahman, and R. Vaculin, "Blockchain enabled AI marketplace: The price you pay for trust," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 2857–2866.
[41] R. Raskar, P. Vepakomma, T. Swedish, and A. Sharan, "Data markets to support AI for all: Pricing, valuation and governance," 2019, *arXiv: 1905.06462*.
[42] E. Kamenica and M. Gentzkow, "Bayesian persuasion," *Amer. Econ. Rev.*, vol. 101, no. 6, pp. 2590–2615, 2011.
[43] A. Smolin, "Disclosure and pricing of attributes," *RAND J. Econ.*, vol. 54, pp. 570–597, 2023. [Online]. Available: https://doi.org/10.1111/1756-2171.12451
[44] D. Bergemann and S. Morris, "Information design: A unified perspective," *J. Econ. Literature*, vol. 57, no. 1, pp. 44–95, 2019.
[45] S. Dughmi, "Algorithmic information structure design: A survey," *SIGecom Exchanges*, vol. 15, no. 2, pp. 2–24, 2017.
[46] W. Mao, Z. Zheng, and F. Wu, "Pricing for revenue maximization in IoT data markets: An information design perspective," in *Proc. IEEE Conf. Comput. Commun.*, 2019, pp. 1837–1845.

Zhenzhe Zheng (Member, IEEE) received the BE degree in software engineering from Xidian University, in 2012, and the MS and PhD degrees in computer science and engineering from Shanghai Jiao Tong University, in 2015 and 2018, respectively. He is an associate professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He has visited the University of Illinois at Urbana-Champaign (UIUC) as a visiting scholar and then a post doc research associate from 2016 to 2019. His research interests include intelligent mobile computing and large-scale decision-making. He is a recipient of NSFC Excellent Young Scholars Program, CCF-Intel Young Faculty Researcher Program Award, the China Computer Federation (CCF) Excellent Doctoral Dissertation Award. He has served as the member of technical program committees of several academic conferences, such as INFOCOM, MobiHoc, KDD, WWW, AAAI, IoTDI and etc. He is a member of the ACM, and CCF.

Weichao Mao (Student Member, IEEE) received the BS degree in computer science from Shanghai Jiao Tong University, in 2019, and the MS degree in electrical and computer engineering from the University of Illinois Urbana—Champaign (UIUC), in 2021. He is currently working toward the PhD degree with the Department of Electrical and Computer Engineering, University of Illinois Urbana—Champaign (UIUC). His research interests include reinforcement learning, game theory, control theory, and multi-agent systems.

Yidan Xing (Student Member, IEEE) received the BS degree in electrical and computer engineering from Shanghai Jiao Tong University, in 2022. She is currently working toward the PhD degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. Her research interests include algorithmic game theory and its applications, and multi-agent systems.

Fan Wu (Member, IEEE) received the BS degree in computer science from Nanjing University, in 2004, and the PhD degree in computer science and engineering from the State University of New York at Buffalo, in 2009. He is a professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He has visited the University of Illinois at Urbana-Champaign (UIUC) as a post doc research associate. His research interests include wireless networking and mobile computing, data management, algorithmic network economics, and privacy preservation. He has published more than 200 peer-reviewed papers in technical journals and conference proceedings. He is a recipient of the first class prize for Natural Science Award of China Ministry of Education, China National Fund for Distinguished Young Scientists, ACM China Rising Star Award, CCFTencent "Rhinoceros bird" Outstanding Award, and CCF-Intel Young Faculty Researcher Program Award. He has served as an associate editor of *IEEE Transactions on Mobile Computing* and *ACM Transactions on Sensor Networks*, an area editor of the *Elsevier Computer Networks*, and as the member of technical program committees of more than 100 academic conferences.