Capitalize Your Data: Optimal Selling Mechanisms for IoT Data Exchange

Qinya Li, Zun Li[‡], Zhenzhe Zheng, *Member, IEEE*, Fan Wu, *Member, IEEE,* Shaojie Tang[†], *Member, IEEE*, Zhao Zhang[§], *Member, IEEE*, and Guihai Chen, *Senior Member, IEEE*

Abstract—More and more IoT data is being traded online in cloud-based data marketplaces due to the fast-growing market demand. Within the current data selling mechanisms, data consumers have difficulties in making purchasing decisions due to uncertain IoT data quality and inflexible pricing interface. To resolve these issues, potential solutions could be to launch data demonstrations and release free sampling data to reduce the uncertainty about data quality, and to charge based on the volume of data actually used to enable flexible pricing. However, there is still no clear understanding of economic benefits of these mechanisms. In this paper, we design the optimal data selling mechanisms for IoT data exchange, and derive the following two results. First, whether to deploy a data demonstration and how much free sampling data to release depend on the extent of data consumers' inaccuracy perceptions for data quality, which varies over a wide range in IoT applications. We found that the data vendor has no incentive to conduct these strategies if data consumers extremely overestimate data quality. Second, although flexible data pricing mechanisms provide convenience for real-time and streaming IoT data exchange, it brings less economic benefits to the data vendor compared with the fixed pricing scheme, which sells the whole data set with a fixed price. We evaluate the optimal selling mechanisms on a real-world Taxi GPS data set, and evaluation results verify the insights derived from our theoretical analysis.

Index Terms—Data exchange, data quality, data pricing, optimal selling mechanism.

1 Introduction

NOWADAYS, data is becoming an important resource in diverse fields, such as finance [1], [2], advertising [3], [4], transportation [5] and etc. With the increasing market demand for data, a number of data vendors have emerged to collect, categorize and trade data on the Internet. For example, Quandl [1] releases financial and economic data for business decision, Factual [3] provides location data for mobile advertising, and Uber [5] publishes traffic data for urban planning. In order to facilitate data sharing and trading over the Internet, several data marketplaces, such as In fochimps [6], Dataexchange [7] and IOTA Data Marketplace [8], have provided centralized platforms for data vendors to sell data and data consumers to purchase the data needed.

The most common method to sell data is via RESTful APIs [1], [9], [10], [11], [12]. Data consumers submit parametrized queries as requests for data. For example, if one wants to purchase data from Yelp [12], she specifies the keywords of interest, such as the name of a restaurant, in the API call, and then Yelp would return the matched events up to a defined API call limit. Typically, the data consumers will be charged based on the total number of API calls.

- *Q. Li, Z. Zheng, F. Wu, G. Chen were with Shanghai Key Laboratory of Scalable Computing and Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. E-mail: {qinyali, zhengzhenzhe}@sjtu.edu.cn, {fwu, gchen}@cs.sjtu.edu.cn*
- *‡Z. Li was with Department of Computer Science and Engineering, University of Michigan, USA. E-mail: lizun@umich.edu*
- *†S. Tang was with Department of Information Systems, University of Texas at Dallas, USA. E-mail: tangshaojie@gmail.com*
- *§Z. Zhao was with College of Mathematics Physics and Information Engineering, Zhejiang Normal University, China. E-mail: zhaozhang@zjnu.cn*
- *Z. Zheng is the corresponding author.*

IoT data commonly is heterogenous, diverse, and with mass data volume. The data quality is uncertain. The current data selling mechanisms impose two problems for IoT data trading: one is uncertain data quality and the other is inflexible pricing interface. In IoT data markets [8], [13], the valuations over data and the decisions for purchasing data highly depend on the data quality, which is diverse and uncertain in most IoT applications. However, data consumers cannot obtain this information before purchasing data, forcing them to make improper purchasing decisions. To resolve this dilemma, some data vendors have deployed data demonstration and released free sampling data, to reveal signals about data quality. The data demonstration provides rough information, *e.g.*, categories, formats, geographic coverage, and etc; while the free sampling data, chosen from the actual data set, have more precise description over the data. The current data selling mechanism is inflexible in the sense that data consumers have to buy the whole data set (or a large number of API calls) even they only need a subset of data, which becomes more severe for real-time and streaming IoT data trading. To tackle this problem, recent work [14], [15], [16] introduced query-based data pricing, in which data consumers issue ad-hoc data queries and are charged based on the data used to answer the queries.

However, data vendors have concerned about these new data selling mechanisms, and hesitate to adopt them in practice. The data vendor does not clearly know market response to data demonstration and free sampling strategies, *e.g.*, whether these mechanisms can increase market demand or revenue? Thus, the data vendor has no idea when to deploy a data demonstration and how many free samples to release. Another unclear question for the data vendor is whether deploying flexible data pricing, such as

query-based pricing, can bring economic benefits, especially revenue increase? The goal of this paper is to answer these questions via rigorous mathematical analysis.

In this paper, we address the above mentioned issues, and design the optimal data selling mechanisms for IoT data exchange.

In order to characterize the market responses to data demonstration and free sampling strategies in an uncertain data quality environment, we first have to model data consumers' perceptions over IoT data quality. Considering that data is one kind of experience goods, data consumers can receive signals about the underlying data quality via watching a data demonstration or receiving free sampling data. With these signals, data consumers can calculate the posterior data quality through Bayesian learning, and then make data purchasing decisions based on this updated perception. Depending on the purchasing conditions, the data vendor can derive a specific demand function (and then an economic objective function), on which we can measure the market response to data demonstration and free sampling strategies. To choose between flexible and fixed data pricing schemes, we explicitly calculate the economic benefits of these two pricing schemes under a discounting valuation model, which captures the decreasing marginal valuations over data sets in practice [17]. We compare these two benefits to evaluate the economic incentive for data vendor to adopt flexible data pricing for IoT data trading.

We summarize the contributions of this paper as follows.

1) We propose a market model for IoT data selling in an uncertain data quality environment. The data consumers' perception over data quality is modeled as a Gaussian distribution, which has been widely used to describe the quality of IoT data. The data vendor can deploy data demo strategy, free strategy, sampling strategy and pure paid strategy to maximize her economic objective, which is a trade-off between revenue and social benefit. We then formulate the problem of designing optimal data selling mechanisms for IoT data exchange.
2) We present a Bayesian learning scheme for data consumers to update their perceptions over data quality, which determine data purchasing decisions. Based on the purchasing conditions of data consumers, we can explicitly express the data demand, and then a specific economic objective function.
3) We start with considering a benchmark case, in which data consumers exactly know the underlying data quality, to shed light on the design rationale of optimal data selling mechanisms. We further investigate the optimal selling mechanisms for the case of uncertain data quality, which is more pervasive in IoT applications. Our results show that when data consumers underestimate data quality too much, the optimal selling mechanism needs to release free samples to enhance data consumers' perception over data quality, attracting them to purchase data. In contrast, the data vendor has no incentive to offer free samples when the extent of overestimation to data quality exceeds a certain threshold.
4) We extend previous results to the flexible pricing with a discounting valuation model, in which data consumers have decreasing marginal valuations over data sets. We further show that the fixed pricing has higher economic benefits than the flexible pricing. Thus, the data vendor has less economic incentive to deploy flexible pricing, which explains the widespread adoption of fixed data pricing in practice.
5) We evaluate the optimal data selling mechanisms on a taxi GPS trace data set. The evaluation results verify our theoretical analysis. Based on evaluation results, we derive two conflict behaviors between the data vendor and data consumers, which demonstrate that market regulations are needed to eliminate these conflicts, facilitating the trading of IoT data.

The rest of this paper is organized as follows. In Section 2, we present our market model for IoT data selling. In Section 3, we determine a specific data demand through a Bayesian learning scheme. We also derive the optimal data selling mechanisms for the cases of certain data quality and uncertain data quality, respectively, in Section 4. We compare the economic benefits of flexible data pricing scheme and fixed data pricing scheme in Section 5. We evaluate the designed data selling mechanisms based on real-world data sets in Section 6. The related work is briefly reviewed in Section 7. We draw our conclusion in Section 8.

2 PRELIMINARIES

In this section, we describe a market model for IoT data trading, and formulate the problem of designing optimal data selling mechanisms from the perspective of a data vendor.

2.1 Market Model

Data Vendor: The data vendor launches an IoT data set for trading, which contains N data packages and is associated with an underlying data quality Q^*. One possible interpretation for the data quality Q^* could be the average accuracy of data packages. For example, the data set could be the GPS traces of cars from one city in a month. The GPS traces in each day, considered as one data package, may have various accuracies due to the noise during data acquisition and data processing. Since data consumers can not know the exact data quality before purchasing data, the data vendor could deploy a data demonstration, or offer free sampling data, to revise the data consumers' perceptions over data quality, For the non-sampling data, the data vendor charges a premium data access price p to extract revenue. In IoT data markets, data vendor determines three decision variables: data demo deployment indicator $\tau \in \{0, 1\}$, size of free sampling data $n \in [0, N]$, and a selling price $p \geq 0$ for the remaining $N - n$ non-sampling data package(s). Given a tuple of specific decision variables (τ^*, n^*, p^*), the data vendor can adopt the following four different data selling mechanisms.

Definition 1 (Data Selling Mechanisms)**.** *By the specific values of (τ^*, n^*, p^*), the data vendor can deploy*
(i) ***data demo strategy*** *if $\tau^* = 1$,*
(ii) ***free strategy*** *if $n^* = N, p^* = 0$,*
(iii) ***sampling strategy*** *if $n^* \in (0, N), p^* > 0$,*
(iv) or pure ***paid strategy*** *if $n^* = 0, p^* > 0$.*

Data Consumer: Data consumers are uncertain about the underlying data quality Q^*, and initially perceive that Q^* follows a Gaussian distribution with a mean $\widehat{Q}$ and a variance $\hat{\sigma}^2$, *i.e.*, $Q^* \sim \mathcal{N}(\widehat{Q}, \hat{\sigma}^2)$. Gaussian distribution has been widely used to model the value of IoT data, such as temperature, noise level and wind level [18], [19], [20], [21]. Data consumers can learn such a common prior belief from numerous exogenous sources, such as reviews, ratings, and "word of mouth". Before purchasing data, data consumers receive signals about data quality from data demo and free sampling data. Let Q_0^E denote the signal from watching a data demo. We further assume

$$Q_0^E \triangleq Q^* + \varepsilon_0, \qquad \text{where } \varepsilon_0 \sim \mathcal{N}(0, \sigma_0^2), \tag{1}$$

meaning that the signal Q_0^E provides noisy information about Q^*. We refer variance σ_0^2 as *demo variance*.

Similarly, sampling data also does not fully reveal Q^*, due to the inherent quality variability from data acquisition and data processing. Specifically, for the ith piece of sampling data, the data consumers experience data quality Q_i^E, which is also a noisy signal of Q^*:

$$Q_i^E \triangleq Q^* + \varepsilon_i, \qquad \text{where } \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{2}$$

Here, σ^2 captures the inherent data quality variability, and we refer it as *experience variance*. It is worth noting that the experience variance σ^2 is smaller than the demo variance σ_0^2, which is further less than the prior variance $\hat{\sigma}^2$, *i.e.*, $\sigma^2 \leq \sigma_0^2 \leq \hat{\sigma}^2$. For convenience of discussion, we assume these three types of variances satisfy the following relation:

$$\frac{\sigma_0^2}{\sigma^2} = \gamma, \quad \frac{\hat{\sigma}^2}{\sigma_0^2} = \gamma, \quad \frac{\hat{\sigma}^2}{\sigma^2} = \gamma^2,$$

where $\gamma > 1$ is referred to as a variance parameter.

After receiving signals from both data demonstration and n free sampling data, data consumers can update their perceptions of data quality in a Bayesian fashion, and get posterior data quality $Q(\tau, n)$, which will be discussed in Section 3

Valuation and Utility: Data consumers normally have large valuations over the data set with high data quality, but they may differ in the way to evaluate data quality. Data consumers integrate the purchased data into various IoT applications [22], [23], and thus could have different valuations for the data set even with the same quality. To capture such heterogeneity, we introduce a parameter θ for diverse preference over data quality, which is uniformly distributed in the interval $[0, 1]$.[1] In the fixed data pricing, each data consumer either purchases the whole data set, or stays with the n free data samples.[2] We normalize the valuation of free sampling data to be zero, and express a data consumer's valuation when purchasing the whole data set as:

$$v(n) \triangleq \theta \times (N - n) \times Q(\tau, n). \tag{3}$$

The valuation over the purchased data packages consists of two components: *private valuation* $\theta \times (N - n)$ and *common valuation* $Q(\tau, n)$. The parameter θ is the type of a data consumer, denoting her private valuation for each piece of non-sampling data. The posterior data quality $Q(\tau, n)$ can be regarded as a "common" valuation for all data consumers, which is derived from the identical data quality learning model. This linear valuation model is the simplest model for IoT data markets, and has been adopted in other markets [24].

Table 1: Key notations

Notation	Definition
τ	The indicator of data demonstration deployment
N	Size of data packages in an IoT data set
n	Size of free sampling data
p	Selling price
Q^*	The underlying data quality
$\mathcal{N}(\widehat{Q}, \hat{\sigma}^2)$	Gaussian distribution of Q^*
$Q(\tau, n)$	The posterior data quality
Q_0^E	The signal from deploying data demonstration
σ_0^2	The demo variance
Q_i^E	The data quality data consumers experienced from the ith piece of sampling data
σ^2	The experience variance that captures the inherent data quality variability
γ	Variance parameter
D	Data demand
v	The valuation of data consumer
u	The utility of data consumer
$\pi(\tau, n, p)$	The economic benefit of the data pricing mechanism
$w(n)$	The social benefit of n free sampling data
K	The maximum volume of non-sampling data packages that data consumers can buy
δ	The discounting factor

The utility of a data consumer is defined as the difference between the valuation over the purchased data and the price p charged by the data vendor:

$$u(p, n) \triangleq v(n) - p. \tag{4}$$

The price p is zero if consumers stay with free sampling data.

Data Demand: Data demand represents the percentage of data consumers buying data set in the market, and is determined by the decision variables τ, n, and p. We denote the data demand by $D(\tau, n, p, \mathbb{E}[Q(\tau, n)])$. We note that the posterior data quality $Q(\tau, n)$ is the private information of consumers, and the data vendor only has an expectation over such information. A feasible data demand satisfies several basic properties. First, the data demand decreases with price, *i.e.*, $\partial D/\partial p < 0$. Second, we require $\partial D/\partial Q > 0$, meaning that data demand depends positively on the expected posterior data quality. Third, the number of free sampling data n has direct and indirect effect on data demand. We derive data demand with respective to n: $\frac{\partial D}{\partial n} + \frac{\partial D}{\partial Q}\frac{\partial Q}{\partial n}$, where the term $\partial D/\partial n$ and the term

1. We can also use other kind of distribution for θ to derive the same results.

2. We will relax this assumption, and consider the flexible data pricing, in which data consumers can purchase any number of data, in Section 5.

$(\partial D/\partial Q) \times (\partial Q/\partial n)$ captures direct and indirect effects, respectively. If $\partial D/\partial n < 0$, then we have $\frac{\partial D}{\partial n} + \frac{\partial D}{\partial Q}\frac{\partial Q}{\partial n} < 0$. It means the data demand decreases with the sample size n. If $\partial D/\partial n > 0$ and $\partial D/\partial n$ is sufficiently large, then the indirect effect will be stronger than the direct effect so that $\frac{\partial D}{\partial n} + \frac{\partial D}{\partial Q}\frac{\partial Q}{\partial n} > 0$, and the data demand increases with the sample size n.

Revenue and Social Benefit: The data vendor adopts different data selling mechanisms from Definition 1 by making a trade-off between revenue and social benefit. Here, we define the revenue as the selling price of non-sampling data multiplies the data demand, *i.e.*, $p \times D(\tau, n, p, \mathbb{E}[Q(\tau, n)])$. The data vendor should also take social benefit of free sampling data into account during data trading. Releasing free sampling data can attract more data consumers, helping to discover the potential applications behind data set. In addition, the released high quality data can also improve the reputation or brand cognition of the data vendor, bringing new revenue in the future. This can be analogous to some kind of "advertising" for the data set. We quantify such advantage of launching free sampling data as the concept of social benefit, and use a general concave function $w(n)$ to represent the social benefit of n free sampling data. The data vendor integrates revenue and social benefit into her optimization objective.

2.2 Problem Formulation

In IoT data markets, the data vendor jointly optimizes the revenue and social benefit from data trading. The data vendor can extract revenue from selling non-sampling data, and obtain social benefit from releasing free sampling data. The data vendor determines three decision variables: τ, n, and p, to maximize the weighted average of revenue and social benefit. We can formulate the design of optimal data selling mechanism in IoT data markets as:

$$\max_{\tau,n,p} \quad \pi(\tau, n, p) \triangleq \alpha p D(\tau, n, p, \mathbb{E}[Q(\tau, n)]) + (1-\alpha)\omega(n) \quad (5)$$
$$s.t. \quad p \geq 0, \quad 0 \leq n \leq N, \quad \tau \in \{0, 1\},$$

where α is a weight parameter, measuring the proportion of revenue in the objective function. We call $\pi(\tau, n, p)$ as the economic benefit of the data pricing mechanism. It is worth to note that the problem of revenue maximization (*i.e.*, $\alpha = 1$) and the problem of social benefit maximization (*i.e.*, $\alpha = 0$) are nested within such formulation.

3 Data Demand Determination

To determine the data demand, we start with describing a Bayesian learning scheme for data consumers to update their perceptions over data quality after receiving signals from data demonstration and free sampling data. As discussed in Section 2.1, data consumers initially have a common prior Gaussian distribution $\mathcal{N}(\widehat{Q}, \hat{\sigma}^2)$ for Q^*. The posterior perception, *i.e.*, the posterior Gaussian distribution $\mathcal{N}(Q(\tau, n), \sigma^2(\tau, n))$, after receiving the data demo signal Q_0^E in (1) and n sampling data signals $\{Q_i^E, 1 \leq i \leq n\}$ in (2), can be given by the standard Bayesian analysis:

$$Q(\tau, n) = \tau \frac{1}{\sigma_0^2 S_n} Q_0^E + \frac{1}{\sigma^2 S_n} \sum_{i=1}^{n} Q_i^E + \frac{1}{\hat{\sigma}^2 S_n} \widehat{Q}, \quad (6)$$

$$\sigma^2(\tau, n) = \frac{1}{\tau \frac{1}{\sigma_0^2} + n \frac{1}{\sigma^2} + \frac{1}{\hat{\sigma}^2}}, \quad (7)$$

where $S_n = 1/\sigma^2(\tau, n)$. Equation (6) describes that the posterior data quality $Q(\tau, n)$ is a weighted average of the prior data quality and the received signals. We note that $Q(\tau, n)$ is a random variable across data consumers, because data consumers may receive different signals from data demonstration and free sampling data. This learning model is simple but appealing, as it captures the heterogeneity across data consumers in perceived data quality, even they start with the identical prior distribution. Equation (7) describes how data consumer's uncertainty over data quality declines after she has received a set of accumulated signals, implying that the greater extent of updating, *e.g.*, deploying data demo or increasing the size of free sampling data, the more accurate the posterior data quality. In the limit, the perceived quality $Q(\tau, n)$ converges to the underlying data quality Q^*.

We next derive a specific demand function based on the purchasing behaviors of data consumers with the above Bayesian learning scheme. A data consumer will buy the data set if the utility in (4) from purchasing the whole data set is non-negative, *i.e.*,

$$\theta \times (N - n) \times Q(\tau, n) - p \geq 0. \quad (8)$$

We recall that data demand is defined as the fraction of consumers that purchase the data set, *i.e.*, the consumers with θ that satisfies (8). Combining with the assumption that θ is uniformly distributed in $[0, 1]$, we can express the data demand as a function of the three decision variables, τ, n, and p

$$D(\tau, n, p) = \max\left\{0, 1 - \frac{p}{(N-n)\mathbb{E}[Q(\tau, n)]}\right\}. \quad (9)$$

It is worth noting that the data vendor uses the expected posterior data quality $\mathbb{E}[Q(\tau, n)]$ rather than the posterior data quality $Q(\tau, n)$. This is because the posterior data quality is private information of data consumers, and is unknown to the data vendor. We now calculate such expected posterior data quality. We assume that the data vendor knows the common prior quality distribution $\widehat{Q}$ and $\hat{\sigma}^2$, data quality Q^*, demo variability σ_0^2 and experience variability σ^2. This information can be learned by conducing standard market research, such as survey. According to equations (1) and (2), we have $\mathbb{E}\left[Q_0^E\right] = Q^*$ and $\mathbb{E}\left[Q_i^E\right] = Q^*$ for all $1 \leq i \leq n$. Together with (6), we can derive

$$\mathbb{E}[Q(\tau, n)] = \tau \frac{1}{\sigma_0^2 S_n} Q^* + \frac{1}{\sigma^2 S_n} n Q^* + \frac{1}{\hat{\sigma}^2 S_n} \widehat{Q}, \quad (10)$$

We substitute this expected data quality into (9), and obtain the data demand function. We note that this demand function satisfies the basic properties in Section 2.1.

4 Optimal Data Selling Mechanisms

In this section, we design the optimal data selling mechanisms for two cases. We first analyze the benchmark case, in which consumers exactly know the underlying data quality Q^*. In this case, the data demonstration and sampling mechanisms do not affect consumers' perceptions over data quality. For convenience of discussion, we allow the size of free sampling n to be any real number in $(0, N]$, which is justified when N is large. We specify the concave function $w(n)$ of social benefit as $\beta \log(n+1)$, where β is a weighted parameter.

Certain Data Quality: When data consumers know the data quality Q^*, the data demand in (9) is

$$D(n,p) = \max\left\{0, 1 - \frac{p}{(N-n)Q^*}\right\}. \tag{11}$$

In this case, data demonstration does not affect the value of objective, and thus the data vendor only has to determine the free sampling size n and the selling price p to maximize

$$\pi(n,p) = \alpha p\left(1 - \frac{p}{(N-n)Q^*}\right) + (1-\alpha)\beta\log(n+1). \tag{12}$$

According to the first-order condition, the optimal selling price $p^*(n)$ for a given sampling size n is:

$$p^*(n) = \frac{(N-n)Q^*}{2}. \tag{13}$$

We substitute the above optimal price into the objective function in (12), and obtain

$$\pi(n) = \alpha \times \frac{Q^*}{4} \times (N-n) + (1-\alpha) \times \beta \times \log(n+1). \tag{14}$$

The corresponding derivative of $\pi(n)$ is

$$\pi'(n) = -\frac{\alpha Q^*}{4} + (1-\alpha)\frac{\beta}{n+1}. \tag{15}$$

The optimal sampling size n^* satisfies the first-order condition:

$$\pi'(n^*) = 0 \;\Rightarrow\; n^* = \frac{4\beta}{\lambda Q^*} - 1, \tag{16}$$

where $\lambda \triangleq \alpha/(1-\alpha)$ is the ratio of weight parameters for revenue and social benefit. Substituting the optimal n^* into (13), we can express the optimal price with λ and Q^*:

$$p^* = \frac{(N+1)\lambda Q^* - 4\beta}{2\lambda}. \tag{17}$$

We can observe that the parameters λ and Q^* have opposite effects on the optimal price in (17) and the optimal sampling size in (16). Specifically, the optimal price p^* increases in λ and Q^*, while the optimal sampling size n^* decreases in λ and Q^*. Furthermore, p^* increases in the size of data set N, while n^* is independent on N.

The following theorem characterizes the optimal data selling mechanism in the setting with certain data quality Q^*.

Theorem 1. *When data consumers know IoT data quality Q^*, there are two cut-off values for the weight ratio λ,* i.e.,

$$\underline{\lambda} = \frac{4\beta}{Q^*(N+1)}, \;\; \textit{and} \;\; \overline{\lambda} = \frac{4\beta}{Q^*},$$

such that

▷ *if $\lambda \le \underline{\lambda}$, the data vendor would deploy free strategy,* i.e., *$n^* = N$ and $p^* = 0$.*

▷ *if $\underline{\lambda} < \lambda < \overline{\lambda}$, the data vendor would deploy sampling strategy,* i.e., *$n^* = \frac{4\beta}{\lambda Q^*} - 1, p^* = \frac{(N+1)\lambda Q^* - 4\beta}{2\lambda}$.*

▷ *if $\lambda \ge \overline{\lambda}$, the data vendor would launch paid strategy,* i.e., *$n^* = 0$, $p^* = \frac{NQ^*}{2}$.*

Proof. The data vendor determines the optimal free sampling size n^* to maximize $\pi(n)$ in (14). There are three possible solutions for this optimization problem, *i.e.*, an interior solution given in (16), two corner solutions $n^* = N$ and $n^* = 0$, which correspond to the three data selling strategies, respectively. The derivative function $\pi'(n)$ in (15) decreases with n. Thus, for a corner solution involving $n^* = N$, the Karush-Kuhn-Tucker conditions require that

$$\pi'(N) \ge 0 \quad \Rightarrow \quad \lambda \le \frac{4\beta}{Q^*(N+1)}.$$

At the other extreme, when $n^* = 0$, the Karush-Kuhn-Tucker conditions imply that

$$\pi'(0) \le 0 \quad \Rightarrow \quad \lambda \ge \frac{4\beta}{Q^*}.$$

It is easy to check that in the condition $\frac{4\beta}{Q^*(N+1)} < \lambda < \frac{4\beta}{Q^*}$, the optimization problem has only one unique interior solution $n^* = \frac{4\beta}{\lambda Q^*} - 1$.

Substituting n^* into (13), we can derive the corresponding optimal selling prices in these three cases. □

Uncertain Data Quality: When data consumers are uncertain about the data quality Q^*, we have derived the expected data demand in (9). Then, the data vendor determines τ, n, and p to maximize

$$\pi(\tau,n,p) = \alpha \times p \times \left(1 - \frac{p}{(N-n)\mathbb{E}[Q(\tau,n)]}\right) + (1-\alpha) \times \beta \log(n+1), \tag{18}$$

where $\mathbb{E}[Q(\tau,n)]$ is the expected posterior data quality in (10). The data vendor can decide whether to deploy a data demo by simply comparing the values of solutions when τ is 1 and 0, respectively. In the following discussion, we set $\tau = 0$, and focus on the determination of n and p.

Similar to the benchmark case, we can obtain the optimal price function with respective to the sampling size n

$$p^*(n) = \frac{(N-n)\mathbb{E}[Q(0,n)]}{2}. \tag{19}$$

The expected posterior data quality in (10) becomes

$$\mathbb{E}[Q(0,n)] = \frac{nQ^*}{\sigma^2 S_n} + \frac{\widehat{Q}}{\hat{\sigma}^2 S_n} = \frac{nQ^*}{n + \frac{\sigma^2}{\hat{\sigma}^2}} + \frac{\widehat{Q}}{n\frac{\hat{\sigma}^2}{\sigma^2} + 1} = \frac{\gamma^2 nQ^* + \widehat{Q}}{\gamma^2 n + 1} = Q^* - \frac{Q^* - \widehat{Q}}{n\gamma^2 + 1}, \tag{20}$$

where $S_n = 1/\sigma^2(0,n) = n/\sigma^2 + 1/\sigma^2$. Substituting $p^*(n)$ and $\mathbb{E}[Q(0,n)]$ back into the objective function in (18), we can rewrite it as

$$\pi(n) = \frac{\alpha}{4}(N-n)\left(Q^* - \frac{Q^* - \widehat{Q}}{\gamma^2 n + 1}\right) + (1-\alpha)\beta\log(n+1).$$

The derivative of $\pi(n)$ is

$$\pi'(n) = \frac{\alpha}{4}\left(\frac{(Q^* - \widehat{Q})(\gamma^2 N + 1)}{(\gamma^2 n + 1)^2} - Q^*\right) + (1-\alpha)\frac{\beta}{n+1}. \quad (21)$$

We observe that objective function $\pi(n)$ has different properties when $\widehat{Q}$ and Q^* have different relations. Specifically, if data consumers underestimate data quality, *i.e.*, $\widehat{Q} < Q^*$, $\pi'(n)$ is monotonically decreasing, and thus $\pi(n)$ is strictly concave. When data consumers overestimate data quality, *i.e.*, $\widehat{Q} \geq Q^*$, the property of $\pi(n)$ is a bit complicated: the revenue term decreases with n and is convex, while the social benefit term decreases with n but is concave. We characterize the optimal n^* and p^* when consumers underestimate and overestimate data quality in Theorem 2 and Theorem 3, respectively. When $\tau = 1$, we can get the similar conclusions. To avoid repetition, we do not describe again.

Theorem 2. *In the case that data consumers underestimate IoT data quality,* i.e., $\widehat{Q} \leq Q^*$, *there are two thresholds*

$$\underline{\lambda} = \frac{4\beta}{(N+1)}\frac{(\gamma^2 N+1)}{(\gamma^2 N Q^* + \widehat{Q})} \quad \textit{and} \quad \overline{\lambda} = \frac{4\beta}{\gamma^2 N(\widehat{Q} - Q^*) + \widehat{Q}},$$

such that

• *Case A: data quality gap* $\frac{\widehat{Q}}{Q^*}$ *satisfies* $\frac{\gamma^2 N}{1+\gamma^2 N} < \frac{\widehat{Q}}{Q^*} \leq 1$,

▷ *if* $\lambda \leq \underline{\lambda}$, *the data vendor would deploy free strategy,* i.e., $n^* = N$ *and* $p^* = 0$.

▷ *if* $\underline{\lambda} < \lambda < \overline{\lambda}$, *the data vendor would deploy sampling strategy,* i.e., $n^* = \arg\{\pi'(n) = 0\}$ *and* $p^*(n^*)$.

▷ *if* $\lambda \geq \overline{\lambda}$, *the data vendor would launch paid strategy,* i.e., $n^* = 0$, $p^* = \frac{N\widehat{Q}}{2}$.

• *Case B: data quality gap* $\frac{\widehat{Q}}{Q^*}$ *satisfies* $\frac{\widehat{Q}}{Q^*} \leq \frac{\gamma^2 N}{1+\gamma^2 N}$,

▷ *if* $\lambda \leq \underline{\lambda}$, *the data vendor would deploy free strategy,* i.e., $n^* = N$ *and* $p^* = 0$.

▷ *if* $\lambda > \underline{\lambda}$, *the data vendor would deploy sampling strategy,* i.e., $n^* = \arg\{\pi'(n) = 0\}$ *and* $p^*(n^*)$.

Proof. In the case of underestimate data quality, *i.e.*, $\widehat{Q} \leq Q^*$, the objective function $\pi(n)$ is strictly concave, because $\pi'(n)$ is a decreasing function. For such concave maximization problem, the optimal solution may stay at the interior point $n^* = \arg\{\pi'(n) = 0\}$, or the two extreme points: $n^* = N$ and $n^* = 0$. These three solutions correspond to the three data selling strategies. We derive the conditions of these three solutions by distinguishing the following two cases.

• Case A: when $\frac{\gamma^2 N}{1+\gamma^2 N} Q^* < \widehat{Q} \leq Q^*$, the analysis is similar to that in Theorem 1. Considering that $\pi'(n)$ is decreasing with respective to n, if the optimal solution is the corner solution $n^* = N$, the Karush-Kuhn-Tucker (KKT) conditions imply that

$$\pi'(N) \geq 0 \Rightarrow \lambda \leq \underline{\lambda} = \frac{4\beta}{(N+1)}\frac{(\gamma^2 N+1)}{(\gamma^2 N Q^* + \widehat{Q})}.$$

When the optimal solution stays at the other extreme point, $n^* = 0$, the KKT conditions require that

$$\pi'(0) \leq 0 \Rightarrow \lambda \geq \overline{\lambda} = \frac{4\beta}{\gamma^2 N(\widehat{Q} - Q^*) + \widehat{Q}}.$$

We note that this derivation holds when the denominator of $\overline{\lambda}$ is positive, *i.e.*, $\frac{\gamma^2 N}{1+\gamma^2 N} Q^* < \widehat{Q}$.

If $\underline{\lambda} < \lambda < \overline{\lambda}$, the concave maximization problem has one unique interior solution $n^* = \arg\{\pi'(n) = 0\}$.

• Case B: when $\widehat{Q} \leq \frac{\gamma^2 N}{1+\gamma^2 N} Q^*$, we have

$$\pi'(0) = \frac{\alpha}{4}((Q^* - \widehat{Q}) \times (\gamma^2 N + 1) - Q^*) + \beta(1-\alpha) \geq 0,$$

meaning that the optimal solution would not be at point $n^* = 0$. Similarly, we can obtain the conditions for $n^* = N$ and $n^* = \arg\{\pi'(n) = 0\}$ are $\lambda \leq \underline{\lambda}$ and $\lambda > \underline{\lambda}$, respectively. □

The above result is consistent with the insight from Theorem 1, and Theorem 2 reduces to Theorem 1 if data consumers have correct estimations over the data quality, *i.e.*, $\widehat{Q} = Q^*$.

In contrast to the underestimate case, it is complicated to characterize the conditions for the data vendor's optimal sampling and pricing decisions analytically in the overestimate case. Nevertheless, we have the following result.

Theorem 3. *In the scenario data consumers overestimate IoT data quality,* i.e., $\widehat{Q} > Q^*$, *the objective function* $\pi(n)$ *is neither concave nor convex. The optimal sampling size is* $n^* = \arg\max\{\pi(0), \pi(n_1^*), \pi(n_2^*), \pi(n_3^*), \pi(N)\}$, *where* n_1^*, n_2^*, *and* n_3^* *are three interior solutions obtained by solving* $\pi'(n) = 0$. *The corresponding optimal selling price is* $p^*(n^*)$ *given by (19).*

The proof for this theorem is straightforward. From standard optimization theory [25], for a differentiable function, a global maximum either must be a local extrema (stationary point) or must lie on the boundary of the domain. We will use specific parameters derived from a real-world data set to show how to determine the optimal mechanism for this case in Section 6.

5 Extensions to Discounting Setting

In previous section, data consumers have inflexible purchasing options: either staying with free data samples or buying the whole data set. In this section, we consider a flexible data selling scenario, in which consumers are allowed to buy any data subset. We derive the optimal mechanisms under various settings. We further show the result that the data vendor has no economic incentive to adopt the flexible pricing scheme.

We extend the market model by introducing a discounting valuation function [17]. This function is motivated by the observation that data consumers always have decreasing marginal valuations over the data set in practice, which are also known as the law of marginal utility in economics. Specifically, the valuation for buying $k \in [0, K]$ non-sampling data packages is defined as

$$\bar{v}(k) = \theta(1 + \delta + \cdots + \delta^{k-1}) \times Q(\tau, n),$$

where K is the maximum volume of non-sampling data packages that data consumers can buy and δ is the discounting factor. In this extended discounting model, the private valuation of k non-sampling data packages is $\theta \times \frac{1-\delta^k}{1-\delta}$, rather than $\theta \times k$ in (3). The common valuation remains to be the posterior data quality, *i.e.*, $Q(\tau, n)$. Given a unit price p_0, the utility of purchasing k data packages becomes

$$\bar{u}(k) = \bar{v}(k) - k \times p_0, \quad k \in [0, K].$$

We now derive the market demand of exactly buying k data packages. The data consumer chooses to buy k data packages if and only if $k = \arg\max \bar{u}(k')$, which is equivalent to $\bar{u}(k) \geq \bar{u}(k-1)$ and $\bar{u}(k) \geq \bar{u}(k+1)$ in the discrete domain. Thus, the marginal type θ_k can be obtained by setting $\bar{u}(k) = \bar{u}(k-1)$. By simple calculation, we can get

$$\theta_k = \frac{p_0}{\mathbb{E}[Q(\tau,n)] \times \delta^{k-1}}. \tag{22}$$

Then, the market demand for buying k data packages is

$$D_k(p_0) = \begin{cases} \theta_{k+1} - \theta_k, & k = 1, 2, \ldots, K-1 \\ 1 - \theta_K & k = K \end{cases} \tag{23}$$

We note that the demand could not be negative, and then we have an additional constraint: $\theta_K \leq 1$.

With the demand $D_k(p_0)$ for each possible k, we can determine the optimal unit price p_0^*. In discounting valuation setting, the objective function in (5) becomes

$$\begin{aligned} \pi(n, p_0) &= \alpha \sum_{k=1}^{K} k p_0 \times D_k(p_0) + (1-\alpha) \times \beta \log(n+1) \\ &= \alpha \left[K \times p_0 \times (1 - \theta_K) + \sum_{k=1}^{K-1} k \times p_0 \times (\theta_{k+1} - \theta_k) \right] \\ &\quad + (1-\alpha) \times \beta \log(n+1) \\ &= \alpha \left[K p_0 - \frac{p_0^2}{\mathbb{E}[Q(\tau,n)]} \frac{1 - \left(\frac{1}{\delta}\right)^K}{1 - \frac{1}{\delta}} \right] + (1-\alpha)\beta \log(n+1) \end{aligned} \tag{24}$$

We derive $\pi(n, p_0)$ with respect to p_0, and set it to be zero. We then get the optimal price function with a certain n

$$p_0^*(n) = \frac{\mathbb{E}[Q(\tau,n)]}{2} \times \frac{K\left(1 - \frac{1}{\delta}\right)}{1 - \left(\frac{1}{\delta}\right)^K}. \tag{25}$$

Plugging this optimal price back to (24), we get

$$\pi(n) = \alpha \frac{\mathbb{E}[Q(\tau,n)]}{4} \frac{K^2\left(1 - \frac{1}{\delta}\right)}{1 - \left(\frac{1}{\delta}\right)^K} + (1-\alpha)\beta \log(n+1) \tag{26}$$

We note that the constraint $\theta_K \leq 1$ determine the value of K and affect the feasible range of n. Substituting (25) into (22) with $k = K$, we can get

$$\theta_K = \frac{K\left(1 - \frac{1}{\delta}\right)}{2\delta^{K-1} \times \left(1 - \left(\frac{1}{\delta}\right)^K\right)} = \frac{K(1-\delta)}{2 \times (1 - \delta^K)}.$$

We can check that θ_K increases with K, and thus for a given δ, there exists a $K^*(\delta)$ such that $\theta_{K^*(\delta)} \leq 1$ and $\theta_{K^*(\delta)+1} > 1$. Considering that $K = \min\{K^*(\delta), N-n\}$, we further distinguish two cases:

► Case A: $K^*(\delta) \leq N-n$. We then have $K = K^*(\delta)$ and $n \in [0, N - K^*(\delta)]$. The objective function in (26) becomes

$$\pi_1(n) = \alpha \frac{\mathbb{E}[Q(0,n)]}{4} \frac{(K^*(\delta))^2\left(1 - \frac{1}{\delta}\right)}{1 - \left(\frac{1}{\delta}\right)^{K^*(\delta)}} + (1-\alpha)\beta \log(n+1). \tag{27}$$

The derivative of this objective function is

$$\pi_1'(n) = \alpha \frac{\gamma^2(Q^* - \widehat{Q})}{(\gamma^2 n + 1)^2} \times C(\delta) + (1-\alpha)\frac{\beta}{n+1},$$

where $C(\delta) = \frac{(K^*(\delta))^2\left(1-\frac{1}{\delta}\right)}{4 \times \left(1 - \left(\frac{1}{\delta}\right)^{K^*(\delta)}\right)}$ is a constant, and is only related to δ. We use the following theorem to characterize the optimal data selling mechanism in this scenario.

Theorem 4. *In the case that data consumers underestimate the data quality,* i.e., $\widehat{Q} \leq Q^*$, *the optimal data selling mechanism is* $n^* = N - K^*(\delta)$ *and* $p_0^*(n^*, K^*(\delta))$. *In the case that data consumers overestimate the data quality,* i.e., $\widehat{Q} > Q^*$, *the optimal data selling mechanism is* $n^* = \arg\max\{\pi(0), \pi(n_a^*), \pi(n_b^*), \pi(N - K^*(\delta))\}$ *and* $p_0^*(n^*, K^*(\delta))$, *where* n_a^* *and* n_b^* *are two interior solutions obtained by solving* $\pi'(n) = 0$.

The proof is straightforward, and similar to that in Theorem 2. Due to the limitation of space, we omit the proof here.

► Case B: $K^*(\delta) > N - n$. We then have $K = N - n$ and $n \in [\max\{0, N - K^*(\delta)\}, N]$. The objective function becomes

$$\pi_2(n) = \alpha \frac{\mathbb{E}[Q(0,n)]}{4} \frac{(N-n)^2\left(1 - \frac{1}{\delta}\right)}{1 - \left(\frac{1}{\delta}\right)^{N-n}} + (1-\alpha)\beta \log(n+1) \tag{28}$$

To maximize $\pi_2(n)$, it is not possible to obtain a closed form solution. We chose specific values for the parameters, and derive the optimal n^* and p^* in Section 6.

We now consider whether data vendor has economic incentive to deploy the above flexible pricing scheme. We first derive the objective of the fixed pricing scheme, in which data consumers have to choose between buying the whole data set or staying at free samples, in the discounting setting. Similar to (22), we can get the marginal type

$$\theta_K = \frac{K \times p_0}{\mathbb{E}[Q(\tau,n)] \times \sum_{i=0}^{K-1} \delta^i}, \tag{29}$$

and the market demand for buying the whole data set is

$$D_K(p_0) = \max\left\{0, 1 - \frac{K \times p_0}{\mathbb{E}[Q(\tau,n)] \times \sum_{i=0}^{K-1} \delta^i}\right\}.$$

We substitute the demand into the objective function in (5)

$$\pi_3(n, p_0) = \alpha K p_0 \left(1 - \frac{K p_0}{\mathbb{E}[Q(\tau,n)] \sum_{i=0}^{K-1} \delta^i}\right) + (1-\alpha)\beta log(n+1). \tag{30}$$

We derive $\pi_3(n, p_0)$ with p_0, and obtain the optimal unit price

$$p_0^* = \frac{\mathbb{E}[Q(\tau,n)]}{2 \times K \times \sum_{i=0}^{K-1} \delta^i}.$$

Substituting p_0^* into (29), we get that the marginal type is $\theta_K = 1/2$, and thus the demand is $D_K(p_0^*) = 1/2$. Putting p_0^* back to $\pi_3(n, p_0)$ in (30), we get

$$\pi_3(n) = \alpha \frac{\mathbb{E}[Q(0,n)]}{4} \sum_{i=0}^{N-n-1} \delta^i + (1-\alpha)\beta \log(n+1). \tag{31}$$

Here, we use the fact $\pi_3(n)$ increases with K and $K \leq N - n$.

We compare $\pi_3(n)$ of the fixed pricing scheme with the objective $\pi_1(n)$ and $\pi_2(n)$ of the flexible pricing scheme. For $\pi_1(n)$ in (27), we have the following relation

$$\frac{(K^*(\delta))^2\left(1 - \frac{1}{\delta}\right)}{1 - \left(\frac{1}{\delta}\right)^{K^*(\delta)}} \leq \frac{(N-n)^2}{\sum_{i=0}^{N-n-1} \delta^{-i}} \leq \sum_{i=0}^{N-n-1} \delta^i. \tag{32}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2021.3113387, IEEE Transactions on Mobile Computing

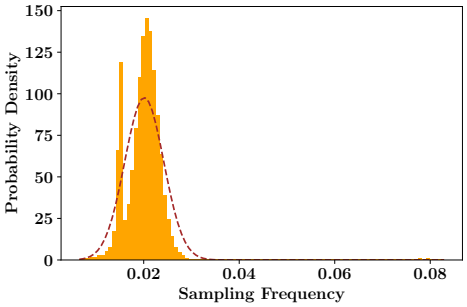


Figure 1: The probability density function of sampling frequency of taxi GPS data set.

The first inequality follows from the monotonicity of function $G(K) = K(1-\delta)/(1-\delta^K)$, and the second inequality follows from the Cauchy-Schwarz inequality. By this result, we can derive that $\pi_1(n) \leq \pi_3(n)$ for any given n. For $\pi_2(n)$ in (28), we can use the second inequality in (32) to get the similar result that $\pi_2(n) \leq \pi_3(n)$ for any n. From the above analysis, we have $\max\{\pi_1(n_1^*), \pi_2(n_2^*)\} \leq \pi_3(n_3^*)$, where n_1^*, n_2^* and n_3^* are the optimal sampling sizes in the corresponding scenarios, respectively. Thus, we can conclude that the economic objective of the flexible pricing scheme is less than that of the inflexible pricing scheme. Our result demonstrates that bundling mechanism [26] could be more profitable in IoT data markets. We characterize this result in the following theorem.

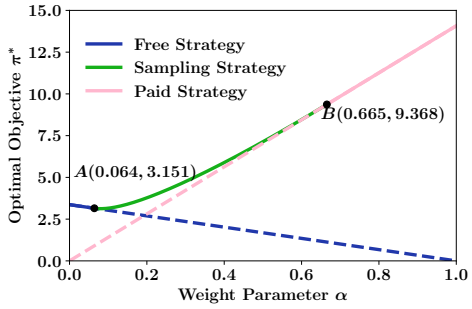
Theorem 5. *In discounting valuation setting, the fixed data pricing scheme has more economic benefit, compared with the flexible data pricing scheme. Thus, the data vendor has no economic incentive to launch the flexible data pricing scheme.*

6 Evaluation Results

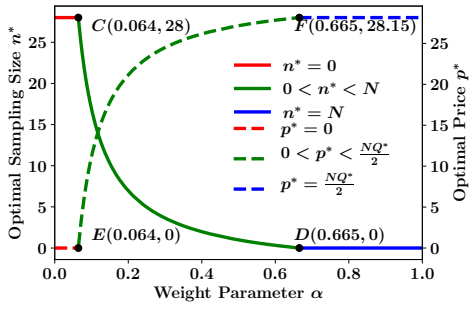
In this section, we report the evaluation results of the designed optimal data selling mechanisms on a real-world GPS trace dataset collected from Shanghai taxis in 2007 [27].

Taxi GPS Dataset: The data set consists of $N = 28$ data packages, where each represents one day of data collected in February 2007. Each data package involves around 2000 files, each of which is collected by one taxi on the corresponding day. Each file further contains around 2000 messages, which records the information of date, time, taxi ID, GPS location and whether there are passengers in taxi.

We adopt *sampling frequency*, the number of messages recorded every second, as data quality in our evaluation. Due to the unreliable wireless communication, there is high data loss during IoT data acquisition, and thus the metric of sampling frequency is critical for data consumers. For the above data set, we can calculate the sampling frequency of each file. The data quality of each data package is defined as the average sampling frequency of files within this data package. The data quality of the whole data set, *i.e.*, Q^*, is the average data quality of all data packages. We assume the data quality of files are independent identical random variables. According to Central Limit Theorem, the data quality of the whole data set follows a normal distribution. We illustrate the probability density function of the sampling frequency of our data set in Figure 1. We can calculate



(a) The optimal objective value π^*.



(b) Optimal n^*, Optimal p^*.

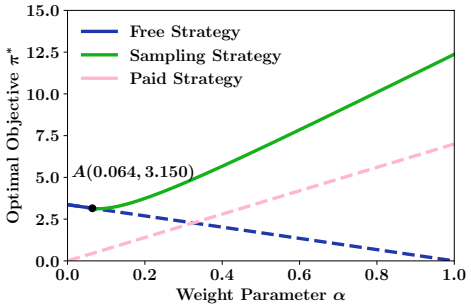
Figure 2: Optimal data selling strategy (π^*, n^*, p^*) in certain data quality case.

the data quality Q^* as 0.02011 message per second, with a variance 2.6075×10^{-5}. For convenience of discussion, we normalize the mean to $Q^* = 2.011$ and the variance to $\sigma_0^2 = 2.6075 \times 10^{-4}$.

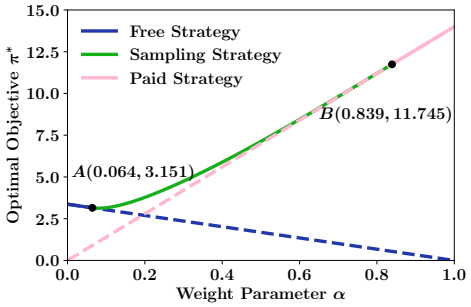
In the following discussion, we investigate the optimal data selling mechanisms under certain data quality and uncertain data quality settings, respectively. Since it is straightforward to check whether launching data demonstration is optimal, we omit the evaluation results of data demo strategy.

6.1 Certain Data Quality

In Figure 2(a), given different weight parameters α, the blue, green, red lines correspond to the optimal π^* for free, sampling, paid strategies, respectively. We use solid lines to denote the optimal data selling strategy. The evaluation results confirm our analysis in Theorem 1. With $N = 28$, $Q^* = 2.011$ and $\beta = 1$, the two cut-off values for the weight ratio λ are $\underline{\lambda} = 0.0686$ and $\overline{\lambda} = 1.989$, and the corresponding weight parameters are $\underline{\alpha} = 0.0642$ and $\overline{\alpha} = 0.665$, respectively. We can observe from Figure 2(a) that if α is less than $\underline{\alpha}$, *i.e.*, $\lambda \leq \underline{\lambda}$, free strategy is the optimal strategy; for an intermediate level of α, *i.e.*, $\underline{\alpha} < \alpha < \overline{\alpha}$, the sampling strategy that jointly considers revenue and social benefit is optimal. If α is greater than $\overline{\alpha}$, the paid strategy becomes optimal. We also denote the two turning points, *i.e.*, the tangent points of sampling line with free line and paid line, as points A and B in Figure 2(a). Figure 2(b) shows the optimal sampling size n^* and price p^* with different weight parameters α. The reason for this trend is that the data vendor prefers to generate revenue and cares less about social benefit when α becomes large.

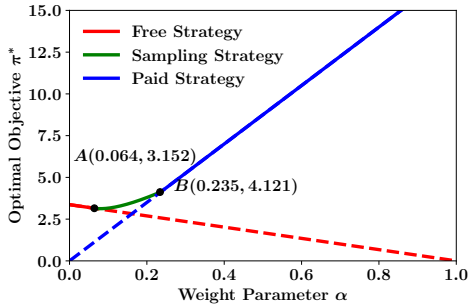


(a) Large Underestimate.

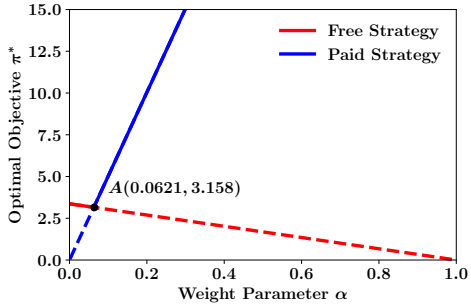


(b) Slight Underestimate.

Figure 3: Optimal π^* in underestimate data quality case.



(a) Moderate Overestimate.



(b) Large Overestimate.

Figure 4: Optimal π^* in overestimate data quality case.

6.2 Uncertain Data Quality

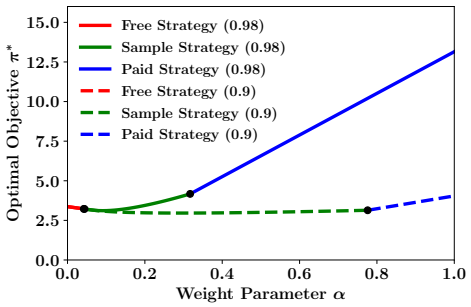
We first consider the case that data consumers underestimate data quality, *i.e.*, $\widehat{Q} \leq Q^*$. We set $\gamma = 2$ and $\beta = 1$ in this set of evaluation. The threshold of data quality gap $\frac{\gamma^2 N}{1+\gamma^2 N}$ is 0.991. We recall that in Theorem 2 there are two cases: large underestimate $\frac{\widehat{Q}}{Q^*} \leq 0.991$ and slight underestimate $0.991 < \frac{\widehat{Q}}{Q^*}$. For the case of large underestimate, we set $\widehat{Q}$ to be 1, and plot π^* in Figure 3(a). From the figure, we can find that the paid strategy would no longer be optimal when consumers underestimate data quality too much. In this case, data vendor has to offer free samples to enhance consumers' perceptions over data quality, attracting them to purchase data set.

For the case of slight underestimation, we set $\widehat{Q}$ to be 2, and plot the optimal π^* in Figure 3(b). We observe that all the three different mechanisms could have chance to be the optimal mechanism, which is similar to that in certain data quality case. One interesting observation is that the paid strategy, which does not provide any free sample, could still be the optimal strategy in some scenarios (when α locates in $[\overline{\alpha}, 1]$). This implies that the data vendor may not release free samples to revise the perceptions of data consumers if the extent of underestimate is not too large.

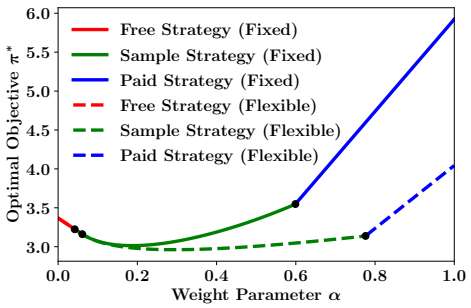
We then evaluate the optimal data selling strategy when data consumers overestimate data quality, and report π^* for the cases of moderate overestimate $\widehat{Q} = 2.5$ and large overestimate $\widehat{Q} = 7.173$ in Figure 4(a) and Figure 4(b), respectively. From Theorem 3, the optimal n^* should be chosen from five candidates. In our evaluation setting, there is only one particular valid candidate among n_1^*, n_2^*, n_3^*, and the other two are either not real number or fall out of $[0, N]$. As shown in Figure 4(a), the interval of α, in which sampling strategy is optimal, *i.e.*, $[\underline{a}, \overline{a}]$, becomes small if the extent of overestimate increases, and reduces to empty if $\widehat{Q}$ exceeds the threshold 7.173. When data consumers overestimate data quality, the data vendor has less incentive to offer free sampling to revise their perceptions, and would like to charge more data packages to extract revenue. Figure 4(b) shows that when the data vendor cares much about social benefit, she would adopt free strategy; otherwise, she would just deploy the paid strategy towards those optimistic consumers to extract high revenue. Sampling strategy would no longer be optimal in this case. Based on these discussions, we can derive the first conflict between data consumers and the data vendor: *the data vendor would not like to release free samples to revise data consumers' mistaken perceptions over data quality in the extreme overestimate case.*

6.3 Discounting Valuation

Following the principle in Section 5, we derive the optimal data selling mechanism and the optimal objective π^* under two different discounting factors $\delta = 0.98$ and $\delta = 0.9$. For a fixed δ, we can observe the similar results for certain data quality case and uncertain data quality case. Here, we only report the evaluation results of the overestimate data quality case with two different discounting factors in Figure 5(a). From Figure 5(a), we can find that the sampling strategy could be optimal in more scenarios when δ is smaller. This is because the data vendor would extract less revenue from charging non-sampling data if consumers have larger discount, *e.g.*, $\delta = 0.9$, and she would like to release more free samples to obtain social benefit in this case. Figure 5(a) also shows that the objective value π^* in the case of $\delta = 0.98$ is significantly larger than that in the case of $\delta = 0.9$, which



(a) $\delta = 0.90$ vs. $\delta = 0.98$.



(b) Flexible vs. Fixed.

Figure 5: Optimal π^* in overestimate data quality case with discounting valuation.

demonstrates that the discounting factor has a high impact on the revenue.

In Figure 5(b), we describe the optimal objective values π^* of the flexible pricing and the fixed pricing when δ is 0.9. From Figure 5(b), we can see that the fixed pricing outperforms the flexible pricing in all cases. which is consistent with the results in Theorem 5. We can derive the second conflict between data consumers and the data vendor: *although data consumers can benefit from the flexible pricing scheme, the data vendor has no economic incentive to deploy such scheme.* To facilitate the sustainable and healthy trading of IoT data, it is necessary to deploy market regulations to eliminate these two conflicts.

7 Related Work

In this section, we briefly review the related work about data markets and pricing mechanisms for information good.

Data Market: Different types of data, *e.g.*, personal data, IoT data and image data, have been collected and monetized by online service providers [28], [29], [30], [31], [32]. The seminal paper of data marketplaces outlines key challenges and potential research opportunities in this direction [33]. Koutris *et al.* [14], [15] designed a query-based data pricing framework to replace the current inflexible data pricing. Jung *et al.* [34] proposed a set of countable protocols for big data trading among dishonest consumers. In paper [35], Li *et al.* adopted information entropy to price data. Mehta *et al.* designed pricing policies for the data set with row-column format [36]. Agarwal *et al.* proposed matching mechanism to efficiently buy and sell training data for machine learning tasks [37]. Theses approaches determine the price of data based on information and determinacy. However, the focus of our work is the widespread data APIs pricing [16], which determines price only based on the number of API calls. Our results in Theorem 5 shows that the flexible pricing, such as the query-based pricing, achieves less economic benefit, compared with the fixed price mechanism.

There are other issues related to data sharing and trading, such as privacy preserving [38], [39], [40], data quality management [41], revenue sharing [42], [43], and data usage policies [44].

Information Pricing: Pricing information or digital goods have been widely studied in economics [17], [45], [46]. The book [47] distilled the pricing rules for information services. There are two effective mechanisms for pricing information services in the literature. One is bundling, which sells a large number of information goods for a fixed price. Geng *et al.* provided guidelines to bundling design in the case that consumers have decreasing valuations [17]. This discounting valuation model is similar to that considered in this work. Our results demonstrate that bundling is also a profitable selling mechanism in the uncertain data quality environment. The other strategy is versioning, which provides multiple versions for one product to satisfy the diverse demands of data consumers. As observed in [45], manufactures may intentionally damage their goods to enable price discrimination, leading to Pareto improvement. Bhargava and Choudhary derived the optimal versioning condition [46]. We observe that in practical data markets, the data vendor also launches different versions for data commodity, such as different numbers of available API calls. In our further work, we would investigate the effect of versioning on designing data selling mechanism.

Information pricing is also a well-explored subject IoT network and wireless network [48], [49], [50], [51], [52]. Niyato *et al.* [48] studied the economics of IoT and presented the information economics approaches. Finally, they proposed an economic model based on game theory to study the price competition of IoT sensing services. Alsheikh *et al.* [49] studied data pricing in IoT data markets from a machine learning perspective. They presented IoT market models and optimal pricing schemes of selling IoT services for standalone sales or bundled sales. In standalone sales, they maximized the profit of service providers by optimizing the size of bought data and service subscription fees, while in bundled sales, they aimed to maximize the total profit of cooperative service providers. Wu *et al.* [53] captured the unique economic characteristics of IoT data and presented a novel data model from the information design perspective. They also proposed data pricing mechanisms to maximize their revenue. Niyato *et al.* [54] designed a smart data pricing approach to achieve flexible and efficient data management in IoT. Moreover, they proposed a pricing mechanism to determine the data price for IoT service providers. In addition, some surveys summarized the research status of data pricing and pricing models in IoT [55], [56].

For the rising of data marketplace, survey [57] discusses a lot related issues. [14], [58] work out the QueryMarket system intending to address the inflexibility problem, while [59] purpose an arbitrage-free pricing scheme to deal with the simplicity issues. More details of query-based pricing with API can be seen in [15], [60]. Data market is

tightly related to cloud computing [33] as well as privacy issues [30], [61], [62]. We also investigate several works [63], [64] where data market appearing in mobile devices which is also another interesting topic regarding data marketing.

8 Conclusion

In this paper, we have considered the optimal data selling mechanisms for IoT data exchange. By modeling IoT data quality as a Gaussian random variable and adopting Bayesian learning scheme to update perceptions over data quality, we can obtain a specific data demand function, and then derive the optimal data selling mechanisms for the scenarios when data consumers underestimate, correctly estimate and overestimate the data quality. Our theoretical analyses and evaluation results show that the data vendor would not release free sampling data for optimistic data consumers to revise their incorrect perceptions. Furthermore, the data vendor has no economic incentive to adopt flexible pricing schemes, which explains the current widely adopted fixed pricing schemes in data markets.

Acknowledgment

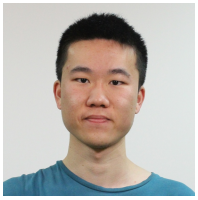
This work was supported in part by National Key RD Program of China No. 2020YFB1707900, in part by China NSF grant No. 62025204, 62072303, 61972252, 61902248, and 61972254, in part by Alibaba Group through Alibaba Innovation Research Program, in part by Shanghai Science and Technology fund 20PJ1407900, and in part by Tencent Rhino Bird Key Research Project. 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] "Quandl," https://www.quandl.com/.
[2] "Xignite," http://www.xignite.com/.
[3] "Factual," http://www.factual.com/.
[4] "Aggdata," https://www.aggdata.com/.
[5] "Uber," https://www.uber.com/.
[6] "Infochimps." http://www.infochimps.com/.
[7] "Big data exchange." http://www.bigdataexchange.com/.
[8] "Internet of Things Data Marketplace by IOTA," https://data.iota.org/.
[9] "Windows azure marketplace." https://datamarket.azure.com/.
[10] "Datasift: Pylon facebook api." http://datasift.com/products/pylon-for-facebook-topic-data/.
[11] "Gnip." https://gnip.com/products/realtime/firehose/.
[12] "Yelp." https://www.yelp.com/developers/display_requirements/.
[13] "Databroker dao," https://databrokerdao.com/.
[14] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with querymarket," in *SIGMOD*, 2013.
[15] ——, "Query-based data pricing," *Journal of the ACM*, vol. 62, no. 5, pp. 43:1–43:44, 2015.
[16] P. Upadhyaya, M. Balazinska, and D. Suciu, "Price-optimal querying with data apis," *Proceedings of the VLDB Endowment*, pp. 1695–1706, 2016.
[17] X. Geng, M. B. Stinchcombe, and A. B. Whinston, "Bundling information goods of decreasing value," *Management Science*, vol. 51, no. 4, pp. 662–667, 2005.
[18] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *VLDB*, 2004.
[19] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *Mobihoc*, 2017.
[20] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, pp. 235–284, 2008.
[21] W. Du, Z. Xing, M. Li, B. He, L. H. C. Chua, and H. Miao, "Optimal sensor placement and measurement of wind for water quality studies in urban reservoirs," in *IPSN*, 2014.
[22] S. Wang, T. He, D. Zhang, Y. Liu, and S. H. Son, "Towards efficient sharing: A usage balancing mechanism for bike sharing systems," in *WWW*, 2019.
[23] G. Ranjan, H. Zang, Z.-L. Zhang, and J. Bolot, "Are call detail records biased for sampling human mobility?" *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 16, no. 3, pp. 33–44, Dec. 2012.
[24] A. Smolin, "Disclosure and pricing of attributes," https://ssrn.com/abstract=3047028, Tech. Rep., 2017.
[25] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[26] Y. Bakos and E. Brynjolfsson, "Bundling information goods: Pricing, profits, and efficiency," *Management Science*, vol. 45, no. 12, pp. 1613–1630, 1999.
[27] "Suvnet data set collected by shanghai jiao tong university," http://wirelesslab.sjtu.edu.cn/taxi_trace_data.html.
[28] W. Mao, Z. Zheng, and F. Wu, "Pricing for revenue maximization in iot data markets: An information design perspective," in *INFOCOM*, 2019.
[29] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe, "Money walks: A human-centric study on the economics of personal mobile data," in *UbiComp*, 2014.
[30] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira, "Your browsing behavior for a big mac: Economics of personal information online," in *WWW*, 2013.
[31] J.-M. Bohli, C. Sorge, and D. Westhoff, "Initial observations on economics, pricing, and penetration of the internet of things market," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 2, pp. 50–55, 2009.
[32] L. Zhang, Y. Li, X. Xiao, X.-Y. Li, J. Wang, A. Zhou, and Q. Li, "Crowdbuy: Privacy-friendly image dataset purchasing via crowdsourcing," in *INFOCOM*, 2018.
[33] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," *Proceedings of the VLDB Endowment*, vol. 4, no. 12, pp. 1482–1485, 2011.
[34] T. Jung, X. Y. Li, W. Huang, J. Qian, L. Chen, J. Han, J. Hou, and C. Su, "Accounttrade: Accountable protocols for big data trading against dishonest consumers," in *INFOCOM*, 2017.
[35] X. Li, J. Yao, X. Liu, and H. Guan, "A first look at information entropy-based data pricing," in *ICDCS*, 2017.
[36] S. Mehta, M. Dawande, G. Janakiraman, and V. Mookerjee, "How to sell a dataset? pricing policies for data monetization," in *EC*, 2019.
[37] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *EC*, 2019.
[38] F. Li, Z. Sun, A. Li, B. Niu, H. Li, and G. Cao, "Hideme: Privacy-preserving photo sharing on social networks," in *INFOCOM*, 2019.
[39] M. Wu, D. Ye, J. Ding, Y. Guo, R. Yu, and M. Pan, "Incentivizing differentially private federated learning: A multi-dimensional contract approach," *IEEE Internet of Things Journal*, 2021.
[40] D. Ye, R. Yu, M. Pan, and Z. Han, "Federated learning in vehicular edge computing: A selective model aggregation approach," *IEEE Access*, vol. 8, pp. 23 920–23 935, 2020.
[41] T. Luo, J. Huang, S. S. Kanhere, J. Zhang, and S. K. Das, "Improving iot data quality in mobile crowd sensing: A cross validation approach," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5651–5664, 2019.
[42] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *ICML*, 2019.
[43] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in *AISTATS*, 2019.
[44] P. Upadhyaya, M. Balazinska, and D. Suciu, "Automatic enforcement of data use policies with datalawyer," ser. SIGMOD, 2015.
[45] R. J. Deneckere and R. Preston McAfee, "Damaged goods," *Journal of Economics & Management Strategy*, vol. 5, no. 2, pp. 149–174, 1996.

[46] H. K. Bhargava and V. Choudhary, "Research note—when is versioning optimal for information goods?" *Management Science*, vol. 54, no. 5, pp. 1029–1035, 2008.
[47] C. Shapiro and H. R. Varian, *Information rules: a strategic guide to the network economy*. Harvard Business Press, 1998.
[48] D. Niyato, X. Lu, P. Wang, D. I. Kim, and Z. Han, "Economics of internet of things: An information market approach," *IEEE Wireless Communications*, vol. 23, no. 4, pp. 136–145, 2016.
[49] M. A. Alsheikh, D. T. Hoang, D. Niyato, D. Leong, P. Wang, and Z. Han, "Optimal pricing of internet of things: A machine learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 4, pp. 669–684, 2020.
[50] D. Niyato, M. A. Alsheikh, P. Wang, D. I. Kim, and Z. Han, "Market model and optimal pricing scheme of big data and internet of things (iot)," in *2016 IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.
[51] V. Haghighatdoost, S. Khorsandi, and H. Ahmadi, "Fair pricing in heterogeneous internet of things wireless access networks using crowdsourcing," *IEEE Internet of Things Journal*, 2020.
[52] A. Ghosh and S. Sarkar, "Pricing for profit in internet of things," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 2, pp. 130–144, 2018.
[53] W. Mao, Z. Zheng, and F. Wu, "Pricing for revenue maximization in iot data markets: An information design perspective," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1837–1845.
[54] 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 Network*, vol. 30, no. 2, pp. 18–25, 2016.
[55] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *Acm computing surveys (csur)*, vol. 46, no. 2, pp. 1–37, 2013.
[56] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Data collection and wireless communication in internet of things (iot) using economic analysis and pricing models: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2546–2590, 2016.
[57] F. Schomm, F. Stahl, and G. Vossen, "Marketplaces for data: an initial survey," *ACM SIGMOD Record*, vol. 42, no. 1, pp. 15–26, 2013.
[58] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Querymarket demonstration: Pricing for online data markets," ser. VLDB, 2012.
[59] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *Proceedings of the VLDB Endowment*, vol. 7, no. 9, pp. 757–768, 2014.
[60] P. Upadhyaya, M. Balazinska, and D. Suciu, "Price-optimal querying with data apis," *Proceedings of the VLDB Endowment*, vol. 9, no. 14, pp. 1695–1706, 2016.
[61] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, "For sale : Your data: By : You," ser. HotNets, 2011.
[62] M. Mun, S. Hao, N. Mishra, K. Shilton, J. Burke, D. Estrin, M. Hansen, and R. Govindan, "Personal data vaults: A locus of control for personal data streams," ser. Co-NEXT, 2010.
[63] S. Ha, S. Sen, C. Joe-Wong, Y. Im, and M. Chiang, "Tube: Time-dependent pricing for mobile data," pp. 247–258, 2012.
[64] J. Staiano, N. Oliver, B. Lepri, R. de Oliveira, M. Caraviello, and N. Sebe, "Money walks: A human-centric study on the economics of personal mobile data," in *UbiComp*, 2014.

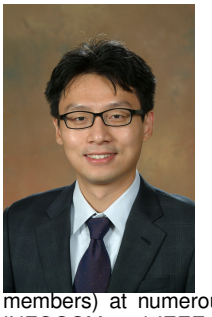
Qinya Li received her B.S. degree in Computer Science and Engineering from Northeastern University, P.R.China in 2015, and the Ph.D. degree in Computer Science and Engineering from Shanghai Jiao Tong University in 2020. Her research interests include mobile computing, mobile crowdsourcing, and algorithmic game theory and its applications.



Zun Li is pursuing a PhD degree in computer science at University of Michigan, Ann Arbor. His research focuses on the interfaces between AI, economics and complex systems. He had been working as a software engineer intern at Google. He received a B.S. degree in computer science from Shanghai Jiaotong University.

Zhenzhe Zheng is an assistant professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He received the B.E. in Software Engineering from Xidian University, in 2012, and the M.S. degree and the Ph.D. degree in Computer Science and Engineering from Shanghai Jiao Tong University, in 2015 and 2018, respectively. He has visited the University of Illinois at Urbana-Champaign (UIUC) as a Post Doc Research Associate from 2018 to 2019. His research interests include game theory and mechanism design, networking and mobile computing, and online marketplaces. He is a recipient of the China Computer Federation (CCF) Excellent Doctoral Dissertation Award 2018, Google Ph.D. Fellowship 2015 and Microsoft Research Asia Ph.D. Fellowship 2015. He has served as the member of technical program committees of several academic conferences, such as MobiHoc, AAAI, MSN, IoTDI and etc. He is a member of the ACM, IEEE, and CCF. For more information, please visit https://zhengzhenzhe220.github.io/

Fan Wu is a professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University. He received his B.S. in Computer Science from Nanjing University in 2004, and Ph.D. in Computer Science and Engineering from the State University of New York at Buffalo in 2009. 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, CCF-Tencent "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 Elsevier Computer Networks, and as the member of technical program committees of more than 100 academic conferences. For more information, please visit http://www.cs.sjtu.edu.cn/~fwu/.



Shaojie Tang is currently an Associate Professor of Naveen Jindal School of Management at University of Texas at Dallas. He received his PhD in computer science from Illinois Institute of Technology in 2012. His research interest includes social networks, mobile commerce, game theory, e-business and optimization. He received the Best Paper Awards in ACM MobiHoc 2014 and IEEE MASS 2013. He also received the ACM SIGMobile service award in 2014. Tang served in various positions (as chairs and TPC members) at numerous conferences, including ACM MobiHoc, IEEE INFOCOM and IEEE ICNP. He is an editor for INFORMS Journal on Computing.

Zhao Zhang received her PhD from Xinjiang University in 2003. She worked in Xinjiang University from 1999 to 2014, and now is a professor in the Department of Computer Science, Zhejiang Normal University. Her main interest is in combinatorial optimization, especially approximation algorithms for NP-hard problems which have their background in networks.

Guihai Chen earned his B.S. degree from Nanjing University in 1984, M.E. degree from Southeast University in 1987, and Ph.D. degree from the University of Hong Kong in 1997. He is a distinguished professor of Shanghai Jiao Tong University, China. He had been invited as a visiting professor by many universities including Kyushu Institute of Technology, Japan in 1998, University of Queensland, Australia in 2000, and Wayne State University, USA during September 2001 to August 2003. He has a wide range of research interests with focus on sensor networks, peer-to-peer computing, high-performance computer architecture and combinatorics. He has published more than 200 peer-reviewed papers, and more than 120 of them are in well-archived international journals such as IEEE Transactions on Parallel and Distributed Systems, Journal of Parallel and Distributed Computing, Wireless Networks, The Computer Journal, International Journal of Foundations of Computer Science, and Performance Evaluation, and also in well-known conference proceedings such as HPCA, MOBIHOC, INFOCOM, ICNP, ICPP, IPDPS and ICDCS.