Trading Data in the Crowd: Profit-Driven Data Acquisition for Mobile Crowdsensing

Zhenzhe Zheng, *Student Member, IEEE*, Yanqing Peng, Fan Wu, *Member, IEEE*, Shaojie Tang, *Member, IEEE*, and Guihai Chen, *Senior Member, IEEE*

***Abstract*— As a significant business paradigm, data trading has attracted increasing attention. However, the study of data acquisition in data markets is still in its infancy. Mobile crowdsensing has been recognized as an efficient and scalable way to acquire large-scale data. Designing a practical data acquisition scheme for crowd-sensed data markets has to consider three major challenges: crowd-sensed data trading format determination, profit maximization with polynomial computational complexity, and payment minimization in strategic environments. In this paper, we jointly consider these design challenges, and propose VENUS, which is the first profit-driVEN data acqUiSition framework for crowd-sensed data markets. Specifically, VENUS consists of two complementary mechanisms: VENUS-PRO for profit maximization and VENUS-PAY for payment minimization. Given the expected payment for each of the data acquisition points, VENUS-PRO greedily selects the most "cost-efficient" data acquisition points to achieve a sub-optimal profit. To determine the minimum payment for each data acquisition point, we further design VENUS-PAY, which is a data procurement auction in Bayesian setting. Our theoretical analysis shows that VENUS-PAY can achieve both strategy-proofness and optimal expected payment. We evaluate VENUS on a public sensory data set, collected by Intel Research, Berkeley Laboratory. Our evaluation results show that VENUS-PRO approaches the optimal profit, and VENUS-PAY outperforms the canonical second-price reverse auction, in terms of total payment.**

***Index Terms*— Data marketplace, mobile crowdsensing, auction theory.**

I. Introduction

THE past few years have witnessed the proliferation of smart devices in people's daily lives. The ubiquitous sensors embedded in pervasive smart devices incessantly generate tremendous volumes of sensed data by seamlessly monitoring a diverse range of human activities and environment phenomena. However, currently, most of operators exclusively analyze the collected data for their own application purposes, which introduces a serious barrier for the wide availability of crowd-sensed data, resulting in a number of isolated data islands. Recognizing the great benefit of data sharing [5], several open platforms, such as Terbine [50], Thingful [51] and Thingspeak [52], have emerged to enable crowd-sensed data to be exchanged on the web, aiming to unlock the potential economic values underlying the crowd-sensed data.

However, due to lack of efficient data acquisition scheme, the amounts of crowd-sensed data in these platforms are very limited, which has significantly suppressed the increasing market demand for data. The success of data markets highly relies on the sufficient amounts of data for trading. On one hand, the data broker needs to aggregate various types of data from exogenous data sources to satisfy the diverse demand of data consumers. On the other hand, the data broker has to periodically supply fresh data into data markets, because the data become less accurate, and even useless, when the contextualized environments evolve over time. Mobile crowdsensing have been recognized as a highly efficient and scalable way to collect large-scale data [32], [58]. For example, Thingspeak [52] has recently launched a crowdsensing platform to collect crowd-sensed data. In mobile crowdsensing, the data providers consume their own physical resources, and spend manual effort in collecting data. Thus, the data broker should offer sufficient payments to incentivize data providers to contribute data. The frugal data broker always wants to procure enough data with a minimum payment, which can be formulated as the problem of *payment minimization*.

In data markets, the ultimate goal of the data broker is to maximize profit, which is defined as the difference between the revenue generated from selling data (possibly data-based services) and the expenditure on data acquisition. Although the data broker can obtain a large revenue by providing high quality data services, she has to disburse expensive expenditure to collect enough data, such that the data services maintain at a high quality level. Therefore, in order to maximize profit, the data broker should make a trade-off between revenue and data acquisition expenditure, which can be formulated as a *profit maximization* problem. Although a number of data acquisition mechanisms with different optimization objectives have been developed in the literature [10], [11], [32], [58], few of them considered the monetary profit produced by trading

Manuscript received April 30, 2016; revised September 30, 2016; accepted November 28, 2016. Date of publication January 26, 2017; date of current version March 31, 2017. This work was supported in part by the State Key Development Program for Basic Research of China (973 Project) under Grant 2014CB340303, in part by the China NSF under Grant 61672348, Grant 61672353, Grant 61422208, Grant 61472252, Grant 61272443, and Grant 61133006, in part by the Shanghai Science and Technology Fund under Grant 15220721300, in part by the CCF-Tencent Open Fund, and in part by the Scientific Research Foundation for the Returned Overseas Chinese Scholars. The work of Z. Zheng was supported by a Google Ph.D. Fellowship and a Microsoft Asia Ph.D. Fellowship. The work of S. Tang was supported by the China NSF under Grant 61473109.

Z. Zheng, Y. Peng, F. Wu, and G. Chen are with the Shanghai Key Laboratory of Scalable Computing and Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zhengzhenzhe@sjtu.edu.cn; yanqing.sjtu@gmail.com; fwu@cs.sjtu.edu.cn; gchen@cs.sjtu.edu.cn).

S. Tang is with the Department of Information Systems, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: tangshaojie@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2017.2659258

the crowd-sensed data in the market, which deviates from the goal of the data broker. To fill this gap, we propose a profit-driven data acquisition mechanism. We summarize three major challenges in designing a practical profit-driven data acquisition mechanism for crowd-sensed data markets.

The first design challenge is to determine the crowd-sensed data trading format, considering both the characteristics of crowd-sensed data and diverse market demand for data. To evaluate the profit of data in the market, we have to identify the specific data trading format, which is still an open problem in both economics and computer science communities. The crowd-sensed data is normally uncertain and has complex correlation, making the crowd-sensed data quite different from the traditional information good [3], [35], and introducing additional difficulties in determining the data trading format. On one hand, the crowd-sensed data is incomplete, imprecise, and erroneous, making it improper to directly feed raw data into the data market. On the other hand, the crowd-sensed data may be correlated in multiple dimensions, and has rich sematic information behind such correlation [10], resulting in that separately selling pieces of data becomes meaningless. Furthermore, we should determine the data trading format aligned with the diverse market demand, meaning that data consumers, from different market segments, would request for the crowd-sensed data with different quality levels. Enabling data consumers to express their diverse market demands would incur a heavy burden of designing a concise and simple data trading format. Due to various types of uncertain factors, complex correlation and diverse market demand for data, it is nontrivial to determine an appropriate and flexible crowd-sensed data trading format.

Yet, another design challenge is the hardness of maximizing profit in a complicated data market environment. From the definition, the value of profit rests on the attained revenue from data trading and the distributed expenditure on data acquisition. Due to the special cost structure of data,[1] the prices of data should be linked to data consumers' valuations over the data, rather than the production cost. Thus, we can express the revenue with the data consumers' valuation distributions. However, it is hard to analyze the property of the revenue (and then the profit), because the valuations always follow complicated distributions in practical market environments. Furthermore, even if we can figure out the maximum expected revenue, finding the minimum expenditure on data acquisition can be proven to NP-Hard, and is normally computationally intractable. Therefore, in order to approach the optimal profit of data trading, we have to overcome the complicated formats of valuation distributions and the high computational complexity in solving the problem of acquisition expenditure minimization.

The last design challenge is to simultaneously guarantee both strategy-proofness and minimum payment in data procurement auctions. As competitive bidding leading to a lower disbursed payment, the data broker would conduct data procurement auctions to determine minimized payments for data providers. Since the data providers are rational and selfish, they always tend to misreport their private data collection costs, if doing so can increase their utilities. Such a selfish behavior inevitably hurts the other data providers' utilities, and significantly increase the data broker's data acquisition expenditure. Therefore, a strategy-proof data procurement mechanism is desirable in such strategic environment. However, it is extremely difficult in simultaneously achieving both strategy-proofness and optimum in auction theory [2], [17], [44], [46]. In forward auctions with Bayesian valuation setting, one of the mature techniques to guarantee strategy-proofness and optimal revenue is to reserve the trading items in the instances that all the bids are below a selected reserve price [39]. This reserve price-based technique does not work in the context of data procurement auctions, because the data broker has to purchase one piece of data from data providers in all the instances. New pricing techniques have to be developed to derive new theoretical results in procurement auctions. The previous works have also proved some negative results about revenue maximization (payment minimization) in strategy-proof auction design. In Bayesian valuation setting, Ronen and Saberi claimed that no deterministic polynomial time ascending auction can achieve an approximation ratio better than $3/4$ in terms of revenue maximization [46]. When the costs of data providers are completely private, no strategy-proof auction mechanism can give any performance guarantee on the payment [2], [17].

In this paper, by jointly considering the above three design challenges, we conduct an in-depth study on the profit-aware data acquisition design for crowd-sensed data markets. To probe in the benefit of model-based data trading format, we build a statistical model upon the raw data, to capture data uncertainty and complex correlation among data. We regard the resulting statistical model as an information commodity, and further partition the commodity into multiple versions with different quality levels, to satisfy the diverse market demand of data consumers. Secondly, we propose VENUS-PRO to decompose the problem of profit maximization into revenue maximization and data acquisition expenditure minimization. Given data consumers' valuation distributions, VENUS-PRO adopts a post-pricing mechanism to determine the trading price for each version, and then calculates the maximum expected revenue for each version. Considering the high computational complexity for the optimum, VENUS-PRO obtains a sub-optimal data acquisition expenditure, assuming that the payments for data acquisition points are given in advance. Combing with the calculated revenue and data acquisition expenditure, VENUS-PRO achieves a constant approximation ratio in terms of profit maximization. Finally, to determine the minimum payment for each of data acquisition points, we propose a strategy-proof and optimal data procurement auction mechanism, namely VENUS-PAY, in Bayesian setting, wherein the data providers' costs are drawn from publicly-known probability distributions.

We summarized our contributions in this paper as follows.

- First, we present a system model, including a data trading model and a data purchasing model, for crowd-sensed data markets. For the data trading model, we build a joint

[1]Data have a fixed production cost, and tend to induce negligible marginal costs for reproduction.

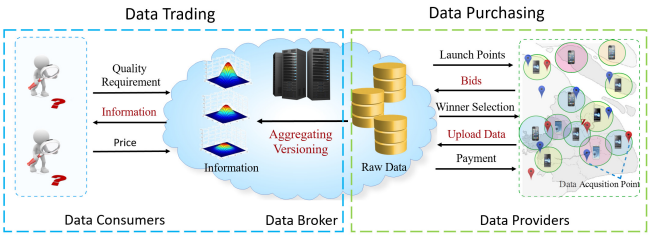


Fig. 1. Crowd-Sensed data market.

probability distribution to capture data uncertainty and complex data correlation, and adopt a versioning technique to satisfy the diverse market demand. We also model the data purchasing process as a reverse auction in Bayesian environment.

• Second, we design a profit-driven data acquisition mechanism, namely VENUS-PRO. Given the data consumers' valuation distributions and expected payment of each of data acquisition points, we propose a post-pricing mechanism and a greedy data acquisition point selection algorithm to achieve a sub-optimal profit.

• Third, we further consider the problem of payment minimization for data acquisition, and propose VENUS-PAY, which is a data procurement auction in Bayesian setting, achieving both strategy-proofness and optimal payment.

• Finally, we evaluate the performance of VENUS-PRO and VENUS-PAY based on a real-world public sensed data set. Our evaluation results show that VENUS-PRO obtains a near-optimal profit, while VENUS-PAY performs better than the classical second-price reverse auction.

The rest of this paper is organized as follows. In Section II, we present a system model for the crowd-sensed data market. Given the expected payment for each data acquisition point, we design VENUS-PRO in Section III. In Section IV, we propose VENUS-PAY to determine the minimum payments of data acquisition points. We present our evaluation results in Section V, and review the related work in Section VI. Finally, we conclude the paper in Section VII.

II. Preliminaries

In this section, we describe data trading model and data purchasing model for profit-driven data acquisition in crowd-sensed data markets.

A. Data Trading Model

As illustrated by Figure 1, in a crowd-sensed data market, the data broker wants to exchange the real-time information about the environment phenomena of a monitoring region for some profit, the data consumers would like to pay for this information to facilitate their data driven services, and the data providers want to earn payments for their contributed data. The data broker virtually deploys several *Data Acquisition Points* to approximately represent the phenomena of the monitoring region. The data acquisition points can be regarded as some kind of Points of Interest (PoIs), on which the data broker wants to collect necessary sensed data to train the real-time information. According to the required quality of service (QoS) in specific crowd-sensed applications, the data broker can determine the quantity, density and physical locations of data acquisition points by exploiting some machine learning techniques, such as active learning [8]. Generally, if the data broker aims to extract accurate knowledge of the monitoring region, she would deploy fine-grained data acquisition points, but at the same time she has to disburse high payments for purchasing raw data from data providers. For example, indoor location service provider needs high precise and multi-dimensional sensed data, such as position, size, coordinates and orientation information of indoor landmark objectives, to reconstruct the map of indoor floor plan [15]. Due to the complex indoor environment, the service provider needs to deploy dense data acquisition points (at each store entrance) to collect sufficient sensed data, incurring a high data acquisition expenditure. In another scenario, the government can simply deploy sparse data acquisition points along the major roads or around shopping malls to roughly profile the noise level of metropolis [41], [42]. We denote the L Data Acquisition Points by $\mathcal{Y} = \{y_1, y_2, \cdots, y_L\}$. We associate a discrete random variable X_y for each data acquisition point $y \in \mathcal{Y}$, representing the possible measurements of the monitoring environment phenomena, and associate a set of discrete random variables X_Y with a set of data acquisition points $Y \subseteq \mathcal{Y}$. We note that the random variables may be correlated in multiple dimensions, *e.g.*, the temperatures of geographically proximate locations are likely to change synchronously [10]. We use the following major notations to define the data trading model.

1) Information Commodity: In the crowd-sensed data market, the information commodity is the joint distribution of random variables $X_{\mathcal{Y}}$ over T time slots. We do not restrict the joint distribution to any specific format, such that the joint distribution can capture different types of uncertainty and complex correlation among data. The *probability mass function* $p\left(\mathbf{x}_{\mathcal{Y}}\right)$ assigns a probability for a possible valuation vector $\mathbf{x}_{\mathcal{Y}} = (x_1, x_2, \cdots, x_L)$ to the random variables $X_{\mathcal{Y}}$. We can use historical data and expert knowledge to construct a rough prior probability mass function, and update it using the new observations from the data acquisition module. Suppose that we observe values $\mathbf{x}_O$ for the selected random variables $X_O \subseteq X_{\mathcal{Y}}$, we can use Bayes' rule to condition our joint probability mass function $p(\mathbf{x}_{\mathcal{Y}})$ on these observations:

$$p(\mathbf{x}_{\overline{\mathcal{Y}}} | X_O = \mathbf{x}_O) = \frac{p(\mathbf{x}_{\overline{\mathcal{Y}}}, \mathbf{x}_O)}{p(\mathbf{x}_O)}, \tag{1}$$

where $X_{\overline{\mathcal{Y}}} = X_{\mathcal{Y}} \backslash X_O$ is the set of unobserved random variables. The posterior distribution is more certain than the prior distribution. Here, we use *entropy* to quantify the uncertainty of a distribution, considering its concise expression and nice properties, *e.g.*, monotonicity and submodularity.[2] Specifically, the *conditional entropy* of the unobserved normal random

[2]We can also use some other alternative metrics, such as Kullback-Leibler divergence and Mutual Information [9], to quantify the uncertainty of the distributions. However, Kullback-Leibler divergence is too complex to be used in data trading, and mutual information does not satisfy neither monotonicity nor submodularity, which significantly increases the complexity of profit maximization.

variables $X_{\overline{\mathcal{Y}}}$, after observing the selected random variables $X_{\mathcal{O}}$ is:

$$H(\overline{\mathcal{Y}}|\mathcal{O}) = -\sum_{\substack{\mathbf{x}_{\overline{\mathcal{Y}}} \in \text{dom } X_{\overline{\mathcal{Y}}} \\ \mathbf{x}_{\mathcal{O}} \in \text{dom } X_{\mathcal{O}}}} p(\mathbf{x}_{\overline{\mathbf{Y}}}, \mathbf{x}_{\mathcal{O}}) \log p(\mathbf{x}_{\overline{\mathbf{Y}}}|\mathbf{x}_{\mathcal{O}}). \quad (2)$$

As time passes by, our belief about the observations of the random variables $X_{\mathcal{O}}$ will be "spread out", increasing the uncertainty of the distribution. Thus, the data broker has to periodically collect new observations to maintain the entropy of the information commodity at a low level.

2) Version: In the crowd-sensed data market, data consumers may have diverse quality requirements over the information commodity, resulting in different valuations for the information commodity. To satisfy the diverse quality requirements of data consumers, the data broker would launch multiple versions for the information commodity, where a version is a posterior distribution with some selected observation random variables. We define the quality of the version $p(\mathbf{x}_{\overline{\mathcal{Y}}}|\mathbf{x}_{\mathcal{O}})$ as a function of its conditional entropy $H(\overline{\mathcal{Y}}|\mathcal{O})$:

$$Q \triangleq 1 - \frac{H(\overline{\mathcal{Y}}|\mathcal{O})}{H(\mathcal{Y})} = \frac{H(\mathcal{O})}{H(\mathcal{Y})}.$$

The second part of the equation holds given the property $H(\overline{\mathcal{Y}}|\mathcal{O}) = H(\mathcal{Y}) - H(\mathcal{O})$. From this definition, the quality of the version $p(\mathbf{x}_{\overline{\mathcal{Y}}}|\mathbf{x}_{\mathcal{O}})$ is directly proportional to the *joint entropy* of the selected random variables $H(X_{\mathcal{O}})$, implying that the version will have a high quality if we choose the random variables with large joint entropy to observe. Generally, the data consumers with various quality requirements would have different willingness to pay. In order to extract revenue from these heterogeneous data consumers, the data broker leverages the technique of differential pricing [54] by selling the different versions of an information commodity at different prices.

By conducting the standard market technique, such as survey, the data broker can determine the K candidate versions and a quality vector $\mathbf{Q} = (Q_1, Q_2, \cdots, Q_K)$, where $Q_i < Q_k$, for all $1 \leq i < k \leq K$.[3] We define the quality gap between two successive versions k and $k+1$ as: $\Delta_k \triangleq Q_{k+1} - Q_k$, and denote the minimum value of all the Δ's by $\Delta_{min} \triangleq \min_k\{\Delta_k\}$. In practice, the quality gap would be large enough to distinguish two successive versions, and we assume that $\Delta_{min} > \max_i H(\{i\})/H(\mathcal{Y})$. We adopt the post-pricing mechanism for data trading because of its convenience and popularity in practice. In the context of data trading, the post-pricing mechanism has several advantages compared with other trading formats, such as auction mechanisms. For example, the post-pricing mechanism can guarantee the robust economic properties, such as strategy-proofness, and handle the dynamical features of markets in a concise way. The data broker only has to determine a take-it-or-leave price, which is independent on the valuations and arrival sequence of data consumers. Specifically, the data broker assigns a *price* p_k for the kth version, and denote the price menu for all the K versions by $\mathbf{p} = (p_1, p_2, \cdots, p_K)$. We discuss the determination of the optimal price for each version in Section III.

3) Data Consumers: There are N single-minded data consumers in the data market. Each data consumer is interested in only one version of the information commodity, and has a valuation over this version.[4] We assume that the version preference of the data consumers follows a distribution with a probability mass function $g(k)$, meaning that the data consumers have a probability $g(k)$ to choose the kth version as her interested version. Therefore, there are $N_k = N \times g(k)$ data consumers, who are interested in the kth version, in expectation. For the kth version, we assume that the valuations of the N_k data consumers are drawn from a distribution with *cumulative distribution function* $V_k(x)$. We denote the vector of all valuation distributions by $\mathbf{V} = (V_1(\cdot), V_2(\cdot), \cdots, V_K(\cdot))$. By learning the historical transactions, the data broker can obtain the knowledge of the distribution $g(k)$ and the vector $\mathbf{V}$. This assumption is also called as Bayesian assumption in economic literature [29].

4) Revenue: If the data consumer's valuation is larger than the trading price of the kth version p_k, then she would purchase this version. In this case, the data broker would receive a revenue of p_k. The expected revenue of selling the kth version to the N_k data consumers is:

$$r_k = N \times g(k) \times (1 - V_k(p_k)) \times p_k. \quad (3)$$

When the data broker creates the kth version, despite of obtaining the revenue r_k, she can also extract revenue r_i from each of the version $1 \leq i < k$ lower than k. This is because the data broker can degrade a high quality version to lower ones without inducing additional costs, *e.g.*, simply adding artificial noises or using less observations. Thus, the expected cumulative revenue of the kth version should be $R_k = \sum_{i=1}^{k} r_i$.

B. Data Purchasing Model

Since the collected data becomes less accurate over time, the data broker has to periodically supply fresh data into the market. In the data trading model, the data broker would select different sets of data acquisition points to generate different versions of the information commodity. For each selected data acquisition point in one specific time slot, if the previous observation has been expired, meaning that it is not accurate enough to represent the environment phenomena, the data broker would purchase one new observation from a pool of active data providers. According to the "freshness" of the current observation and the accuracy requirement of each data acquisition point, the data broker determines the time slots to launch the data purchasing process for each data acquisition point. Thus, we can assume that the data collection procedures for different data acquisition points in different time slots are independent, so we focus on the data purchasing process for one data acquisition point in one specific time

[3]The determination of the number of versions and the corresponding quality for each version is beyond the scope of this paper. Several previous works [4], [43], [53], [54] shed light on possible solutions for this problem.

[4]We initialize the examination of data trading model with a simple purchasing behaviour model for data consumers, and leave more complex models to our future works.

slot in the following discussion.[5] We model the process of data purchasing as a single-item *data procurement auction*, in which the data broker, also called as an auctioneer or a buyer, wants to buy one piece of data from m_y competitive data providers, also called as suppliers or sellers, for the data acquisition point $y \in \mathcal{O}$ in one specific time slot.[6] The item being auctioned is the right to supply data, which can be considered as one kind of scarce resource. Auction mechanism is believed to be an effective way to allocate the scarce resource, because in procurement auction, the data broker can discover the true collection cost of data providers, and exploit the competition among the data providers to reduce the procurement payment. In a direct-revelation data procurement auction, the data providers simultaneously declare their bids to the data broker, who thereafter makes a decision on winner determination and payment to winner. We use some useful notations to define the data procurement auction model.

1) Data Provider: We denote the data providers by $\mathcal{M}_y = \{1, 2, \cdots, m_y\}$. Each data provider $i \in \mathcal{M}_y$ has a data collection cost c_i, which is private information to her, and is known as *type* in mechanism design. We consider a Bayesian setting, in which cost c_i is drawn from a publicly-known distribution $F_i(x)$ with a density function $f_i(x)$ in the range $[\underline{c}_i, \overline{c}_i]$. Let $\mathbf{F} = (F_1(\cdot), F_2(\cdot), \cdots, F_{m_y}(\cdot))$ denote the cost distributions of all data providers. We assume that the cost distributions are independent, but is not necessary to be identical.

Each data provider $i \in \mathcal{M}_y$ declares a bid b_i to the data broker, meaning that she requests for a compensation of at least b_i to cover her cost. Since data providers are rational and selfish, they may not truthfully reveal their costs, *i.e.*, the bids may not necessarily be equal to the cost c_i. We denote the cost and bidding profile of all data providers by $\mathbf{c} = (c_1, c_2, \cdots, c_{m_y})$ and $\mathbf{b} = (b_1, b_2, \cdots, b_{m_y})$, respectively. After collecting bidding profile, the auctioneer selects a winner, and determines payments for data providers. That is, a data procurement auction has two major components:

- *Selection Rule*: Choose a feasible selection rule $\mathbf{x}(\mathbf{b}) = (x_1(\mathbf{b}), x_2(\mathbf{b}), \cdots, x_{m_y}(\mathbf{b}))$ as a function of bidding profile $\mathbf{b}$. $x_i(\mathbf{b}) = 1$ if data provider i is the winner; otherwise $x_i(\mathbf{b}) = 0$.
- *Payment Rule*: Determine a payment vector $\mathbf{w}(\mathbf{b}) = (w_1(\mathbf{b}), w_2(\mathbf{b}), \cdots, w_{m_y}(\mathbf{b}))$, where $w_i(\mathbf{b})$ is the payment for the data provider i when the bidding profile is $\mathbf{b}$.

Data provider $i \in \mathcal{M}_y$ has a quasi-linear utility $u_i(\mathbf{b})$ on the bidding profile $\mathbf{b}$, which is defined as the difference between payment and cost: $u_i(\mathbf{b}) \triangleq w_i(\mathbf{b}) - c_i \times x_i(\mathbf{b})$.

2) Expected Payment: The data broker determines T time slots to collect data for the data acquisition point $y \in \mathcal{Y}$, and the expected accumulated payment is:

$$\Omega_y = T \times \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}\left[\sum_{i=1}^{m_y} w_i(\mathbf{c})\right], \tag{4}$$

where $\mathbf{c} \sim \mathbf{F}$ means that the expectation is over the cost distributions. We use $\mathbf{\Omega} = \{\Omega_y | y \in \mathcal{Y}\}$ to denote the expected payments of all data acquisition points $\mathcal{Y}$.

C. Problem Statement

In data market, the data broker faces two closely relevant optimization problems: *Profit Maximization* and *Payment Minimization*. We formulate these two problems as follows.

1) Profit Maximization: Although the data broker can obtain large revenue by launching a version with high quality, at the same time, she has to disburse expensive expenditure to select more data acquisition points. Hence, the data broker prefers to select the "profitable" and "cheap" data acquisition points $\mathcal{O}$ to observe, such that the obtained profit $\Phi(\mathcal{O})$ is maximum. We define the profit as the difference between the obtained revenue and the disbursed expenditure:

$$\Phi(\mathcal{O}) \triangleq R(\mathcal{O}) - S(\mathcal{O}), \tag{5}$$

where $R(\mathcal{O})$ is the revenue generated by the selected data acquisition points $\mathcal{O}$, and $S(\mathcal{O})$ is the data acquisition expenditure for the data acquisition points $\mathcal{O}$, *i.e.*, $S(\mathcal{O}) = \sum_{y\in\mathcal{O}} \Omega_y$. We note that $R(\mathcal{O})$ is equal to the revenue R_{k^*} of the version k^*, which is the highest version that the selected points $\mathcal{O}$ can reach, *i.e.*, $k^* \leftarrow \arg\max_k \{(H(\mathcal{O})/H(\mathcal{Y})) \geq Q_k\}$. We can state the problem of profit maximization as: selecting a subset of data acquisition points $\mathcal{O}^*$, such that the obtained profit is maximized, *i.e.*, $\mathcal{O}^* = \arg\max_{\mathcal{O}\subseteq\mathcal{Y}} (R(\mathcal{O}) - S(\mathcal{O}))$. In profit maximization, we assume that the expected payment Ω_y for each data acquisition point $y \in \mathcal{Y}$ is known in advance, and determine this expected payment in payment minimization module. Given the minimum payment for data acquisition points and the optimal revenue for each version, the profit maximization problem is actually a data acquisition point selection problem.

2) Payment Minimization: For each of data acquisition points, the frugal data broker always wants to purchase one piece of data with the lowest payment. In Bayesian environment, the data broker intends to design a data procurement auction that achieves the lowest expected payment Ω_y, where the expectation is with respect to the cost distributions $\mathbf{F}$.

The problems of profit maximization and payment minimization are closely related. On one hand, the output of payment minimization is the input of profit maximization. The minimum expected payment on each of data acquisition points, determined by the payment minimization procedure, directly affects the data acquisition point's probability of being selected in the profit maximization subroutine. On the other hand, the result of profit maximization, *i.e.*, the selected data acquisition points, also has an impact on payment minimization. The solution to profit maximization is to make a trade-off between the revenue extracted from data trading and expenditure for data acquisition. The intuitive idea behind the solution is to select the data acquisition points with a lower payment and a

[5]Considering the dependence among data purchasing processes in temporal dimensions would significantly increase the complexity of designing the data procurement auction. In dynamic data acquisition scenario, we should model the interdependent data processes as repeated procurement auctions or generalized online auctions, in which the data providers can participate in the data purchasing process in successive time slots. As the data providers can repeatedly interact with the data broker, there are more strategic behaviours for data providers to manipulate the auction to further increase their long-term utilities [1]. We will relax this assumption in our future work.

[6]Our results can be extended to more flexible auction formats, *e.g.*, multi-unit auctions or combinatorial auctions, adapting to the scenario, where one data acquisition point needs multiple observations to guarantee fault tolerance.

high revenue contribution to final revenue. Under this selection rule, the data acquisition points with a large payment and a significantly high contribution would also have chances to be selected. In the long term, the data acquisition points with a large expected payment would attract more data providers, increasing the competition among bidders, and thus decreasing the payment in the end.

III. VENUS-PRO

In this section, we propose a profit-driven data acquisition mechanism, namely VENUS-PRO, for profit maximization in the crowd-sensed data market.

A. Detailed Design

VENUS-PRO consists of two components: Revenue Determination and Acquisition Expenditure Calculation.

1) Revenue Determination: In contrast to physical goods, information commodity, regarded as one kind of digital goods, has a different cost structure, *i.e.*, a fixed cost of production, *e.g.*, a substantial data acquisition expenditure, but negligible marginal costs, *i.e.*, a lower cost of producing an additional copy. Under this special cost structure, the prices of the information commodity should be linked to the valuations of data consumers rather than the production costs. Thus, given the data consumers' valuation distribution of the kth version: $V_k(x)$, the expected revenue of the kth version with a trading price p is: $(1 - V_k(p)) \times p$. Therefore, the data broker can set the optimal price p_k for the kth version by

$$p_k \leftarrow \arg\max_p \left[(1 - V_k(p)) \times p\right].$$

Under this optimal pricing strategy, the expected revenue for the kth version is:

$$R_k = \sum_{i=1}^{k} r_i = \sum_{i=1}^{k} \left[N \times g(i) \times (1 - V_i(p_i)) \times p_i\right]. \quad (6)$$

Given the optimal revenue for each version and the minimum payment for each of data acquisition points, the profit maximization problem is actually a data acquisition point selection problem: selecting a subset of data acquisition points $\mathcal{O}^*$, such that the obtained profit is maximized, *i.e.*, $\mathcal{O}^* = \arg\max_{\mathcal{O}\subseteq\mathcal{Y}} (R(\mathcal{O}) - S(\mathcal{O}))$. Considering that the version preference distribution $g(i)$ and the valuation distributions $V_i(p)$ can be quite complicated in the practical data market, the specific format of the revenue $R(\mathcal{O})$ (or R_{k^*}) (and then the profit $\Phi(\mathcal{O})$) has unclear properties, making it difficult to directly solve the above profit maximization problem by adopting the classical optimization problem. In the following, we overcome this issue by transforming profit maximization problem to a solvable expenditure minimization problem, and designing an approximation algorithm for it.

2) Acquisition Expenditure Calculation: In the practical market, the commodity normally has a constant number of versions as a result of the trade-off between efficiency and complexity [47], [53]. It is obvious that the optimal number of the version for an information commodity is equal to the number of types of data consumers in the data market. Data consumers from different market segments have significantly different valuations for one data set, because they use the data set in diverse application scenarios. As the possible applications for one data set may be large, the data broker may not have a clear idea of the exact number of types of data consumers. For the market with no obvious market segments, some previous works [20], [48] in marking suggest that the optimal number of versions is three: a high-end version, a middle version, and a low-end version, which are also called as Goldilocks pricing in the literature. Furthermore, the maximum number of versions in practical data marketplaces is always small, *e.g.*, in Windows Azure Data Marketplace [56], the maximum number versions is no more than ten, and in Quandl [45], a financial and economic data trading platform, the data vendor offers most of data sets with four versions. Under this observation, our basic idea to solve the profit maximization is to enumerate the profit of each version and select the maximum one as the result. Specifically, in order to calculate the profit of the kth version, by the definition of profit, we should know the possible highest revenue R_k and its corresponding lowest data acquisition expenditure $S(\mathcal{O})$. Although we can exactly calculate the highest revenue R_k by Equation (6), it is nontrivial to figure out the minimum data acquisition expenditure $S(\mathcal{O})$. We now formulate the problem of expenditure minimization for the kth version as follows.

Problem: *Expenditure Minimization for the kth version*
Objective: Minimize $S(\mathcal{O})$
Subject to:

$$\frac{H(\mathcal{O})}{H(\mathcal{Y})} \geq Q_k, \quad \mathcal{O} \subseteq \mathcal{Y}.$$

Here, the data broker attempts to select a set of acquisition points $\mathcal{O}$ with the lowest acquisition expenditure, to satisfy the quality requirement of the kth version.

Unfortunately, the problem of expenditure minimization can be proven to NP-Hard by reducing from the general set covering problem [57]. Considering the computational intractability of the expenditure minimization problem, we present an alternative solution with a greedy acquisition points selection algorithm, to achieve a near-optimal expenditure in polynomial time. To this end, we take the advantage of the submodularity of entropy function $H(\cdot)$. We first give the definition of submodular function.

Definition 1 (Submodular Function): Let X be a finite set. A function $f : 2^X \mapsto \mathbb{R}$ is a submodular function if for any $\mathcal{A} \subseteq \mathcal{B} \subseteq X$ and $x \in X \setminus \mathcal{B}$:

$$f(\mathcal{A} \cup \{x\}) - f(\mathcal{A}) \geq f(\mathcal{B} \cup \{x\}) - f(\mathcal{B}).$$

We show the entropy $H(\cdot)$ is submodular and non-decreasing.

Lemma 1: The entropy function $H : 2^{\mathcal{Y}} \mapsto \mathbb{R}$ is submodular, non-decreasing, and non-negative.

Proof: To prove the submodularity, we first introduce an interesting property of entropy: the "information never hurts" principle [9]: $H(y|\mathcal{O}) \leq H(y)$, for any $y \in \mathcal{Y}$ and $\mathcal{O} \subseteq \mathcal{Y}$, *i.e.*, in expectation, observing the random variables $X_{\mathcal{O}}$ cannot increase the uncertainty about the random variable X_y. Since the marginal entropy increase can be expressed as $H(y \cup \mathcal{O}) - H(\mathcal{O}) = H(y|\mathcal{O})$, the submodularity of the entropy

Algorithm 1 Greedy Data Acquisition Point Selection

Input: A set of data acquisition points $\mathcal{Y}$, a set of expected payments $\mathbf{\Omega}$, the quality Q_k of the kth version.

Output: The set of selected random variables $\mathcal{O}_T$.

1 $t \leftarrow 0$; $\mathcal{O}_0 \leftarrow \varnothing$;
2 **do**
3 $\quad t \leftarrow t+1$;
4 $\quad \pi_t \leftarrow \arg\min_{y \in \mathcal{Y} \setminus \mathcal{O}_{t-1}} \left\{ \frac{\Omega_y}{H(y|\mathcal{O}_{t-1})} \right\}$;
5 $\quad \mathcal{O}_t \leftarrow \mathcal{O}_{t-1} \cup \{\pi_t\}$;
6 **while** $\frac{H(\mathcal{O}_t)}{H(\mathcal{Y})} < Q_k$ **and** $t < L$;
7 $T \leftarrow t$;
8 **return** $\mathcal{O}_T$;

function is simply the consequence of the information never hurt principle: for any set $\mathcal{O} \subseteq \mathcal{O}' \subseteq \mathcal{Y}$, we have:

$$H(y \cup \mathcal{O}) - H(\mathcal{O}) = H(y|\mathcal{O}) \geq H(y|\mathcal{O}') = H(y \cup \mathcal{O}') - H(\mathcal{O}').$$

Contrary to the differential entropy, which can be negative, in the discrete case, the entropy is non-negative, *i.e.*, $H(y \cup \mathcal{O}) - H(y) = H(y|\mathcal{O}) \geq 0$ for any set $\mathcal{O} \subseteq \mathcal{Y}$. Furthermore, $H(\varnothing) = 0$. This demonstrates that the entropy function $H(\cdot)$ is non-negative and non-degreasing. □

By Lemma 1, the expenditure minimization is a submodular set covering problem. Greedy approach is a nature fit for submodular optimization [34], [57]. One nature greedy rule is to select the most "cost-efficient" acquisition point in each iteration, *i.e.*, the acquisition point with a lower expected payment and high marginal entropy. This simple and efficient heuristic rule has been widely adopted in the other variations of submodular set covering problems [14], [16], [55], [57]. Other greedy naive greedy rules, such as keep choosing the acquisition points with minimum expected payment, or keep choosing the acquisition points with the highest marginal entropy can experience arbitrary bad results in some extreme cases. We now describe the greedy data acquisition point selection algorithm for the problem of expenditure minimization in Algorithm 1. We use $\mathcal{O}_t$ to denote the set of selected acquisition points until the tth iterations, and initialize $\mathcal{O}_0$ to be $\varnothing$. In the tth iteration, the data broker will select the data acquisition $y \in \mathcal{Y} \setminus \mathcal{O}_{t-1}$ that has the smallest $\Omega_y / H(y|\mathcal{O}_{t-1})$, where $H(y|\mathcal{O}_{t-1}) = H(y \cup \mathcal{O}_{t-1}) - H(\mathcal{O}_{t-1})$ represents the marginal entropy of the acquisition point y, given the currently selected acquisition points $\mathcal{O}_{t-1}$.[7] Let π_t denote the acquisition point selected in the tth iteration. This selection process iterates until the normalized joint entropy of the selected acquisition points $H(\mathcal{O}_t)/H(\mathcal{Y})$ is higher than the quality of the kth version Q_k, or there are no more acquisition points to select (Lines 2 to 6). Algorithm 1 outputs the set of T acquisition points $\mathcal{O}_T$ as the result. Since we have to check each unselected acquisition point in each iteration, and there are at most L iterations, the time complexity of Algorithm 1 is $O(L^2)$, where L is the number of acquisition points.

[7] In large scale mobile crowdsensing systems, it is hard to compute the exact conditional entropy. We can adopt sampling approaches [33] to efficiently calculate such approximate conditional entropy within a tolerant error bound.

Algorithm 2 VENUS-PRO for Profit Maximization

Input: A vector of valuation distributions $\mathbf{V}$, a version preference distribution $g(k)$, a set of random variables $\mathcal{Y}$, a set of expected payments $\mathbf{\Omega}$, a quality vector $\mathbf{Q}$.

Output: A pair of profit and selected version (Φ^*, k^*).

1 $\Phi^* \leftarrow 0$; $k^* \leftarrow 0$;
2 **for** $k = 1$ *to* K **do**
3 $\quad p_k \leftarrow \arg\max_p [(1 - V_k(p)) \times p]$;
4 **for** $k = 1$ *to* K **do**
5 $\quad R_k \leftarrow \sum_{i=1}^{k} [N \times g(i) \times (1 - V_i(p_i)) \times p_i]$;
6 $\quad \mathcal{O}_k \leftarrow GDY\_ALG(\mathcal{Y}, \mathbf{\Omega}, Q_k)$;
7 $\quad S(\mathcal{O}_k) \leftarrow \sum_{y \in \mathcal{O}_k} \Omega_y$;
8 $\quad \Phi_k \leftarrow R_k - S(\mathcal{O}_k)$;
9 $\quad$ **if** $\Phi_k > \Phi^*$ **then**
10 $\quad\quad \Phi^* \leftarrow \Phi_k$; $k^* \leftarrow k$;
11 **return** (Φ^*, k^*);

Combining revenue calculation with acqusition expenditure determination, we describe the detailed steps of VENUS-PRO in Algorithm 2. VENUS-PRO first calculates the optimal trading price for each version (Lines 2 to 3). Using this optimal trading price strategy, VENUS-PRO can figure out the maximum expected revenue R_k for each version k (Line 5). Although VENUS-PRO cannot obtain the minimum expenditure, it can get the approximate one by running Algorithm 1 (GDY_ALG for short) (Lines 6 to 7). Upon obtaining the expected revenue R_k and the approximate expenditure $S(\mathcal{O}_k)$, VENUS-PRO can calculate the approximate profit $\Phi_k = R_k - S(\mathcal{O}_k)$ for each version k. Among these K candidate versions, VENUS-PRO chooses the one with the maximum approximate profit as the final result (Lines 9 to 10). Since VENUS-PRO calls GDY_ALG algorithm K times, the computational complexity of VENUS-PRO is $O(KL^2)$, where K is the number of candidate versions.

B. Analysis

In this section, we analyze the approximation ratio of VENUS-PRO. We first present the performance guarantee of the greedy algorithm (*i.e.*, Algorithm 1) for the problem of data acquisition expenditure minimization.

Theorem 1: We use $\mathcal{O}^*$ to denote the optimal set of acquisition points for the problem of acquisition expenditure minimization. If Algorithm 1 is applied, we are guaranteed to obtain:

$$\frac{S(\mathcal{O}_T)}{S(\mathcal{O}^*)} \leq 1 + \log_e \beta,$$

where $\beta = \min\{\beta_1, \beta_2, \beta_3\}$, and $\beta_1 = \frac{H'(\mathcal{Y}) - H'(\varnothing)}{H'(\mathcal{Y}) - H'(\mathcal{O}_{T-1})}$, $\beta_2 = \max_{y \in \mathcal{O}_T, 1 \leq t \leq T} \left\{ \frac{H'(y|\mathcal{O}_0)}{H'(y|\mathcal{O}_t)} : H'(y|\mathcal{O}_t) > 0 \right\}$, $\beta_3 = \frac{\theta'_T}{\theta'_1}$. The function $H'(\cdot)$ is defined by $H'(\mathcal{O}) \triangleq \min\{H(\mathcal{O}), H(\mathcal{Y}) \times Q_k\}$, and $\theta'_t = \frac{\Omega_{\pi_t}}{H'(\pi_t|\mathcal{O}_{t-1})}$.

Proof: Designing greedy algorithms with good approximation factors for the problem of submodular set covering have been widely studied in submodular optimization literature [21], [22], [34], [57]. We can reduce the problem of acquisition expenditure minimization to a special submodular covering maximization: $\min_{\mathcal{O}\subseteq\mathcal{Y}}\left\{S(\mathcal{O}) : H'(\mathcal{O}) = H'(\mathcal{Y})\right\}$, by introducing a new submodular and non-decreasing function $H'(\mathcal{O}) = \min\{H(\mathcal{O}), H(\mathcal{Y}) \times Q_k\}$. It is easy to check that such special submodular covering formulation is equivalent to the original formulation described in Section III-A. Wolsey analyzed the approximation ratio of the greedy algorithm for this special submodular covering maximization problem in [57]. Thus, using the similar analysis technique, we can immediately obtain the approximation ratio of Algorithm 1 for the problem of acquisition expenditure minimization. □

Before presenting the approximation ratio of VENUS-PRO, we show a useful lemma.

Lemma 2: In the expenditure minimization for the version k, the optimal solution $\mathcal{O}^*$ satisfies $Q_k \leq H(\mathcal{O}^*)/H(\mathcal{Y}) < Q_{k+1}$.

Proof: We first show that for any random variable $i \in \mathcal{O}^*$, we have $H(\mathcal{O}^*_{-i})/H(\mathcal{Y}) < Q_k$, where $\mathcal{O}^*_{-i} = \mathcal{O}^*\backslash\{i\}$. If this inequality does not hold, *i.e.*, $H(\mathcal{O}^*_{-i})/H(\mathcal{Y}) \geq Q_k$, we can get a better solution $\mathcal{O}^*_{-i}$ for the problem of expenditure minimization. This is because $\mathcal{O}^*_{-i}$ is a feasible solution when $H(\mathcal{O}^*_{-i})/H(\mathcal{Y}) \geq Q_k$, and the expenditure of random variables $\mathcal{O}^*_{-i}$ is certainly less than that of $\mathcal{O}^*$ *i.e.*, $S(\mathcal{O}^*_{-i}) \leq S(\mathcal{O}^*)$. Therefore, $\mathcal{O}^*_{-i}$ is better than the optimal solution $\mathcal{O}^*$, which makes a contradiction.

We now prove $H(\mathcal{O}^*)/H(\mathcal{Y}) < Q_{k+1}$ by contradiction. Assume that $H(\mathcal{O}^*)/H(\mathcal{Y}) \geq Q_{k+1}$. On one hand, combining with $H(\mathcal{O}^*_{-i})/H(\mathcal{Y}) < Q_k$, we have:

$$\frac{H(\mathcal{O}^*) - H(\mathcal{O}^*_{-i})}{H(\mathcal{Y})} > Q_{k+1} - Q_k = \Delta_k \geq \Delta_{min} > \frac{H(\{i\})}{H(\mathcal{Y})}.$$

On the other hand, according to the submodularity of entropy function, we have:

$$H(\mathcal{O}^*) - H(\mathcal{O}^*_{-i}) \leq H(\{i\}) - H(\varnothing) = H(\{i\}).$$

Here, we get a contradiction, and thus $H(\mathcal{O}^*)/H(\mathcal{Y}) < Q_{k+1}$. It is obvious that $Q_k \leq H(\mathcal{O}^*)/H(\mathcal{Y})$. Therefore, we can conclude that $Q_k \leq H(\mathcal{O}^*)/H(\mathcal{Y}) < Q_{k+1}$. □

We now show that VENUS-PRO can achieve sub-optimal profit for each version. We use Φ^*_k and Φ_k to denote the optimal profit and the approximate profit of the version k, respectively. It is obvious that $R_k \geq S(\mathcal{O}^*)$, and we further assume that the revenue should be larger than the approximate expenditure, *i.e.*, $R_k \geq (1 + \log_e \beta_k)S(\mathcal{O}^*) \geq S(\mathcal{O}_T)$. Otherwise, the data broker may get negative approximate profit, and she would not sell the information commodity.

Lemma 3: For each version k, we have the following relation between the optimal profit and the approximate profit:

$$\frac{\Phi^*_k}{\Phi_k} \leq \frac{\xi_k - 1}{\xi_k - 1 - \log_e \beta_k},$$

where ξ_k denote the ratio between the revenue and the expenditure, i.e.*, $\xi_k = R_k/S(\mathcal{O}^*)$, and $\xi_k \geq 1 + \log_e \beta_k$.*

Proof: By Lemma 2, the highest version that the optimal solution $\mathcal{O}^*$ can achieve is exactly the version k. By the definition of profit, we have:

$$\Phi^*_k = R_k - S(\mathcal{O}^*). \tag{7}$$

Furthermore, by Theorem 1, we have:

$$\Phi_k = R_k - S(\mathcal{O}_T) \geq R_k - (1 + \log_e \beta_k)S(\mathcal{O}^*). \tag{8}$$

Together with Equations (7) and (8), we have:

$$\frac{\Phi^*_k}{\Phi_k} \leq \frac{R_k - S(\mathcal{O}^*)}{R_k - (1 + \log_e \beta_k)S(\mathcal{O}^*)} \leq \frac{\xi_k - 1}{\xi_k - 1 - \log_e \beta_k}.$$ □

Based on Lemma 3, We now present the approximation ratio of VENUS-PRO.

Theorem 2: For the profit maximization, the approximation factor of VENUS-PRO is $\max_k \left\{\frac{\xi_k - 1}{\xi_k - 1 - \log_e \beta_k}\right\}$.

Proof: We use APX and OPT to denote the profit obtained by VENUS-PRO and the optimal solution, respectively. VENUS-PRO selects the maximum approximate profit as the final result, *i.e.*, $APX = \max_k\{\Phi_k\}$, and the optimal profit is the maximum profit of all versions, *i.e.*, $OPT = \max_k\{\Phi^*_k\}$. Using Lemma 3, we immediately have:

$$\frac{OPT}{APX} = \frac{\max_k \Phi^*_k}{\max_k \Phi_k} \leq \max_k \frac{\Phi^*_k}{\Phi_k} \leq \max_k \left\{\frac{\xi_k - 1}{\xi_k - 1 - \log_e \beta_k}\right\}.$$ □

IV. VENUS-PAY

In VENUS-PRO, we assumed that the expected payment for each acquisition point is known. In this section, we determine this payment by designing an optimal and strategy-proof data procurement auction, namely VENUS-PAY. We first briefly review related solution concepts used in this section from game theory. Secondly, we prove a useful theorem in the context of procurement auctions: *expected payment is equal to expected virtual social welfare*, extending the main results of seminal Myerson's work [39]. Thirdly, combining this theorem with the knowledge of cost distributions, we calculate the value of expected minimum payment before conducting the data procurement auction, and regard such obtained payments as the inputs of VENUS-PRO. Finally, we designed VENUS-PAY to realize this expected minimum payment.

A. Solution Concepts

A strong solution concept from game theory is *dominant strategy*.

Definition 2 (Dominant Strategy [13]): Strategy s_i is player i's dominant strategy, if for any strategy $s'_i \neq s_i$ and any other players' strategy profile $\mathbf{s}_{-i}$: $u_i(s_i, \mathbf{s}_{-i}) \geq u_i(s'_i, \mathbf{s}_{-i})$.

The concept of dominant strategy is the basis of *incentive-compatibility* (**IC**), which means that there is no incentive for any player to lie about her private information, and thus revealing truthful information is the dominant strategy for every player. An accompanying concept is *individual-rationality* (**IR**), which means that every player participating in the game expects to gain no less utility than staying outside. As the utility of not participating in the game is normally zero, the individual-rationality requires that the utility of each player

should be non-negative. We now can introduce the definition of *Strategy-Proof Mechanism*.

Definition 3 (Strategy-Proof Mechanism [38]): A mechanism is strategy-proof when it satisfies both incentive-compatibility and individual-rationality.

According to Myerson's theorem [39], a single-parameter procurement auction, in which bidders have single private information, *i.e.*, data collection cost in this paper, is strategy-proof when its selection rule is monotone.

Theorem 3 (Myerson's Theorem [39]): A single parameter procurement auction is strategy-proof if and only if:

Monotone Selection: A selection rule $\mathbf{x}$ *is monotone if for every bidder* i *and bids* $\mathbf{b}_{-i}$ *by the other bidders, the selection rule* $x_i(b_i', \mathbf{b}_{-i})$ *to* i *is non-increasing in its bid* b_i'.

In the seminal paper [39], Myerson proved that for strategy-proof procurement auctions, the monotone selection rule implies a unique payment calculation rule:

$$w_i(b_i, \mathbf{b}_{-i}) = -\int_{b_i}^{\bar{c}} z \times \frac{d}{dz} x_i(z, \mathbf{b}_{-i})\, dz. \tag{9}$$

Here, we assume that $x_i(z, \mathbf{b}_{-i})$ is differentiable.[8]

According to Theorem 3, we can transform the design of a strategy-proof mechanism, guaranteeing the properties of incentive-compatibility (**IC**) and individual-rationality (**IR**) to the search for the monotone selection rule with good performance guarantee.

Another standard solution concept from game theory is Nash Equilibrium (NE). A strategy profile $\mathbf{s}^*$ is a Nash Equilibrium of a game, if for any player i and any strategy $s_i \neq s_i^*$, $u_i(s_i, \mathbf{s}_{-i}^*) \geq u_i(s_i', \mathbf{s}_{-i}^*)$. However, NE does not provide an ideal solution to the problem of data procurement. There are two reasons: (1) NE is not a very strong solution concept. Specifically, when in NE, the game player has incentives to keep her equilibrium strategy only under the assumption that all the other players are also keeping their equilibrium strategies. Without this assumption, NE no longer provides incentives for game player. (2) More importantly, NE usually has no guarantee on system performance, which means that the system performance is not optimized. When the system converge to one of NEs, the corresponding performance, such as social welfare or revenue, may be lower. In contrast to NE, the Dominant Strategy Equilibrium (DSE) or the strategy-proofness ensures every player has incentive to use the equilibrium strategy, regardless of the other players' strategies. We show that when the data procurement auction converge to DSE, the optimal expected payment is also achieved. Thus, the DSE-based auction mechanism is more desirable than the above NE-based mechanism in data acquisition process.

B. Design Rationale

The data procurement auction, in which the data broker wants to purchase one observation from a pool of competitive data providers with a minimum expected payment, can be considered as a reversed version of the Myerson auction [39]. The main result in Myerson auction (*i.e.*, maximizing expected revenue can be reduced to maximizing expected virtual social welfare) collapses in our context of data procurement auction. We theoretically prove a powerful theorem: *minimizing expected payment is equal to minimizing expected virtual social welfare*. This theorem is the basis of designing the reverse auction with the goal of payment minimization. Before proving this theorem, we formally define *virtual cost* in data procurement auctions.

Definition 4 (Virtual Cost): In a data procurement auction, the virtual cost of the data provider $i \in \mathcal{M}_y$ *with the cost* c_i *drawn from* F_i *is defined as:*

$$\varphi_i(c_i) \triangleq c_i + \frac{F_i(c_i)}{f_i(c_i)}. \tag{10}$$

As the previous works [32], [39], we assume that the cost distribution $F_i(\cdot)$ is regular, *i.e.*, the virtual cost function $\varphi_i(c_i)$ is a strictly increasing function, which is met by most of the distribution functions, such as uniform distributions, exponential distributions, and lognormal distributions.

We prove our main result for data procurement auction.

Theorem 4: In a data procurement auction, the expected payment is equal to the expected virtual social welfare, i.e.*,*

$$\mathbf{E}_{\mathbf{c}\sim\mathbf{F}}\left[\sum_{i=1}^{m_y} w_i(\mathbf{c})\right] = \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}\left[\sum_{i=1}^{m_y} \varphi_i(c_i) \times x_i(\mathbf{c})\right].$$

Proof: We fix the costs of the other data providers as $\mathbf{c}_{-i}$, and consider the expected payment of the provider i:

$$\begin{aligned}
\mathbf{E}_{c_i\sim F_i}\left[w_i(\mathbf{c})\right] &= \int_{\underline{c}}^{\overline{c}} w_i(\mathbf{c}) f_i(c_i)\, dc_i \\
&= \int_{\underline{c}}^{\overline{c}} \left[-\int_{c_i}^{\bar{c}} z \times \frac{d}{dz} x_i(z, \mathbf{c}_{-i})\, dz\right] f_i(c_i)\, dc_i \\
&= -\int_{\underline{c}}^{\overline{c}} \left[\int_{\underline{c}}^{z} f_i(c_i)\, dc_i\right] z \times \frac{d}{dz} x_i(z, \mathbf{c}_{-i})\, dz \\
&= -\int_{\underline{c}}^{\overline{c}} F_i(z) \times z \times \frac{d}{dz} x_i(z, \mathbf{c}_{-i})\, dz
\end{aligned} \tag{11}$$

In the first equation, we exploit the independence of cost distributions, *i.e.*, the fixed value of $\mathbf{c}_{-i}$ has no impact on the distribution F_i. The second equation comes from the Myerson's payment formula (9). We reverse the integration order in the third equation. We adopt the method of integration by parts to make the integral a more interpretable form.

$$\begin{aligned}
(11) &= \underbrace{-F_i(z) \times z \times x_i(z, \mathbf{c}_{-i})\big|_{\underline{c}}^{\overline{c}}}_{=0-0} \\
&\quad + \int_{\underline{c}}^{\overline{c}} x_i(z, \mathbf{c}_{-i}) \times \left(F_i(z) + z f_i(z)\right) dz \\
&= \int_{\underline{c}}^{\overline{c}} \left(z + \frac{F_i(z)}{f_i(z)}\right) x_i(z, \mathbf{c}_{-i}) f_i(z)\, dz \\
&= \int_{\underline{c}}^{\overline{c}} \varphi_i(c_i) \times x_i(c_i, \mathbf{c}_{-i}) f_i(c)\, dc \\
&= \mathbf{E}_{c_i\sim F_i}\left[\varphi_i(c_i) \times x_i(\mathbf{c})\right].
\end{aligned}$$

Finally, we have the equation for every bidder i and a cost vector $\mathbf{c}_{-i}$: $\mathbf{E}_{c_i\sim F_i}\left[w_i\ (\mathbf{c})\right] = \mathbf{E}_{c_i\sim F_i}\left[\varphi_i(c_i) \times x_i\ (\mathbf{c})\right]$.

[8] Actually, by standard advanced calculus, the same formula holds for an arbitrary monotone selection function, including piecewise constant function, for a suitable interpretation of the derivative and the corresponding integral.

We recall that $\mathbf{c}_{-i}$ is a vector of acquisition points, and we take the expectation, with respect to $\mathbf{c}_{-i}$, of both sides of this equation to obtain: $\mathbf{E}_{\mathbf{c}\sim\mathbf{F}}[w_i\ (\mathbf{c})] = \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}[\varphi_i(c_i) \times x_i\ (\mathbf{c})]$. Applying linearity of expectations twice, we can get:

$$\begin{aligned}\mathbf{E}_{\mathbf{c}\sim\mathbf{F}}\left[\sum_{i=1}^{m_y} w_i\ (\mathbf{c})\right] &= \sum_{i=1}^{m_y} \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}[w_i\ (\mathbf{c})] \\ &= \sum_{i=1}^{m_y} \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}[\varphi_i(c_i) \times x_i\ (\mathbf{c})] \\ &= \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}\left[\sum_{i=1}^{m_y} \varphi_i(c_i) \times x_i\ (\mathbf{c})\right].\end{aligned}$$

We refer to $\sum_{i=1}^{m_y} \varphi_i(c_i) \times x_i\ (\mathbf{c})$ as the *virtual social welfare* with a cost profile $\mathbf{c}$. Thus, we have proved our claim. □

By Theorem 4, we will always choose the data provider with the lowest virtual cost as the winner. The minimum expected payment for the data acquisition point $y \in \mathcal{Y}$ can be expressed as: $\Omega_y = T \times \mathbf{E}_{\mathbf{c}\sim\mathbf{F}}[\min_i \varphi_i(c_i)]$. We can calculate this optimal payment with the knowledge of cost distributions $\mathbf{F}$. For easy illustration, we introduce some notations. Let $\varphi_{min}(\mathbf{c})$ denote the minimum virtual cost for a given cost profile $\mathbf{c}$, *i.e.*, $\varphi_{min}(\mathbf{c}) \triangleq \min_i \varphi_i(c_i)$. We denote the cumulative distribution function of $\varphi_i(c_i)$ by $G_i(z)$, which can be derived from the cost distribution $F_i(x)$, *i.e.*, $G_i(z) = F_i(\varphi_i^{-1}(z))$. Let $G_{min}(z)$ denote the cumulative distribution function of $\varphi_{min}(\mathbf{c})$:

$$\begin{aligned}G_{min}(z) &= \mathbf{Pr}(\varphi_{min}(\mathbf{c}) \le z) = 1 - \mathbf{Pr}(\varphi_{min}(\mathbf{c}) > z) \\ &= 1 - \prod_{i=1}^{m_y} \mathbf{Pr}(\varphi_i(c_i) > z) \\ &= 1 - \prod_{i=1}^{m_y} (1 - G_i(z)).\end{aligned}$$

The payment Ω_y for each acquisition point $y \in \mathcal{Y}$ is:

$$\begin{aligned}\Omega_y &= T \times \mathbf{E}_{\varphi_{min}(\mathbf{c})\sim G_{min}}[\varphi_{min}(\mathbf{c})] \\ &= T \times \int_{\underline{\varphi}}^{\overline{\varphi}} z \times g_{min}(z)\, dz \\ &= T \times \left(\overline{\varphi} - \int_{\underline{\varphi}}^{\overline{\varphi}} G_{min}(z)\, dz\right), \qquad (12)\end{aligned}$$

where $g_{min}(z)$ is the probability density function of the random variable $\varphi_{min}(\mathbf{c})$, and the range $[\underline{\varphi}, \overline{\varphi}]$ is the support of the random variable $\varphi_{min}(\mathbf{c})$, $\underline{\varphi} \triangleq \min_i \varphi_i(\underline{c}_i)$ and $\overline{\varphi} \triangleq \max_i \varphi_i(\overline{c}_i)$. We adopt the method of integration by parts in the last part of Equation (12).

How should we design a selection rule $\mathbf{x}$ to realize this optimal payment? We have no control over the cost distributions $\mathbf{F}$ or the virtual cost functions $\varphi_i(c_i)$, so the natural approach is to design the selection rule $\mathbf{x}(\mathbf{c})$, such that the achieved virtual social welfare is minimum for every possible input cost profile $\mathbf{c}$. With this observation, we design an optimal procurement auction in next subsection.

Algorithm 3 VENUS-PAY for Payment Minimization

Input: The number of data providers m_y, a bidding profile $\mathbf{b}$, a vector of cost distributions $\mathbf{F}$, a vector of corresponding probability density function $\mathbf{f}$.

Output: A pair of selection result and payment result $(\mathbf{x}(\mathbf{b}), \mathbf{w}(\mathbf{b}))$.

1 $\mathbf{x}(\mathbf{b}) \leftarrow \mathbf{0}$; $\mathbf{w}(\mathbf{b}) \leftarrow \mathbf{0}$;
2 // Winner Selection
3 **for** $i = 1$ *to* m_y **do**
4 $\quad \varphi_i(b_i) \leftarrow b_i + \frac{F_i(b_i)}{f_i(b_i)}$;
5 $i^* \leftarrow \arg\min_i \varphi_i(b_i)$;
6 $x_{i^*}(\mathbf{b}) \leftarrow 1$;
7 // Payment Calculation
8 $\hat{i} \leftarrow \arg\min_{i \neq i^*} \varphi_i(b_i)$;
9 $w_{i^*}(\mathbf{b}) \leftarrow \min\left(\varphi_{i^*}^{-1}(\varphi_{\hat{i}}(b_{\hat{i}})), \overline{c}_{i^*}\right)$;
10 **return** $(\mathbf{x}(\mathbf{b}), \mathbf{w}(\mathbf{b}))$;

C. Detailed Design

VENUS-PAY consists of two major components: Winner Selection and Payment Calculation. We depict the pseudo-code of VENUS-PAY in Algorithm 3.

1) Winner Selection: After collecting the bids $\mathbf{b}$, the data broker calculates the virtual bid for each data provider by Equation (10) (Lines 3 to 4). Based on Theorem 4, minimizing the expected payment is equal to minimizing the expected virtual social welfare. Thus, in order to minimize the expected payment, the data broker chooses the data provider with the lowest virtual bid as the winner, *i.e.*, $x_{i^*}(\mathbf{b}) = 1$ for $i^* = \arg\min_i \{\varphi_i(b_i)\}$ (Lines 5 to 6). We note that selecting the data provider with the lowest bid does not lead to an optimal data procurement auction. We break the tie following any bid-independent rule, *e.g.*, the lexicographic order of data provider's ID. Due to the regularity of cost distributions, this winner selection rule is monotone.

Lemma 4: The selection rule in VENUS-PAY is monotone.

Proof: To prove the monotonicity of the selection rule, we have to show that any winning data provider i^* will still be selected as a winner when she decreases her cost, $c'_{i^*} \le c_{i^*}$. Since the cost distribution is regular, *i.e.*, the virtual cost function is a strictly increasing function, the virtual cost of c'_{i^*} will not be larger than that of c_{i^*}, *i.e.*, $\varphi_{i^*}(c'_{i^*}) \le \varphi_{i^*}(c_{i^*})$. Therefore, the data provider i^* will still be the winner, and the selection rule is monotone. □

2) Payment Calculation: By Theorem 3, the monotone selection rule implies a unique payment calculation rule. We note that in the data procurement auction, the selection rule $\mathbf{x}$ is a piecewise constant monotone function, meaning that $x_i(b_i, \mathbf{b}_{-\mathbf{i}})$ slump from 1 to 0 at some threshold point. In this case, the payment calculated by Myerson's formula (9) is equal to critical bid, which is defined as follows.

Definition 5 (Critical Bid): The critical bid $cr(i)$ for data provider $i \in \mathcal{M}_y$ is a threshold such that if i bids lower than $cr(i)$, she wins; otherwise she loses.

The critical bid of the winner i^* can be calculated by the following steps. If the data provider i^* still wants to be the

winner in the auction, her virtual cost $\varphi_{i^*}(b_{i^*})$ must be lower than the minimum virtual cost of the remaining data providers, *i.e.*, $\varphi_{i^*}(b_{i^*}) \leq \min_{i \neq i^*} \varphi_i(b_i)$; otherwise she will lose the auction. Since the virtual cost function is a strictly increasing function, there must exist a largest bid that satisfies the above winning condition. Considering that this largest bid may be larger than $\overline{c}_{i^*}$, we set the critical bid of the winner i^* as:

$$cr(i^*) \triangleq \min\left(\varphi_{i^*}^{-1}\left(\varphi_{\hat{i}}\left(b_{\hat{i}}\right)\right), \overline{c}_{i^*}\right), \tag{13}$$

where $\hat{i} = \arg\min_{i \neq i^*} \varphi_i(b_i)$. Finally, the payment of the winner i^* is set as her critical bid $cr(i^*)$, and the payments of the losers are zero (Lines 8 to 9).

By Lemma 4 and Theorem 3, we have following theorem.

Theorem 5: VENUS-PAY is a strategy-proof data procurement auction.

Proof: We first show that data provider $i \in \mathcal{M}_y$ cannot obtain a higher utility by bidding untruthfully. We discuss the analysis in the following two cases.

- The data provider i wins the auction and gets an utility $u_i \geq 0$ when bidding truthfully, *i.e.*, $b_i = c_i$. Suppose the data provider still wins the auction when she cheats the bid, *i.e.*, $b_i' \neq c_i$. The utility of the data provider remains the same, because the payment is unchanged. If the data provider loses the auction when she cheats the bid, her utility is zero, which is not better than the non-negative utility when bidding truthfully.
- The data provider i loses the auction when bidding truthfully, resulting in the utility to be zero. If she still loses when bidding untruthfully, her utility cannot be changed. We consider the scenario, in which she cheats the bid $b_i' \neq c_i$ and wins the auction. We represent the virtual bids $\varphi_i(b_i')$ and $\varphi_i(b_i)$ when the data consumer i bids truthfully and untruthfully, respectively. According to the winner selection principle, we have $\varphi_i(b_i') \leq \varphi_i(cr(i)) \leq \varphi_i\ (b_i)$. As the virtual cost function $\varphi_i(\cdot)$ is strictly increasing with respective to the declared bid. We can get $b_i' \leq cr(i) \leq b_i$. Her utility now becomes non-positive:

$$u_i' = cr(i) - c_i \leq b_i - c_i = c_i - c_i = 0.$$

From the above analysis of two cases, we can see that the data provider i cannot increases her utility by bidding any other value than c_i, and thus bidding truthfully is a dominant strategy for each data provider. Therefore, VENUS-PAY satisfies incentive compatibility.

We next prove that VENUS-PAY satisfies the property of individual rationality. On one hand, data provider's utility is zero if she loses in the auction. On the other hand, winning data provider gets utility:

$$u_i = cr(i) - c_i = \min\left(\varphi_i^{-1}(\varphi_{\hat{i}}(b_{\hat{i}})), \bar{c}_i\right) - b_i,$$

where $\varphi_{\hat{i}}(b_{\hat{i}})$ is the virtual bid of the critical bidder $\hat{i}$, *i.e.*, $\hat{i} = \arg\min_{i \neq i^*} \varphi_i(b_i)$. On one hand, since the data provider i is a winner and $\varphi_i(\cdot)$ is a strictly increasing function, we have $\varphi_i(b_i) \leq \varphi_{\hat{i}}(b_{\hat{i}})$ and then $b_i \leq \varphi_i^{-1}(\varphi_{\hat{i}}(b_{\hat{i}}))$. On the other hand, $\bar{c}_i$ is the upper bound of the bid, *i.e.*, $b_i \leq \bar{c}_i$. Combining these two equalities, we have $b_i \leq \min\left(\varphi_i^{-1}(\varphi_{\hat{i}}(b_{\hat{i}})), \bar{c}_i\right)$, implying that the utility of the data provider is always non-negative in this scenario. Therefore, we can conclude that VENUS-PAY satisfies individual rationality.

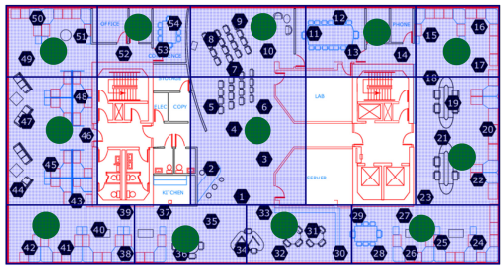


Fig. 2. Map of Intel Berkeley Lab deployment, with the placement of 54 sensors shown in dark hexagons. The green circles represent data acquisition points in the 12 regions.

Since VENUS-PAY satisfies both incentive compatibility and individual rationality, according to Definition 3, VENUS-PAY is a strategy-proof mechanism. □

Remark: We note that VENUS-PAY with i.i.d bidders and regular cost distributions $\mathbf{F}$ is simply the conventional second-price auction. In this scenario, all the cost distributions $\mathbf{F}$ reduce to a common distribution F, and thus the virtual cost functions $\varphi_i(c_i)$ are the same for all bidders. The bidder with the lowest virtual cost is also the bidder with the lowest cost. Furthermore, according to the expression of critical bid (Equation (13)), the payment of the winner is exactly the second lowest bid. In the asymmetric case, the cost distributions are non-identical, but still independent and regular. VENUS-PAY does not generally resemble any auctions used in practice. In next section, we will show that VENUS-PAY outperforms the second price auction in the asymmetric case.

V. Evaluation Results

In this section, we present the evaluation results of VENUS based on a real-world sensed data set.

A. Sensed Data Set

We first introduce the public data set collected by Intel Research, Berkeley Lab [19]. As shown in Figure 2, researchers deployed 54 Mica2Dot sensors in the lab to measure multiple environmental phenomena, *e.g.*, light, humidity, temperature and voltage readings. in a real time manner. We tailor the data set, and focus on the temperature measurements from all the 54 sensor nodes at 30 seconds intervals between February 28th, 2004 to April 5th, 2004. We discretize the collected data into 5 bins of 3 degrees Celsius each. We artificially partition the lab into 12 non-overlapped regions, and virtually deploy one data acquisition point in each of region, to represent the average of the readings measured by the sensors located in the corresponding region. We can use the collected samples to build a joint distribution over the 12 data acquisition points. For some selected random variables X_O, we can calculate its probability mass function $p(X_O)$ by projecting the joint distribution $p(X_\mathcal{Y})$ over X_O. With Equations (1) and (2), we can calculate the conditional distribution and the corresponding conditional entropy for any selected random variables.

B. Evaluation Setup

We set the number of versions as $K = 8$, and set the corresponding quality vector as $\mathbf{Q} = (0.25, 0.38, 0.53, 0.63, 0.70, 0.78, 0.89, 0.98)$. We fix the number of data consumers as $N = 500$ through the evaluation[9]. In order to examine the performance of VENUS under different data consumers' purchasing behavior models, we choose two typical version preference distributions and two common valuation distributions. Specifically, we adopt the *Poisson* distributions with two different parameters (parameter λ can be either 3 or 8) to be the version preference distributions. We set two types of valuation distributions as follows.

► *Uniform Distribution:* The data consumers' valuations on the kth version are uniformly distributed over a range $[2k, 4k]$.

► *Gaussian Distribution:* The data consumers' valuations over the kth version are drawn from a Gaussian distribution with mean $3k$ and variance 3, and with a lower bound 0.

Upon these distributions, we can obtain four different behavior models, and use "Poisson-λ, Uniform (or Gaussian)" to denote the Poisson distribution with parameter λ and the uniform (or Gaussian) valuation distribution. While the uniform distribution describes that data consumers have diverse valuations over the same set of data, the Gaussian distribution can capture the scenario that data consumers have similar valuations located in a centralized range. We note that the parameter setting of the valuation distributions guarantees that the valuation over the high version is always larger than that of the lower version.

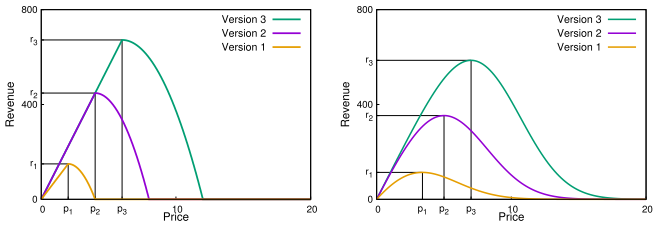
We evaluate the performance of VENUS-PAY for one randomly selected data acquisition point in a single time slot, *i.e.*, $T = 1$. The number of data providers for this data acquisition point increases from 6 to 10 sequentially. We set the cost distribution of the data providers as exponential distributions. Specifically, the cost of data provider i is drawn from the exponential distribution:

$$f_i(x) = \begin{cases} \alpha_i e^{-\alpha_i x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where the support of the distribution is $(0, \infty)$. We further consider identical cost distributions and non-identical cost distributions scenarios in VENUS-PAY. In the identical case, we set the parameters α to be 0.001 for all data providers, and in non-identical case, we set $\alpha_i = e^i/10^5$ for the data provider i. All the results of performance are averaged over 1000 runs.

1) Performance of VENUS-PRO: We implement VENUS-PRO, and compare its performance with the optimal algorithm and random algorithm. We obtain the optimal profit, denoted by "OPT", using the brute-force search method for the problem of profit maximization. The "OPT" result is served as the reference of upper bound of profit. In "Random" algorithm, we first randomly select a version as the result, and then use a set of random data acquisition points to satisfy the quality requirement of the selected version. The existing works about data acquisition for mobile crowdsensing focused on other optimization objectives, such as cost minimization [32] and data quality [23], and ignored the revenue extracted from data trading in the market. Thus, we do not compare VENUS-PRO with the existing data acquisition mechanisms.

[9]It is worth emphasizing that all parameters can be different from the ones used here. Considering that the evaluation results of using different parameters are identical, we only show the results for these parameters in this paper.



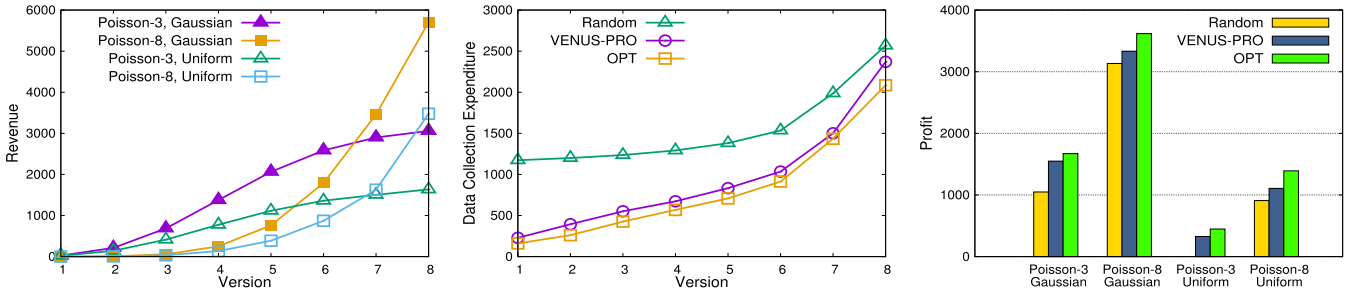
(a) Uniform Valuation Distribution. (b) Gaussian Valuation Distribution.

Fig. 3. Revenue curves for the first three versions when the version preference distribution is Poisson distribution with parameter 3.

With the knowledge of data consumers' purchasing behavior models, we can plot the revenue curve r_k in Equation (3) with respective to price, for the first three versions. We set the version preference distribution as the Poisson distribution with parameter 3. We further examine the evaluation result of uniform valuation distribution and Gaussian valuation distribution in Figure 3(a) and Figure 3(b), respectively. We obtain the optimal expected revenue of each version by calculating the optimal price point of its corresponding revenue curve. After that, we can plot the optimal revenue for all of the versions in Figure 4(a). From Figure 4(a), we can see that the data broker obtains a large revenue by providing high version for the information commodity. This is because the high version can satisfy the data consumers with high quality requirements, extracting additional revenue from these data consumers. From Figure 4(a), we can also see that the revenue function with respective to version has different trends and properties, under different purchasing behavior models. This result demonstrates that the property of revenue function is indeed complex, and is hard to analyze in complicated data markets in terms of diverse purchasing behavior models, making directly solving the problem of profit maximization infeasible.

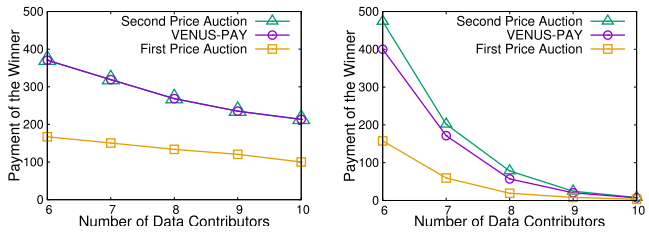
In order to calculate the data acquisition expenditure, we have to know the minimum expected payment at each data acquisition point. In this set of experiments, we assume that data providers have the identical cost distributions, and the number of data providers at each data acquisition point is randomly chosen from [6] and [17]. With this information, we can calculate the minimum expected payment for each data acquisition point by Equation (12). For a fixed version, VENUS-PRO applies the greedy algorithm, *i.e.*, Algorithm 1, to calculate an approximate data acquisition expenditure, and the result is shown in Figure 4(b). From Figure 4(b), we can see that the expenditure becomes large when high version is provided in the data market. The reason is that we need to select more data acquisition points to assure the high quality requirement. We can also see that VENUS-PRO always outperforms the Random algorithm, and approaches to the optimum.

Based on the obtained revenue (*i.e.*, Figure 4(a)) and the acquisition expenditure (*i.e.*, Figure 4(b)), we can plot the



(a) Revenue (b) Data Acquisition Expenditure (c) Profit

Fig. 4. Performance of VENUS-PRO, OPT, and Random.



(a) Identical Cost Distributions. (b) Non-Identical Cost Distributions.

Fig. 5. Performance of VENUS-PAY, First-Price Auction, and Second-Price Auction.

profits under different data consumers' purchasing behavior models in Figure 4(c). The result shows that VENUS-PRO is very close to the optimum in all the four representative purchasing behavior models, demonstrating the efficiency of VENUS-PRO for the crowd-sensed data trading. We note that in the "Poisson-3, Uniform" model, the expected profit of the random algorithm is negative, so the data broker would not sell the information commodity. We omit the profit of the random algorithm in this case.

2) Performance of VENUS-PAY: We now present the evaluation results of VENUS-PAY. We implement VENUS-PAY, and compare its performance with two classical auctions: first-price auction and second-price auction. The non-truthful first-price auction always disburse less payment compared to VENUS-PAY, because VENUS-PAY overpays the winner to guarantee strategy-proofness. We compare the results of first-price auction and VENUS-PAY to illustrate the system performance degradation caused by the requirement of strategy-proofness.

By varying the number of data providers, we collect a set of performance results, as illustrated in Figure 5. In Figure 5(a), VENUS-PAY has the same performance as the second-price auction in the identical cost distribution scenario. This coincides with our analysis that VENUS-PAY reduces to the second-price auction when the costs of data providers follow the same distribution. Figure 5(b) shows the evaluation results in the non-identical scenario. From Figure 5(b), we can see that VENUS-PAY outperforms the second-price auction, which does not take advantage of the knowledge of cost distributions. This result demonstrates that exploiting the information of cost distributions can reduce the expected payment to some extent. In both identical and non-identical cases, the result of VENUS-PAY is close to that of the first-price auction, denoting that VENUS-PAY sacrifices limited performance to satisfy the strategy-proofness. Although the first price auction always achieves the lowest disbursed payment, we can not apply it to the context of data procurement, because it has not any guarantee on economic properties.

VI. Related Work

In this section, we briefly review related work.

A. Data Market Design

In recent years, designing pricing mechanisms for online data markets attracts increasing interests, especially from database research community [3], [30], [31], [35]. These previous works mainly focused on designing computationally efficient and economic-robust data pricing mechanisms, achieving two important axioms, *i.e.*, arbitrage-free and discount-free [31]. Koutris *et al.* [30] showed that the prices of a large class of queries can be computed using an ILP solver. Later Lin and Kifer [35] designed an arbitrage-free pricing function for arbitrary query formats. However, these works did not consider the problem of data acquisition in data marketplaces. While there are a number of pricing mechanisms for different kinds of network services [27], [36], [37], [49], they cannot be directly applied into data marketplaces due to the unique characteristics of crowd-sensed data in terms of cost structure and uncertain feature.

B. Mobile Crowdsensing

Recently, mobile crowdsensing has emerged as a new paradigm to generate collective knowledge about phenomena at interested regions. Data acquisition is a critical component in mobile crowdsensing system, and various incentive mechanisms have been proposed to motivate mobile users to contribute data [7], [11], [18], [23], [25], [32], [58]. Yang *et al.* [58] applied Stackelberg game and reverse auction theory to design incentive mechanisms for two basic data acquisition models. Chen *et al.* [6] considered the network effect in user recruitment in crowdsoucing. Cheung *et al.* [7] designed an asynchronous and distributed algorithm to recruit mobile users for time sensitive tasks. Considering the locations of tasks and the movements of mobile users, He *et al.* [18] proposed two incentive mechanisms based on discounted-reward TSP algorithm and bargaining theory. Inspired by opportunistic networks, Karaliopoulos *et al.* [25] examined a practical crowdsensing scenario, in which mobile users can play the roles of both data collectors and data transmitter.

Karaliopoulos *et al.* [24] studied the payment distribution problem in light of learning the user profiles. Our data acquisition model for the problem of payment minimization is similar to Koutsopoulos [32], in which he designed an optimal reverse auction, achieving Bayesian Nash equilibrium. In this paper, we proposed an optimal data procurement auction with the guarantee of strategy-proofness, which is a stronger solution concept than Bayesian Nash equilibrium. Furthermore, we built a data trading model to capture the economic value of data in the market. Thus, our ultima goal is to extract maximum profit from data trading, which is different from the objective of minimizing the expected payment in [32].

C. Auction Mechanism Design

Myerson [39] initially studied the optimal single-item forward auction, and proved that the prevalent reserve-price-based auctions can achieve maximum expected revenue. In contrast to the forward auction, few of works studied the procurement auction design. Procurement auctions, introduced to computer science already in [40], were at first studied to minimize social welfare [12], [40], which is different from our objective of payment minimization. Recently, researchers have studied the problem of payment optimization under different definitions of frugality ratio, which measures the amount by which an auction "overpays" [26], [28]. By extending Myerson's seminal work, we designed the first strategy-proof and optimal procurement auction, and applied it into a new context of Internet economic system, *i.e.*, crwod-sensed data markets.

VII. Conclusion

In this paper, by jointly considering the problems of profit maximization and payment minimization, we have proposed the first framework of profit-driven data acquisition, namely VENUS, in the crowd-sensed data marketplace. Given the expected payment for each data acquisition point, we have proposed VENUS-PRO to achieve a sub-optimal profit. In order to determine the minimum payment for each data acquisition point, we have designed VENUS-PAY, which is an optimal, strategy-proof data procurement auction in Bayesian setting. We have implemented VENUS, and evaluated its performance on a real-world data set. Our evaluation results have shown that VENUS-PRO approaches the optimal profit, and VENUS-PAY outperforms the canonical second-price auction in terms of payments.

Acknowledgment

The authors would like to thank Wenxin Li for helpful discussions in the procurement auction design in Section III. 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] K. Amin, A. Rostamizadeh, and U. Syed, "Learning prices for repeated auctions with strategic buyers," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, South Lake Tahoe, CA, USA, Dec. 2013, pp. 1169–1177.

[2] A. Archer, C. Papadimitriou, K. Talwar, and É. Tardos, "An approximate truthful mechanism for combinatorial auctions with single parameter agents," in *Proc. 14th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, Baltimore, MD, USA, Jan. 2003, pp. 205–214.

[3] M. Balazinska, B. Howe, and D. Suciu, "Data markets in the cloud: An opportunity for the database community," *Proc. VLDB Endowment*, vol. 4, no. 12, pp. 1482–1485, 2011.

[4] H. K. Bhargava and V. Choudhary, "Research note—When is versioning optimal for information goods?" *Manage. Sci.*, vol. 54, no. 5, pp. 1029–1035, 2008.

[5] J.-M. Bohli, C. Sorge, and D. Westhoff, "Initial observations on economics, pricing, and penetration of the Internet of Things market," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 2, pp. 50–55, Apr. 2009.

[6] Y. Chen, B. Li, and Q. Zhang, "Incentivizing crowdsourcing systems with network effects," in *Proc. 35th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[7] M. H. Cheung, R. Southwell, F. Hou, and J. Huang, "Distributed time-sensitive task selection in mobile crowdsensing," in *Proc. 16th ACM Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Hangzhou, China, Jun. 2015, pp. 157–166.

[8] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, May 1994.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[10] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *Proc. 13th Int. Conf. Very Large Data Bases (VLDB)*, Toronto, ON, Canada, Aug. 2004, pp. 588–599.

[11] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1701–1709.

[12] J. Feigenbaum, C. Papadimitriou, R. Sami, and S. Shenker, "A BGP-based mechanism for lowest-cost routing," *Distrib. Comput.*, vol. 18, no. 1, pp. 61–72, Jul. 2005.

[13] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA, USA: MIT Press, 1991.

[14] T. Fujito, "Approximation algorithms for submodular set cover with applications," *IEICE Trans. Inf. Syst.*, vol. 83, no. 3, pp. 480–487, Mar. 2000.

[15] R. Gao *et al.*, "Jigsaw: Indoor floor plan reconstruction via mobile crowdsensing," in *Proc. 20th Int. Conf. Mobile Comput. Netw. (MobiCom)*, Maui, HI, USA, Sep. 2014, pp. 249–260.

[16] L. Gargano and M. Hammar, "A note on submodular set cover on matroids," *Discrete Math.*, vol. 309, no. 18, pp. 5739–5744, Sep. 2009.

[17] A. V. Goldberg and J. D. Hartline, "Collusion-resistant mechanisms for single-parameter agents," in *Proc. 16th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, Vancouver, BC, Canada, Jan. 2005, pp. 620–629.

[18] S. He, D.-H. Shin, J. Zhang, and J. Chen, "Toward optimal allocation of location dependent tasks in crowdsensing," in *Proc. 33rd IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Toronto, ON, Canada, Apr. 2014, pp. 745–753.

[19] *Intel Research Berkeley Sensor Network Data*. (2004). [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html

[20] I. Simonson and A. Tversky, "Choice in context: Tradeoff contrast and extremeness aversion," *J. Marketing Res.*, vol. 29, no. 3, pp. 281–295, 1992.

[21] R. K. Iyer and J. Bilmes, "Submodular optimization with submodular cover and submodular knapsack constraints," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, South Lake Tahoe, CA, USA, Dec. 2013, pp. 2436–2444.

[22] R. Iyer, S. Jegelka, and J. A. Bilmes, "Fast semidifferential-based submodular function optimization," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Atlanta, GA, USA, Jun. 2013, pp. 855–863.

[23] H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu, "Quality of information aware incentive mechanisms for mobile crowd sensing systems," in *Proc. 16th ACM Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Hangzhou, China, Jun. 2015, pp. 167–176.

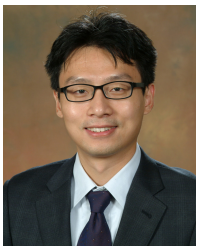
[24] M. Karaliopoulos, I. Koutsopoulos, and M. Titsias, "First learn then earn: Optimizing mobile crowdsensing campaigns through data-driven user profiling," in *Proc. 17th ACM Symp. Mobile Ad Hoc Netw. Comput. (MobiHoc)*, Paderborn, Germany, Jul. 2016, pp. 271–280.

[25] M. Karaliopoulos, O. Telelis, and I. Koutsopoulos, "User recruitment for mobile crowdsensing over opportunistic networks," in *Proc. 34th IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr. 2015, pp. 2254–2262.
[26] A. R. Karlin and D. Kempe, "Beyond VCG: Frugality of truthful mechanisms," in *Proc. 46th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2005, pp. 615–624.
[27] G. S. Kasbekar and S. Sarkar, "Spectrum pricing games with spatial reuse in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 1, pp. 153–164, Jan. 2012.
[28] D. Kempe, M. Salek, and C. Moore, "Frugal and truthful auctions for vertex covers, flows and cuts," in *Proc. 51th Annu. Symp. Foud. Comput. Sci. (FOCS)*, Las Vegas, NV, USA, Oct. 2010, pp. 745–754.
[29] G. Koop, D. J. Poirier, and J. L. Tobias, *Bayesian Econometric Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[30] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with QueryMarket," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, New York, NY, USA, Jun. 2013, pp. 613–624.
[31] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *J. ACM*, vol. 62, no. 5, Nov. 2015, Art. no. 43.
[32] I. Koutsopoulos, "Optimal incentive-driven design of participatory sensing systems," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1402–1410.
[33] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *Proc. 21st Conf. Uncertainty Artif. Intell. (UAI)*, Edinburgh, Scotland, Jul. 2005, pp. 324–331.
[34] A. Krause and C. Guestrin, "A note on the budgeted maximization on submodular functions," Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CALD-05-103, 2005.
[35] B.-R. Lin and D. Kifer, "On arbitrage-free pricing for general data queries," *Proc. VLDB Endowment*, vol. 7, no. 9, pp. 757–768, May 2014.
[36] R. T. B. Ma, D. M. Chiu, J. C. S. Lui, V. Misra, and D. Rubenstein, "Internet economics: The use of shapley value for ISP settlement," *IEEE/ACM Trans. Netw.*, vol. 18, no. 3, pp. 775–787, Jun. 2010.
[37] P. Marbach and R. Berry, "Downlink resource allocation and pricing for wireless networks," in *Proc. 21st Annu. IEEE Conf. Comput. Commun. (INFOCOM)*, New York, NY, USA, Jun. 2002, pp. 1470–1479.
[38] A. Mas-Colell, M. D. Whinston, and J. R. Green, *Microecon. Theory*. London, U.K.: Oxford Univ. Press, 1995.
[39] R. B. Myerson, "Optimal auction design," *Math. Oper. Res.*, vol. 6, no. 1, pp. 58–73, 1981.
[40] N. Nisan and A. Ronen, "Algorithmic mechanism design (extended abstract)," in *Proc. 31st Annu. Symp. Theory Comput. (STOC)*, Atlanta, GA, USA, 1999, pp. 129–140.
[41] *NoiseMap: A Research Project at Technische Universität Darmstadt*. [Online]. Available: https://www:tk:informatik:tu-darmstadt:de/de/research/smarturban-networks/noisemap/
[42] *NoiseTube: A Research Project at the Sony Computer Science Laboratory in Paris*. (2008). [Online]. Available: http://www:noisetube:net/
[43] A. Odlyzko, "Paris metro pricing for the Internet," in *Proc. ACM Symp. Electron. Commerce (EC)*, Denver, CO, USA, Oct. 1999, pp. 140–147.
[44] C. Papadimitriou, M. Schapira, and Y. Singer, "On the hardness of being truthful," in *Proc. 49th Annu. Symp. Foud. Comput. Sci. (FOCS)*, Philadelphia, PA, USA, Oct. 2008, pp. 250–259.
[45] *Quandl*. (2011). [Online]. Available: https://www.quandl.com/
[46] A. Ronen and A. Saberi, "On the hardness of optimal auctions," in *Proc. 43rd Annu. Symp. Foud. Comput. Sci. (FOCS)*, Vancouver, BC, Canada, Oct. 2002, pp. 396–405.
[47] C. Shapiro and H. R. Varian, "Versioning: The smart way to sell information," *Harvard Bus. Rev.*, vol. 107, no. 6, pp. 106–114, 1998.
[48] G. E. Smith and T. T. Nagle, "Frames of reference and buyers' perception of price and value," *California Manage. Rev.*, vol. 38, no. 1, pp. 98–116, 1995.
[49] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Pricing for demand shaping and proactive download in smart data networks," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 3189–3194.
[50] *Terbin*. (2013). [Online]. available: http://www.terbine.com
[51] *Thingful*. (2014). [Online]. Available: https://thingful.net/
[52] *Thingspeak*. (2010). [Online]. Available: https://thingspeak.com/
[53] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. V. Vazirani, "How many tiers?: Pricing in the Internet transit market," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, Toronto, ON, Canada, Aug. 2011, pp. 194–205.
[54] H. R. Varian, "Versioning information goods," Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep., 1997.
[55] P.-J. Wan, D.-Z. Du, P. Pardalos, and W. Wu, "Greedy approximations for minimum submodular cover with submodular cost," *Comput. Optim. Appl.*, vol. 45, no. 2, pp. 463–474, Mar. 2010.
[56] *Windows Azure Data Marketplace*. [Online]. Avbailable: https://datamarket:azure:com/browse/data.
[57] L. A. Wolsey, "An analysis of the greedy algorithm for the submodular set covering problem," *Combinatorica*, vol. 2, no. 4, pp. 385–393, Dec. 1982.
[58] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proc. 18th Int. Conf. Mobile Comput. Netw. (Mobicom)*, Aug. 2012, pp. 173–184.

Zhenzhe Zheng (S'16) is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. His research interests include algorithmic game theory, resource management in wireless networking, and data center. He is a student member of ACM and CCF.

Yanqing Peng received the B.Eng. degree in computer science and engineering from Shanghai Jiao Tong University, in 2016. He is currently pursuing the Ph.D. degree with the School of Computing, University of Utah, USA. His research interests include wireless networking, data center networking, algorithmic game theory, and large-scale data management.

Fan Wu (M'14) received the B.S. degree in computer science from Nanjing University, in 2004, and the Ph.D. degree in computer science and engineering from The State University of New York at Buffalo, in 2009. He is currently 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, as a Post-Doctoral Research Associate. He has published over 100 peer-reviewed papers in technical journals and conference proceedings. His research interests include wireless networking and mobile computing, algorithmic game theory and its applications, and privacy preservation. He is a recipient of the first-class prize for Natural Science Award of China Ministry of Education, the NSFC Excellent Young Scholars Program, the ACM China Rising Star Award, the CCF-Tencent Rhinoceros Bird Outstanding Award, the CCF-Intel Young Faculty Researcher Program Award, and the Pujiang Scholar. He has served as the Chair of CCF YOCSEF Shanghai, on the Editorial Board of *Elsevier Computer Communications*, and as the member of the Technical Program Committees of over 60 academic conferences.



Shaojie Tang (M'15) received the Ph.D. degree in computer science from the Illinois Institute of Technology, in 2012. He is currently an Assistant Professor with the Naveen Jindal School of Management, The University of Texas at Dallas. 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 the IEEE MASS 2013. He also received the ACM SIGMobile Service Award in 2014. He served in various positions (as chairs and TPC members) at numerous conferences, including the ACM MobiHoc and the IEEE ICNP. He is an Editor of the *Elsevier Information Processing in the Agriculture* and the *International Journal of Distributed Sensor Networks*.

Guihai Chen (SM'16) received the B.S. degree from Nanjing University, in 1984, the M.E. degree from Southeast University, in 1987, and the Ph.D. degree from The University of Hong Kong, in 1997. He has been invited as a Visiting Professor to many universities, including Kyushu Institute of Technology, Japan, in 1998, The University of Queensland, Australia, in 2000, and Wayne State University, USA, from 2001 to 2003. He is a Distinguished Professor with Shanghai Jiaotong University, China. He has published over 200 peer-reviewed papers, and over 120 of them are in well-archived international journals, such as the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the *Journal of Parallel and Distributed Computing*, *Wireless Network*, *The Computer Journal*, the *International Journal of Foundations of Computer Science*, and *Performance Evaluation*. His research interests include sensor network, peer-to-peer computing, high-performance computer architecture, and combinatorics. He is also in well-known conference proceedings, such as HPCA, MOBIHOC, INFOCOM, ICNP, ICPP, IPDPS, and ICDCS.