High-Confidence Computing 5 (2025) 100232

Contents lists available at ScienceDirect

High-Confidence Computing

journal homepage: www.sciencedirect.com/journal/high-confidence-computing

Research article

Federated data acquisition market: Architecture and a mean-field based data pricing strategy

Jiejun Hu-Bolz [a,*], Martin Reed [b], Kai Zhang [b], Zelei Liu [c], Juncheng Hu [d]

[a] Data Science Institute, Lancaster University Leipzig, Leipzig 04109, Germany
[b] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK
[c] Unicom (Shanghai) Industrial Internet Co., Ltd., Shanghai 200050, China
[d] College of Computer Science and Technology, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 28 November 2023
Revised 30 January 2024
Accepted 14 March 2024
Available online 22 April 2024

Keywords:
Federated data acquisition market
Pricing strategy
Data trading

ABSTRACT

With the increasing global mobile data traffic and daily user engagement, technologies, such as mobile crowdsensing, benefit hugely from the constant data flows from smartphone and IoT owners. However, the device users, as data owners, urgently require a secure and fair marketplace to negotiate with the data consumers. In this paper, we introduce a novel federated data acquisition market that consists of a group of local data aggregators (LDAs); a number of data owners; and, one data union to coordinate the data trade with the data consumers. Data consumers offer each data owner an individual price to stimulate participation. The mobile data owners naturally cooperate to gossip about individual prices with each other, which also leads to price fluctuation. It is challenging to analyse the interactions among the data owners and the data consumers using traditional game theory due to the complex price dynamics in a large-scale heterogeneous data acquisition scenario. Hence, we propose a data pricing strategy based on mean-field game (MFG) theory to model the data owners' cost considering the price dynamics. We then investigate the interactions among the LDAs by using the distribution of price, namely the mean-field term. A numerical method is used to solve the proposed pricing strategy. The evaluations demonstrate that the proposed pricing strategy efficiently allows the data owners from multiple LDAs to reach an equilibrium on data quantity to sell regarding the current individual price scheme. The result further demonstrates that the influential LDAs determine the final price distribution. Last but not least, it shows that cooperation among mobile data owners leads to optimal social welfare even with the additional cost of information exchange.

1. Introduction

Mobile crowdsensing (MCS) is a crowd sourcing technique that is empowered by the development of communication technologies and devices, such as smartphones and the Internet of Things (IoT). Participants in the MCS contribute private information, also known as personally identifiable information to the MCS task initiators [1]. The MCS task initiators can use the information to infer even more facts about the participants and predict the behaviour of other individuals or groups. MCS participants are becoming more and more aware that their data has more value than it first seems even though there is usually a reward to stimulate their participation. However, the cost of data proliferation is so low that MCS task initiators can easily trade the data to an advertisement company, for example. The data collecting process is crucial to both the participants and the task initiators. However, the current mechanisms, typically involving implicit consent from the MCS data owners by using a service that is also managed by the MCS task initiator (data consumer) are far from ideal for both parties. The end-users, *data owners*, are exploited by the data consumers (i.e., MCS task initiators, the technology companies, and advertisement companies); while if the data consumers are limited to their direct customers they cannot benefit from wider data gathering. *Consequently, a proper marketplace for the data owners and data consumers is urgently required to protect the data owners' welfare and allow data consumers access to a wider pool of data.*

In MCS applications, a data owner is interchangeable with a participant. For a data acquisition framework, direct negotiation between a massive number of individual data owners and the data consumer is not realistic. Thus, we propose a federated data acquisition market, where a *data union* interlinks the data consumer and the *local data aggregators* (LDA) of data owners. LDAs represent the data owners, they provide not only additional aggregation and encryption but also crucially, the collective

* Corresponding author.
E-mail address: J.hu14@lancaster.ac.uk (J. Hu-Bolz).

https://doi.org/10.1016/j.hcc.2024.100232

negotiation power of the data owners. The proposed federated data acquisition market is flexible at different scales: for a geographically constrained MCS task, the LDAs can be hybrid base stations; for large-scale data acquisition tasks, the LDAs can be mobile carriers, or other network service providers with a group of subscribers as data owners.

The proposed federated data acquisition market aims to enable large-scale data collection, data diversity, and privacy-preserving data acquisition. Meanwhile, it provides a regulated marketplace for data owners and data consumers, which also empowers a new business model for data owners and consumers. However, the federated data acquisition market brings a new challenge to data consumers. For example, localised data may not serve well in machine learning algorithms i.e., localised data could be an *isolated data island* [2]. To accumulate more knowledge, the data consumer will purchase data from multiple LDAs during a period of time. The data consumers make an individual offer to the data owners to gain access to the data. It is natural for the data owners to gossip about the individual unit price to see if their data is worth more than others. Hence, the individual unit price fluctuates with the data demand quantity and the rest of the offers. One LDA consists of several data owners who decide the quantity of data to sell regarding the unit price and the influence of fellow data owners from intra-, or even inter-LDA(s). Inevitably, the data owners will share information to cooperate against the data consumer, which ensures their data is fairly traded. To investigate the pricing strategy in a large-scale and time-dependent scenario, we, therefore, design the pricing strategy based on the mean-field game theory (MFG) [3].

Game theory is an effective tool to analyse the interactions (such as competition and cooperation) between participants [4]. However, it is challenging to analyse large-scale problems when the number of participants approaches infinity in a dynamic environment. Every individual is constantly generating data and involved in the data collection process. Additionally, the proposed federated data acquisition market is facing the dynamic problem that there are not only interactions within a local data aggregator but also with other adjoining local data aggregators. MFG is an ideal tool to investigate both the optimal strategy of the data owners and the evolution of the pricing regarding the interactions. MFG was first proposed by Lasry et al. [3]. It utilises both the Hamilton–Jacobi–Bellman (HJB) equation and the Fokker–Planck–Kolmogorov (FPK) equation to capture the optimal strategy and the state evolution, respectively. The mean-field term is crucial in MFG, which describes the probability density of the players' states. We will explain the details of the MFG-based pricing strategy in Section 5.

In the proposed marketplace for an MCS task, the data consumer first notifies all the data owners of their individual price schemes. Individual price schemes value different data at different times for different data owners to stimulate the willingness of data trading. Once data owners receive the offers, it is natural for them to compare the offers with other data owners first before deciding the quantity of data/privacy to sell. For example, if data owner A talks to data owner B and finds out that B's unit price is higher, then A chooses to reduce the data quantity for trading, and vice versa. This natural tendency of gossip in the marketplace leads to unit price fluctuation, namely *cooperation*. The data owners seek the optimal data quantity during a period of trading time with a specific data consumer considering the data sharing cost, information exchange cost, and the income of data trading. According to the analysis, we model this problem as a partial differential equation (PDE)-constrained optimisation problem, which aims to minimise the cost of the data owners during trading with the data consumer. Since data consumers collect data from different LDAs to achieve better data diversity, data owners are influenced not only by the data owners within the LDA, but also by the ones from other LDAs. Hence, we generalise the problem to multiple LDAs. We adopt a numerical method, namely the finite difference method with gradient descent, to solve the proposed pricing strategy. The main contributions of this paper are summarised as follows:

- We design a federated data acquisition market, which provides a solution for large-scale data trading. We design the data trading workflows among the data consumer, Data Union, and data owners.
- We propose a mean-field-based pricing strategy to capture the dynamics in the proposed market. The mathematical model first considers the dynamic of price, information exchange (namely cooperation) cost, and data trading income of a single data owner when interacting within the LDA. Then, we generalise the model into multiple LDAs by using the distribution of the unit price in each LDA.
- We adopt a numerical method to solve the mathematical model. The simulation demonstrates first, that through the cooperation among the data owners, the optimal data quantity reaches equilibrium. Second, the influential local data aggregators show a dominant effect over the other local data aggregators. Third, cooperation of data owners leads to higher *social welfare*.[1] Last, but not least, the proposed algorithm achieves good efficiency.

In the following, we first review related works on data marketplaces in Section 2. Then, in Section 4, we introduce the architecture and the workflow of the proposed federated data acquisition marketplace. Next, we provide the considered mathematical model of the proposed pricing strategy in Section 5. We, then, solve the model using the numerical method in Section 6. Our solution is evaluated extensively in Section 7 to get an understanding of the price evolution and optimal data quantity. Finally, we draw attention to the future data marketplace in Section 8.

2. Overview of the data marketplace

Smart data pricing (SDP) [5] has been introduced as an approach to solve network resource management, pricing, and allocation issues in the computer science realm. Researchers tend to use economic approaches to analyse data and digital products as common commodities. Data and digital product marketplaces are usually equivalent to traditional marketplaces, i.e., monopoly, duopoly, oligopoly, and a competitive market. The data pricing strategies according to the data quantity [1,6], quality [7,8], privacy [9,10], and learning performance [11] have been widely investigated.

Quantity driven pricing strategies: In a data trading market, it is widely accepted that the income of a data owner is proportional to the amount of data it owns. IoT applications are commonly enabled by pervasive sensors to contribute data. Data quantity is one of the essential factors to the accuracy of the applications. In [1] a monthly-pay and instant-pay sensory data pricing strategy based on data quantity was proposed to ensure a constant data contribution from the monthly/instant-pay sensors. Opting in as monthly- or instant-pay participation grants flexibility in the MCS participation to some extent. However, the data and digital products markets desire more dynamic pricing strategies. In [6], the buyers' social relationships and network resources of the buyers are studied in the digital product (mobile data plan) marketplace. By leveraging the network effect of the buyers, a dynamic pricing strategy is proposed to maximise the utilities of the seller and the buyers.

[1] *i.e.*, benefit to all the data owners

Quality driven pricing strategies: Though data quantity intuitively plays a vital role in the data marketplace, it is not sufficient for an elaborate pricing strategy to only consider one feature. Data quality is defined variously in different applications. In IoT applications, the response time of data collection is vital. To reflect the freshness of the data, an auction-based profit-driven data acquisition in the MCS, namely VENUS [7], was proposed. Data quality can also be verified by machine learning algorithms. [8] defined data quality based on the size, completeness, types, and combinations of data in the dataset, while also incorporating the data owners' willingness to sell.

Privacy driven pricing strategies: The value of privacy can be defined as the difference between the original data and the data after the ε-differential privacy operation, as defined in [9]. This metric considers both the privacy leakage (negative) and the network effect (positive) in the MCS; it also proposes a reverse "privacy" auction, which implicitly provides a solution for the privacy pricing strategy. However, it is not very easy to define the value of private data. For example, the value may vary from different use cases, timing, and cognition. Therefore, it remains an open question.

Learning performance-driven pricing strategies: With machine learning development, the role of data sets has drawn more and more attention since it is essential for algorithmic training [12]. Yoon et al. [13] propose a data valuation scheme aided by reinforcement learning. This work studied the value of datasets by proposing joint learning of the predictor and the value/weights of the data elements in the data set. This paper aims to provide a guideline for future data collection. Yu et al. [11] introduce a data pricing strategy in federated learning to enable fairness awareness among data owners.

3. Rethinking the existing market

In the previous works, data and digital product marketplaces are usually equivalent to traditional marketplaces, i.e., monopoly, duopoly, oligopoly, and a competitive market. However, the rigid market structures do not fit the current world, where heterogeneous individuals/devices are constantly generating data distributively. In this paper, the proposed federated data acquisition market distributively collects data from the local aggregators. Local data aggregators represent the local data owners whose data is limited due to geographical/logical partition. Local data aggregators can be seen as isolated data islands, which compels the data consumer to purchase data sets from different local data aggregators to achieve a better outcome. This vertical market structure and pricing strategy can ultimately add resiliency in digital product markets and empower data owners to gain benefits and control their contributions.

There exist a few works related to mean-field-based data pricing strategies. Wang et al. [14] proposed a dynamic pricing strategy to ensure the freshness of the information. Deep learning was adopted by [15] to solve the equilibrium of a double auction market and was modelled by a mean-field game. Different from the previous works, this paper elevates the status of the data owners in the data marketplace by providing a natural habitat, namely the proposed federated data acquisition framework, which enables cooperation among the data owners to maximise social welfare. Compared to the related works that assume the IoT devices contribute data willingly, the proposed framework considers privacy-concerning individuals as data owners who naturally care about privacy and exchange opinions with their social connections. The main advantages of the proposed pricing strategy include: first, it enables a customised price plan for every data owner; second, a mean-field based pricing strategy reveals unit pricing evolution when data owners cooperate with each other, which can direct the data consumers towards a better individual reward in the future. Then, cooperation enables better social welfare among the data owners. Fourth, the MFG-based pricing strategy does not require the details of all the information exchanges among the data owners, which further preserves the privacy of the data owners. Last but not least, the algorithm is fairly lightweight and thus suitable for a real-time scenario.

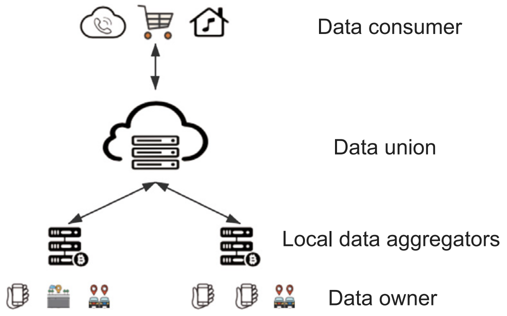


Fig. 1. The architecture of federated data acquisition framework.

4. Federated data acquisition market

In this section, we introduce the architecture and the workflow of the proposed market as depicted in Fig. 1. Additionally, we propose new use cases empowered by the new marketplace.

4.1. Structure and entities

Data Consumer: According to historical data, a data consumer is aware of the required key features of the data sets [13]. The data consumer communicates with the Data Union to acquire data sets with key features. It demands data sets from different data owners to support the MCS tasks with individual price schemes. Data sets need to be diverse to ensure the best learning outcome.

Data Union: The Data Union acts as a broker in the proposed market. It is a trusted third entity that assists in the negotiation and trading between the data consumers and the LDAs. The Data Union provides first verification of the identity and legitimacy of the data consumers before they launch the data trading requirements to the data owners. This mitigates the negative impact of malicious data consumers who may try to obtain the data without paying by requesting a deposit. The Data Union at the same time monitors the price dynamics of the LDAs and the data consumer's requirements. The statistical information is useful in the future trading process. Additionally, the LDAs only need to reveal the statistics instead of the detailed data information to protect privacy.

LDAs with data owners: We introduce LDAs to enable a feasible implementation of the proposed market, since a fully distributed negotiation with the data owners and consumers is not trivial. LDAs can collect, sort, and anonymise data from the data owners. Additionally, LDAs represent data owners and interact with the Data Union. The concept of LDAs can also assist the modelling process, which allows a natural division of a large number of data owners. Note that LDAs and the data owners can be both geographical and virtual: geographical LDAs (such as mobile base stations) classify the data owner according to a physical location, i.e., community, campus; virtual LDAs (such as mobile carriers and network service provider) host the data from their subscribers who are in geographically diverse locations. Though the Data Union and LADs provide essential functions and enable the proposed market, in this paper, we focus on modelling the interaction between the data consumers and owners.

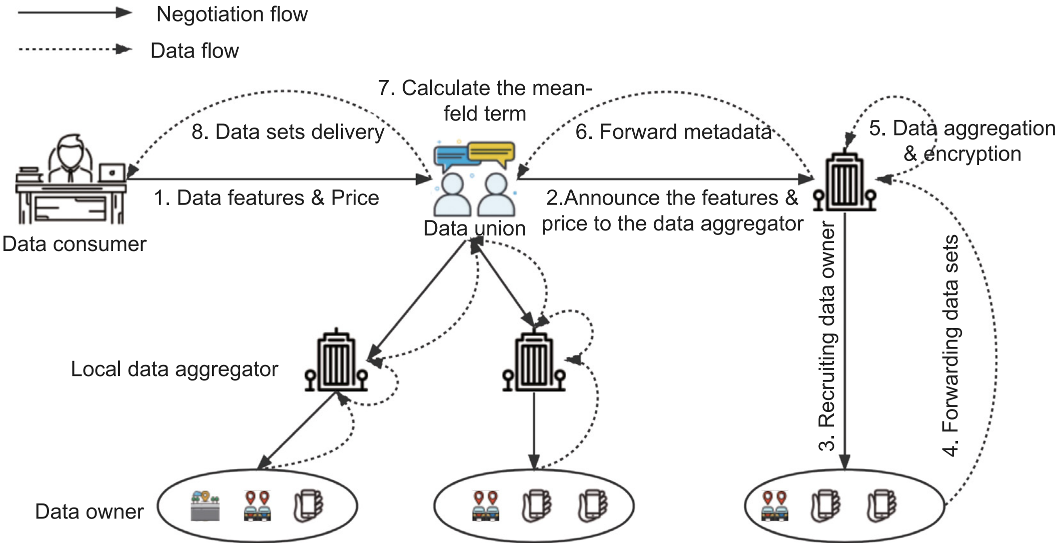


Fig. 2. Workflow of the federated data acquisition market: 1. Data consumers request the data features and offer individual unit prices; 2. Data Union receives the request and announces it to the LDAs; 3. LDAs start to recruit data owners in their domain; 4. Data owners collect and forward the data sets after gossipping and evaluating the given price; 5. LDAs calculate the distribution of individual unit prices; 6. forward the information to the Data Union; 7. Data Union checks the statistics and negotiates with the data consumer; 8. Data Union and data consumer reach an agreement, and data is delivered.

4.2. Workflow of the federated data acquisition market

In Fig. 2, we illustrate the workflow of the proposed market: first, the data consumers propose the request to the Data Union; then, the Data Union announces the request to the data owners through the corresponding LDAs; third, LDAs recruit data owners and gather the data sets. They sort the raw data sets according to the features and then send the metadata, i.e., the features and their quantity, to the Data Union. Additionally, LDAs compute the distribution of price and coordinate with the Data Union. Last but not least, the Data Union gathers data from the LDAs and saves the statistics for future negotiation with the data consumers.

4.3. New market empowering new use cases

The new market has the potential to regulate data trading, which also empowers new use case scenarios. For example, 2020 brought new data collection and workflows in the context of the Covid-19 pandemic. To evaluate the policy before deploying it in real life, a *digital city twin* [16] is proposed by simulating cities individually according to local data. Machine learning shows that the digital city twin desires joint data from different individuals and cities to evaluate various policies. The proposed data market can assist the policy-making procedure on a large scale and help to form a global treaty of crisis containment. Furthermore, the proposed federated data acquisition market follows and will strengthen, the General Data Protection Regulation (GDPR). The data union can serve as a proxy of GDPR in Europe, which can also deploy a series of branch local data aggregators. This approach scales up the federated data acquisition market to fulfil the global data market scenario. More importantly, the proposed data structure and the pricing strategy allow technology focuses on human behaviours, such as cooperation, homophily, and information exchanges (see Table 1).

5. Data pricing strategy: problem formulation

In this section, we first introduce the pricing strategy of one data owner interacting with its neighbours within one local data aggregator. Then, we generalise the problem into multiple LDAs interacting with each other. Last, we solve the pricing strategy for the general case using mean-field game theory. Note that we assume that all the data owners are distributed randomly and that they can interact freely with their contacts.

Table 1
Notation and descriptions.

Description	Parameter
Data owners	$i \in \mathcal{V} = \{1, \ldots, N\}$
Neighbours of i	N_i
Mean-field term	m_i
Cost function	L_i
Data quantity of i	q_i
Weights of price dynamics	β_1, β_2
Brownian factors	σ, W_i
Value function	v_i
Unit price of i	p_i
Weights of cost	$\alpha_1, \alpha_2, \alpha_3$
Drift term	f_i
Aggregated mean-field term	m

5.1. Within one local data aggregator

According to the architecture of the proposed market, there exists the interaction, namely information exchange, of the data owners. Human nature is such that individuals tend to influence each other by communication; the influence of one individual to another is essentially information exchange. We consider data owner $i \in \mathcal{V} = \{1, \ldots, N\}$ in a LDA. The data owner decides the quantity of data, q_i, to sell based on the unit price, p_i, offered by the data consumer. The data price fluctuates during the gossip, i.e., when a data owner knows its neighbours are offered higher prices, it will lower the quantity of data to sell. We define data owner i's neighbours at time t as $N_i(t)$. The notations and decriptions are in Table 1. The interaction of i and its neighbours is time-dependent, which leads to the dynamic of the unit price. Hence, we have

$$dp_i(t) = [\beta_1 N_i(t) - p_i(t) + \beta_2 q_i(t)]dt + \sigma dW(t) \tag{1}$$

where β_1, β_2 are positive weights; σ and W_i use Brownian movement [17] to model the randomness of the price fluctuation. As shown in Eq. (1), the price dynamics not only relate to the neighbours' prices but also the quantity of personal data, which indicates the concern of privacy loss. The data owner aims to minimise the cost during the trade by choosing an optimal quantity p_i during the time in the data market. The cost includes the loss of privacy, information exchange cost with the neighbours, and the gain from trading the data. Hence, we have the cost function of data owner i as

$$Li(q,p,N_i,t)=\frac{1}{2}\alpha_1 q_i^2(t)-\alpha_2 p_i(t)q_i(t)+\alpha_3(p_i(t)-N_i(t))^2 \tag{2}$$

where α_1, α_2, and α_3 are positive weights. The first term of the above reflects the cost of sharing data and is quadratic such that when the cost is used later the law of diminishing returns applies [18]. The second term is the income from selling data. The third term is the information exchange cost, i.e., when the offered price is identical to the neighbours, there is no information exchange cost. All the variables are time-dependent, which indicates that the data consumers make individual offers to each data owner during a period of time to obtain rich data. The accumulated cost of data owner i trading with a certain data consumer during a period of time is defined as

$$J_i=\int_0^T[\frac{1}{2}\alpha_1 q_i^2(t)-\alpha_2 p_i(t)q_i(t)+\alpha_3(p_i(t)-N_i(t))^2]dt \tag{3}$$

Data owner i's cost is constrained by the price fluctuation caused by information exchange defined in Eq. (1). Hence, we define the minimisation problem of data owner i in its own LDA.

$$\min_{q_i} J_i=\int_0^T[\frac{1}{2}\alpha_1 q_i^2(t)-\alpha_2 p_i(t)q_i(t)+\alpha_3(p_i(t)-N_i(t))^2]dt \tag{4}$$
$$\text{s.t.}\quad dp_i(t)=[\beta_1 N_i(t)-p_i(t)+\beta_2 q_i(t)]dt+\sigma dW(t)$$

The problem in (4) is a PDE-constrained minimisation problem, which aims to find the optimal data quantity for the data owner i. We can solve this using a Lagrangian function and Karush–Kuhn–Tucker (KKT) conditions [19]. We will introduce the solution of the general case in the next section.

5.2. Multiple local data aggregators

Since there are multiple LDAs with multiple data owners in it, it is not trivial to consider the interaction between each data owner like the traditional game theory [20]. Therefore, to capture the states of the neighbours, we introduce the mean-field term, m_i, of data aggregator i, namely the probability density of the price. Mean-field term was first introduced in physics to represent the behaviour of systems of large numbers of particles when the number of particles approaches infinity. Considering the neighbour's neighbours in the proposed market, the interaction can be captured by the mean-field term. All the data owners aim to minimise the cost. Hence, the data owner i can also represent the LDA i. The expected cost function of LDA i is defined as

$$J_i=\int_0^T m_i[\frac{1}{2}\alpha_1 q_i^2(t)-\alpha_2 p_i(t)q_i(t)+\alpha_3(p_i(t)-m)^2]dt \tag{5}$$

where m is the aggregated mean-field term, which is defined as $m=\sum_{j\in\mathcal{M}_i} m_j(t)$, where $\mathcal{M}_i$ is a set of i's neighbours. Note the first m_i indicates the expected cost of the LDA i. We can also refine the price dynamic

$$dp_i(t)=[\beta_1 m-p_i(t)+\beta_2 q_i(t)]dt+\sigma_i dW_i(t)$$
$$dp_i(t)=f_i dt+\sigma_i dW_i(t) \tag{6}$$

where f_i is the drift term, which is the influence of the other LDAs on LDA i during the gossip. The price state p_i of LDA i is affected by the drift term and the randomness. To model the evolution of price states of multiple LDAs, we introduce the *Fokker–Planck–Kolmogorov* (FPK) equation from statistical mechanics, which describes the evolution of the particle velocity probability density under the influence of forces [21]. In our case, the FPK function of the LDA's price probability density dynamics over time is related to the combined effect of the expected dynamic change rate (defined by drift term) and the randomness, which is shown in the equation below:

$$\frac{\partial m_i}{\partial t}=-div(f_i m_i)+\frac{\sigma^2}{2}\frac{\partial {m_i}^2}{\partial p_i^2} \tag{7}$$

where $m_i(p_i,t)$ is a function of the price state and time; and $div(f_i m_i)$ represents the divergence of the probability density of all of the other prices with respect to the current price probability density m_i and drift term. In Eq. (7), the divergence term can be derived as $div(f_i m_i)=\frac{\partial f_i}{\partial p_i}m_i+f_i\frac{\partial m_i}{\partial p_i}$. We propose the LDA i's minimisation problem with the constraint of the dynamic of the probability density of the price state

$$\min_{q_i} J_i=\int_0^T m_i[\frac{1}{2}\alpha_1 q_i^2-\alpha_2 p_i q_i+\alpha_3(p_i-m)^2]dt \tag{8}$$
$$\text{s.t.}\quad \frac{\partial m_i}{\partial t}=-div(f_i m_i)+\frac{\sigma^2}{2}\frac{\partial {m_i}^2}{\partial p_i^2}$$

The problem in (8) is a PDE-constrained optimisation problem. Hence, we write the Lagrangian function

$$\mathcal{L}_i=\int_{p\in P}\int_0^T m_i[\frac{1}{2}\alpha_1 q_i^2-\alpha_2 p_i q_i+\alpha_3(p_i-m)^2]dtdp$$
$$+\int_{p\in P}\int_0^T v_i\left(\frac{\partial m_i}{\partial t}+div(f_i m_i)-\frac{\sigma_i^2}{2}\frac{\partial {m_i}^2}{\partial p^2}\right)dtdp \tag{9}$$

where v_i is value function $v_i(p_i,t)=\inf_{q\in Q} J_i$, which is the value of the cost attained by the optimal data quantity. To obtain the optimal quantity q_i, it needs to satisfy the KKT conditions

$$\frac{\partial\mathcal{L}_i}{\partial q_i}=0,\quad \frac{\partial\mathcal{L}_i}{\partial m_i}=0,\quad \frac{\partial\mathcal{L}_i}{\partial v_i}=0 \tag{10}$$

First, we solve $\frac{\partial\mathcal{L}_i}{\partial m_i}=0$

$$\frac{\partial\mathcal{L}_i}{\partial m_i}=L_i+m_i\frac{\partial L_i}{\partial m_i}+\frac{\partial v_i}{\partial m_i}\left(\frac{\partial m_i}{\partial t}+\frac{\partial f_i}{\partial p_i}m_i+f_i\frac{\partial m_i}{\partial p_i}-\frac{\sigma_i^2}{2}\frac{\partial^2 m_i}{\partial {p_i}^2}\right)$$
$$\frac{\partial\mathcal{L}_i}{\partial m_i}=\frac{\partial v_i}{\partial t}+L_i+m_i+\frac{\partial L_i}{\partial m_i}+\frac{\partial v_i}{\partial p_i}\frac{\partial f_i}{\partial m_i}m_i+\frac{\partial v_i}{\partial p_i}f_i-\frac{\sigma_i^2}{2}\frac{\partial^2 v_i}{\partial {p_i}^2}$$
$$-\frac{\partial v_i}{\partial t}=L_i+m_i+\frac{\partial L_i}{\partial m_i}+\frac{\partial v_i}{\partial p_i}\frac{\partial f_i}{\partial m_i}m_i+\frac{\partial v_i}{\partial p_i}f_i-\frac{\sigma_i^2}{2}\frac{\partial^2 v_i}{\partial {p_i}^2} \tag{11}$$

We note that Eq. (11) is effectively the HJB equation [22], which is used to solve the optimal control with respect to the cost function in the control theory. Then, we solve $\frac{\partial\mathcal{L}_i}{\partial v_i}=0$

$$\frac{\partial m_i}{\partial t}+\frac{\partial(f_i m_i)}{\partial p_i}-\frac{\sigma_i^2}{2}\frac{\partial {m_i}^2}{\partial p_i^2}=0 \tag{12}$$

Finally, we solve the optimal data quantity, following $\frac{\partial\mathcal{L}_i}{\partial q_i}=0$

$$\frac{\partial\mathcal{L}_i}{\partial q_i}=m_i(\alpha_1 q_i-\alpha_2)+\beta_2\frac{\partial v_i}{\partial p_i}m_i=0$$
$$q_i^*=(\alpha_2 p_i-\beta_2\frac{\partial v_i}{\partial p_i})/\alpha_1 \tag{13}$$

We can use Eq. (11), (12), and (13) jointly to solve the optimal data quantity and the price evolution among multiple LDAs. A numerical method is adopted to obtain the final solution.

Algorithm 1: Solving the mean-field game

```
Input : $M_i^0, Q_i^0, V_i^T$
while $iter \leq I$ or $Err \geq \epsilon$ do
    for $i = 0, \dots, N$ do
        for $t = 0, \dots, T$ do
            if $U_i^t = 0$ then
                $M_i^{t+1} = M_i^t$
            Solve mean field term using Eq. (16)
    for $i = 0, \dots, N$ do
        for $t = T, \dots, 0$ do
            Solve adjoint variable using Eq. (15)
    for $i = 0, \dots, N$ do
        for $t = 0, \dots, T$ do
            Update control using $Q_i^t \leftarrow \alpha Q_i^t + (1-\alpha)\frac{\partial L_i}{\partial u_i}$
    Compute $err$
for $i = 0, \dots, N$ do
    for $t = 0, \dots, T$ do
        Update the optimal data quantity using Eq. (17)
```

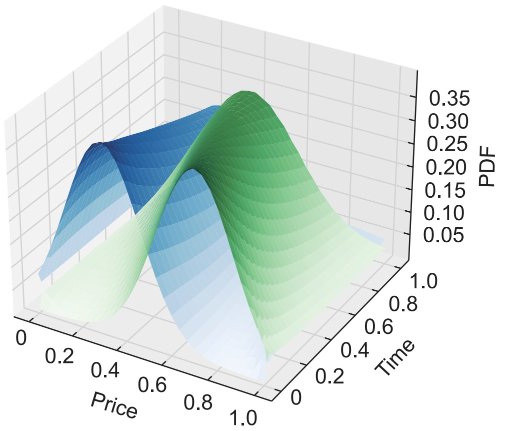


Fig. 3. A full picture of the price dynamic from $t = 0$ to T: the blue price dynamic is LDA 1 with $\mu_1 = 0.3$ and the green price dynamic is LDA 2 with $\mu_2 = 0.7$.

6. Solution of the MFG pricing strategy

As observed in Eq. (11) the HJB consists of partial differential derivations of the value function v_i with respect to time t and price p_i. Hence, we adopt the *finite difference method* (FDM) [23] to solve the above partial differential equations. The key idea of FDM is to *"replace derivatives in a differential equation with approximations"* [24]. This method can be described as the "Forward in Time, Centred in Space" (FTCS) scheme. By this method, the derivative can be replaced with

$$\frac{\partial F(x,t)}{\partial t} = \frac{F_i^{t+1} - \frac{1}{2}(F_{i+1}^t + F_{i-1}^t)}{\Delta t}$$
$$\frac{\partial F(x,t)}{\partial x} = \frac{F_{i+1}^t - F_{i-1}^t}{2\Delta x}$$
$$\frac{\partial^2 F(x,t)}{\partial x^2} = \frac{F_{i+1}^t - 2F_i^t + F_{i-1}^t}{4\Delta x^2} \tag{14}$$

where $F(x,t)$ is a function of variable x and t (such as the mean-field term m_i and the value function v_i), Δx is the step size, i is the number of the step.

To solve (11), (12), and (13), we first discretise the time interval $[0, T]$ and price space $[0, 1]$. We define X and Y as the step size of time and state, respectively. The step sizes of time and state space are denoted as $\Delta t = \frac{T}{X}$ and $\Delta p = \frac{1}{Y}$, respectively. Before applying the FTCS scheme, we define the discrete mean-field term, value function, and data quantity as M_i^t, V_i^t, and Q_i^t, respectively, where i is the ith price state and t is the tth time step. We obtain the value function v_i and the mean-field term m_i

$$V_i^{t-1} = \frac{1}{2}(V_{i+1}^t + V_{i-1}^t) + \Delta t A + \frac{\Delta t}{2\Delta p}(V_{i+1}^t - V_{i-1}^t)B - \frac{\Delta t \sigma^2}{8\Delta p^2}(V_{i+2}^t - V_i^t + V_{i-1}^t) \tag{15}$$

$$M_i^{t+1} = \frac{1}{2}(M_{i+1}^t + M_{i-1}^t) - \frac{\Delta t}{2\Delta p}(f_{i+1}^t M_{i+1}^t - f_{i-1}^t M_{i-1}^t) - \frac{\Delta t \sigma^2}{8\Delta p^2}(M_{i+2}^t - M_i^t + M_{i-1}^t) \tag{16}$$

$$Q_i^t = (\alpha_2 p_i - \beta_2 \frac{(V_{i+1}^t - V_{i-1}^t)}{2\Delta p})/\alpha_1 \tag{17}$$

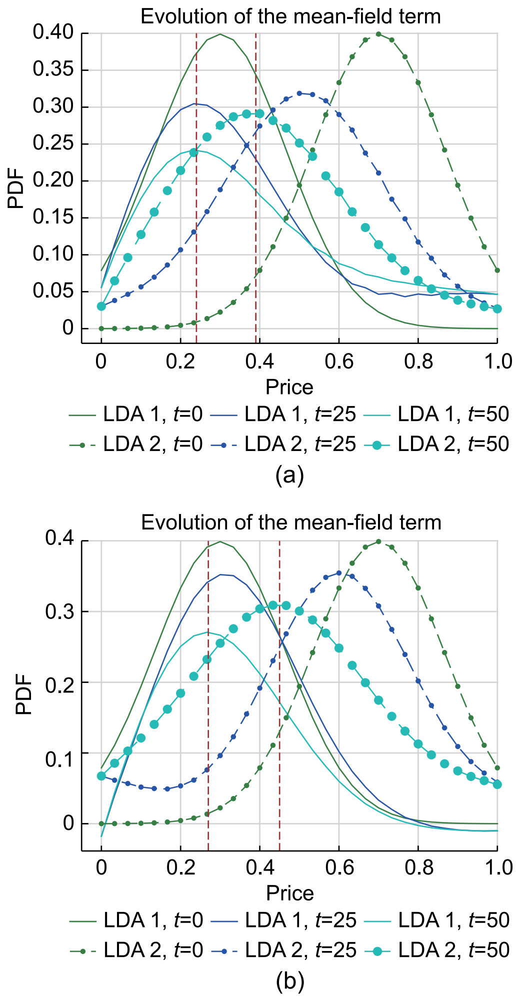
where $A = L_i + m_i + \frac{\partial L_i}{\partial m_i}$ and $B = \beta_1 m_i + f_i$. (15) and (16) are solved by backward iteration and forward iteration, respectively. We then initiate the discrete mean-field term at time 0 as M_i^0, the value function at time T, V_i^T, and the data quantity at time 0, Q_i^T, for all the price states. The joint solution of the pricing strategy is shown in Algorithm 1, where error $\epsilon = 0.01$. We update the data quantity according to the equation in Algorithm 1 to obtain the mean-field term and the value function, where α is the learning rate. The iteration runs until the error reaches the satisfaction or the maximum iteration limitation. We obtain the value function V_i^t and the mean-field term M_i^t at all the states and time steps at the end of the iteration. Then, the optimal data quantity can be calculated using Eq. (17).

7. Simulation

In this section, we evaluate the proposed data pricing strategy in the federated data acquisition market. We assume that there are two LDAs interacting with each other, and the data owners within each LDA can freely interact with data owners in the same LDA and the other LDA. The two LDAs follow the normal distribution with means 0.3 and 0.7. We assume the LDAs have the same diffusion parameter $\sigma = 0.002$ to simplify the problem. The choice of the time steps and the state steps follow the Courant–Friedrichs–Lewy (CFL) condition [25] to ensure convergence.

We first investigate the price dynamic when two LDAs interact with each other, as shown in Fig. 3. The blue and green surfaces represent LDA 1 and LDA 2, respectively. At the beginning of the interaction, the mean of LDA 1's price distribution is $\mu_1 = 0.3$, and the mean of LDA 2's price distribution is $\mu_2 = 0.7$. While the interaction progresses, the data owners exchange information with each other, which forms cooperation. At the end of the interaction, it demonstrates that the final price distributions of both LDA 1 and 2 differ strongly from the beginning due to the cooperation. In addition, the price distribution reaches equilibrium. This indicates that the LDAs can reach an agreement under cooperation. In the current setting, the optimal price distribution is with a mean of about 0.3.

We then evaluate the pricing dynamic over time and influence factor β_1 in Fig. 4. The influence factor β_1 in Eq. (1) indicates how influential the data owners neighbours are, i.e., when interacting with the influential neighbour LDA (i.e., bigger β_1), the price distribution evolves towards to the influential LDA's distribution. In Fig. 4, we illustrate how LDA 2 with different influence factors β_1 affects the pricing distribution of LDA 1 over time. We sampled $t = 0, 25, 50$ from the beginning $t = 0$ till the end of the



(a)

(b)

Fig. 4. Price evolution with respect to different influence factors β_1: how different influence factors of LDA 2 affect LDA 1's price dynamics. (a) LDA 2 low influence factor β_1 = 0.1. (b) LDA 2 high influence factor β_1 = 1.

interaction $t = T$ for both LDAs. With $\beta_1 = 0.1$ in Fig. 4(i), LDA 1 and 2 move to the optimal price distribution as shown in Fig. 3. With $\beta_1 = 1$ in Fig. 4(ii), first LDA 1 and 2 both show resistance to reach optimal price distribution; second, LDA 1 shows a tendency to move to LDA 2's price distribution. This is because, with a higher influential factor of LDA 2, LDA 1 values the information from LDA 2 more and tends to act the same.

In Fig. 5, we then demonstrate the optimal data quantity with respect to the data owners' privacy sensitivity. α_1 in the cost function is the privacy cost weight, i.e., higher α_1 leads to a higher cost of privacy, which also represents the data owner's privacy sensitivity. We sample the optimal data quantity with respect to price every 10 time steps from $t = 0$ to $t = 75$. It first shows that the optimal data quantities of both LDAs reach equilibrium due to cooperation, as shown in the bold blue lines. Then, the optimal data quantity in Fig. 5(i) is higher than it is in Fig. 5(ii) with respect to unit price. This is due to the sensitivity of privacy: higher sensitivity leads to higher cost to sell the data. We further consider the impact of the neighbour LDA with $\beta_1 = 1.5$ in Fig. 5(iii). Compared to Fig. 5(i) and (ii), the data quantity fluctuates for a longer period to settle down due to the influential LDA. However, the optimal data quantity follows the same pattern as in previous cases. In Fig. 5, we examine the internal factor (i.e., privacy sensitivity) and external factor (i.e., influential LDAs) and demonstrate that the internal factors drive the final data quantity for trading and the external factors affect the duration of reaching stability.

We also compare the social welfare of the proposed pricing strategy in Fig. 6. Social welfare is defined by using the aggregated cost of data owner in LDAs. Additionally, we defined natural cooperation where data owners from different LDAs interact with each other freely, "no cooperation" (NC) where the gossip does not exist. Note that cost is negative when the price reaches a certain level. This means there is profit for the data owners to sell the data. It demonstrates that cooperation leads to a higher cost compared to the 'no cooperation' scenario, which is due to the extra cost of information exchange. However, the cooperation leads to higher social welfare for the data owners, which proves to be a leverage when negotiating with the data consumers. The proposed pricing strategy leads to higher social welfare in the higher-price states.

Last but not least, in Fig. 7, we show that the proposed pricing strategy converges under an acceptable number of iterations when the error is set to 0.01 in the current setting. However, in practice, the number of iterations to reach satisfactory convergence highly depends on the parameters of the finite difference method. Consequently, when implementing the technique it is necessary to monitor the convergence rather than simply relying upon a fixed number of iterations.

8. Challenges in the future data market

In this paper, we propose a novel federated data acquisition system with local data aggregators and a data union presenting the data owners to negotiate with the data consumer. Despite the advantages of the proposed data pricing strategy, we still face challenges in future practice.

Sensible decision making of human beings: Though this work introduces agents, i.e., data unions, to represent data owners, the ideal solution still remains to be solved. The agents usually charge markups to serve as a broker. Data owners should be able to decide the value of the data and trade directly with the data consumer in a distributed fashion. This brings a challenge to data owners to make a sensible decision regarding the data value [26]. Indeed, multiple factors affect the data value as it can vary for example: according to different applications; usage of the data consumer; or even different application preferences of the data owner. Consequently, data owners may need nudges or cues to understand the data value as it is hard for human beings to evaluate the true value of the data and trade with the data consumers.

Regulation and law: With the establishment of the General Data Protection Regulation (GDPR) in Europe in 2016, data privacy has been widely discussed. Data protection not only brings positive effects, such as clear rights for the data subject but also leads to some drawbacks [27]. For example, it will be harder for high-tech companies to provide customised services. The government is not able to have a macro view of society during resource allocation, such as vaccine distribution. The flexibility of data trading is still required in the regulation.

New market structure: As we mentioned, the structures of the data market are limited. One of the promising directions is a fully distributed market structure aided by distributed ledger technology (DLT) [28]. It allows peer-to-peer interactions between the data owner and the data consumer. More importantly, with the increase of smart cities and IoT applications, microtransactions can also be implemented via DLT. Additionally, an incentive mechanism is desirable in balancing the power [29] and creating a healthy marketplace for information systems.

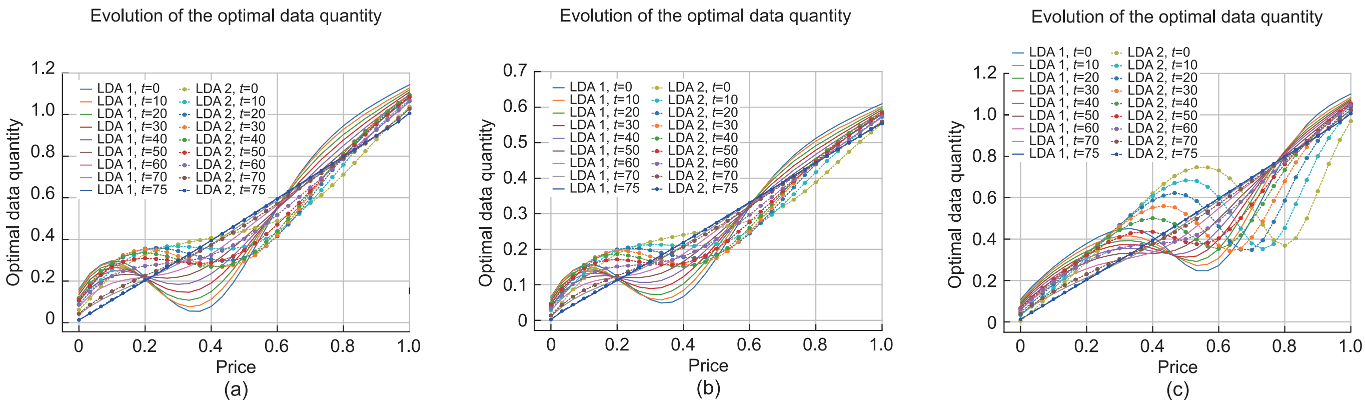


Fig. 5. Optimal data quantity evolution with respect to privacy cost weight α_1: the solid lines and the dotted lines are the data quantity evolution of LDA 1 and LDA 2, respectively. The final optimal data quantity is shown in the bold blue lines. (a) LDAs with α_1 = 0.5 (setting: α_2 = 0.5, α_3 = 0.1, β_1 = 0.1, β_2 = 0.3.). (b) Privacy sensitive LDAs with α_1 = 0.9 (setting: α_2 = 0.5, α_3 = 0.1, β_1 =0.1, β_2 =0.3.). (c) LDAs with influential neighbours β_1 = 1.5 (setting: α_1 = 0.5, α_2 = 0.5, α_3 = 0.1, β_2 =0.3).

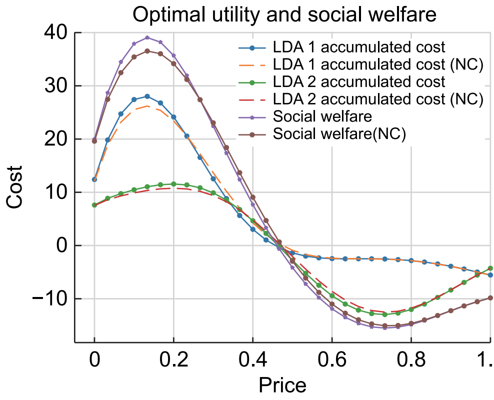


Fig. 6. Accumulated optimal cost of the populations and the social welfare (NC represents "no cooperation").

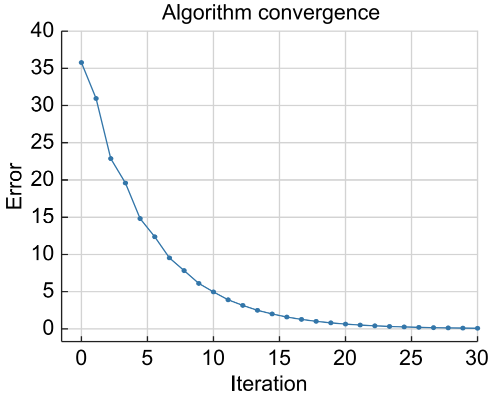


Fig. 7. Algorithm convergence with error set as 0.01.

9. Conclusion

This paper proposes a novel federated data acquisition framework with a data pricing strategy based on mean-field game theory. We first analyse the current data trading solution and then introduce the architecture of the federated data acquisition framework and its pricing strategy. For evaluation, we demonstrate the evolution of the price distributions of multiple local data aggregators. Additionally, the proposed pricing strategy enables data owners to achieve optimal social welfare with equilibrium in the data quantity that is traded. Last but not least, we have presented challenges in the future data market. For our future work, we will consider a pricing strategy with multiple data consumers and owners in the proposed federated data acquisition market.

CRediT authorship contribution statement

Jiejun Hu-Bolz: Conceptualization, Methodology, Formal analysis, Writing – original draft, Funding acquisition. **Martin Reed:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision. **Kai Zhang:** Methodology, Validation, Formal analysis. **Zelei Liu:** Writing – review & editing, Visualization. **Juncheng Hu:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported within the project TRACE-V2X, which has received funding from the European Union's HORIZON-MSCA-2022-SE-01-01 under grant agreement (101131204).

References

[1] J. Hu, K. Yang, K. Wang, K. Zhang, A blockchain-based reward mechanism for mobile crowdsensing, IEEE Trans. Comput. Soc. Syst. 7 (1) (2020) 178–191.

[2] Z. Xiong, Z. Cai, D. Takabi, W. Li, Privacy threat and defense for federated learning with non-i.i.d. Data in aIoT, IEEE Trans. Ind. Inform. 18 (2) (2022) 1310–1321, http://dx.doi.org/10.1109/TII.2021.3073925.

[3] J.-M. Lasry, P.-L. Lions, Mean field games, Japanese J. Math. 2 (1) (2007) 229–260.

[4] M.J. Osborne, et al., An introduction to game theory, vol. 3, (3) Oxford university press New York, 2004.

[5] D. Niyato, D.T. Hoang, N.C. Luong, P. Wang, D.I. Kim, Z. Han, Smart data pricing models for the internet of things: a bundling strategy approach, IEEE Netw. 30 (2) (2016) 18–25.

[6] Z. Xiong, D. Niyato, P. Wang, Z. Han, Y. Zhang, Dynamic pricing for revenue maximization in mobile social data market with network effects, IEEE Trans. Wireless Commun. 19 (3) (2020) 1722–1737.

[7] Z. Zheng, Y. Peng, F. Wu, S. Tang, G. Chen, Trading data in the crowd: Profit-driven data acquisition for mobile crowdsensing, IEEE J. Sel. Areas Commun. 35 (2) (2017) 486–501.

[8] H. Oh, S. Park, G.M. Lee, H. Heo, J.K. Choi, Personal data trading scheme for data brokers in IoT data marketplaces, IEEE Access 7 (2019) 40120–40132.

[9] M. Zhang, L. Yang, X. Gong, J. Zhang, Privacy-preserving crowdsensing: Privacy valuation, network effect, and profit maximization, in: 2016 IEEE Global Communications Conference, GLOBECOM, IEEE, 2016, pp. 1–6.

[10] H. Jin, L. Su, H. Xiao, K. Nahrstedt, Incentive mechanism for privacy-aware data aggregation in mobile crowd sensing systems, IEEE/ACM Trans. Netw. 26 (5) (2018) 2019–2032.

[11] H. Yu, Z. Liu, Y. Liu, T. Chen, M. Cong, X. Weng, D. Niyato, Q. Yang, A fairness-aware incentive scheme for federated learning, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 393–399.

[12] X. Xu, Z. Wu, C.-S. Foo, B.K.H. Low, Validation free and replication robust volume-based data valuation, Proc. NeurIPS (2021).

[13] J. Yoon, S. Arik, T. Pfister, Data valuation using reinforcement learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 10842–10851.

[14] X. Wang, L. Duan, Dynamic pricing and mean field analysis for controlling age of information, IEEE/ACM Trans. Netw. (2022) 1–13, http://dx.doi.org/10.1109/TNET.2022.3174114.

[15] B. Shi, Z. Song, J. Xu, Trading and pricing sensor data in competing edge servers with double auction markets, J. Sens. 2021 (2021).

[16] J. Pang, Y. Huang, Z. Xie, J. Li, Z. Cai, Collaborative city digital twin for the COVID-19 pandemic: A federated learning solution, Tsinghua Sci. Technol. 26 (5) (2021) 759–771, http://dx.doi.org/10.26599/TST.2021.9010026.

[17] M. Csörgő, Brownian motion—Wiener process, Canad. Math. Bull. 22 (3) (1979) 257–279.

[18] R.W. Shephard, R. Färe, The law of diminishing returns, in: Production Theory, Springer, 1974, pp. 287–318.

[19] D. Tabak, B.C. Kuo, Optimal control by mathematical programming, SRL Publishing Company, 1971.

[20] M. Reed, J. Hu, N. Thomos, M. Al-Naday, K. Yang, A dynamic service trading in a DLT-assisted industrial IoT marketplace, IEEE Trans. Netw. Serv. Manag. (2022).

[21] H. Risken, Fokker-planck equation, in: The Fokker-Planck Equation, Springer, 1996, pp. 63–95.

[22] S. Peng, A generalized dynamic programming principle and Hamilton-Jacobi-Bellman equation, Stochastics 38 (2) (1992) 119–134.

[23] J. Kiusalaas, Numerical methods in engineering with Python 3, Cambridge University Press, 2013.

[24] W.F. Ames, Numerical methods for partial differential equations, Academic Press, 2014.

[25] R. Courant, K. Friedrichs, H. Lewy, On the partial difference equations of mathematical physics, IBM J. Res. Dev. 11 (2) (1967) 215–234.

[26] A. Acquisti, L. Brandimarte, G. Loewenstein, Privacy and human behavior in the age of information, Science 347 (6221) (2015) 509–514.

[27] Z. Cai, X. Zheng, J. Wang, Z. He, Private data trading towards range counting queries in internet of things, IEEE Trans. Mob. Comput. 22 (8) (2023) 4881–4897, http://dx.doi.org/10.1109/TMC.2022.3164325.

[28] J. Hu, M. Reed, N. Thomos, M.F. Al-Naday, K. Yang, A dynamic service trading in a DLT-assisted industrial IoT marketplace, IEEE Trans. Netw. Serv. Manag. (2022).

[29] A. Pentland, A. Lipton, T. Hardjono, Building the new economy: Data as capital, MIT Press, 2021.