Share: Stackelberg-Nash based Data Markets

Yuran Bi*, Jinfei Liu†, Chen Zhao*, Junyi Zhao*, Kui Ren*, Li Xiong‡
*Zhejiang University, {stellabyr, zhaochen49, junyizhao, kuiren}@zju.edu.cn
†Zhejiang University, ZJU-Hangzhou Global Scientific and Technological Innovation Center, jinfeiliu@zju.edu.cn
‡Emory University, lxiong@emory.edu

***Abstract*—With the prevalence of data-driven intelligence, data markets with various data products are gaining considerable interest as a promising paradigm for commoditizing data and facilitating data flow. In this paper, we present Stackelberg-Nash based Data Markets (*Share*) to first realize a demand-driven incentivized data market with absolute pricing. We propose a three-stage Stackelberg-Nash game to model trading dynamics which not only optimizes the profits of all selfish participants but also adapts to the common *buyer-broker-sellers* market flow and solves the seller selection problem based on sellers' inner competition. We define Stackelberg-Nash Equilibrium and use backward induction to solve the equilibrium. For inner Nash equilibrium, we apply the conventional direct derivation approach and propose a novel mean-field based method along with provable approximation guarantees for complicated cases where direct derivation fails. Experiments on real datasets verify the effectiveness and efficiency of *Share*.**

I. Introduction

Data products (e.g., query services, aggregate statistics, and machine learning models) have paved the way for a variety of data-driven tasks in diverse industries. High-performance data products require a large amount of high-quality data. While there is a wealth of data generated from different sources, they are highly dispersed, which brings significant challenges to data aggregation. Besides, there is a gap between data supply and demand, and data suppliers or demanders usually lack the necessary resources and techniques to survey the vast data sources and turn data into data products. Thus, despite the increasingly available and enriched data, the wealth of data is far from being fully exploited. As one of the most important topics in Boston Database Meeting 2023 [1], data markets have been demonstrated as a promising paradigm to commoditize data and connect data suppliers and demanders [2], [3].

Motivations. A typical data (product) market consists of three parties: buyers, brokers, and sellers [4]–[6]. Buyers propose demands for data products and pay for them; brokers facilitate the transactions between buyers and sellers (and take charge of manufacturing data products from data); sellers offer data with different quality and sell data to brokers in exchange for compensation. We use two motivating examples to further specify our targeted settings.

Vehicle Example. An automaker (e.g., Ford Motors Company) wants to get insight into users' purchase preferences of vehicles to decide investments. Ford **(buyer)** turns to McKinsey, a consulting company, and proposes a series of queries $\mathbf{Q}$, each with distinctive conditions specified and the corresponding purchase intention to be answered, e.g., *whether a female customer in Texas would most likely buy fuel vehicles, pure electric vehicles, or hybrid vehicles*. To answer the queries, McKinsey **(broker)** needs to gather sales data of different vehicle types and produce a data product $\mathbf{M}$, e.g., *data statistics aggregating total sales of each vehicle type after screening location and gender*. McKinsey buys data from multiple vehicle retailers **(sellers)** who own sales data of various vehicles. Retailers sell data with different qualities responding to different compensations.

Health Example[1]. A biopharmaceutical company, e.g., Pfizer (who spends \$12 million to buy health data from a variety of sources including IMS Health as reported by Scientific America [7]), wants to get insight into the effects of their released COVID-19 vaccination for further development. Pfizer **(buyer)** turns to a healthcare consulting firm with a series of queries $\mathbf{Q}$ searching for the arising adverse reactions, e.g., *select the reported nausea within three months after the vaccination in America*. To answer the queries, the consulting firm **(broker)** needs to gather realistic health data to produce a data product $\mathbf{M}$, e.g., *data aggregation listing the symptomatic description of nausea with the time and location of vaccine inoculation filtered*. The consulting firm collects data from healthcare companies **(sellers)**, e.g., the aforementioned IMS Health which owns and sells de-identified prescription data, medical claims, and electronic medical records. While the health data is anonymized to comply with privacy laws, IMS Health may still suffer from risks of medical disputes and thus enhance privacy preservation using techniques such as perturbation which contributes to different quality of the provided health data.

Ford or Pfizer can benefit from the data product (accessing it via querying and answering) and in the meantime needs to pay for it; McKinsey or the healthcare consulting firm gains by selling the product (answering the queries) after spending resources to buy the data and produce the product; and the vehicle retailers or the healthcare companies sell data for compensations while suffering from costs (majorly the privacy loss incurred from data). They are driven to join the data market by the profit they can earn, and thus a general data market is considered where all three parties are *selfish*, i.e., have their own *revenue* and *cost*, and aim to maximize their

Jinfei Liu is the corresponding author.

[1] While data sharing in healthcare can bring societal benefits, many issues such as privacy and ethics need to be considered. Here we aim to show the motivation of *Share* supported by real cases, leaving the other issues outside the scope of the paper.

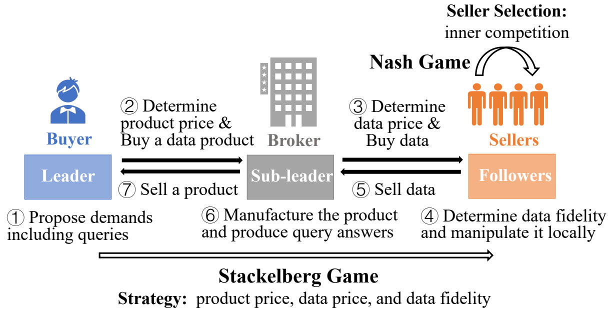

Seller Selection:
inner competition
Nash Game
Buyer
Broker
Sellers
② Determine product price & Buy a data product
③ Determine data price & Buy data
Leader
Sub-leader
Followers
⑦ Sell a product
⑤ Sell data
① Propose demands including queries
⑥ Manufacture the product and produce query answers
④ Determine data fidelity and manipulate it locally
Stackelberg Game
Strategy: product price, data price, and data fidelity

Fig. 1. The framework of *Share*.

profit (the difference between *revenue* and *cost*). Moreover, how they act in the market affects each other. If Ford sets a low price for the demanded product in pursuit of profit, McKinsey may pay little to buy the data to recover costs. Getting low compensations, the vehicle retailers offer poor-quality data, inducing a low-performance product and in turn, harming the profit of Ford. Therefore, **it is a critical research issue to design an incentive mechanism for data markets (especially data pricing) that can encourage three selfish yet interdependent parties to participate in data trading and thus invigorate the market.**

While all three parties need to maximize their profits, they play different roles in the market flow. Data sellers such as vehicle retailers likely do not regard data selling as the main business. Rather, data transactions are likely initiated by data buyers (demanders) such as Ford and Pfizer as in our motivating examples. For other examples, Perkins School for the Blind demands patient data to understand how to identify cerebral visual impairment [8], and logistics companies ask for foot-traffic data streams to forecast future inventory demand [4]. In such demand-driven scenarios, a single transaction serves one specific demand (e.g., personalized services) and thus buyers can be considered as orientating the market in turn (coming one at a time) as practiced in [4]. A data buyer proposes its demand to a data broker in the market, e.g., McKinsey, which then buys data from data sellers. While there may exist multiple competitive brokers, we consider one broker and put emphasis on its interactions with the other two parties. Concerning the limited data owned by one single seller, multiple data sellers (e.g., numerous vehicle retailers) are considered, each having a dataset that can together contribute to solving the buyer's demand. Therefore, to facilitate data trading in scenarios like the illustrated examples, **we focus on demand-driven data markets with one buyer, one broker, and multiple sellers.**

Many recent works [4]–[6], [9], [10] on data markets have emphasized various aspects of the market design, yet not targeted the demand-driven market with incentives for all three parties. Therefore, it is tempting to ask: **how to build a well-functioning data market with an incentivized pricing mechanism, which can satisfy the profit needs of all selfish participants and adapt to the demand-driven scenarios.**

Challenges and Contributions. We summarize three challenges $(\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3)$ faced in constructing the demand-driven data market with incentives for all three parties, and propose a feasible solution, **S**tackelberg-Nas**h** based D**a**ta Ma**r**k**e**ts (*Share*) utilizing game theory as in Fig. 1. The detailed workflow is presented in Section III-B.

Existing works on data markets vary in design goals and typically address one aspect or one party's need, such as product quality optimization for buyers [11], revenue maximization for sellers [5], social welfare maximization [12], or market protection from strategic participants [13], but fail to realize profit optimization for all parties. Pricing constitutes a key mechanism in incentive market construction, but the lack of market practices and pricing references makes data pricing far from trivial (similar to petroleum pricing at the early stage), especially an absolute price compared to the relative one which is determined in comparison with other data, e.g., by Shapley value [14] as in [4], [6], [10]. Therefore, the first challenge is $(\mathcal{C}_1)$: **How to design a pricing mechanism for data markets to realize absolute pricing and maximize the profits of all three parties.** To solve this challenge, we adopt game theory which can support the multi-objective incentive mechanism design in data markets. The interactions of the three entities are modeled as a game, in which each participant can achieve its profit-maximization goal by making its optimal strategy. Moreover, absolute prices of data are modeled as strategies and directly determined in the game process with the involvement of all the parties, which further encourages their participation.

Though efforts have been made to satisfy buyers' needs (e.g., utility demand and purchase budget in [4], [6]), no existing paradigm can well adapt to the demand-driven data market with the market flow in order, i.e., first demanded and initiated by the buyer, then translated and transmitted through the broker, and finally received and realized by the sellers. Therefore, the second challenge is $(\mathcal{C}_2)$: **How to encode the *buyer-broker-sellers* market flow to cater to the targeted demand-driven data market.** To solve this challenge, we formalize the interactions among three parties as a multi-stage dynamic game and adapt Stackelberg game [15] which can deal with the sequential order of participants, by regarding the buyer as the leader, the broker as the sub-leader, and sellers as followers. As shown in Fig. 1, the buyer first announces what data product it demands and determines the product price based on its profit-maximization goal; the broker then tries to buy data from sellers and decides the data price; each seller then chooses what data quality to provide.

Since there are multiple sellers, it is critical to select the *best* data (with the highest data quality) from the sellers to meet the buyer's product demand and in the meantime satisfy the broker's resource constraints, which is referred to as *seller selection problem*. Many existing works made the buyer [11] or broker [6], [16] responsible for seller selection, which not only requires the buyer/broker's capability of learning the data quality but also limits the sellers' ability to choose their provided data quality according to the data price. Hence, the third challenge is $(\mathcal{C}_3)$: **How to model the seller selection problem to select the best set of data for trading.** To solve this challenge, we consider the inner competition among sellers which can make the winners the selected sellers without the assistance of the buyer or broker. Sellers are allowed to manipulate their provided data quality by Local Differential Privacy (LDP) [17] to compete for the selling quantity of data. We model the inter-seller competition as a Nash game [18] because of its advantage in modeling sellers' equal positions, and find the desired Nash equilibrium by applying direct derivation and proposing a mean-field based approximation for complex cases when direct derivation fails.

Our goal is not to cover all data markets nor to address all critical issues in real-world data trading, but rather to propose an incentive mechanism for data markets anchored in a demand-driven scenario, which is meaningful in practice but has not been studied yet. Although borrowed from existing methodologies in game theory, *Share* attempts to contribute to data management research by innovatively adapting promising theories to data market construction with both 1) data-specific problem formulations considering data quality and privacy cost and 2) generally applicable equilibrium-solving solutions. The major contributions are summarized below.

- We present *Share*, an incentivized data market framework with an absolute pricing mechanism based on a three-stage Stackelberg-Nash game, which is the first to satisfy all-party profit maximization in demand-driven scenarios.
- We apply Nash game for the seller selection problem, which formulates sellers' inner competition and incorporates data selection into the three-party game process.
- We define Stackelberg-Nash Equilibrium in data markets and derive it by backward induction. To solve the inner Nash game, we apply direct derivation as well as design a novel mean-field method for complex cases, for which error analysis is presented.
- We conduct experiments on real and synthetic datasets to verify the effectiveness and efficiency of *Share*.

Organization. Section II provides the related work. Section III presents the data market framework based on Stackelberg-Nash game. Section IV constructs a market instance for which approaches to deriving the equilibrium and the trading dynamics are presented. Section V reports the experimental results while Section VI draws a conclusion.

II. Related Work

A. Data Market

Data markets trade data in direct (raw data [19], [20]) or indirect forms (derived data products, e.g., queries [9], [21] and models [4]–[6], [10]). While research on data market evolves in myriad directions including data mining [22]–[27], data storage [28], [29], and data security [30], [31], we explore the data market design emphasizing on the incentives by formalizing trading (pricing) mechanisms. Related research problems are reviewed below.

In terms of profit maximization for all parties, few studies can provide a thorough solution. [6] established a model marketplace with the needs of buyers and sellers considered, but assumed that the broker is neutral without its profit consideration and determines model prices only for single objective optimization, i.e., revenue maximization for sellers. [16] studied multiple objective optimization for crowdsensing data trading. Nevertheless, the specific characteristics of crowdsensing data (e.g., sensing time) limit the extension to the general data market. Diversified data products, typical privacy issues, and latent interrelations among participants should be considered. In *Share*, by combining multiple game mechanisms, we formulate for-all profit-maximization data markets with unrestricted data and data products, privacy consideration for data sellers, as well as inner competition modeling for seller selection.

As for data pricing, several surveys [32]–[34] claimed fundamental principles and reviewed the evolution of pricing models. In terms of absolute pricing, [6] provided absolute prices for data models, which, however, highly rely on the survey results and can't be adjusted dynamically. [4] applied Myerson's payment rule to determine absolute model payment but allocated relative compensations to sellers in proportion to their contributions based on Shapley value. While auction [35] can be a promising way for absolute price discovery and has been widely adopted in data pricing [12], [13], rarely can every party be included in the price determination. In *Share*, we propose a feasible absolute pricing mechanism for both data and data products with all three parties involved.

Many works looked at the seller (data) selection problem. One strand of research lied in data acquisition. For example, [11] dug into how a buyer purchases data under a budget to improve machine learning models, yet without addressing the market design issues including data pricing, revenue allocation, as well as the strategic actions of sellers and brokers. Within the data market design scope, [6] made brokers choose datasets to maximize Shapley coverage of the trained model. [16] used a combinatorial multi-armed bandit mechanism for brokers to select sellers. However, the selection results directly affect the profits of sellers, and therefore the seller selection problem is closely correlated to the profit maximization problem for sellers and should not be considered separately. Seller selection can be seen as the spontaneous process of the inner competition among sellers, proactively determined by their strategies rather than passively conducted by the buyer or broker. In *Share*, the seller selection problem is formalized as the inter-seller Nash game, which is a part of the incentive mechanism for profit optimization of all participants.

Data marketplaces have been practiced, e.g., Snowflake [36] and AWS [37], which involve products centrally listed with posted pricing, and focus more on scalability than incentives. Instead, *Share* provides a parameterized solution to guide autonomous data trading among profit-seeking participants by giving full play to their pricing power, which can act as an alternative framework catering to strong incentive needs.

B. Game Theory

(Non-cooperative) Game theory provides a tool for analyzing the interplay among individuals with conflicting objectives and has been widely used in various situations. Nash [18] accurately described Nash equilibrium as a solution concept for simultaneous-move games. Many researchers used Nash game as a powerful tool to formulate and solve problems with simultaneous interactions [38], [39]. Instead, Stackelberg game [15] features sequential actions, which was first used to formulate the determining process for oligopoly firms producing homogeneous products and has been further applied to many practical situations with hierarchical organizations, e.g., security game [40] and crowdsensing [41].

Since Nash proposed his theory, many researchers have sought algorithms for finding Nash equilibrium. [42] showed complexity results of deriving Nash equilibrium and [43] further studied the complexity of computing a mixed Nash equilibrium. In terms of solving Stackelberg game, backward induction approach, an iterative technique to derive dynamic game equilibrium, is often used [16], [41], [44]. In fact, deriving Stackelberg equilibrium with complete information can be formulated as a bilevel optimization problem [45].

In *Share*, we adopt Stackelberg game for the focused demand-driven data markets because it captures the sequential actions of participants and can thus adapt to the *buyer-broker-sellers* market flow while maintaining the profit maximization for all parties. Moreover, we first adopt Nash game for the seller selection problem since Nash game models the simultaneous-move interaction among equals and can be used for the inner competition among data sellers, which can select sellers based on their strategies.

III. MARKET FRAMEWORK: PARTICIPANTS, MECHANISM, AND EQUILIBRIUM

We crystallize the market participants by profit functions in Section III-A, formulate the market mechanism in Section III-B, and define the market equilibrium in Section III-C. For reference, Table I summarizes the frequently used notations. To set up, we clarify the assumptions and the problem below.

Assumptions. The key assumptions in *Share* are outlined below including the general settings of the market and the detailed roles of participants, which are motivated in Section I supported by practical examples.

- *Buyer-broker-sellers market.* A demand-driven data market is considered with one buyer, one broker, and multiple sellers acting in order.
- *Profit-seeking participants.* All the participants want to maximize their own profits.
- *Complete information.* The profit functions of participants are considered available as assumed in Stackelberg game and Nash game [15].

In terms of participants, the detailed assumptions concerning their roles are set as follows.

- **Buyer.** To fulfill a data-driven task, buyer $\mathcal{B}$ asks for a data product from broker $\mathcal{A}$. Buyer $\mathcal{B}$ gets access to the product via queries $\mathbf{Q}$ (either transactional or analytical) and claims its required product performance (the answer accuracy[2]), notated as ν. Note that the trading would fail if the demand of buyer $\mathcal{B}$ is not satisfied, e.g., the product cannot give answers for the asked 100 queries, or some answer has an accuracy lower than expected. Buyer $\mathcal{B}$ gains utility from the product (query answers) while giving the payment to broker $\mathcal{A}$.

TABLE I
FREQUENTLY USED NOTATIONS.

	Notation	Definition
Buyer $\mathcal{B}$	$\mathbf{Q}$	demanded query to be solved
	ν	demanded product performance
	p^M	basic price of data product
	θ_1, θ_2	parameters of concern on each attribute
	ρ_1, ρ_2	parameters of sensitivity to each attribute
	$\mathbf{U}(\cdot)$	utility function of the product
	$\mathbf{PB}(\cdot)$	payment function between buyer and broker
	$\Phi(\cdot)$	profit function of the buyer
Broker $\mathcal{A}$	N	total data quantity
	p^D	basic price of data
	σ_k	parameters of manufacturing cost
	$\mathbf{C}(\cdot)$	cost function of manufacturing data product
	$\Omega(\cdot)$	profit function of the broker
Seller $\mathcal{S}_i$	i	index of seller
	m	total number of sellers
	τ_i	data fidelity
	ϵ_i	parameter in local differential privacy
	χ_i	sold data quantity
	λ_i	parameter of privacy sensitivity
	$\mathbf{L}_i(\cdot)$	privacy loss function
	$\mathbf{PS}_i(\cdot)$	payment function between broker and seller
	$\Psi_i(\cdot)$	profit function of the seller
Data	D_i	seller $\mathcal{S}_i$'s raw dataset
	D_i^t	seller $\mathcal{S}_i$'s provided dataset
	D^t	whole dataset for manufacturing
	q_i^D	dataset quality provided by seller $\mathcal{S}_i$
	q^D	total quality of dataset for manufacturing
	q^M	data product quality
	ω_i	weight of seller $\mathcal{S}_i$'s dataset

- **Broker.** Broker (Arbiter) $\mathcal{A}$ wants to make profits by bridging the transactions between buyers and sellers. To answer queries $\mathbf{Q}$ with demand ν, broker $\mathcal{A}$ needs to buy N data records (limited by its computation resources) from sellers to make the data product (in any needed form from statistics aggregating to model training) which incurs costs (e.g., computing cost). Then, broker $\mathcal{A}$ sells the product to buyer $\mathcal{B}$ in exchange for payment.
- **Sellers.** A large number of sellers $\{\mathcal{S}_i | i = 1, 2, ..., m\}$ exist in the market. Each seller $\mathcal{S}_i$ owns dataset D_i and wants to sell it for profit. Seller $\mathcal{S}_i$ applies perturbation to its data utilizing a privacy scheme to manipulate the data quality locally and sells χ_i processed data records to broker $\mathcal{A}$. The data quantity χ_i, with $\sum_{i=1}^{m} \chi_i = N$, is to be decided by the market mechanism, and for any required number $\chi_i \in \mathbb{N}^+$, $|D_i| \geq \chi_i$, indicating that each seller has enough data for the trading. Seller $\mathcal{S}_i$ receives compensation from broker $\mathcal{A}$ while suffering from the (privacy) cost for the data it sells.

Problem Statement. The problem is to establish an incentive mechanism for data markets under the above assumptions. The *input* is the profit functions $\Phi(\cdot)$, $\Omega(\cdot)$, and $\Psi_i(\cdot)$ of buyer $\mathcal{B}$, broker $\mathcal{A}$, and each seller $\mathcal{S}_i$ while the *output* is a strategy profile jointly decided by participants, $\langle p^M, p^D, \boldsymbol{\tau} \rangle$

[2] We use *accuracy* in a broad sense which, independent of our model, can be measured in distinctive ways depending on the query task (e.g., the precision of the returned purchase intention on vehicles, or the completeness of the aggregated nausea cases).

that designates product price p^M, data price p^D, and data fidelity $\boldsymbol{\tau}$ for data trading, so as to achieve the *goal* that the profits of all participants are maximized at an equilibrium.

A. Market Participants

The profits $\Phi(\cdot)$, $\Omega(\cdot)$, and $\Psi_i(\cdot)$ are defined below including the function templates and desired properties, which can be instantiated in various forms. A market instance with every function instantiated will be constructed in Section IV and feasible approaches are proposed to solve its equilibrium.

1) Profit Function of Buyer: When buyer $\mathcal{B}$ comes to the market and asks for a data product, it cares about its *revenue*, the utility it can get from the product, and its *cost*, the payment it should give to the broker.

Revenue. The revenue of buyer $\mathcal{B}$ is the utility gained from the product. The performance of the product itself (embodied in the accuracy of query answers which is specified in demand ν) affects the utility. Moreover, the quality of data used to make the product contributes to the utility. While the answer accuracy only indicates how the product performs under a certain testing environment (related to specific validation datasets), dataset quality measures how good *raw materials* are, making the judgment of product utility more stable and less sensitive to various application scenarios. The dataset quality is measured as the total quality of datasets contributed by all sellers, $q^D = \sum_{i=1}^{m} q_i^D$, where q_i^D is the dataset quality seller $\mathcal{S}_i$ provides. Intuitively, the dataset quality is related to the intrinsic characteristic of data (e.g., the number of features) and the contribution of the data to the product which will be captured into the weights of sellers in Section IV-A. Besides, the dataset quality can be manipulated by sellers through the provided data fidelity τ_i and data quantity χ_i. The dataset quality is thus notated as $q_i^D = g(\chi_i, \tau_i)$ where $g(\cdot)$ is positively correlated with τ_i and χ_i and will be instantiated in Section IV-A. Data fidelity τ_i is determined by the perturbation added by seller $\mathcal{S}_i$, measured as the privacy level of LDP mechanism (see more in Section III-A3) while χ_i is determined by sellers' inner competition on $\boldsymbol{\tau}$ (see more in Section III-B). Combining both dataset quality and product performance, the gained utility is quantified by a function $\mathbf{U}(q^D, \nu)$ following the law of diminishing marginal utility [46] in economics, as instantiated in Section IV-A.

Cost. The *cost* of buyer $\mathcal{B}$ is the payment to broker $\mathcal{A}$. Based on the above analysis, we define $q^M = h(q^D, \nu)$ to objectively represent the quality of the data product which depends on both data quality q^D and product performance ν, and $h(\cdot)$ will be instantiated in Section IV-A. Also, p^M is defined as the basic price of q^M (the product price), and the payment for the product can be formulated as the function $\mathbf{PB}(p^M, q^M)$ positively correlated to the basic price and the product quality, which will be instantiated in Section IV-A.

Profit. The profit $\Phi(\cdot)$ of buyer $\mathcal{B}$ is the difference between the quantification of utility and the payment to broker $\mathcal{A}$.

$$\Phi\left(p^M, \boldsymbol{\tau}\right) = \mathbf{U}\left(q^D, \nu\right) - \mathbf{PB}(p^M, q^M). \tag{1}$$

2) Profit Function of Broker: When broker $\mathcal{A}$ receives the demand from buyer $\mathcal{B}$, it cares about its *revenue*, i.e., the payment from buyer $\mathcal{B}$, and its *cost* consisting of the compensations to sellers to buy the data and the manufacturing cost in the process of producing the data product.

Revenue. The *revenue* of broker $\mathcal{A}$ is the payment from buyer $\mathcal{B}$, i.e., $\mathbf{PB}(p^M, q^M)$ (the *cost* of buyer $\mathcal{B}$).

Cost. The *cost* of broker $\mathcal{A}$ is the sum of 1) the compensations to sellers and 2) the manufacturing cost. Broker $\mathcal{A}$ needs to pay each seller $\mathcal{S}_i$ compensation according to its provided data quality, which is formulated as function $\mathbf{PS}(p^D, q_i^D)$ and instantiated in Section IV-A. Here p^D describes the basic price of data (the data price) similar to the product price p^M. Broker $\mathcal{A}$ also needs to consume some resources to make the product. Different manufacturing consumption would be induced if processing data with different sizes or producing a product with different performance. Therefore, cost function $\mathbf{C}(N, \nu)$ is formulated related to total data size N and product performance ν and will be instantiated in Section IV-A.

Profit. The profit $\Omega(\cdot)$ of broker $\mathcal{A}$ is defined as the received payment from buyer $\mathcal{B}$ minus the compensations to sellers and the manufacturing cost as follows.

$$\Omega\left(p^M, p^D, \boldsymbol{\tau}\right) = \mathbf{PB}(p^M, q^M) - \sum_{i=1}^{m} \mathbf{PS}(p^D, q_i^D) - \mathbf{C}\left(N, \nu\right). \tag{2}$$

3) Profit Function of Seller: When seller $\mathcal{S}_i$ gets the purchase request for data from broker $\mathcal{A}$, it cares about its *revenue*, the compensation from broker $\mathcal{A}$ and its *cost* coming mostly from its privacy loss.

Revenue. The *revenue* of seller $\mathcal{S}_i$ is the compensation from broker $\mathcal{A}$, i.e., $\mathbf{PS}(p^D, q_i^D)$ (one part of the *cost* of broker $\mathcal{A}$).

Cost. The *cost* of seller $\mathcal{S}_i$ is mainly the privacy loss incurred based on data fidelity τ_i it provides (we ignore other costs of collecting, processing, and packaging data which can be formulated as a constant and integrated into the profit function). Data fidelity τ_i is determined and manipulated by seller $\mathcal{S}_i$ through LDP mechanism which provides a well-justified measurement tool to simultaneously capture noise level (fidelity) and privacy level (cost). Hence, τ_i is defined as $f\left(\epsilon_i\right)$ where ϵ_i represents the privacy level in standard LDP. We conclude the following characteristics $f\left(\cdot\right)$ should satisfy concerning the marginal trend and boundary conditions, and the instantiation will be shown in Section IV-A.

1. The data has fidelity $\tau_i = 0$ when $\epsilon_i = 0$ which means the data is random.
2. Larger ϵ_i, higher τ_i, since less noise is added to data.
3. τ_i increases slower as ϵ_i becomes larger because very little noise is being added and further decreasing noise does not make a significant difference to data fidelity anymore. On the other hand, when ϵ_i is very small, i.e., with extremely large noise, increasing ϵ_i can significantly increase data fidelity. Besides, τ_i cannot increase perpetually and should be upper bounded.

Bigger τ_i means better fidelity of data and more privacy loss for seller $\mathcal{S}_i$. We quantify such loss by function $\mathbf{L}_i(\cdot)$ which is positively related to τ_i. It's intuitive that the cost function should not only increase but also increase faster for higher τ_i, which corresponds to the principle of increasing marginal cost [46] in economics. Moreover, the privacy cost would increase as more data is sold (larger χ_i). Specific function $\mathbf{L}_i(\tau_i)$ will be elaborated in Section IV-A.

Profit. The profit $\Psi_i(\cdot)$ of seller $\mathcal{S}_i$ is the difference between the compensation and the quantification of privacy loss.

$$\Psi_i\left(p^D, \tau_i\right) = \mathbf{PS}(p^D, q_i^D) - \mathbf{L}_i\left(\tau_i\right). \tag{3}$$

B. Market Mechanism

In *Share*, the three entities take strategies in order. We first present the market workflow. Then we specify the strategies of buyer $\mathcal{B}$, broker $\mathcal{A}$, and each seller $\mathcal{S}_i$, respectively. Based on the strategies, the market mechanism is proposed.

Market Workflow. The market workflow is shown in Fig. 1. ① Buyer $\mathcal{B}$ puts forward the demand for a product including the queries and the required performance. ② Buyer $\mathcal{B}$ determines the product price to buy the data product from broker $\mathcal{A}$. ③ Broker $\mathcal{A}$, acting as the bridge for the transaction between the buyer and m sellers, determines the data price to buy the data from sellers. ④ Each seller chooses what data (strictly speaking, data fidelity) to sell, and conducts corresponding privacy perturbation locally. ⑤ Sellers sell the protected datasets to broker $\mathcal{A}$ in exchange for the compensations. ⑥ Using the dataset bought from sellers, broker $\mathcal{A}$ manufactures the product. ⑦ Broker $\mathcal{A}$ sells the product to buyer $\mathcal{B}$. After buyer $\mathcal{B}$ receives the product via query answers and gives payment to broker $\mathcal{A}$, the transaction is finished.

Buyer's Strategy. Buyer $\mathcal{B}$ makes its strategy first, which is to determine the product price p^M, in order to maximize its profit by considering the desired utility of the product and stimulating the responses of the broker and sellers, i.e., what data price and data fidelity broker $\mathcal{A}$ and sellers would provide according to p^M.

Broker's Strategy. Broker $\mathcal{A}$ takes its strategy second, which is to determine data price p^D, in order to maximize its profit given p^M by stimulating the sellers' responses, i.e., what data fidelity each seller would provide according to p^D.

Seller's Strategy. Sellers make their strategies last. The strategy of each seller $\mathcal{S}_i$ is to determine data fidelity τ_i to maximize its profit by balancing the revenue of selling data and the cost of the privacy loss given the data price p^D.

Meanwhile, the inner competition among m sellers should be considered. Given the data price p^D, if seller $\mathcal{S}_i$ provides data with higher fidelity τ_i, more quantity would likely be sold. If other sellers provide better fidelity, less data quantity of seller $\mathcal{S}_i$ could be chosen. The data quantity that each seller $\mathcal{S}_i$ can sell is thus formalized as $\chi_i(\tau_i, \boldsymbol{\tau_{-i}})$ (see the instance in Section IV-A). Each seller competes for the quantity of data that can be sold by manipulating the data fidelity while balancing the compensation and the privacy cost. We define such inner competition among sellers as a Nash game. Seller $\mathcal{S}_i$ determines its strategy τ_i simultaneously with each other to maximize its own profit which is also affected by other sellers' strategies $\boldsymbol{\tau_{-i}}$. Nash equilibrium would be achieved where no seller can increase its profit by unilaterally changing its strategy with all other sellers' strategies fixed. The data quantity χ_i sold by each seller $\mathcal{S}_i$ can be calculated according to the equilibrium state, treated as the seller selection results.

Note that if one participant finds that its maximized profit is below zero, it will quit since it can gain no benefit from participating in the data trading, which guarantees the individual rationality [47] of participants. If it is the buyer or the broker who quits the trading or all sellers simultaneously get negative profits and quit, the current transaction would fail and a new transaction would be initiated. Otherwise, the remaining participants would continue and finish the transaction. Since it is easy to deal with the quit situation, we focus on the more common case and assume that all participants can get non-negative profits in the following discussions.

Three-Stage Stackelberg-Nash Game. Strategies of buyer $\mathcal{B}$, broker $\mathcal{A}$, and sellers $\mathcal{S}_i$ $(i = 1, 2, ..., m)$ constitute the strategy profile $\left\langle p^M, p^D, \boldsymbol{\tau} \right\rangle$ of data markets. Such a profile determines market trading rules including selling at what price for both data product (p^M) and data (p^D), what data (data fidelity) to sell $(\boldsymbol{\tau})$, as well as how to select sellers (the calculated $\boldsymbol{\chi} = (\chi_1, \chi_2, ...\chi_m)$ based on $\boldsymbol{\tau}$). The market mechanism is formulated as a three-stage Stackelberg-Nash game, where buyer $\mathcal{B}$ is the leader, broker $\mathcal{A}$ is the sub-leader, and m sellers act as the followers. Each of them tries to maximize its own profit by determining its optimal strategy variable. The three-stage Stackelberg-Nash game is defined as follows.

Definition 1 (Three-Stage Stackelberg-Nash Game): The game consists of three stages for buyer, broker, and sellers.
Stage 1 Buyer $\mathcal{B}$: $p^{M^*} = \arg\max_{p^M} \Phi\left(p^M, \boldsymbol{\tau}(p^D(p^M))\right)$.
Stage 2 Broker $\mathcal{A}$: $p^{D^*} = \arg\max_{p^D} \Omega\left(p^M, p^D, \boldsymbol{\tau}(p^D)\right)$.
Stage 3 Seller $\mathcal{S}_i$: $\tau_i^* = \arg\max_{\tau_i} \Psi_i\left(p^D, \boldsymbol{\tau}\right), i = 1, 2, ..., m$.

The above three-stage Stackelberg-Nash game involves both sequentiality and simultaneity. Sequentiality indicates the order in market flow, i.e., driven by demand, the data trading proceeds with buyer $\mathcal{B}$ acting first, broker $\mathcal{A}$ taking its strategy second, and sellers making their strategies last. Simultaneity indicates the equal positions of m sellers who take strategy simultaneously in their inner Nash game.

C. Market Equilibrium

In the above game, our objective is to find an optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$, by which each participant can maximize its own profit. Meanwhile, the optimal solution must satisfy some equilibrium so that no one is willing to adopt other strategies, which indicates market stability, making our design feasible. We define a Stackelberg-Nash Equilibrium (SNE) in data markets.

Definition 2 (Stackelberg-Nash Equilibrium): An optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$ constitutes a Stackelberg-Nash Equilibrium (SNE) if and only if the following set of inequalities is satisfied.

$$\Phi\left(p^{M^*}, \boldsymbol{\tau}^*(p^{D^*}(p^{M^*}))\right) \geq \Phi\left(p^{M}, \boldsymbol{\tau}^*(p^{D^*}(p^{M}))\right),$$
$$\Omega\left(p^{M}, p^{D^*}(p^M), \boldsymbol{\tau}^*(p^{D^*})\right) \geq \Omega\left(p^{M}, p^{D}, \boldsymbol{\tau}^*(p^{D})\right),$$
$$\Psi_i\left(p^{D}, \boldsymbol{\tau}^*(p^{D})\right) \geq \Psi_i\left(p^{D}, \boldsymbol{\tau_{-i}}^*(p^{D}), \tau_i\right), i=1,2,...,m.$$

SNE indicates that each participant takes its optimal strategy which maximizes its own profit in a demand-driven data market with *buyer-broker-sellers* sequence. No one can add its own profit by unilaterally changing its strategy.

IV. Market Construction: Equilibrium Solving and Trading Dynamics

In this section, we first instantiate a data market by specifying each function template in Section IV-A and then derive the market equilibrium by backward induction in Section IV-B. We describe the market dynamics in Section IV-C.

A. Market Instance

In terms of profit functions of participants, we claim the basic properties that should be satisfied in Section III, based on which we give instances below following certain practices. Other alternatives can be adopted based on real cases.

1) Profit instantiation of Buyer $\mathcal{B}$: In terms of the utility of buyer $\mathcal{B}$, it has been analyzed that combining product performance and dataset quality to measure product utility can make the quantification of product utility more comprehensive. Since the dataset quality q_i^D of each seller is positively correlated with the number of tuples χ_i and the fidelity of each tuple τ_i, we instantiate $q_i^D = g(\chi_i, \tau_i)$ as $\chi_i\tau_i$ based on the intuition that the multiplication reflects the fidelity of the whole dataset. As mentioned before, other inherent factors of data may also contribute to the data quality, which have been studied and are complementary to our work [48], and we focus on the effect sellers can exert on the data instead of the intrinsic characteristics which can be further formulated into q_i^D as a constant. Based on the instance of q^D, we define the utility of a data product as the weighted sum of the utility of the dataset quality and the utility of the product performance, which are further formulated as the logarithmic functions following utility theory [49] in economics.

$$\mathbf{U}\left(q^D, \nu\right) = \theta_1 \ln\left(1 + \rho_1 q^D\right) + \theta_2 \ln\left(1 + \rho_2 \nu\right). \quad (4)$$

Here θ_1 and θ_2 satisfy $\theta_1, \theta_2 \in (0,1), \theta_1 + \theta_2 = 1$, which measure the relative significance of the two for buyer $\mathcal{B}$. In our example, if dataset quality q^D plays a greater role than product performance ν in the decision-making of the automaker, the automaker may set $\theta_1 = 0.7$ and $\theta_2 = 0.3$. $\rho_1 > 0$ and $\rho_2 > 0$ refer to buyer $\mathcal{B}$'s sensitivity to these two attributes respectively. More sensitive, more utility added when the attribute gets better. For example, if higher dataset quality can bring the automaker much more utility, its ρ_1 would be big, meaning that the automaker is highly sensitive to the quality of production materials.

In terms of the payment of buyer $\mathcal{B}$, we instantiate $q^M = h(q^D, \nu)$ as $q^D\nu$ since it is positively correlated to q^D and ν, and specify $\mathbf{PB}(p^M, q^M) = p^M q^M$, which borrows from common sense that the payment for goods is equal to the unit (basic) price multiplied by the quantity (quality).

2) Profit instantiation of Broker $\mathcal{A}$: In terms of the first part of the cost of broker $\mathcal{A}$, the payment to each seller $\mathcal{S}_i$ is similarly instantiated as the product of the basic price p^D and the dataset quality p^D, i.e., $\mathbf{PS}(p^D, q_i^D) = p^D q_i^D$.

In terms of the second part of the cost of broker $\mathcal{A}$, we adopt a widely used transcendental logarithmic function for the manufacturing cost because of its adaptability to varied economies of scale and manufacturing strategy (e.g., how to allocate computing resources) according to the work [50]. Here $\sigma_k, k \in \{0,1,2,3,4,5\}$ are the parameters of the translog cost function which can be fitted by broker $\mathcal{A}$ based on the actual manufacturing procedure.

$$\mathbf{C}(N, \nu) = \exp\left(\sigma_0 + \sigma_1 \ln(N) + \sigma_2 \ln(\nu) + \frac{1}{2}\sigma_3 \ln^2(N) + \frac{1}{2}\sigma_4 \ln^2(\nu) + \sigma_5 \ln(N) \cdot \ln(\nu)\right). \quad (5)$$

3) Profit instantiation of Seller $\mathcal{S}_i$: In terms of the privacy cost of seller $\mathcal{S}_i$, we first instantiate data fidelity τ_i and data quantity χ_i. We choose an inverse trigonometric function form as $f(\cdot)$ which satisfies the characteristics stipulated in Section III-A3 and give the following definition of τ_i.

$$\tau_i = f(\epsilon_i) = \frac{2}{\pi}\operatorname{arcsec}(w_i\epsilon_i + 1),\ \epsilon_i \in [0, \infty), \quad (6)$$

which leads to $\tau_i \in [0, 1)$. Additionally, $\tau_i = 1$ when no noise is added. Thus $\tau_i \in [0, 1]$.

We then instantiate the quantity χ_i of data that can be sold by seller $\mathcal{S}_i$ as proportional to the data fidelity τ_i it provides.

$$\chi_i = N\frac{\omega_i\tau_i}{\sum_{j=1}^{m}\omega_j\tau_j}, \quad (7)$$

where $\omega_1, \omega_2, ..., \omega_m$ refer to the weights of sellers' data, which are maintained by the broker. Such weights reflect the historical performance of each seller's data in past deals (implying the verifiable data value). The broker would update these weights after each round of transactions. For example, new weights can be updated based on the contributions of sellers to the data product in the current transaction. One of the evaluation methods for the data contribution is by Shapley value [14], which is adopted in this market instance and implemented in our experiments.

Based on the instances of data fidelity τ_i and data quantity χ_i, we adopt a widely used quadratic function for the cost of seller $\mathcal{S}_i$. Here $\lambda_i > 0$ is seller $\mathcal{S}_i$'s privacy sensitivity. In our health example, the privacy loss of IMS Health corresponds to the negative impact of data exposure, and $\mathbf{L}_i(\cdot)$ quantifies the economic estimation of the impact (e.g., legal expenses or amends to patients).

$$\mathbf{L}_i\left(\tau_i\right) = \lambda_i(\chi_i\tau_i)^2. \quad (8)$$

B. Solving Equilibrium: Backward Induction

To determine the optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^*\right\rangle$, we adopt the backward induction approach [51]. We first investigate Stage 3 to solve Nash equilibrium among sellers and derive the expression of each seller's optimal strategy

$\tau_i^*, i = 1, 2, ..., m$ (Eq. 12) for any given data price p^D in Section IV-B1. We explore two methods, direct derivation and an approximate method using the mean-field state which can deal with complicated cases. Next, we consider Stage 2 to determine the expression of the optimal strategy p^{D^*} (Eq. 16) of broker $\mathcal{A}$ for any given product price p^M in Section IV-B2. In this process, the expression of $\tau_i^*, i = 1, 2, ..., m$ solved from Nash game can be used as sellers' optimal reactions to p^D. Then, we back to Stage 1 to find the value (rather than the expression) of buyer $\mathcal{B}$'s optimal strategy p^{M^*} (Eq. 17) based on the optimal reactions of the broker as well as sellers in Section IV-B3. After that, we can get the value of the optimal strategy p^{D^*} by substituting p^{M^*} into the result (Eq. 16) in Stage 2. Finally, we can compute the value of each seller's optimal strategy τ_i^* by substituting p^{D^*} into the result (Eq. 12) in Stage 3. Till now, the complete strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$ has been determined. The detailed deduction is presented as follows.

1) Expression of $\boldsymbol{\tau}^*$ in Stage 3: We present two approaches to derive the expression of τ_i^* for sellers, direct derivation and a mean-field based approximation method for large numbers of sellers and complicated profit function forms that can hardly be solved by direct derivation.

Direct Derivation. By substituting Eqs. 7, 8 into Eq. 3, we get each seller's profit

$$\begin{aligned}\Psi_i\left(p^D, \tau_i\right) &= p^D \chi_i \tau_i - \lambda_i (\chi_i \tau_i)^2 \\ &= p^D \cdot N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} - \lambda_i \left(N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} \right)^2 .\end{aligned}$$

Ψ_i is correlated to not only seller $\mathcal{S}_i$'s strategy τ_i but also other sellers' strategies $\tau_j, j \neq i$ because of the inner competition formulated as Nash game among sellers. As we discussed before, each seller aims to maximize its own profit. Therefore, we derive each of the first-order derivatives for m sellers' profit functions and let each of them equal to zero, thus getting m equations. The equation for seller $\mathcal{S}_i$ is

$$p^D \frac{\partial \left(N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} \right)}{\partial \tau_i} - 2\lambda_i \cdot N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} \cdot \frac{\partial \left(N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} \right)}{\partial \tau_i} = 0. \tag{9}$$

If $\frac{\partial \left(N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} \right)}{\partial \tau_i} = 0$, it is an all-zero solution, which does not meet our problem situation, so we can directly eliminate $\frac{\partial \left(N \frac{\omega_i \tau_i^2}{\sum_{j=1}^m \omega_j \tau_j} \right)}{\partial \tau_i}$, and then get

$$p^D \sum_{i=1}^m \omega_i \tau_i - 2N \lambda_j \omega_j \tau_j^2 = 0, \quad j = 1, 2, ..., m, \tag{10}$$

where each $\mathcal{S}_i$'s equation not only relates to its own strategy τ_i but also contains other sellers' strategies, requiring us to solve m simultaneous equations together. Finding that

$$2N\lambda_1 \omega_1 \tau_1^2 = 2N\lambda_2 \omega_2 \tau_2^2 = ... = 2N\lambda_m \omega_m \tau_m^2 = p^D \sum_{i=1}^m \omega_i \tau_i. \tag{11}$$

By adding all m equations in Eq. 10, we get

$$m p^D \sum_{i=1}^m \omega_i \tau_i - 2N \sum_{i=1}^m \lambda_i \omega_i \tau_i^2 = 0.$$

Using τ_1 to indicate other τ_i $(i = 2, 3, ..., , m)$ from Eq. 11,

$$m p^D \tau_1 \sum_{i=1}^m \sqrt{\frac{\lambda_1 \omega_1 \omega_i}{\lambda_i}} - 2Nm\lambda_1 \omega_1 \tau_1^2 = 0.$$

Therefore,

$$\tau_1^* = \frac{p^D}{2N\sqrt{\omega_1 \lambda_1}} \sum_{i=1}^m \sqrt{\frac{\omega_i}{\lambda_i}},$$

and using Eq. 11 again, we get all sellers' optimal strategies

$$\tau_i{}^* = \frac{p^D}{2N\sqrt{\omega_i \lambda_i}} \sum_{j=1}^m \sqrt{\frac{\omega_j}{\lambda_j}}, i = 1, 2, ..., m. \tag{12}$$

Note that the second-order derivative $\frac{\partial^2 \Psi_i\left(p^D, \tau_i\right)}{\partial \tau_i^2} < 0$, so these solutions can maximize each seller's profit.

Mean-field based Approximate Method. It is theoretically feasible that the optimal $\boldsymbol{\tau}$ can be derived by directly using the derivation method for each seller's profit function and then solving m simultaneous equations as above. However, for complicated function forms (e.g., more complicated loss function rather than the used one), since the number of sellers m can be quite large in practice, it may be difficult to derive analytical expressions by solving a large number of simultaneous equations each with complex forms. Specifically, the m equations are highly coupled, i.e., each with all $\tau_i, i = 1, 2, ..., m$, and eliminating the similar terms to simplify the equations as we did in Eq. 9 is not always feasible. Therefore, we propose an approximate method that makes each equation with a single τ_i and independent from others. Note that the approximate approach is proposed to deal with the case where direct derivation would fail rather than to improve the efficiency. Thus we take a different privacy loss function form for the sellers as an example where the direct derivation is not practically feasible in order to illustrate the mean-field method. Specifically, we replace Eq. 8 with $\mathbf{L}_i\left(\tau_i\right) = \lambda_i \chi_i \tau_i{}^2$.

The approximation is based on the mean-field theory [52], which deals with situations that involve a great number of agents, i.e., sellers in our context. When there are a great number of sellers in Nash game, it is reasonable to expect that a single seller has a *tiny* (infinitesimal) influence on the equilibrium and is affected by other sellers through a mean-field state, which we formulate as the weighted mean of all sellers' strategies, $\overline{\tau}$.

$$\overline{\tau} = \frac{\sum_{i=1}^m \omega_i \tau_i}{m}. \tag{13}$$

The mean-field state $\overline{\tau}$ indicates the overall data fidelity provided by sellers at equilibrium and is not intensively affected by the data fidelity from one specific seller.

Using the new privacy loss function, the profit function of seller $\mathcal{S}_i$ in Eq. 3 is changed into

$$\Psi_i\left(p^D, \tau_i\right) = p^D (\chi_i \tau_i) - \lambda_i \chi_i \tau_i{}^2. \tag{14}$$

Using $\overline{\tau}$, χ_i can be simplified as $N\frac{\omega_i \tau_i}{m\overline{\tau}}$. Since $\overline{\tau}$ is not strongly affected by specific τ_i, we can easily derive the first-order derivative of each seller's profit function $\Psi_i\left(p^D, \tau_i\right)$ with respect to τ_i and let them equal to zero.

$$p^D \cdot N\frac{\omega_i \tau_i^2}{m\overline{\tau}} - \lambda_i \cdot N\frac{\omega_i \tau_i^3}{m\overline{\tau}} = 0, \quad i = 1, 2, ..., m.$$

We derive $\mathcal{S}_i$'s optimal strategy

$$\tau_i^* = \frac{2p^D}{3\lambda_i}, i = 1, 2, ..., m. \tag{15}$$

Note that the second-order derivative $\frac{\partial^2 \Psi_i\left(p^D, \tau_i\right)}{\partial \tau_i^2} < 0$, so these solutions can maximize each seller's profit.

Error Analysis. We use fixed $\overline{\tau}$ to replace $\frac{\sum_{i=1}^m \omega_i \tau_i}{m}$ when deriving the derivatives. Such replacement is an approximation and its error depends on the form of the profit function. We analyze the error bound of the mean-field approach.

Theorem 1: The exact weighted mean of all sellers' strategies by the direct derivation is defined as $\overline{\tau}^{DD}$, and the approximated one by the mean-field method is $\overline{\tau}^{MF}$. The error is $\overline{\tau}^{DD} - \overline{\tau}^{MF}$. Consider the case that the privacy loss function is $\mathbf{L}_i\left(\tau_i\right) = \lambda_i \chi_i {\tau_i}^2$. When the number of sellers m is large and by scaling $\omega_1, \omega_2, ..., \omega_m$ such that $\frac{\omega_i}{\lambda_i} \leq \frac{1}{p^D m^2}$, we get

$$-\frac{1}{6m^2} < \overline{\tau}^{DD} - \overline{\tau}^{MF} < \frac{1}{m} - \frac{2}{3m^2}.$$

Note that what makes sense is the proportional relationship among $\omega_i, i = 1, 2, ..., m$, allowing us to arbitrarily scale them.

Proof 1: By applying direct derivation to Eq. 14 and considering it as a quadratic equation about τ_i, we can get the expression of τ_i^* w.r.t. $\Sigma_{\boldsymbol{\tau}_{\neg i}} = \sum_{j=1, j\neq i}^m \omega_j \tau_j$. Through multiple times of scaling and simplifying, both the upper and lower bounds of $\overline{\tau}^{DD} = \frac{1}{m}\sum_{i=1}^m \omega_i \tau_i^*$ are derived, i.e., $\frac{1}{m}\sum_{i=1}^m \frac{p^D \omega_i}{2\lambda_i} < \overline{\tau}^{DD} < \frac{1}{m}\left(\sum_{i=1}^m \sqrt{p^D \frac{\omega_i}{\lambda_i}}\right)^2$. With the condition further applied, the approximation error is bounded. See more details in the complete version [53].

Through the above error analysis, we draw the following empirical conclusion: by scaling the value of ω_i $(i = 1, 2, ..., m)$ to satisfy $\frac{\omega_i}{\lambda_i} \leq \frac{1}{p^D m^2}$, the error of the mean-field approximation method will be bounded in an acceptable range and decrease with increasing m when m is very large. When m approaches infinity, the error is approximately zero. This result is in line with the mean-field theory [52]. When the number of sellers m is big, our proposed mean-field method appears reasonable in terms of error.

2) Expression of p^{D^*} in Stage 2: We give the expression of p^{D^*} for the broker by direct derivation. The detailed derivation is presented in the complete version [53].

$$p^{D^*} = \frac{\nu p^M}{2}. \tag{16}$$

Note that the second-order derivative $\frac{\partial^2 \Omega\left(p^M, p^D, \boldsymbol{\tau}\right)}{\partial {p^D}^2} < 0$, so the solution can maximize the broker's profit.

3) Value of p^{M^*} in Stage 1: We also use direct derivation in this stage, and by using the results in Sections IV-B1 and IV-B2, we can directly derive the value rather than the expression of p^{M^*} for the buyer. The detailed derivation is presented in the complete version [53].

$$p^{M^*} = \frac{-\mathrm{c}_2 + \sqrt{\mathrm{c}_2^2 + 4\mathrm{c}_1^2 \mathrm{c}_2}}{2\mathrm{c}_1 \mathrm{c}_2}, \tag{17}$$

where $\mathrm{c}_1 = \frac{\rho_1 \nu}{4}\sum_{i=1}^m \frac{1}{\lambda_i}$ and $\mathrm{c}_2 = \frac{\nu^2}{2\theta_1}\sum_{i=1}^m \frac{1}{\lambda_i}$. Note that the second-order derivative $\frac{\partial^2 \Phi\left(p^M, \boldsymbol{\tau}\right)}{\partial {p^M}^2} = -\frac{\theta_1 \mathrm{c}_1^2}{(1 + \mathrm{c}_1 p^M)^2} - \theta_1 \mathrm{c}_2 < 0$, so the solution can maximize the buyer's profit.

Getting p^{M^*}, we can determine the optimal value of p^{D^*} by substituting p^{M^*} into Eq. 16 and each seller's optimal value of τ_i^* by substituting p^{D^*} into Eq. 12. Till now, the complete optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$ has been determined, based on which the market transaction can be conducted.

4) Equilibrium Analysis: We prove the existence and uniqueness of SNE in *Share*.

Theorem 2: The complete optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$ determined by backward induction approach uniquely constitutes SNE.

Proof 2: The existence and uniqueness of SNE can be deduced by the property that the strategy space is a convex and compact subspace of Euclidean space while the profit functions are concave [54]. Take the buyer as an example. It can be justified that the buyer's maximum profit is obtained only at p^{M^*} derived by direct derivation due to the strictly concave property of $\Phi(\cdot)$ w.r.t. p^M, leading to the first inequation in Definition 2 holding uniquely at p^{M^*}. Similar results apply to the second inequation for the broker. In terms of the sellers, a unique Nash equilibrium can be justified similarly and the third inequation holds for every seller only at τ_i^*. Therefore, SNE exists in our mechanism as the set of inequalities in Definition 2 can be satisfied at $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$, while any other strategy profile cannot satisfy the three inequations simultaneously and constitute SNE. More details can be found in [53].

C. Complete Data Trading Dynamics

We summarize the complete dynamics of data markets in Alg. 1 with the above equilibrium-solving process integrated.

The first phase is *Initialization*. Each party reports its input parameters which can be fitted based on the function shape and historical data through parameter estimation techniques [55]. The buyer sets appropriate parameters $\theta_1, \theta_2, \rho_1, \rho_2$ for its utility function and proposes queries $\mathbf{Q}$ with performance parameter ν which need to be solved by the product (Line 2). Note that the product is not restricted in forms decided by the broker while the access channel for data buyers is uniformly set as queries. The broker crystallizes the size N of data it can handle, determines $\sigma_k, k \in \{0, 1, 2, 3, 4, 5\}$ for its cost function, and maintains the weights $\omega_i, i = 1, 2, ..., m$ of sellers' datasets (Line 3). To decide the real weights before the first transaction, the broker can use dummy buyers to iterate several times where Shapley value can be used to evaluate

Algorithm 1: Data trading dynamics.

```
1  %% Initialization;
2  From the current buyer $\mathcal{B}$, demanded queries $\mathbf{Q}$ and parameters
     $\nu, \theta_1, \theta_2, \rho_1, \rho_2$ are provided;
3  From broker $\mathcal{A}$, $N, \sigma_k (k \in \{0, 1, 2, 3, 4, 5\}), \omega_i (i = 1, 2, ..., m)$
     are given;
4  From existing $m$ sellers, each seller $\mathcal{S}_i$ decides $\lambda_i$;
5  %% Strategy Decision;
6  Through three-stage Stackelberg-Nash game, the optimal strategy
     profile $\langle p^{M*}, p^{D*}, \boldsymbol{\tau}^* \rangle$ is determined by the buyer, the broker,
     and sellers, respectively;
7  %% Data Transaction;
8  The quantity of data each seller can sell, $\boldsymbol{\chi}^*$, is calculated according
     to Eq. 7;
9  for each seller $\mathcal{S}_i, i = 1, 2, ...m$ do
10     Randomly pick ${\chi_i}^*$ data pieces from its dataset $D_i$;
11     Calculate ${\epsilon_i}^*$ from the strategy ${\tau_i}^*$ according to Eq. 6;
12     Conduct LDP with ${\epsilon_i}^*$ on its ${\chi_i}^*$-sized dataset, and then give
         the protected $D_i^t$ to broker $\mathcal{A}$;
13 Broker $\mathcal{A}$ gets data from sellers to form dataset $D^t$ for production
     and pays compensation $\mathbf{PS}_i^*$ to each seller;
14 %% Product Production;
15 Broker $\mathcal{A}$ then uses $D^t$ to produce the data product as well as
     computes the answers to queries $\mathbf{Q}$;
16 After manufacturing the product, broker $\mathcal{A}$ updates $\omega_1, \omega_2, ..., \omega_m$
     (might scale down or normalized as needed) based on the
     contribution to the product from each seller's $D_i^t$;
17 %% Product Transaction;
18 Broker $\mathcal{A}$ gives the product to buyer $\mathcal{B}$ (by returning the query
     answers), and meantime buyer $\mathcal{B}$ pays $\mathbf{PB}^*$ to broker $\mathcal{A}$.
```

the sellers' datasets. Sellers give their privacy sensitivity $\lambda_i, i = 1, 2, ..., m$ (Line 4).

The second phase is *Strategy Decision*. Using the strategy mechanism, buyer $\mathcal{B}$, broker $\mathcal{A}$, and each seller $\mathcal{S}_i$ give product price p^{M*}, data price p^{D*}, and data fidelity $\boldsymbol{\tau}_i^*$ in order according to Eqs. 17, 16, 12, respectively (Line 6).

Then *Data Transaction* between the broker and sellers begins. The data quantity chosen from each seller can be calculated according to Eq. 7 (Line 8). Each seller randomly picks χ_i^*-sized dataset (Line 10) and pre-processes it for privacy protection based on ${\epsilon_i}^*$ calculated from Eq. 6 (Lines 11-12). After that, seller $\mathcal{S}_i$ gives its protected dataset D_i^t to the broker in exchange for compensation $\mathbf{PS}_i^*$ (Line 13).

The next phase is *Product Production*. The broker collects the data as D^t and uses it to make the product (Line 15). Moreover, the weights of sellers' datasets are updated by the broker based on their corresponding contributions to the data product (Line 16). We give one update formula based on Shapley value as an example: $\omega_i' = 0.2\omega_i + 0.8\mathcal{SV}_i$, where $\mathcal{SV}_i$ is the Shapley value of D_i^t to the product and the coefficient 0.8 indicates to what extent the historical performance of data can be useful for the current task. The updated weights ω_i' can be used in the subsequent transaction.

The last phase is *Product Transaction* between the broker and the buyer. The broker gives the product (strictly speaking, the query answers based on the product) to the buyer and the buyer pays $\mathbf{PB}^*$ to the broker (Line 18). So far, the current data transaction among buyer $\mathcal{B}$, broker $\mathcal{A}$, and sellers $\mathcal{S}_i, i = 1, 2, ..., m$ has finished. When the next buyer comes, the next transaction will start.

V. Experiments

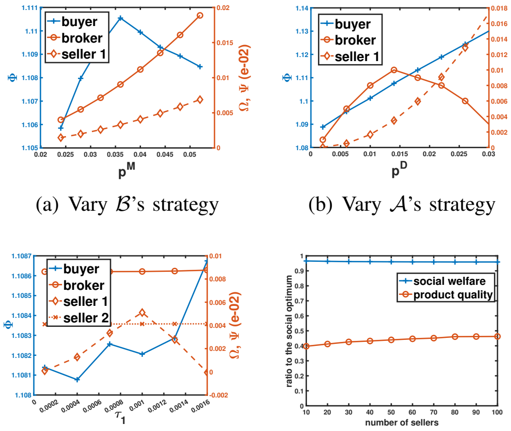
In this section, we present experimental studies validating the effectiveness and efficiency of *Share*. We first describe our experiment setups in Section V-A. Sections V-B and V-C show the results verifying the effectiveness and efficiency of *Share*, respectively. Section V-D shows the effects of the main parameters used in *Share*.

A. Experiment Setup

We conduct experiments on a machine with an Intel Core i7-11700KF running Ubuntu with 64GB memory. The Shapley value is calculated based on Monte Carlo Method [56]. Laplace mechanism [57], a technique for achieving LDP, is applied to each record to adjust data fidelity for each seller.

Datasets. We use a real dataset, Combined Cycle Power Plant (CCPP) [58], which contains 9,568 data points with four features. The buyer's demanded query task is to get the prediction of net hourly electrical energy output given the specific conditions of features, which can be understood as the analytical query extensively used in the analytical database. A linear regression model is considered as the data product that the broker manufactures to serve the queries and explained variance is used to measure the performance. We randomly choose a training dataset (the data of sellers) with a size of 9,000, and the 568 data records left are used for validation (based on which the queries of the buyer are generated). In the real world, the datasets of sellers can be the same in quality (which makes it easy to randomly choose sellers to buy data) or vary in quality, which is the case we deal with in *Share*. To simulate the distinction in data quality, we first sort data by quality measured by Shapley value, which indicates the contribution of each data record to regression. Then by distributing data in decreasing quality over sellers, each seller owns 90 data records with different quality. Besides the real dataset, we augment CCPP through replication and Gaussian noise $\mathcal{N}(0, 0.1^2)$ injection to generate a synthetic dataset with a size of $1,000,000$ to test the efficiency of *Share*.

Parameter Settings. Our parameters include the number of sellers m, the total data quantity N, the required explained variance ν, the individual parameters of each party's profit (i.e., buyer $\mathcal{B}$'s $\theta_1, \theta_2, \rho_1, \rho_2$ related to utility, broker $\mathcal{A}$'s cost parameters $\sigma_k, k \in \{0, 1, 2, 3, 4, 5\}$, and seller $\mathcal{S}_i$'s privacy sensitivity $\lambda_i, i = 1, 2, ..., m$), and the initial weights $\omega_i, i = 1, 2, ..., m$ of sellers. We set $m = 100$, $N = 500$, and $\nu = 0.8$. The utility parameters of the buyer are set as $\theta_1 = 0.5, \theta_2 = 0.5, \rho_1 = 0.5, \rho_2 = 250$ (in order to balance the impacts of product performance and dataset quality). The cost parameters of the broker are related to the practical manufacturing situation and are set as default values $\sigma_0 = 1 \times 10^{-3}, \sigma_1 = -2, \sigma_2 = -3, \sigma_3 = 1 \times 10^{-3}, \sigma_4 = 2 \times 10^{-3}, \sigma_5 = 1 \times 10^{-3}$. Sellers' $\lambda_i, i = 1, 2, ..., m$ are picked randomly in $(0, 1)$. $\omega_1, \omega_2, ..., \omega_m$ are initially generated by using a dummy buyer to iterate the mechanism which takes five times to stabilize the profits. We consider buyer $\mathcal{B}$ as a general buyer coming after several transactions have finished. Shapley values of sellers' datasets can be calculated after regression to update the weights for the next transaction.



(a) Vary $\mathcal{B}$'s strategy

(b) Vary $\mathcal{A}$'s strategy

(c) Vary $\mathcal{S}_1$'s strategy

(d) Social efficacy

Fig. 2. Effectiveness.

Measurement Metrics. Four main indexes are adopted to evaluate the mechanism concerning both effectiveness and efficiency. We will show the results of using direct derivation for equilibrium solving. The mean-field approach (used when direct derivation fails) performs the same on the metrics.

- *Profit.* The profits of the buyer, the broker, and sellers.
- *Social welfare.* The effect on the overall benefit, notated as a function of $\boldsymbol{\tau}$, i.e., $\mathbf{SW}(\boldsymbol{\tau}) = \mathbf{U}(\boldsymbol{\tau}) - \sum_{i=1}^{m} \mathbf{L}_i(\tau_i) - \mathbf{C}$. Note that the social welfare measures the total profits, correlated to the collected data (more precisely, data fidelity $\boldsymbol{\tau}$) yet independent of payments (prices) which only circulate among participants. The optimum social welfare can be derived by solving the social welfare maximization problem $\max_{\boldsymbol{\tau}} \mathbf{SW}$ and used to measure the social welfare level of our mechanism which is represented as the ratio to the social optimum.
- *Product quality.* The quality of the data product manufactured by the collected data with distinctive fidelity.
- *Runtime.* The time cost of executing the mechanism which evaluates the efficiency.

B. Effectiveness

Given the parameters set above, the optimal strategy $\left\langle p^{M*}, p^{D*}, \boldsymbol{\tau}^* \right\rangle$ is determined according to solutions in Eqs. 17, 16, and 12. We verify the optimality (profit maximization) for all participants as well as the implied equilibrium in Figs. 2(a)-(c). The social welfare and product quality are then evaluated with different numbers of sellers in Fig. 2(d). We also investigate the effect of the inner Nash game in Table II.

Fig. 2(a) shows the profit maximization of buyer $\mathcal{B}$. By changing its strategy p^M while maintaining the rest, the peak of the buyer's profit $\Phi(\cdot)$ appears when its optimal strategy $p^{M*} = 0.036$ determined in Eq. 17 is adopted (the monetary unit can adjust with how the utility/cost function is mapped into money). Whatever strategy the buyer chooses except p^{M*}, it will get a lower profit when all other participants' strategies are fixed. We can also observe the change in the profits of the broker and the seller (with seller $\mathcal{S}_1$ taken as a representative of sellers) shown in the other two lines. Specifically, with growing p^M, the broker can gain more profit, which can further add the compensations and thus the profits for sellers.

Fig. 2(b) shows the profit maximization of broker $\mathcal{A}$. By changing its strategy p^D while maintaining the rest, the peak of the broker's profit $\Omega(\cdot)$ verifies that its optimal strategy $p^{D*} = 0.014$ solved by Eq. 16 uniquely guarantees its profit maximization while any other p^D would lead to an inferior profit. The change in the profits of the buyer and seller can be also observed. Specifically, the growing p^D brings more compensations to sellers, adding their profits. Due to more compensations, the dataset quality from sellers can therefore be improved, which causes the rise of the buyer's profit.

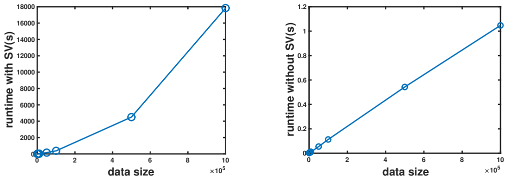
Fig. 2(c) shows the profit maximization of seller $\mathcal{S}_1$ (taken as a representative). By changing its strategy τ_1 while maintaining the rest, the peak of seller $\mathcal{S}_1$'s profit Ψ_1 verifies that the optimal strategy $\tau_1^* = 0.001$ solved by Eq. 12 achieves maximum profit and unilaterally changing its strategy promises no more profit. Combined with the results analyzed above, it can be concluded that SNE is reached since no participant can improve its profit by individually manipulating its strategy, and the optimality of all the participants is achieved. The other lines show the change in profits of other participants, indicating the transparency of the inner Nash game of Stage 3 to the upper stages and the dilution of the individual effect by the large number of sellers. See more analyses in [53].

Fig. 2(d) shows the ratio results of social welfare and data product quality, collectively referred to as social efficacy, compared to the optimum ones derived from the social welfare maximization problem. The proposed mechanism can achieve extremely high (over 95%) social welfare. As the number of sellers rises, the social welfare slightly decreases, which implies that more strategic gaming among participants exacerbates the social inefficiency in terms of overall profits. The product quality performs inferior to the socially optimal result due to the selfish profit-seeking behaviors of participants, which, however, would still outperform the baselines as shown in the following justification of Nash game. Better products can be acquired when more sellers (thus, more data) engage, implying the significance of data circulation.

Inner Nash Game. To verify the effectiveness of using Nash game to formulate Stage 3, we implement the mechanism compared to the baselines of using Random and Average strategies to select data. Table II shows the product quality q^M, the profit results of the buyer Φ, the broker Ω, and the average level of sellers $\bar{\Psi}$, as well as the social welfare $\mathbf{SW}$ (represented as the ratio to the optimum) through Nash game, Random, and Average respectively with the parameters kept the same. We can observe that the Nash-based seller selection outperforms the baselines in terms of the data product quality, the profits of all three parties, and the social welfare. The product quality based on the data selected by Nash game is the highest, which explicitly shows the effectiveness of the seller selection results. In terms of individual profits, not only data sellers benefit from their inner benign competition, the profits of both the buyer and the broker also increase, which indicates the advantage of the inner Nash game to

TABLE II
COMPARING NASH WITH OTHERS.

	Nash	Random	Average
q^M	6.379658	2.013382	2.022149
Φ	3.255229	2.995894	2.996831
Ω	0.099928	0.031531	0.031669
$\bar{\Psi}$	0.000500	0.000154	0.000161
SW	96%	90%	90%



(a) Runtime with $\mathcal{SV}$ (b) Runtime without $\mathcal{SV}$

Fig. 3. Efficiency.

the upper stages. In terms of social benefits, the Nash game modeling achieves the highest social welfare compared to the optimal one, implying its effectiveness in data markets both individually and collectively.

Note that the buyer's profit is much more than the broker's and sellers', which is in line with the property of Stackelberg game (in favor of the leader) and consistent with the desired effect in demand-driven markets. The buyer as the transaction initiator can create value using the demanded product and gain long-term benefits (e.g., the huge revenue Ford earns owing to the business decision based on the acquired query answers), while the broker or the sellers make a profit from the one-shot transaction which is relatively lower.

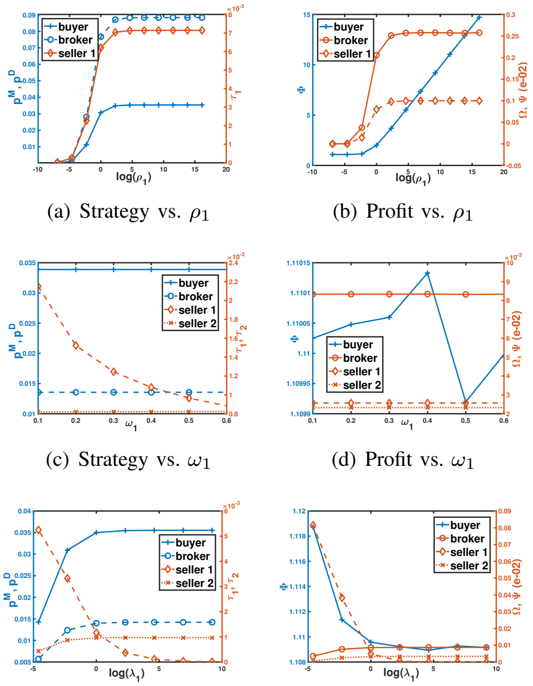
C. Efficiency

Fig. 3(a) and Fig. 3(b) show the runtime of the proposed data trading algorithm with and without Shapley value to update weights. We use the synthetic dataset with $1,000,000$ data records and adjust the number of sellers m from 5 to $10,000$ while fixing the other parameters and the average number of data records chosen from each seller as 100. Fig. 3(a) shows that the runtime grows as m goes higher but with an acceptable rate. Even when $m = 10,000$, it does not take too much time. While our mechanism contains a time-consuming part to calculate Shapley values, Fig. 3(b) shows that our mechanism without Shapley value calculation can run very fast with a linear time complexity.

D. Parameter Influence

In this section, we make sensitivity analyses of the major parameters in our mechanism and investigate how the parameters affect the strategies and profits of the three parties.

Fig. 4(a) and Fig. 4(b) present the effect of ρ_1 on strategies and profits. Note that ρ_1 is a parameter of the buyer's sensitivity to dataset quality, which objectively reflects the relationship between dataset quality and product utility. Fig. 4(a) shows that too small of a ρ_1 can hardly lead to effective markets because of the buyer's indifference to the data. When ρ_1 reaches a certain level, all the strategies stay the same and the market reaches equilibrium. The influence of ρ_1 is limited within the utility for the buyer and can no longer disturb the equilibrium, which may be due to common sense that the dataset quality cannot increase unlimitedly and with sharper sensitivity to the data, higher prices wouldn't bring about better data anymore. Figs. 4(b-f) show other parameter effects, which are detailedly analyzed in the complete version [53].



(a) Strategy vs. ρ_1 (b) Profit vs. ρ_1

(c) Strategy vs. ω_1 (d) Profit vs. ω_1

(e) Strategy vs. λ_1 (f) Profit vs. λ_1

Fig. 4. Effect of Parameters.

VI. CONCLUSION AND FUTURE WORK

We presented *Share*, the first demand-driven incentivized data market framework. The profit maximization for all participants and the *buyer-broker-sellers* market flow are fulfilled by considering the mutual interaction among three parties as a three-stage Stackelberg game, in which the absolute pricing for data is also realized. We addressed the seller selection problem by considering the inter-seller competition as a Nash game. To derive the Stackelberg-Nash Equilibrium, backward induction is used, and a novel mean-field approximation with provable guarantees is proposed. Our proposed data market framework performs well on real and synthetic datasets in terms of both effectiveness and efficiency.

Our work opens up many interesting research questions, e.g., how to accommodate multiple buyers and how to support complex costs of sellers across transactions, which can be promising future directions.

VII. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by the National Key RD Program of China (2021YFB3101100), NSFC grants (62102352, U23A20306), and NSF grant (CNS-2125530).

References

[1] "Boston database meeting," 2023, https://www.linkedin.com/posts/seemohan_45-worldwide-database-researchers-brainstorm-activity-7121547469573824512-hsB-.

[2] J. Pei, R. C. Fernandez, and X. Yu, "Data and AI model markets: Opportunities for data and model sharing, discovery, and integration," *Proc. VLDB Endow.*, vol. 16, no. 12, pp. 3872–3873, 2023. [Online]. Available: https://www.vldb.org/pvldb/vol16/p3872-pei.pdf

[3] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data market platforms: Trading data assets to solve data problems," *Proc. VLDB Endow.*, vol. 13, no. 11, pp. 1933–1947, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol13/p1933-fernandez.pdf

[4] A. Agarwal, M. Dahleh, and T. Sarkar, "A marketplace for data: An algorithmic solution," in *Proceedings of the 2019 ACM Conference on Economics and Computation*. ACM, 2019, pp. 701–726.

[5] L. Chen, P. Koutris, and A. Kumar, "Towards model-based pricing for machine learning in a data marketplace," in *SIGMOD*, P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, Eds. ACM, 2019, pp. 1535–1552. [Online]. Available: https://doi.org/10.1145/3299869.3300078

[6] J. Liu, J. Lou, J. Liu, L. Xiong, J. Pei, and J. Sun, "Dealer: An end-to-end model marketplace with differential privacy," *Proc. VLDB Endow.*, vol. 14, no. 6, pp. 957–969, 2021. [Online]. Available: http://www.vldb.org/pvldb/vol14/p957-liu.pdf

[7] A. Tanner, "How data brokers make money off your medical records," *Scientific American*, vol. 314, no. 2, pp. 26–29, 2016.

[8] "Mckinsey's case," 2023, https://www.mckinsey.com/about-us/new-at-mckinsey-blog/mckinsey-pro-bono-effort-helps-shed-light-on-hidden-cause-of-child-blindness.

[9] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," in *PODS*. ACM, 2012, pp. 167–178.

[10] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. Spanos, and D. Song, "Efficient task-specific data valuation for nearest neighbor algorithms," *Proceedings of the VLDB Endowment*, vol. 12, no. 11, pp. 1610–1623, 2019.

[11] Y. Li, X. Yu, and N. Koudas, "Data acquisition for improving machine learning models," *Proc. VLDB Endow.*, vol. 14, no. 10, pp. 1832–1844, 2021. [Online]. Available: http://www.vldb.org/pvldb/vol14/p1832-li.pdf

[12] X. Cao, Y. Chen, and K. J. R. Liu, "Data trading with multiple owners, collectors, and users: An iterative auction mechanism," *IEEE Trans. Signal Inf. Process. over Networks*, vol. 3, no. 2, pp. 268–281, 2017. [Online]. Available: https://doi.org/10.1109/TSIPN.2017.2668144

[13] R. C. Fernandez, "Protecting data markets from strategic buyers," in *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Z. Ives, A. Bonifati, and A. E. Abbadi, Eds. ACM, 2022, pp. 1755–1769. [Online]. Available: https://doi.org/10.1145/3514221.3517855

[14] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[15] H. Von Stackelberg, *Market structure and equilibrium*. Springer Science & Business Media, 2010.

[16] B. An, M. Xiao, A. Liu, X. Xie, and X. Zhou, "Crowdsensing data trading based on combinatorial multi-armed bandit and stackelberg game," in *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 2021, pp. 253–264.

[17] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*. IEEE Computer Society, 2013, pp. 429–438. [Online]. Available: https://doi.org/10.1109/FOCS.2013.53

[18] J. F. Nash Jr, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[19] Bloomberg, "https://www.bloomberg.com/professional/product/market-data/," 1981.

[20] DAWEX, "https://www.dawex.com/en/," 2015.

[21] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Toward practical query pricing with querymarket," in *SIGMOD*. ACM, 2013, pp. 613–624.

[22] M. Lei, X. Zhang, L. Chu, Z. Wang, P. S. Yu, and B. Fang, "Finding route hotspots in large labeled networks," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 6, pp. 2479–2492, 2021. [Online]. Available: https://doi.org/10.1109/TKDE.2019.2956924

[23] Y. Xu, C. Ma, Y. Fang, and Z. Bao, "Efficient and effective algorithms for generalized densest subgraph discovery," *Proc. ACM Manag. Data*, vol. 1, no. 2, pp. 169:1–169:27, 2023. [Online]. Available: https://doi.org/10.1145/3589314

[24] K. Huang, H. Hu, Q. Ye, K. Tian, B. Zheng, and X. Zhou, "TED: towards discovering top-k edge-diversified patterns in a graph database," *Proc. ACM Manag. Data*, vol. 1, no. 1, pp. 51:1–51:26, 2023. [Online]. Available: https://doi.org/10.1145/3588736

[25] M. Chen, Y. Zhao, Y. Liu, X. Yu, and K. Zheng, "Modeling spatial trajectories with attribute representation learning (extended abstract)," in *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2023, pp. 3813–3814. [Online]. Available: https://doi.org/10.1109/ICDE55515.2023.00333

[26] H. Xie, Y. Fang, Y. Xia, W. Luo, and C. Ma, "On querying connected components in large temporal graphs," *Proc. ACM Manag. Data*, vol. 1, no. 2, pp. 170:1–170:27, 2023. [Online]. Available: https://doi.org/10.1145/3589315

[27] L. Chu, Z. Wang, J. Pei, Y. Zhang, Y. Yang, and E. Chen, "Finding theme communities from database networks," *Proc. VLDB Endow.*, vol. 12, no. 10, pp. 1071–1084, 2019. [Online]. Available: http://www.vldb.org/pvldb/vol12/p1071-chu.pdf

[28] J. Wang and Q. Zhang, "Disaggregated database systems," in *Companion of the 2023 International Conference on Management of Data, SIGMOD/PODS 2023, Seattle, WA, USA, June 18-23, 2023*, S. Das, I. Pandis, K. S. Candan, and S. Amer-Yahia, Eds. ACM, 2023, pp. 37–44. [Online]. Available: https://doi.org/10.1145/3555041.3589403

[29] R. Wang, J. Wang, P. Kadam, M. T. Özsu, and W. G. Aref, "dlsm: An lsm-based index for memory disaggregation," in *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2023, pp. 2835–2849. [Online]. Available: https://doi.org/10.1109/ICDE55515.2023.00217

[30] Y. Zhang, Q. Ye, R. Chen, H. Hu, and Q. Han, "Trajectory data collection with local differential privacy," *Proc. VLDB Endow.*, vol. 16, no. 10, pp. 2591–2604, 2023. [Online]. Available: https://www.vldb.org/pvldb/vol16/p2591-chen.pdf

[31] R. Du, Q. Ye, Y. Fu, H. Hu, J. Li, C. Fang, and J. Shi, "Differential aggregation against general colluding attackers," in *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*. IEEE, 2023, pp. 2180–2193. [Online]. Available: https://doi.org/10.1109/ICDE55515.2023.00169

[32] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang, "Data pricing in machine learning pipelines," *Knowl. Inf. Syst.*, vol. 64, no. 6, pp. 1417–1455, 2022. [Online]. Available: https://doi.org/10.1007/s10115-022-01679-4

[33] J. Pei, "A survey on data pricing: from economics to data science," *IEEE Trans. Knowl. Data Eng.*, 2021.

[34] J. Pei, F. Zhu, Z. Cong, X. Luo, H. Liu, and X. Mu, "Data pricing and data asset governance in the AI era," in *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, F. Zhu, B. C. Ooi, and C. Miao, Eds. ACM, 2021, pp. 4058–4059. [Online]. Available: https://doi.org/10.1145/3447548.3470818

[35] R. P. McAfee and J. McMillan, "Auctions and bidding," *Journal of economic literature*, vol. 25, no. 2, pp. 699–738, 1987.

[36] (Snowflake) https://www.snowflake.com/en/data-cloud/marketplace/.

[37] (AWS Data Exchange) https://aws.amazon.com/data-exchange/.

[38] A. Nagurney and P. Dutta, "Supply chain network competition among blood service organizations: a generalized nash equilibrium framework," *Ann. Oper. Res.*, vol. 275, no. 2, pp. 551–586, 2019. [Online]. Available: https://doi.org/10.1007/s10479-018-3029-2

[39] Y. Zhao, K. Zheng, J. Guo, B. Yang, T. B. Pedersen, and C. S. Jensen, "Fairness-aware task assignment in spatial crowdsourcing: Game-theoretic approaches," in *37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021*. IEEE, 2021, pp. 265–276. [Online]. Available: https://doi.org/10.1109/ICDE51399.2021.00030

[40] A. Sinha, F. Fang, B. An, C. Kiekintveld, and M. Tambe, "Stackelberg security games: Looking beyond a decade of success," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed. ijcai.org, 2018, pp. 5494–5501. [Online]. Available: https://doi.org/10.24963/ijcai.2018/775

[41] M. Xiao, Y. Xu, J. Zhou, J. Wu, S. Zhang, and J. Zheng, "Aoi-aware incentive mechanism for mobile crowdsensing using stackelberg game," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications, New York City, NY, USA, May*

17-20, 2023. IEEE, 2023, pp. 1–10. [Online]. Available: https://doi.org/10.1109/INFOCOM53939.2023.10229079

[42] V. Conitzer and T. Sandholm, "Complexity results about nash equilibria," in *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, G. Gottlob and T. Walsh, Eds. Morgan Kaufmann, 2003, pp. 765–771. [Online]. Available: http://ijcai.org/Proceedings/03/Papers/111.pdf

[43] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a nash equilibrium," *SIAM Journal on Computing*, vol. 39, no. 1, pp. 195–259, 2009.

[44] L. Xie, S. Meng, W. Yao, and X. Zhang, "Differential pricing strategies for bandwidth allocation with LFA resilience: A stackelberg game approach," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 4899–4914, 2023. [Online]. Available: https://doi.org/10.1109/TIFS.2023.3299181

[45] A. Sinha, P. Malo, and K. Deb, "A review on bilevel optimization: From classical to evolutionary approaches and applications," *IEEE Trans. Evol. Comput.*, vol. 22, no. 2, pp. 276–295, 2018. [Online]. Available: https://doi.org/10.1109/TEVC.2017.2712906

[46] A. Marshall, *Principles of economics: unabridged eighth edition*. Cosimo, Inc., 2009.

[47] T. Roughgarden, "Algorithmic game theory," *Communications of the ACM*, vol. 53, no. 7, pp. 78–86, 2010.

[48] X. Ding, H. Wang, D. Zhang, J. Li, and H. Gao, "A fair data market system with data quality evaluation and repairing recommendation," in *Web Technologies and Applications: 17th Asia-Pacific Web Conference, APWeb 2015, Guangzhou, China, September 18-20, 2015, Proceedings 17*. Springer, 2015, pp. 855–858.

[49] D. V. Winterfeldt and G. W. Fischer, "Multi-attribute utility theory: models and assessment procedures," *Utility, probability, and human decision making*, pp. 47–85, 1975.

[50] L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental logarithmic utility functions," *The American Economic Review*, vol. 65, no. 3, pp. 367–383, 1975.

[51] T. Başar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1998.

[52] J.-M. Lasry and P.-L. Lions, "Mean field games," *Japanese journal of mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[53] "Complete version link," 2023, https://github.com/StellaBYR/Data-Markets-with-Stackelberg-Nash-Equilibria-Complete-/.

[54] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[55] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image and vision Computing*, vol. 15, no. 1, pp. 59–76, 1997.

[56] J. Castro, D. Gómez, and J. Tejada, "Polynomial calculation of the shapley value based on sampling," *Computers & OR*, vol. 36, no. 5, pp. 1726–1730, 2009. [Online]. Available: https://doi.org/10.1016/j.cor.2008.04.004

[57] C. Dwork, "Differential privacy," in *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, ser. Lecture Notes in Computer Science, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., vol. 4052. Springer, 2006, pp. 1–12. [Online]. Available: https://doi.org/10.1007/11787006_1

[58] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml