Share: Stackelberg-Nash based Data Markets

Jinfei Liu, Yuran Bi
Zhejiang University
{jinfeiliu,stellabyr}@zju.edu.cn

Chen Zhao, Junyi Zhao
Zhejiang University
{zhaochen49,junyizhao}@zju.edu.cn

Li Xiong
Emory University
lxiong@emory.edu

Kui Ren
Zhejiang University
kuiren@zju.edu.cn

ABSTRACT

With the prevalence of data-driven intelligence, data markets with various data products are gaining considerable interests as a promising paradigm for commoditizing data and facilitating data flow. In this paper, we present **S**tackelberg-Nas**h** based D**a**ta Ma**rke**ts (*Share*) to first realize the absolute-pricing data market construction for a buyer-leading and multi-seller market with selfish participants. We propose a three-stage Stackelberg-Nash game to model trading dynamics which not only optimizes the profits of all selfish participants but also ensures the buyer' priority 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 propose both conventional direct derivation and a novel mean-field based method for complicated cases along with provable approximation guarantees. Experiments on real and synthetic datasets verify the effectiveness and efficiency of *Share*.

PVLDB Reference Format:
Jinfei Liu, Yuran Bi, Chen Zhao, Junyi Zhao, Li Xiong, and Kui Ren. Share: Stackelberg-Nash based Data Markets. PVLDB, 14(6): 957-969, 2021.
doi:10.14778/3447689.3447700

1 INTRODUCTION

Data products (e.g., machine learning models, aggregate statistics, and query services) have paved the way for a variety of data-driven tasks in many different industries. High-performance data products require a large amount of high-quality data. While there are 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. Recent studies and practices [1, 8, 24, 26, 32] have demonstrated data markets as a promising paradigm to commoditize data and connect data suppliers and demanders.

A typical data market consists of three parties: buyers, brokers, and sellers [1, 8, 32, 55]. Buyers propose demands for data products and pay for them; brokers facilitate the transactions between buyers and sellers (as well as take charge of manufacturing data products from data); sellers offer data with different quality and sell data to brokers in exchange for compensation. While many recent works [1, 8, 24, 26, 32] have addressed different aspects of the data market, there are several challenges remained to be addressed. We use a motivating example below to describe the desired properties of buyer-leading and multi-seller data markets.

Motivations. An automaker (e.g., Ford Motors Company) wants to get insight into users' purchase preferences of vehicles to decide future investments to traditional fuel vehicles, pure electric vehicles, or hybrid vehicles. The automaker **(buyer)** turns to McKinsey, a consulting company, and demands for a data product (e.g., data analysis, aggregate statistics, or a machine learning model which we use as an example) based on the real-world purchasing data of vehicles which implies valuable information, e.g., distinctive purchase intentions to different types of vehicles. McKinsey **(broker)** needs to gather data to train the required model. McKinsey buys data from multiple vehicle retailers **(sellers)** who own sales data of various vehicles containing not only the sales records but also the discount policy, profit level, and customer type. Since the data may contain sensitive customer information, sellers can sell data with different quality manipulated by privacy-preserving mechanisms.

The automaker can benefit from the data model and meantime needs to pay for it, McKinsey gains by selling the model after spending resources to buy the data and train the model, and the vehicle retailers sell data for compensations while incurring privacy costs. They are incentivized to join the data market by the profit they can earn. It's safe to extend to a general data market where all three parties are *selfish*, i.e., have their own *revenue* and *cost*, and aim to maximize their own profit (the difference between *revenue* and *cost*). Moreover, how they act in the market affects each other. If the automaker sets a low price to buy the desired model for profit, McKinsey may pay little to buy the training data to recover costs, and therefore the vehicle retailers offer poor-quality data, which causes a low-performance model and in turn harms the profit of the automaker. Therefore, *it's necessary to design data markets that consider the three selfish but interdependent parties and maximize the profits of buyers, brokers, and sellers simultaneously to motivate their participation (Desired Property $\mathcal{P}_1$).*

While all three parties need to maximize their profit, they take different roles in the market flow. Data sellers such as the vehicle retailers typically do not regard data selling as the main business. Rather, in many cases (especially the demand-driven scenarios) data

Proceedings of the VLDB Endowment, Vol. 14, No. 6 ISSN 2150-8097.
doi:10.14778/3447689.3447700

transactions are initiated by data buyers (demanders) such as the automaker instead of being orientated by three parties together. One way to reflect the asymmetric positions of the three parties is to give buyers priority over others. Therefore, *it's necessary to construct data markets with the buyer-leading property ($\mathcal{P}_2$) which takes full account of buyers' leading position.*

Pricing rules play a significant role in the market. The prices for the data (same for the data product) should reflect the actual value of the data. Existing methods such as Shapley value [42] use relative price (as in [1, 32, 45]) which is determined in comparison with other data without an absolute value. In contrast, absolute prices are determined independently on the inherent characteristics of data, which can directly measure the value of data and be applied to data trading. Besides, incentive mechanisms should involve the automaker, McKinsey, and the vehicle retailers in the pricing process rather than forcing them to passively receive certain prices which may discourage their participation. Thus, *it's necessary to formulate data markets in which absolute prices can be directly decided by market participants based on their mutual interactions ($\mathcal{P}_3$).*

Considering such a buyer-leading and multi-seller market and the desired three properties, an important research question is: *how to build a well-functioning data market with an absolute pricing mechanism where both the profit needs for all selfish participants and the leading position for buyers are considered.*

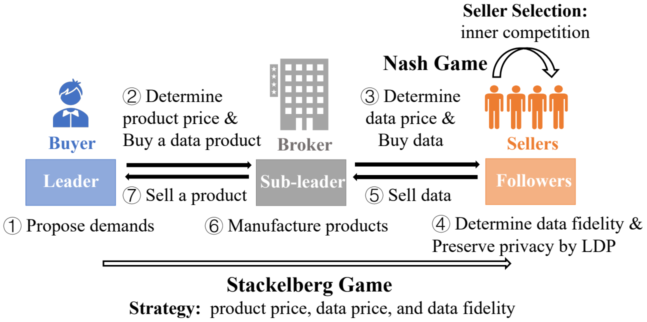


Figure 1: the framework of *Share*.

Challenges and Contributions. We summarize three challenges (C_1, C_2, C_3) faced in constructing the buyer-leading and multi-seller market, and propose a feasible solution, **S**tackelberg-Nas**h** based Dat**a** Ma**rke**ts (*Share*) utilizing game theory, as in Figure 1. The detailed workflow is presented in Section 4.2.

Existing works on data markets vary in design goals and typically address one aspect or one party's need, such as model quality optimization for buyers [21], revenue maximization for sellers [8], social welfare maximization [25], or market protection from strategic participants [20], but fail to realize profit optimization for all parties ($\mathcal{P}_1$). Besides, the lack of market practices and pricing references makes absolute pricing for data ($\mathcal{P}_3$) far from trivial (similar with petroleum pricing at the early stage). Therefore, the first challenge is (C_1): **How to design an incentive mechanism for data markets to realize absolute pricing and to maximize the profits of all three selfish and interdependent participants.** 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 her profit-maximization goal by making her optimal strategy. Moreover, absolute prices of data are modeled as strategies of participants and are directly determined in the game process.

Though efforts have been made to satisfy buyers' needs (e.g., utility demands and purchase budget) in [1, 32], data markets where transactions are initiated by buyers ($\mathcal{P}_2$) are understudied. Furthermore, the decisive influence buyers can have on the price was ignored. Therefore, the second challenge is (C_2): **How to embody the advantages of buyers over other parties in the buyer-leading data markets.** To solve this challenge, we formalize the interactions among three parties as a multi-stage dynamic game, called Stackelberg game [47] in game theory, which can deal with the asymmetric status of participants by regarding buyers as leaders, brokers as sub-leaders, and sellers as followers. As shown in Figure 1, the buyer first announces what data product she demands for and determines the product price based on her profit-maximization goal; the broker then tries to buy data from sellers; each seller then chooses what data quality to provide. Buyers are thus endowed with a dominant position in two aspects: the priority of initiating transactions and the intensive influence on prices as well as the corresponding product quality.

Since there are multiple sellers, it's critical to select the *best* data (with the highest data quality) from the sellers to bring buyers high product utility in order to give priority to buyers ($\mathcal{P}_2$). This is referred to as *seller selection problem.* Many existing works made the broker responsible for seller selection [2, 32], which not only requires the broker's capability of learning the data quality but also limits the sellers' ability of choosing their provided data quality. Hence, the third challenge is (C_3): **How to model the seller selection problem to select the best set of data for data transaction.** To solve this challenge, we consider the inner competition among sellers which can make the winners as the selected sellers without the assistance of brokers or others and allow sellers to manipulate their provided data quality by local differential privacy [15] in addition to the basic data value measured by Shapley value to increase their chances of winning. We further model the inter-seller competition as a Nash game [38] because of its advantage in modeling sellers' equal positions. Nash equilibrium among sellers is preferred, which guarantees that no seller can change her strategy individually to increase her profit and the seller selection result is stable, but is challenging to derive especially in the complex cases when the number of sellers is large and the profit function is complicated. We apply direct derivation and propose a mean-field based approximation for complex cases to find Nash equilibrium.

Our goal in this paper is not to cover all data markets nor to address all critical issues in real-world data trading, but rather to propose a feasible mechanism for the buyer-leading and multi-seller market satisfying the aforementioned desiderata which are meaningful in practice but not studied in existing works. The major contributions are summarized as follows.

- We present *Share*, a buyer-leading and multi-seller data market framework based on a three-stage Stackelberg-Nash game, which is the first work to satisfy the properties of buyer-leading, all-participant profit maximization, and absolute pricing.

- We apply Nash game for the seller selection problem, which formulates sellers' inner competition and incorporates seller selection into the game process among three parties.
- We define Stackelberg-Nash Equilibrium in data markets and derive it by backward induction. To solve 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. The rest of the paper is organized as follows. Section 2 presents the related work. Section 3 provides the preliminaries. Section 4 presents our data market framework with Stackelberg-Nash game and market mechanism. Approaches to deriving the market equilibrium and the market dynamics are presented in Section 5. Section 6 reports the experimental results and findings. Section 7 draws a conclusion and discusses future work.

2 RELATED WORK

In this section, we discuss related work on the data market and game theory.

2.1 Data Market

Data markets trade data either in direct or indirect forms (derived data products). Bloomberg [5], SafeGraph [41], and Dawex [13] implemented data markets where buyers directly purchase data. Koutris et al. [26, 27] proposed query-based data markets which allow buyers to obtain information through querying the database and pay for the query. Recently, model-based data markets [1, 8, 24, 32] have been proposed. Agarwal et al. [1] applied algorithmic game theory for two-sided data-driven machine learning model markets. Dealer [31, 32], an end-to-end model marketplace with differential privacy, more comprehensively considered the needs of buyers and sellers, and also regulated the broker's role as model pricing and model training.

In terms of profit maximization for all parties, however, Liu et al. [32] assumed that the broker is neutral without her own profit consideration and determines model prices only for single goal optimization, i.e., revenue maximization for sellers. An et al. [2], who studied transactions for crowdsensing data, were devoted to multiple goal optimization. Nevertheless, they oversimplified the markets in terms of transaction objects, market participants, and profit function formulations. 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 more reasonable profit functions based on economics theories.

As for data pricing, there are several surveys [10, 39, 40] claiming fundamental principles and reviewing the development and evolution of pricing models. Ghorbani and Zou [22] introduced Shapley value to data valuation. Federated learning based data pricing was developed by Wang et al. [50] while reinforcement learning was adopted by Yoon et al. [53] to price data. In terms of the absolute pricing mechanism, Liu et al. [32] provided absolute prices for data models, which, however, highly rely on the survey results and can't be adjusted dynamically. Agarwal et al. [1] applied Myerson's payment function rule to determine absolute model payment but allocated relative compensations to sellers in proportion to their contributions. In *Share*, we propose a feasible absolute pricing mechanism for both data and data product.

Many works looked at the seller selection problem. Liu et al. [32] made brokers choose datasets to maximize Shapley coverage of the trained model. An et al. [2] 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. In fact, seller selection can be seen as the spontaneous process of the inner competition among sellers, not conducted by the broker or others. In *Share*, the seller selection problem is formalized as the inner Nash game among sellers, which is a part of the incentive mechanism for profit optimization of all participants.

2.2 Game Theory

Game theory provides a decision-making and analysis tool for individual behaviors with conflicting objectives and has been widely used in various situations. Nash accurately described Nash equilibrium [38] and proved its existence in N-player finite non-cooperative game with mixed strategies [37]. Many researchers used Nash game as a powerful tool to formulate and solve problems where there is competition [23, 36]. Stackelberg game was first used to formulate the determining process for oligopoly firms producing homogeneous products [47], and has been further applied to many other practical situations with hierarchical organizations [43, 49, 54]. Besides the original Stackelberg game composed of one leader and one follower, many variants [3, 30, 46] were proposed and investigated. For example, Bansal et al. [3] used a two-stage Stackelberg game to determine prices for Unmanned Aerial Vehicles. Some studies [30, 46] combined Stackelberg game and Nash game together to deal with the problems in non-data markets where both hierarchy and simultaneity exist, but their issues including participant roles, major actions, and optimizing goals in traditional scenarios are quite different from those in data markets.

Since Nash proposed his theory, many researchers have sought algorithms for finding Nash equilibrium. Conitzer et al. [11] showed complexity results of deriving Nash equilibrium and Daskalakis et al. [12] 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 [2, 48, 52]. In fact, establishing Stackelberg equilibrium can be formulated as a bilevel optimization problem [44]. Some studies also combined other techniques into the equilibrium solving problem [3, 29, 46].

In this paper, we design the buyer-leading data markets based on Stackelberg game because Stackelberg game concerns interactions of participants with asymmetric status and can thus realize the desired buyer-leading property while remaining the profit maximization for all parties. Moreover, we first adopt Nash game for the seller selection problem since Nash game models the competition among equals with conflicting profits and can be used for the inner competition among data sellers, which can select sellers based on their optimal strategies.

3 PRELIMINARIES

In this section, we describe local differential privacy, Shapley value, and game theory used in *Share*. For reference, Table 1 summarizes the frequently used notations.

Table 1: The summary of frequently used notations.

	Notation	Definition
Buyer $\mathcal{B}$	N	data quantity for production
	p^M	unit price of data product
	v	product performance
	θ_1, θ_2	parameters of concern on each attribute
	ρ_1, ρ_2	parameters of sensitivity to each attribute
	$\mathrm{U}(\cdot)$	utility function of the product
	$\Phi(\cdot)$	profit function of the buyer
Broker $\mathcal{A}$	p^D	unit price of data
	$\sigma_k, k = 0, 1, ..., 5$	parameters related to cost
	$\mathrm{C}(\cdot)$	cost function of production
	$\Omega(\cdot)$	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
	$\mathrm{L}_i(\cdot)$	privacy loss function
	$\Psi_i(\cdot)$	profit function
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
	v_i	basic value of seller $\mathcal{S}_i$'s dataset

3.1 Local Differential Privacy

Differential Privacy (DP) [16–18] is a framework for privacy protection against the discovery of presence or absence of a record in a dataset via randomization or perturbation mechanisms. In our setting, each seller ensures Local Differential Privacy (LDP) [15] for each record in her dataset by perturbation before providing it to the broker, which protects the privacy of the data and determines the resulting data quality.

Definition 3.1 (Local Differential Privacy). A randomized algorithm $\mathcal{A} : \mathcal{Y} \to \mathcal{Z}$ satisfies ϵ-local differential privacy if and only if for any pairs of input tuples $y, y' \in \mathcal{Y}$, and for any $z \in \mathcal{Z}$, it always holds

$$\mathbb{P}[\mathcal{A}(y) = z] \leq e^{\epsilon} \cdot \mathbb{P}\left[\mathcal{A}\left(y'\right) = z\right], \tag{1}$$

where $\epsilon \geq 0$ is the privacy budget and $\mathbb{P}[\cdot]$ is the probability.

Widely used (local) DP mechanisms include Laplace mechanism [16], Index mechanism [34], and Gaussian mechanism [18]. The smaller ϵ is, the less the privacy loss is, and the worse the dataset quality is. In our context, each seller adopts a privacy scheme satisfying LDP to form her for-sale dataset with a distinctive quality corresponding to the privacy level.

3.2 Shapley Value

Shapley value [42] is an approach to fairly evaluate data importance. It satisfies the four fundamental requirements of fairness in markets, i.e., balance, symmetry, zero element, and additivity. In *Share*, Shapley value is used to measure the contribution of each seller's provided data to the data product by measuring its marginal utility improvement, e.g., the accuracy increase for a classification model or the residual error decrease for a regression model.

Definition 3.2 (Shapley Value). Consider a set of m data sellers such that each seller $\mathcal{S}_i$ owns a dataset D_i $(i = 1, 2, ..., m)$. A coalition $\mathbb{D}$ is a subset of $\{D_1, D_2, ..., D_m\}$. Denote by $\mathbf{u}(\mathbb{D})$ a utility function that represents the performance of a data product manufactured by coalition $\mathbb{D}$ towards a task, e.g., machine learning model accuracy. Shapley value of seller $\mathcal{S}_i$ is defined as follows.

$$\mathcal{SV}_i = \frac{1}{m} \sum_{\mathbb{D} \subseteq \{D_1, D_2, ..., D_m\} \backslash D_i} \frac{\mathbf{u}\left(\mathbb{D} \cup \{D_i\}\right) - \mathbf{u}\left(\mathbb{D}\right)}{\binom{m-1}{|\mathbb{D}|}}. \tag{2}$$

3.3 Nash Game and Stackelberg Game

Nash game [38] and Stackelberg game [47] as well as their corresponding equilibriums are applied in this work. An equilibrium to the game is a stable state in which no rational player would change her strategy to gain more payoff.

Definition 3.3 (Nash Game, Nash Equilibrium). Nash game is a game where two or more players take strategy simultaneously to maximize their own expected payoff. Nash equilibrium is the stable state where no player has anything to gain by changing only one's own strategy. Formally, let T_i be the set of all possible strategies for player i, where $i = 1, 2, ..., n$. Let $\boldsymbol{t}^* = \left(t_i^*, \boldsymbol{t}_{\neg i}^*\right)$ be a strategy profile, a set consisting of one strategy for each player, where $\boldsymbol{t}_{\neg i}^* = (t_1^*, ..., t_{i-1}^*, t_{i+1}^*, ..., t_n^*)$ denotes the $n-1$ strategies of all the players except i. Let $\pi_i\ (t_i, \boldsymbol{t}_{\neg i})$ be player i's payoff as a function of the strategies. The strategy profile $\boldsymbol{t}^*$ is a Nash equilibrium if

$$\pi_i\left(t_i^*, \boldsymbol{t}_{\neg i}^*\right) \geq \pi_i\left(t_i, \boldsymbol{t}_{\neg i}^*\right), \forall t_i \in T_i, i = 1, 2, ..., n. \tag{3}$$

Definition 3.4 (Stackelberg Game, Stackelberg Equilibrium). Stackelberg game is a game where one player (the *leader L*) moves first, and the other player (the *follower F*) can observe the *leader*'s behavior and move sequentially. Stackelberg equilibrium is a refinement of Nash equilibrium used in dynamic games. Similarly, the strategy profile (t_L^*, t_F^*) is a Stackelberg equilibrium if

$$\begin{aligned} \pi_L\left(t_L^*, t_F^*\right) &\geq \pi_L\left(t_L, t_F^*\right), \\ \pi_F\left(t_L^*, t_F^*\right) &\geq \pi_F\left(t_L^*, t_F\right). \end{aligned} \tag{4}$$

4 MARKET FRAMEWORK: PARTICIPANTS, MECHANISM, AND EQUILIBRIUM

In this section, we first describe the market framework from the perspectives of buyers, brokers, and sellers respectively in Section 4.1. Then, we formulate the market mechanism as a three-stage Stackelberg-Nash game in Section 4.2 and define the market equilibrium, Stackelberg-Nash Equilibrium in Section 4.3.

4.1 Market Participants

Considering the data markets composed of three types of participants, i.e., buyers, brokers, and sellers, we define the role of each party in *Share* as follows.

- **Buyer.** Buyer $\mathcal{B}$ wants to fulfill her data-driven task through data trading. Buyer $\mathcal{B}$ needs to purchase a data product from broker $\mathcal{A}$ with her claimed demands including the required product performance indicated as v and data quantity N (for model stability). Buyer $\mathcal{B}$ gains utility from the product while giving the payment to broker $\mathcal{A}$.

- **Broker.** Broker (Arbiter) $\mathcal{A}$ wants to make profits by bridging the transactions between buyers and sellers. Broker $\mathcal{A}$ needs to buy N data records from m sellers and make the product from the data to satisfy the product performance demand v which incurs certain manufacturing costs (e.g., computing resources). Then, broker $\mathcal{A}$ sells the product to buyer $\mathcal{B}$ in exchange for payment.

- **Seller.** Each seller $\mathcal{S}_i, i = 1, 2, ..., m$ owns a dataset D_i and wants to sell it for profit. Seller $\mathcal{S}_i$ needs to preprocess her dataset for quality manipulation and sell χ_i processed data records to broker $\mathcal{A}$. Note that χ_i is to be decided by the market mechanism and $\sum_{i=1}^{m} \chi_i = N$. Seller $\mathcal{S}_i$ receives compensation from broker $\mathcal{A}$ while incurring the privacy cost for the data she sells.

Since there are various types of markets with distinctive characteristics, we focus on the buyer-leading multi-seller market with the following assumptions.

- The market is buyer-leading, and a round of transactions starts when a buyer raises the demand. Buyers orientate the market in turn (coming one at a time) as in work [1], which is also widely applicable in many real-life scenarios where the buyers need personalized service, so only one buyer $\mathcal{B}$ is considered in each round of transactions.
- Only one broker $\mathcal{A}$ is considered in the market. The markets with multiple brokers and the competition among brokers for the same data product are beyond our consideration.
- There are a large number of sellers $\{\mathcal{S}_i | i = 1, 2, ..., m\}$, which is common in the real world (e.g., numerous vehicle retailers in our example). Each seller $\mathcal{S}_i$ has a dataset D_i to participate in the trading which is big enough, i.e., for any required number $\chi_i \in \mathbb{N}^+$ of data records, $|D_i| \geq \chi_i$.
- The market is highly transparent and the profit functions are known public. All the parties are honest, i.e., no two entities would collude to get a better outcome.

Problem Statement. Each of the participants has the *revenue* (gained utility, received payment, or compensation) and the *cost* (payment, manufacturing cost, or privacy loss). All the participants in the market are profit-driven and want to maximize their own profit, i.e., the difference between *revenue* and *cost*. The problem is to find an optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^* \right\rangle$, where buyer $\mathcal{B}$, broker $\mathcal{A}$, and each seller $\mathcal{S}_i$ determine product price p^{M^*}, data price p^{D^*}, and data fidelity τ_i^* respectively such that the profits of all participants are maximized. Detailed definitions of the profit functions of buyers, brokers, and sellers are given below.

4.1.1 *Profit Function of Buyer.*

When buyer $\mathcal{B}$ comes to the market and asks for a data product with the required performance, she cares about her *revenue*, the utility she can get from the product, and her *cost*, the payment she should give to the broker.

Revenue. The revenue of buyer $\mathcal{B}$ is correlated to the product utility which is quantified by a utility function $\mathbf{U}(\cdot)$. For data products, what buyer $\mathcal{B}$ concerns include both 1) product performance v and 2) the quality of the dataset used in the manufacture q^D. Apparently, product performance matters to buyer $\mathcal{B}$. For data models, v can be the explained variance, accuracy, or other indicators measuring model performance. However, it 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 $q^D = \sum_{i=1}^{m} q_i^D = \sum_{i=1}^{m} g(\chi_i, \tau_i)$, where q_i^D is the dataset quality provided by seller $\mathcal{S}_i$ and is positively correlated with data fidelity τ_i and data quantity χ_i. $g(\cdot)$ will be instantiated in Section 5.1.1. τ_i is determined jointly by the basic data value measured by Shapley value and the privacy level of LDP mechanism added by seller $\mathcal{S}_i$. χ_i is determined by sellers' inner competition. The detailed formulations for τ_i and χ_i will be elaborated in Eqs. 9 and 12.

Combining product performance and dataset quality to measure product utility can make the quantification of product utility more comprehensive. According to the utility theory [51] in economics, 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 the law of diminishing marginal utility [33] in economics.

$$\mathbf{U}\left(q^D, v\right) = \theta_1 \ln\left(1 + \rho_1 q^D\right) + \theta_2 \ln\left(1 + \rho_2 v\right). \tag{5}$$

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 model performance v 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 buyer $\mathcal{B}$ much more utility, its ρ_1 would be big, meaning that buyer $\mathcal{B}$ is highly sensitive to the quality of production materials.

Cost. The *cost* of buyer $\mathcal{B}$ is the payment to broker $\mathcal{A}$. $q^M = h(q^D, v)$ is defined to objectively represent the quality of the data product which depends on both dataset quality q^D and product performance v, and $h(\cdot)$ will be instantiated in Section 5.1.2. Also, p^M is defined as the unit price of q^M (or the product price). Therefore, the payment for the product can be formulated as the price times the product quality, i.e., $p^M q^M$, which corresponds to common sense that the payment for goods is equal to the unit price multiplied by the quantity.

Profit. The profit $\Phi(\cdot)$ of buyer $\mathcal{B}$ is the difference between the quantification of the product utility and the payment to the broker as follows. Note that $\boldsymbol{\tau} = (\tau_1, \tau_2, ...\tau_m)$.

$$\Phi\left(p^M, \boldsymbol{\tau}\right) = \mathbf{U}\left(q^D, v\right) - p^M q^M. \tag{6}$$

4.1.2 *Profit Function of Broker.*

When broker $\mathcal{A}$ receives the data product requirements from buyer $\mathcal{B}$, she cares about her *revenue*, the payment from buyer $\mathcal{B}$ and her *cost* which consists 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., $p^M q^M$ (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 compensations to sellers according to their provided dataset quality. The total compensation is defined as $p^D q^D$ which is the product of the total quality of the dataset p^D and the unit price p^D (or the data price). Broker $\mathcal{A}$ also needs to consume some resources to make the product, e.g., computation resources for training models, which is measured by cost function $\mathbf{C}(\cdot)$ related to data size N and product performance v, both significantly affecting the manufacturing cost. According to the work [9], we adopt a widely used transcendental logarithmic cost function as follows because of its adaptability to varied economies of scale and manufacturing strategy (e.g., how to allocate computing resources).

$$\mathbf{C}(N, v) = \exp\left(\sigma_0 + \sigma_1 \ln(N) + \sigma_2 \ln(v) + \frac{1}{2}\sigma_3 \ln^2(N) + \frac{1}{2}\sigma_4 \ln^2(v) + \sigma_5 \ln(N) \cdot \ln(v)\right). \tag{7}$$

where $\sigma_k, k \in \{0, 1, 2, 3, 4, 5\}$ are the parameters of the translog cost function which can be fitted by broker $\mathcal{A}$ based on the actual manufacturing procedure.

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

$$\Omega\left(p^M, p^D, \boldsymbol{\tau}\right) = p^M q^M - \mathbf{C}\left(N, v\right) - p^D q^D. \tag{8}$$

4.1.3 Profit Function of Seller.

When seller $\mathcal{S}_i$ gets the purchase request for data from broker $\mathcal{A}$, she cares about her *revenue*, the compensation from broker $\mathcal{A}$ and her *cost* coming from her privacy loss.

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

Cost. The *cost* of seller $\mathcal{S}_i$ is the privacy loss incurred based on data fidelity τ_i she provides. As discussed earlier, data fidelity τ_i is correlated to the basic value v_i of $\mathcal{S}_i$'s dataset and the privacy level ϵ_i in the data processing. $v_i \geq 0$ is measured by Shapley value indicating the contribution of $\mathcal{S}_i$'s dataset in the last transaction and is calculated by the broker and given to seller $\mathcal{S}_i$ as a fixed parameter. Hence τ_i can be defined as $f\left(\epsilon_i\right)$ where ϵ_i is the privacy parameter in LDP as in Section 3.1. Higher ϵ_i, less noise added, leading to better fidelity τ_i. There are many alternative function forms for $f\left(\cdot\right)$, as long as it follows the diminishing trend of marginal effect. Based on Inada Conditions [33] in economics which can stipulate the marginal change trend and encourage the market equilibrium, we conclude the following characteristics that $f\left(\cdot\right)$ should satisfy.

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

We choose an inverse trigonometric function form as $f(\cdot)$ and give the following definition of τ_i.

$$\tau_i = f(\epsilon_i) = \frac{2}{\pi} \operatorname{arcsec}(v_i \epsilon_i + 1),\ \epsilon_i \in [0, \infty), \tag{9}$$

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

Bigger τ_i means better fidelity of data and more privacy loss for $\mathcal{S}_i$. We quantify such loss by function $\mathbf{L}_i(\cdot)$ which is positively related to τ_i. It's intuitive that the loss function should not only increase but also increase faster for higher τ_i, which corresponds to the principle of increasing marginal cost [33] in economics. Moreover, the loss should be positively related to data quantity χ_i provided by seller $\mathcal{S}_i$. Various function forms can be applied following these principles. For simplicity in the solving process, we adopt a widely used quadratic function as follows.

$$\mathbf{L}_i\left(\tau_i\right) = \lambda_i(\chi_i \tau_i)^2, \tag{10}$$

where $\lambda_i > 0$ is $\mathcal{S}_i$'s privacy sensitivity. In our example, the privacy loss of the vehicle retailer corresponds to the negative impact of data exposure and $\mathbf{L}_i(\cdot)$ quantifies the retailer's economic estimation of the impact.

Profit. The profit of seller $\mathcal{S}_i$ is the difference between the compensation from broker $\mathcal{A}$ and the quantification of the privacy loss as follows.

$$\Psi_i\left(p^D, \tau_i\right) = p^D q_i^D - \mathbf{L}_i\left(\tau_i\right). \tag{11}$$

4.2 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 required product performance and the corresponding indicators. 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 buyer $\mathcal{B}$ and m sellers, determines the data price to buy the data from sellers. ④ Each seller chooses what data, strictly speaking, what 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 $\mathcal{B}$ receives the product and gives payment to $\mathcal{A}$, the transaction is finished.

Buyer's Strategy. Buyer $\mathcal{B}$ makes her strategy first, which is to determine the product price p^M, in order to maximize her 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 her strategy second, which is to determine data price p^D, in order to maximize her profit by considering the given p^M from buyer $\mathcal{B}$ and 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 her profit by balancing the revenue of selling data given the data price p^D and the cost of the privacy loss.

Meanwhile, 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 $\mathcal{S}_i$ could be chosen. Therefore, the data quantity χ_i sold by $\mathcal{S}_i$ can be calculated according to all sellers' $\boldsymbol{\tau}$ as below.

$$\chi_i = N\frac{\omega_i\tau_i}{\sum_{j=1}^{m}\omega_j\tau_j}, \tag{12}$$

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 and can therefore mirror previous buyers' influence on the current buyer to some extent. The broker would update these weights after each round of transactions. For example, new weights can be generated based on both old weights and sellers' contributions to the data product in the last transaction measured by Shapley value as in Section 3.2.

We define such inner competition in sellers as a Nash game. Seller $\mathcal{S}_i$ determines her strategy τ_i simultaneously with each other to maximize her own profit which is also affected by other sellers' strategies. Nash equilibrium would be achieved where no seller can increase her profit by unilaterally changing her strategy with all other sellers' strategies fixed.

Note that if one participant finds that her maximized profit is below zero, she will quit since she can gain no benefit from participating in the data trading, which guarantees the individual rationality [35] 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 the new transaction would be initiated by the next buyer. Otherwise, the remaining participants would continue and finish the transaction. Since it's easy to deal with the quit situation, we focus on the more common and complex 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 $\langle p^M, p^D, \boldsymbol{\tau}\rangle$ 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 her own profit by determining her optimal strategy variable. The three-stage Stackelberg-Nash game is defined as follows.

Definition 4.1 (Three-Stage Stackelberg-Nash Game). The game consists of three stages for buyer $\mathcal{B}$, broker $\mathcal{A}$, and sellers $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_m$, respectively.

Stage 1 Buyer $\mathcal{B}$: $p^{M^*} = \arg\max_{p^M} \Phi\left(p^M, \boldsymbol{\tau}\right)$.

Stage 2 Broker $\mathcal{A}$: $p^{D^*} = \arg\max_{p^D} \Omega\left(p^M, p^D, \boldsymbol{\tau}\right)$.

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 hierarchy and simultaneity. Hierarchy indicates that a certain participant, buyer $\mathcal{B}$ in our context, has some advantages over others that enable her to act first, broker $\mathcal{A}$ takes her strategy second, and sellers make their strategies last, while simultaneity indicates the equal positions of m sellers who take strategy simultaneously in their inner Nash game.

4.3 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 her own profit. Meanwhile, the optimal solution must satisfy some equilibrium so that no one is willing to adopt other strategies, which indicates market stability and sustainability, making our design reasonable. We define a Stackelberg-Nash Equilibrium (SNE) in data markets as follows.

Definition 4.2 (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.

$$\Phi\left(p^{M^*}, \boldsymbol{\tau}^*\right) \geq \Phi\left(p^M, \boldsymbol{\tau}^*\right), \tag{13}$$

$$\Omega\left(p^{M^*}, p^{D^*}, \boldsymbol{\tau}^*\right) \geq \Omega\left(p^{M^*}, p^D, \boldsymbol{\tau}^*\right), \tag{14}$$

$$\Psi_i\left(p^{D^*}, \boldsymbol{\tau}^*\right) \geq \Psi_i\left(p^{D^*}, \boldsymbol{\tau}^*_{\neg i}, \tau_i\right), i = 1, 2, ..., m, \tag{15}$$

where $\boldsymbol{\tau}_{\neg i}$ means other sellers' strategies except $\mathcal{S}_i$'s, i.e., τ_j, $j = 1, 2, ..., m, j \neq i$.

SNE indicates that each participant takes her optimal strategy which maximizes her own profit in a buyer-leading sequence. No one can add her own profit by unilaterally changing her strategy with all other participants' strategies fixed.

5 MARKET CONSTRUCTION: EQUILIBRIUM SOLVING APPROACH AND TRADING DYNAMICS

In this section, we first derive the market equilibrium by backward induction in Section 5.1. Then, we describe the market dynamics in Section 5.2.

5.1 Solving Equilibrium: Backward Induction Approach

To determine the optimal strategy profile $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^*\right\rangle$, we adopt the backward induction approach [4]. 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. 19) for any given data price p^D in Section 5.1.1. 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. 24) of broker $\mathscr{A}$ for any given product price p^M in Section 5.1.2. 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 $\mathscr{B}$'s optimal strategy p^{M^*} (Eq. 26) based on the optimal reactions of the broker as well as sellers in Section 5.1.3. After that, we can get the value of the optimal strategy p^{D^*} by substituting p^{M^*} into the result (Eq. 24) 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. 19) 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.

5.1.1 *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 complex cases with large numbers of sellers and complicated profit function forms.

Direct Derivation. By substituting Eqs. 10,12 into Eq. 11 and instantiating $q_i^D = g(\chi_i, \tau_i)$ as $\chi_i\tau_i$ since q_i^D is positively correlated with χ_i and τ_i, we get each seller's profit

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

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 her 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

$$p^D\frac{\partial\left(N\frac{\omega_i\tau_i^2}{\sum_{j=1}^m \omega_j\tau_j}\right)}{\partial\tau_i} - 2\lambda_i\cdot N\frac{\omega_i\tau_i^2}{\sum_{j=1}^m \omega_j\tau_j}\cdot\frac{\partial\left(N\frac{\omega_i\tau_i^2}{\sum_{j=1}^m \omega_j\tau_j}\right)}{\partial\tau_i} = 0. \quad (16)$$

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

$$\begin{cases} p^D\sum_{i=1}^m \omega_i\tau_i - 2N\lambda_1\omega_1\tau_1^2 = 0 \\ p^D\sum_{i=1}^m \omega_i\tau_i - 2N\lambda_2\omega_2\tau_2^2 = 0 \\ \vdots \\ p^D\sum_{i=1}^m \omega_i\tau_i - 2N\lambda_m\omega_m\tau_m^2 = 0, \end{cases} \quad (17)$$

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

$$2N\lambda_1\omega_1\tau_1^2 = 2N\lambda_2\omega_2\tau_2^2 = ... = 2N\lambda_m\omega_m\tau_m^2 = p^D\sum_{i=1}^m \omega_i\tau_i. \quad (18)$$

By adding all m equations in Eq. 17, we get

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

and using τ_1 to indicate other τ_i $(i = 2, 3, ...,, m)$ from Eq. 18, we get

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

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. 18 again, we get all sellers' optimal strategies τ_i^* as

$${\tau_i}^* = \frac{p^D}{2N\sqrt{\omega_i\lambda_i}}\sum_{j=1}^m\sqrt{\frac{\omega_j}{\lambda_j}}, i = 1, 2, ..., m. \quad (19)$$

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

Mean-field based Approximate Method. It's 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. 16 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. 10 with $\mathbf{L}_i(\tau_i) = \lambda_i\chi_i{\tau_i}^2$.

The approximation is based on the mean-field theory [28], 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's 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, $\overline{\tau}$.

$$\overline{\tau} = \frac{\sum_{i=1}^m \omega_i\tau_i}{m}. \quad (20)$$

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. 11 is changed into

$$\Psi_i\left(p^D, \tau_i\right) = p^D(\chi_i\tau_i) - \lambda_i\chi_i{\tau_i}^2. \quad (21)$$

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.

$$\begin{cases} p^D \cdot N\frac{\omega_1\tau_1^2}{m\overline{\tau}} - \lambda_1 \cdot N\frac{\omega_1\tau_1^3}{m\overline{\tau}} = 0 \\ p^D \cdot N\frac{\omega_2\tau_2^2}{m\overline{\tau}} - \lambda_2 \cdot N\frac{\omega_2\tau_2^3}{m\overline{\tau}} = 0 \\ \vdots \\ p^D \cdot N\frac{\omega_m\tau_m^2}{m\overline{\tau}} - \lambda_m \cdot N\frac{\omega_m\tau_m^3}{m\overline{\tau}} = 0. \end{cases}$$

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

$$\tau_i^* = \frac{2p^D}{3\lambda_i}, i = 1, 2, ..., m. \tag{22}$$

Note that we justify that the second-order derivative $\frac{\partial^2\Psi_i(p^D,\tau_i)}{\partial\tau_i^2} < 0$, 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. In this part, we analyze the error bound of the approximated mean-field approach.

Theorem 5.1. *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 $\mathbf{L}_i(\tau_i) = \lambda_i\chi_i{\tau_i}^2$. 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 ω_i, $i = 1, 2, ..., m$, allowing us to arbitrarily scale them.

Proof. We first calculate the upper bound of $\overline{\tau}^{DD} - \overline{\tau}^{MF}$. By applying direct derivation to Eq. 21, we can get

$$2p^D \sum_{j=1}^m \omega_j\tau_j - p^D\omega_i\tau_i = 3\lambda_i\tau_i \sum_{j=1}^m \omega_j\tau_j - \lambda_i\omega_i{\tau_i}^2.$$

By splitting $\sum_{j=1}^m \omega_j\tau_j$ into $\sum_{j=1,j\neq i}^m \omega_j\tau_j$ and $\omega_i\tau_i$, we obtain a quadratic equation about τ_i by deforming the above formula, and using root formula for the quadratic equation, we can get

$$\tau_i^* = \frac{p^D\omega_i - 3\lambda_i\Sigma_{\tau_{\neg i}} + \sqrt{(3\lambda_i\Sigma_{\tau_{\neg i}} - p^D\omega_i)^2 + 16p^D\lambda_i\omega_i\Sigma_{\tau_{\neg i}}}}{4\lambda_i\omega_i}, \tag{23}$$

where $\Sigma_{\tau_{\neg i}} = \sum_{j=1,j\neq i}^m \omega_j\tau_j$. With the constraint $\frac{\omega_i}{\lambda_i} \leq \frac{1}{p^D m^2}$, we can justify that $3\lambda_i\Sigma_{\tau_{\neg i}} - p^D\omega_i > 0$ when m is very large. Thus according to $\sqrt{x+y} < \sqrt{x} + \sqrt{y}$, we can scale and deform the above formula to get

$$\omega_i\tau_i^* < \frac{\sqrt{16p^D\lambda_i\omega_i\Sigma_{\tau_{\neg i}}}}{4\lambda_i}.$$

Further simplifying and scaling the above formula, we can get

$$\omega_i\tau_i^* < \sqrt{p^D\frac{\omega_i}{\lambda_i}\Sigma_{\tau_{\neg i}}} \leq \sqrt{p^D\frac{\omega_i}{\lambda_i}\sum_{j=1}^m \omega_j\tau_j^*},$$

which applies to all $\tau_i^*, i = 1, 2, ..., m$. Then, by adding m inequalities together and simplifying it, we can obtain

$$\sum_{i=1}^m \omega_i\tau_i^* < \left(\sum_{i=1}^m \sqrt{p^D\frac{\omega_i}{\lambda_i}}\right)^2,$$

and thus

$$\overline{\tau}^{DD} = \frac{1}{m}\sum_{i=1}^m \omega_i\tau_i^* < \frac{1}{m}\left(\sum_{i=1}^m \sqrt{p^D\frac{\omega_i}{\lambda_i}}\right)^2.$$

Additionally, using Eqs. 20 and 22, we can derive $\overline{\tau}^{MF}$ as below.

$$\overline{\tau}^{MF} = \frac{1}{m}\sum_{i=1}^m \frac{2p^D\omega_i}{3\lambda_i}.$$

Then we use $\frac{\omega_i}{\lambda_i} \leq \frac{1}{p^D m^2}$ and get

$$\begin{aligned} \overline{\tau}^{DD} - \overline{\tau}^{MF} &< \frac{1}{m}\left(\sum_{i=1}^m \sqrt{p^D\frac{\omega_i}{\lambda_i}}\right)^2 - \frac{1}{m}\sum_{i=1}^m \frac{2p^D\omega_i}{3\lambda_i} \\ &\leq \frac{1}{m}\left(\sum_{i=1}^m \sqrt{\frac{1}{m^2}}\right)^2 - \frac{1}{m}\sum_{i=1}^m \frac{2}{3m^2} \\ &= \frac{1}{m} - \frac{2}{3m^2}. \end{aligned}$$

Next, we calculate the lower bound. Since $(3\lambda_i\Sigma_{\tau_{\neg i}} - p^D\omega_i)^2 + 16p^D\lambda_i\omega_i\Sigma_{\tau_{\neg i}} > (p^D\omega_i + 3\lambda_i\Sigma_{\tau_{\neg i}})^2$, using Eq. 23 we can get

$$\begin{aligned} \overline{\tau}^{DD} &= \frac{1}{m}\sum_{i=1}^m \omega_i\tau_i^* \\ &> \frac{1}{m}\sum_{i=1}^m \frac{p^D\omega_i - 3\lambda_i\Sigma_{\tau_{\neg i}} + \sqrt{(p^D\omega_i + 3\lambda_i\Sigma_{\tau_{\neg i}})^2}}{4\lambda_i} \\ &= \frac{1}{m}\sum_{i=1}^m \frac{p^D\omega_i}{2\lambda_i}, \end{aligned}$$

and using $\frac{\omega_i}{\lambda_i} \leq \frac{1}{p^D m^2}$ again, we get

$$\begin{aligned} \overline{\tau}^{DD} - \overline{\tau}^{MF} &= \frac{1}{m}\sum_{i=1}^m \omega_i\tau_i^* - \frac{1}{m}\sum_{i=1}^m \frac{2p^D\omega_i}{3\lambda_i} \\ &> \frac{1}{m}\sum_{i=1}^m \frac{p^D\omega_i}{2\lambda_i} - \frac{1}{m}\sum_{i=1}^m \frac{2p^D\omega_i}{3\lambda_i} \geq -\frac{1}{6m^2}. \end{aligned}$$

Therefore, Theorem 5.1 holds. □

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 [28]. When the number of sellers m is big, our proposed mean-field method appears reasonable in terms of error.

5.1.2 *Expression of* p^{D^*} *in Stage 2.*

We use direct derivation to derive the expression of p^{D^*} for the broker.

Direct Derivation. By substituting Eq. 19 into Eq. 12, we get

$$\chi_i{}^* = N\frac{\omega_i\tau_i^*}{\sum_{j=1}^{m}\omega_j\tau_j^*} = N\frac{\sqrt{\frac{\omega_i}{\lambda_i}}}{\sum_{j=1}^{m}\sqrt{\frac{\omega_j}{\lambda_j}}}.$$

Then we get

$$q^{D^*} = \sum_{i=1}^{m}\chi_i^*\tau_i^* = \sum_{i=1}^{m}\frac{p^D}{2\lambda_i}.$$

Since $q^M = h(q^D, v)$ is positively correlated to q^D and v, we instantiate $h(q^D, v)$ as $q^D v$. We get

$$q^{M^*} = q^{D^*}v = \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\lambda_i}p^D v.$$

By substituting q^{D^*} and q^{M^*} into $\mathcal{A}$'s profit function in Eq. 8, we get

$$\Omega\left(p^M, p^D, \boldsymbol{\tau}\right) = p^M\cdot\left(\frac{1}{2}\sum_{i=1}^{m}\frac{1}{\lambda_i}p^D v\right) - \mathrm{C}\left(N, v\right) - p^D\cdot\left(\frac{1}{2}\sum_{i=1}^{m}\frac{1}{\lambda_i}p^D\right).$$

We derive the first-order derivative with respect to p^D and let it equal to 0.

$$\frac{\partial\Omega\left(p^M, p^D, \boldsymbol{\tau}\right)}{\partial p^D} = \frac{1}{2}\sum_{i=1}^{m}\frac{1}{\lambda_i}vp^M - \sum_{i=1}^{m}\frac{1}{\lambda_i}p^D = 0.$$

We can thus get the expression of p^{D^*}

$$p^{D^*} = \frac{vp^M}{2}. \tag{24}$$

We justify that the second-order derivative $\frac{\partial^2\Omega(p^M,p^D,\boldsymbol{\tau})}{\partial {p^D}^2} = -\sum_{i=1}^{m}\frac{1}{\lambda_i} < 0$, so the solution can maximize the broker's profit.

5.1.3 *Value of* p^{M^*} *in Stage 1.*

We also use direct derivation in this stage, and by using the results in Sections 5.1.1 and 5.1.2, we can directly derive the value rather than the expression of p^{M^*} for the buyer.

Direct Derivation. By substituting Eq. 19 and Eq. 24 into $\mathcal{B}$'s profit function in Eq. 6, we can obtain the profit of buyer $\mathcal{B}$

$$\begin{aligned}\Phi\left(p^M, \boldsymbol{\tau}\right) &= \theta_1\ln\left(1+\rho_1 q^{D^*}\right) + \theta_2\ln\left(1+\rho_2 v\right) - p^M q^{M^*}\\ &= \theta_1\ln\left(1+\mathsf{c}_1 p^M\right) + \theta_2\ln\left(1+\rho_2 v\right) - \frac{\mathsf{c}_2\theta_1}{2}{p^M}^2,\end{aligned}$$

where $\mathsf{c}_1 = \frac{\rho_1 v}{4}\sum_{i=1}^{m}\frac{1}{\lambda_i}$ and $\mathsf{c}_2 = \frac{v^2}{2\theta_1}\sum_{i=1}^{m}\frac{1}{\lambda_i}$. Then, we derive the first-order derivative of $\Phi\left(p^M, \boldsymbol{\tau}\right)$ as follows.

$$\frac{\partial\Phi(p^M, \boldsymbol{\tau})}{\partial p^M} = \frac{\theta_1\mathsf{c}_1}{1+\mathsf{c}_1 p^M} - \mathsf{c}_2\theta_1 p^M. \tag{25}$$

By letting $\frac{\partial\Phi(p^M,\boldsymbol{\tau})}{\partial p^M}$ in Eq. 25 equal to zero, we obtain

$$\mathsf{c}_1\mathsf{c}_2\cdot{p^M}^2 + \mathsf{c}_2\cdot p^M - \mathsf{c}_1 = 0.$$

Using the characteristic root method, we find buyer $\mathcal{B}$'s optimal strategy p^{M^*} (after discarding the negative solution).

$$p^{M^*} = \frac{-\mathsf{c}_2 + \sqrt{\mathsf{c}_2^2 + 4\mathsf{c}_1^2\mathsf{c}_2}}{2\mathsf{c}_1\mathsf{c}_2}. \tag{26}$$

We justify that the second-order derivative $\frac{\partial^2\Phi(p^M,\boldsymbol{\tau})}{\partial {p^M}^2} = -\frac{\theta_1\mathsf{c}_1^2}{(1+\mathsf{c}_1 p^M)^2} - \theta_1\mathsf{c}_2 < 0$, so the solution can maximize the buyer's profit.

After that, we can determine the optimal value of p^{D^*} by substituting p^{M^*} into Eq. 24 as well as each seller's optimal value of τ_i^* by substituting p^{D^*} into Eq. 19. 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.

5.1.4 *Equilibrium Analysis.*

In this part, we prove the existence and uniqueness of SNE in *Share*.

Theorem 5.2. *The complete optimal strategy profile* $\left\langle p^{M^*}, p^{D^*}, \boldsymbol{\tau}^*\right\rangle$ *determined by backward induction approach uniquely constitutes SNE.*

Proof.

Existence. For the buyer, when the broker and sellers hold the optimal strategies in Eq. 24 and Eq. 19, the buyer's profit $\Phi\left(p^M, \boldsymbol{\tau}^*\right)$ only changes with p^M. In the process of deriving optimal p^{M^*}, the first-order derivation is set to be 0 and the second-order is strictly less than 0, which means that the maximum profit is obtained at p^{M^*}. Thus, Eq. 13 holds at p^{M^*}. For the broker, when the buyer and sellers hold the optimal strategies, p^{M^*} in Eq. 26 and $\boldsymbol{\tau}^*$ in Eq. 19, the broker's optimal profit can be obtained at p^{D^*} since the profit function is strictly concave and has a single extreme point p^{D^*}. Thus, Eq. 14 holds at p^{D^*}. For the sellers, each seller determines each τ_i^* simultaneously in the same way, and $\tau_i^*, i = 1, 2, ..., m$ are jointly decided. For each seller $\mathcal{S}_i$, her optimal strategy τ_i^* is determined by letting the first-order derivation equal to 0. Since the profit function is strictly concave, the extreme point τ_i^* maximizes seller $\mathcal{S}_i$'s profit if $\tau_i^* \leq 1$. Otherwise, when the extreme point is larger than 1, the optimal value $\tau_i^* = 1$ can also maximize $\mathcal{S}_i$'s profit since the profit function is monotonically increasing in the feasible range of τ_i and maximized at the right endpoint 1. Thus, Eq. 15 holds at τ_i^*. Therefore, it's proved that SNE exists in our mechanism.

Uniqueness. For the buyer, since her profit function is strictly concave, the maximum profit is obtained only at p^{M^*}. Any other value of $p^M \neq p^{M^*}$ will yield an inferior profit. Such result can be also explained by Convex Optimization [6], i.e., the strategy space of p^M is a convex and compact subspace of Euclidean space, and the profit function $\Phi\left(\cdot\right)$ is a convex function of p^M, leading to the unique optimal p^{M^*} that maximizes $\Phi\left(\cdot\right)$. Thus, Eq. 13 holds only at p^{M^*}. For the broker, her profit function is also strictly concave and only has a single extreme point p^{D^*}. Any other value of $p^D \neq p^{D^*}$ will lower the broker's profit. Thus, Eq. 14 holds only at p^{D^*}. For sellers, seller $\mathcal{S}_i$ can only have lower profit by deciding $\forall\tau_i \neq \tau_i^*$. If $\tau_i^* \leq 1$, seller $\mathcal{S}_i$'s profit function is concave and only maximized at τ_i^*. Otherwise, the profit can only be maximized at the right endpoint, $\tau_i^* = 1$, since the profit function is monotonically increasing. For each seller, if she chooses other $\forall \tau_i \neq \tau_i^*$, she can only have lower profit with $\boldsymbol{\tau}_{\neg i}^*$ kept the same. Thus, the unique Nash equilibrium among sellers is achieved and Eq. 15 holds only at τ_i^*. Therefore, it's proved that other strategy profiles except our solution cannot satisfy SNE, which indicates the uniqueness of SNE in our mechanism. □

5.2 Complete Data Trading Dynamics

We summarize the complete dynamics of data markets in Algorithm 1, which integrates the equilibrium solving process in Section 5.1.

The first phase is *Parameter Collection*. We assume that each party can report their specific input parameters which can be fitted based on the historical experiences. The buyer sets appropriate parameters $\theta_1, \theta_2, \rho_1, \rho_2$ for her utility function and proposes demand parameters v and N for the product (Line 2). Note that the form of the product is not restricted and can range from simple data aggregation to deep learning models. The broker determines $\sigma_k, k \in \{0, 1, 2, 3, 4, 5\}$ for her translog cost function and maintains the weights $\omega_i, i = 1, 2, ..., m$ and basic values $v_i, i = 1, 2, ..., m$ of sellers' datasets (Line 3). Specifically, the weights and basic values are initialized to be the same respectively when the market is established. To decide the real weights and basic values before the first transaction, the broker can use dummy buyers to iterate several times, and in each iteration, Shapley values of sellers' datasets are calculated after production and can be used to update the weights and basic values in the next iteration. Sellers give their privacy sensitivity $\lambda_i, i = 1, 2, ..., m$ (Line 4). We assume that participants provide their truthful parameters in line with the practical situation under the supervision of market regulators (e.g., by regular spot-check).

The second phase is *Strategy Decision*. Using the strategy mechanism, buyer $\mathcal{B}$, broker $\mathcal{A}$, and sellers $\mathcal{S}_i, i = 1, 2, ..., m$ give strategies of product price p^{M^*}, data price p^{D^*}, and data fidelity $\boldsymbol{\tau}_i^*$ in order according to Eqs. 26, 24, 19, 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. 12 (Line 8). Next, each seller randomly picks χ_i^*-sized dataset (Line 10) and pre-processes it for privacy protection based on ${\epsilon_i}^*$ calculated from Eq. 9 (Lines 11-12). After that, seller $\mathcal{S}_i$ gives her protected dataset D_i^t to the broker in exchange for the compensation $p^{D^*} q_i^{D^*}$ (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: $\omega_i' = 0.2\omega_i + 0.8\mathcal{SV}_i, i = 1, 2, ..., m$, where $\mathcal{SV}_i$ is the Shapley value of D_i^t to the product, and the updated weights ω_i' can be used in the subsequent transaction. Similarly, the basic values of sellers' datasets are updated by Shapley value, e.g., $v_i = \mathcal{SV}_i$, and are given to sellers as fixed parameters in the next transaction.

The last phase is *Product Transaction* between the broker and the buyer. The broker gives the product to the buyer and the buyer pays $p^{M^*} q^{M^*}$ 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 and can use the updated $\omega_i', i = 1, 2, ..., m$.

Algorithm 1: Data trading dynamics.

```
1  %% Parameter Collection;
2  From the current buyer $\mathcal{B}$, parameters $N, v, \theta_1, \theta_2, \rho_1, \rho_2$ are provided;
3  From broker $\mathcal{A}$, $\sigma_k\,(k \in \{0, 1, 2, 3, 4, 5\}), \omega_i, v_i\ (i = 1, 2, ..., m)$ 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. 12;
9  for each seller $\mathcal{S}_i, i = 1, 2, ...m$ do
10     Randomly pick ${\chi_i}^*$ data pieces from her dataset $D_i$;
11     Calculate ${\epsilon_i}^*$ from the strategy ${\tau_i}^*$ according to Eq. 9;
12     Conduct LDP with ${\epsilon_i}^*$ on her ${\chi_i}^*$-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 $p^{D^*} q_i^{D^*}$ to each seller;
14 %% Product Production;
15 Broker $\mathcal{A}$ then uses $D^t$ to produce the data product;
16 After manufacturing the product, broker $\mathcal{A}$ updates $\omega_1, \omega_2, ..., \omega_m$ and
     $v_1, v_2, ..., v_m$ (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 $\mathcal{B}$, and meantime buyer $\mathcal{B}$ pays
     $p^{M^*} q^{M^*}$ to broker $\mathcal{A}$.
```

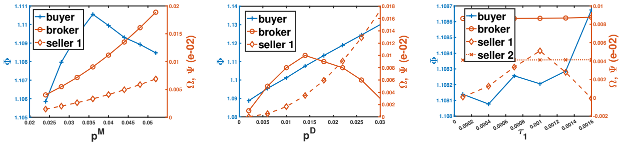
Time Complexity. As seen from Algorithm 1, the phase of *Parameter Collection* costs $O(m)$ since m sellers need to provide $\lambda_i, i = 1, 2, ..., m$. *Strategy Decision* costs $O(m)$ based on the optimal strategy profile. *Data Transaction* costs $O(m + N)$ because each seller needs to form and protect her χ_i-sized dataset, and N data records in total are preprocessed and sold to the broker. The time cost of *Product Production* depends on the exact product type, production mode, and the way of updating the weights $\omega_1, \omega_2, ..., \omega_m$. The last phase of *Product Transaction* takes constant time. Therefore, the complexity of data trading algorithm excluding *Product Production* is $O(m + N)$.

6 EXPERIMENTS

In this section, we present experimental studies validating the effectiveness and efficiency of *Share*. We first describe our experiment setup including the datasets and parameter settings in Section 6.1. Sections 6.2 and 6.3 show the results verifying the effectiveness and efficiency of *Share*, respectively. Section 6.4 shows the effect of the main parameters used in *Share*.

6.1 Experiment Setup

We conduct experiments on a machine with an Intel Core i7-11700KF running Ubuntu with 64GB memory. We choose Linear Regression model as the data product and use explained variance to measure model performance. We will show the results of using direct derivation in the mechanism. Note that the mean-field approach (used when direct derivation fails) performs the same in terms of effectiveness, efficiency, and parameter influence.



(a) Vary $\mathcal{B}$'s strategy (b) Vary $\mathcal{A}$'s strategy (c) Vary $\mathcal{S}_1$'s strategy

Figure 2: Effectiveness.

Datasets. We use a real dataset, Combined Cycle Power Plant (CCPP) [14], which contains 9,568 data points with four features. The Linear Regression task is to predict the net hourly electrical energy output. We randomly choose a training dataset with a size of 9,000, and the 568 data records left are used for validation. 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 model performance v, and individual parameters of each party, i.e., buyer $\mathcal{B}$'s $\theta_1, \theta_2, \rho_1, \rho_2$ related to model 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 $\lambda_i, i = 1, 2, ..., m$. We set $m = 100$, $N = 500$, and $v = 0.8$. 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 quality 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 $(0, 1)$.

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 stimulate the distinction in data quality, we first sort data by quality measured by Shapley value, which indicates the contribution of each data record to model training. The Shapley value is calculated based on Monte Carlo Method [7, 19]. Then by distributing data in decreasing quality over sellers, each seller owns 90 data records with different quality. Laplace mechanism [16] is applied to each record to adjust data fidelity for each seller.

$\omega_1, \omega_2, ..., \omega_m$ and $v_1, v_2, ..., v_m$ are generated by using buyer $\mathcal{B}$ as the 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 model training to update the weights and basic values for the next transaction.

6.2 Effectiveness

We implement the mechanism and unilaterally change the strategies of the buyer, the broker, and sellers respectively to verify the profit maximization of all parties as well as the corresponding equilibrium.

Fig.2(a) shows the results of profits when we change p^M around the optimal strategy p^{M^*} while maintaining the rest. Seller $\mathcal{S}_1$ acts as a representative of sellers. It's found that the peak of the buyer's profit $\Phi(\cdot)$ appears when her optimal strategy $p^{M^*} = 0.036$ determined in SNE 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^*}, she will get a lower profit when all other participants' strategies are fixed. The change of the profits of the broker and the seller is intuitive. Specifically, with growing p^M, the broker can gain more profit, which can further add the compensations for sellers and make their profits higher.

Fig.2(b) shows the results of profits when we change p^D around the optimal strategy $p^{D^*} = 0.014$ while maintaining the rest. Similarly, it is found that the broker cannot increase her profit by unilaterally changing her strategy. The change of the profits of the buyer and seller is also intuitive. 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 results of profits when we change τ_1 around the optimal strategy $\tau_1^* = 0.001$ while maintaining the rest. The first two sellers $\mathcal{S}_1$ and $\mathcal{S}_2$ are chosen as representatives. It is the same that the seller who changes her strategy unilaterally gets no more profit. Even if one seller changes her strategy, the broker can nearly keep her profit as before, benefiting from the *inner* competition among sellers which is formulated as a Nash game. Specifically, the effect of sellers' bounded rationality is almost limited among sellers and is corrected automatically in Stage 3, which signifies the transparency of Stage 3 to the upper stages. The change of the buyer's profit may be due to the effect of the data on the model, which is not always predictable, causing the irregular curve of $\Phi(\cdot)$. In theory, varying τ_1 surely makes differences on other sellers. However, since the number of sellers is large, this effect is *diluted* and negligible, making the profit of $\mathcal{S}_2$ almost unchanged.

Note that the buyer's profit is much more than the broker's and sellers', which is consistent with the buyer-leading property desired in *Share*. The buyer (e.g., the automaker) can create value using the product and gain long-term benefits (e.g., the huge revenue the automaker earns owing to the business decision based on the data model), while the broker or the sellers only make profit from the one-shot transaction which is relatively lower.

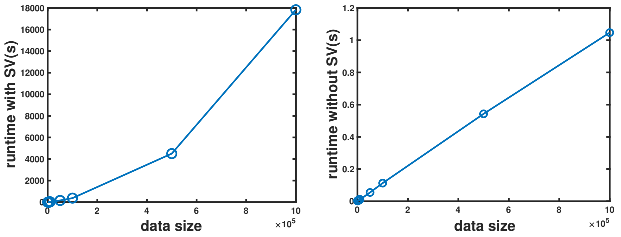
6.3 Efficiency

Fig.3(a) and Fig.3(b) show the runtime of the proposed data trading algorithm with and without Shapley value to update weights, respectively. 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. Note that our mechanism contains an extremely 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, which corresponds to the complexity analysis of Algorithm 1 in Section 5.2.

6.4 Parameter Influence

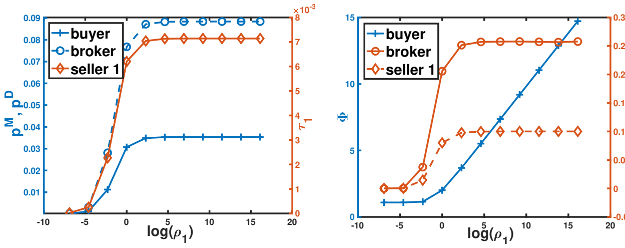
In this section, we make sensitivity analyses on the major parameters in our mechanism and investigate how the parameters affect the strategies and profits of the three parties.

Fig.4(a) and Fig.4(b) present the effect of ρ_1 on strategies and profits, respectively. Note that ρ_1 is a parameter relevant to the



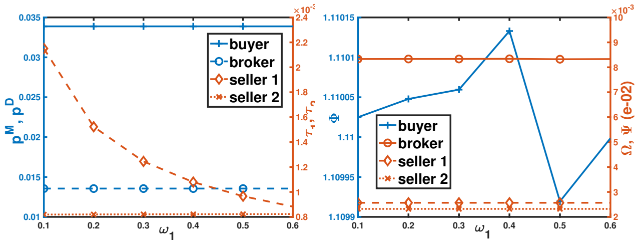
(a) Runtime with $\mathcal{SV}$ (b) Runtime without $\mathcal{SV}$

Figure 3: Efficiency.



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

Figure 4: Effect of ρ_1.

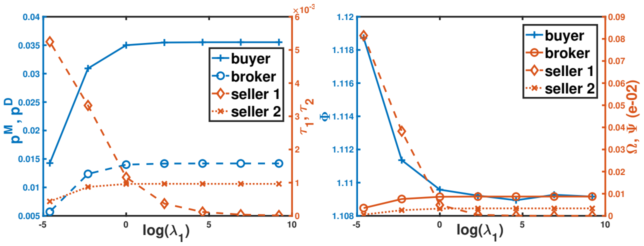


(a) Strategy vs. ω_1 (b) Profit vs. ω_1

Figure 5: Effect of ω_1.

buyer's sensitivity to the dataset quality, which objectively reflects the relationship between the dataset quality and the product utility. Fig.4(a) shows that too small of ρ_1 can hardly lead to effective markets because of the buyer's indifference on 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 market equilibrium, which may be due to common sense that the dataset quality cannot increase unlimitedly and even with sharper sensitivity to the data, higher prices wouldn't bring about better data anymore. Fig.4(b) shows that the profit of the buyer surges as ρ_1 increases because she will get more utility from the raise of the dataset quality. When ρ_1 is big enough, the increase of ρ_1 has little effect on the profits of the broker and sellers, which can be explained by the trend of strategies discussed above.

Fig.5(a) and Fig.5(b) present the effect of ω_1 on strategies and profits, respectively. Note that $\omega_1, \omega_2, ..., \omega_m$ are the weights of sellers' datasets and assess the sellers' data in previous transactions. We select $\mathcal{S}_1$ and $\mathcal{S}_2$ as representatives. Fig.5(a) shows that ω_1 only affects the strategy of the corresponding seller $\mathcal{S}_1$. The strategies of the buyer and the broker remain the same because ω_1 only affects the inner competition among sellers. Since the number of sellers is large, varying ω_1 makes little difference on other sellers, making the strategy of $\mathcal{S}_2$ almost unchanged. Fig.5(b) shows that when ω_1 varies from 0.1 to 0.6, all profits except the buyer's are stable. Once ω_1 gets a non-appropriate value, the data of this seller $\mathcal{S}_1$ won't work as expected and affects the profit of the buyer, leading to the unsmooth curve of $\Phi(\cdot)$.



(a) Strategy vs. λ_1 (b) Profit vs. λ_1

Figure 6: Effect of λ_1.

Fig.6(a) and Fig.6(b) show the effect of $\mathcal{S}_1$'s parameter λ_1 on strategies and profits, respectively. Note that λ_i is related to seller $\mathcal{S}_i$'s privacy sensitivity. Fig.6(a) shows that τ_1 sinks with increasing λ_1 since $\mathcal{S}_1$ will strengthen her data protection if more sensitive to privacy risks. p^M and p^D increase possibly because higher prices may be provided to encourage conservative sellers to offer high-fidelity data in spite of heavy privacy risks. Fig.6(b) shows that λ_1 mainly influences the buyer's and the corresponding seller $\mathcal{S}_1$'s profits. The profit of $\mathcal{S}_1$ decreases because bigger λ_1, more privacy loss $\mathcal{S}_1$ will suffer. The profit of buyer $\mathcal{B}$ dives probably because the seller would enhance the protection on her data when faced with huge privacy risks, thus lowering the data fidelity and further harming the buyer's profit. The profit of the broker remains unchanged because the broker herself does not rely on the data fidelity but just *transfers* the data from the sellers to the buyer.

7 CONCLUSION AND FUTURE WORK

We present *Share*, the first buyer-leading multi-seller data markets with absolute pricing rules based on a three-stage Stackelberg-Nash game. The profit maximization for all participants and the priority of buyers are fulfilled by considering the mutual interaction among three parties (buyers, brokers, and sellers) as a three-stage Stackelberg game, in which the absolute pricing for data is also realized. We address the seller selection problem by considering the inter-seller competition as a Nash game. To derive the equilibrium, backward induction is used. Specifically, to solve the inner Nash game, we propose two methods, direct derivation and a novel mean-field approximation which can address complex cases with provable approximation guarantees. Our proposed data market framework performs well on real and synthetic datasets in terms of both effectiveness and efficiency.

There are also other interesting and practical scenarios and issues in data markets, e.g., how to formulate a market with multiple buyers or streaming buyers, how to support seller-leading data transactions, and how to solve the challenge of parameter fitting for each party due to the deficiency of real-world trading records.

REFERENCES

[1] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. 2019. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. ACM, 701–726.

[2] Baoyi An, Mingjun Xiao, An Liu, Xike Xie, and Xiaofang Zhou. 2021. Crowdsensing Data Trading based on Combinatorial Multi-Armed Bandit and Stackelberg Game. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 253–264.

[3] Gaurang Bansal and Biplab Sikdar. 2021. Security Service Pricing Model for UAV Swarms: A Stackelberg Game Approach. In *2021 IEEE Conference on Computer Communications Workshops, INFOCOM Workshops 2021, Vancouver, BC, Canada, May 10-13, 2021*. IEEE, 1–6. https://doi.org/10.1109/INFOCOMWKSHPS51825.2021.9484577

[4] Tamer Başar and Geert Jan Olsder. 1998. *Dynamic noncooperative game theory*. SIAM.

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

[6] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

[7] Javier Castro, Daniel Gómez, and Juan Tejada. 2009. Polynomial calculation of the Shapley value based on sampling. *Computers & OR* 36, 5 (2009), 1726–1730. https://doi.org/10.1016/j.cor.2008.04.004

[8] Lingjiao Chen, Paraschos Koutris, and Arun Kumar. 2019. Towards Model-based Pricing for Machine Learning in a Data Marketplace. In *SIGMOD*, Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska (Eds.). ACM, 1535–1552. https://doi.org/10.1145/3299869.3300078

[9] Laurits R Christensen, Dale W Jorgenson, and Lawrence J Lau. 1975. Transcendental logarithmic utility functions. *The American Economic Review* 65, 3 (1975), 367–383.

[10] Zicun Cong, Xuan Luo, Jian Pei, Feida Zhu, and Yong Zhang. 2022. Data pricing in machine learning pipelines. *Knowl. Inf. Syst.* 64, 6 (2022), 1417–1455. https://doi.org/10.1007/s10115-022-01679-4

[11] Vincent Conitzer and Tuomas Sandholm. 2003. Complexity Results about Nash Equilibria. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, Georg Gottlob and Toby Walsh (Eds.). Morgan Kaufmann, 765–771. http://ijcai.org/Proceedings/03/Papers/111.pdf

[12] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. 2009. The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39, 1 (2009), 195–259.

[13] DAWEX. 2015. https://www.dawex.com/en/.

[14] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[15] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. 2013. 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, 429–438. https://doi.org/10.1109/FOCS.2013.53

[16] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 4052)*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.). Springer, 1–12. https://doi.org/10.1007/11787006_1

[17] Cynthia Dwork and Jing Lei. 2009. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, Michael Mitzenmacher (Ed.). ACM, 371–380. https://doi.org/10.1145/1536414.1536466

[18] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Found. Trends Theor. Comput. Sci.* 9, 3-4 (2014), 211–407. https://doi.org/10.1561/0400000042

[19] S. Shaheen Fatima, Michael J. Wooldridge, and Nicholas R. Jennings. 2008. A linear approximation method for the Shapley value. *Artif. Intell.* 172, 14 (2008), 1673–1699. https://doi.org/10.1016/j.artint.2008.05.003

[20] Raul Castro Fernandez. 2022. Protecting Data Markets from Strategic Buyers. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 1755–1769. https://doi.org/10.1145/3514221.3517855

[21] Hui Gao, Chi Harold Liu, Jian Tang, Dejun Yang, Pan Hui, and Wendong Wang. 2019. Online Quality-Aware Incentive Mechanism for Mobile Crowd Sensing with Extra Bonus. *IEEE Trans. Mob. Comput.* 18, 11 (2019), 2589–2603. https://doi.org/10.1109/TMC.2018.2877459

[22] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*. PMLR, 2242–2251.

[23] Chandra K. Jaggi, Mamta Gupta, Amrina Kausar, and Sunil Tiwari. 2019. Inventory and credit decisions for deteriorating items with displayed stock dependent demand in two-echelon supply chain using Stackelberg and Nash equilibrium solution. *Ann. Oper. Res.* 274, 1-2 (2019), 309–329. https://doi.org/10.1007/s10479-018-2925-9

[24] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas Spanos, and Dawn Song. 2019. Efficient task-specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment* 12, 11 (2019), 1610–1623.

[25] Changkun Jiang, Lin Gao, Lingjie Duan, and Jianwei Huang. 2018. Data-Centric Mobile Crowdsensing. *IEEE Trans. Mob. Comput.* 17, 6 (2018), 1275–1288. https://doi.org/10.1109/TMC.2017.2763956

[26] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2012. Query-based data pricing. In *PODS*. ACM, 167–178.

[27] Paraschos Koutris, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu. 2013. Toward practical query pricing with QueryMarket. In *SIGMOD*. ACM, 613–624.

[28] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* 2, 1 (2007), 229–260.

[29] Henger Li, Wen Shen, and Zizhan Zheng. 2020. Spatial-Temporal Moving Target Defense: A Markov Stackelberg Game Model. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith (Eds.). International Foundation for Autonomous Agents and Multiagent Systems, 717–725. https://dl.acm.org/doi/abs/10.5555/3398761.3398847

[30] Man Li, Jiahu Qin, Qichao Ma, Wei Xing Zheng, and Yu Kang. 2021. Hierarchical Optimal Synchronization for Linear Systems via Reinforcement Learning: A Stackelberg-Nash Game Perspective. *IEEE Trans. Neural Networks Learn. Syst.* 32, 4 (2021), 1600–1611. https://doi.org/10.1109/TNNLS.2020.2985738

[31] Jinfei Liu, Qiongqiong Lin, Jiayao Zhang, Kui Ren, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. 2021. Demonstration of Dealer: An End-to-End Model Marketplace with Differential Privacy. *Proc. VLDB Endow.* 14, 12 (2021), 2747–2750. http://www.vldb.org/pvldb/vol14/p2747-zhang.pdf

[32] Jinfei Liu, Jian Lou, Junxu Liu, Li Xiong, Jian Pei, and Jimeng Sun. 2021. Dealer: An End-to-End Model Marketplace with Differential Privacy. *Proc. VLDB Endow.* 14, 6 (2021), 957–969. http://www.vldb.org/pvldb/vol14/p957-liu.pdf

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

[34] Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2007), October 20-23, 2007, Providence, RI, USA, Proceedings*. IEEE Computer Society, 94–103. https://doi.org/10.1109/FOCS.2007.41

[35] Roger B Myerson. 1989. Mechanism design. In *Allocation, information and markets*. Springer, 191–206.

[36] Anna Nagurney and Pritha Dutta. 2019. Supply chain network competition among blood service organizations: a Generalized Nash Equilibrium framework. *Ann. Oper. Res.* 275, 2 (2019), 551–586. https://doi.org/10.1007/s10479-018-3029-2

[37] John Nash. 1951. Non-cooperative games. *Annals of mathematics* (1951), 286–295.

[38] John F Nash Jr. 1950. Equilibrium points in n-person games. *Proceedings of the national academy of sciences* 36, 1 (1950), 48–49.

[39] Jian Pei. 2021. A Survey on Data Pricing: from Economics to Data Science. *IEEE Trans. Knowl. Data Eng.* (2021). https://doi.org/10.1109/TKDE.2020.3045927

[40] Jian Pei, Feida Zhu, Zicun Cong, Xuan Luo, Huiwen Liu, and Xin Mu. 2021. 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*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 4058–4059. https://doi.org/10.1145/3447548.3470818

[41] SafeGraph. 2016. https://www.safegraph.com/.

[42] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[43] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. 2018. 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*, Jérôme Lang (Ed.). ijcai.org, 5494–5501. https://doi.org/10.24963/ijcai.2018/775

[44] Ankur Sinha, Pekka Malo, and Kalyanmoy Deb. 2018. A Review on Bilevel Optimization: From Classical to Evolutionary Approaches and Applications. *IEEE Trans. Evol. Comput.* 22, 2 (2018), 276–295. https://doi.org/10.1109/TEVC.2017.2712906

[45] Tianshu Song, Yongxin Tong, and Shuyue Wei. 2019. Profit Allocation for Federated Learning. In *2019 IEEE International Conference on Big Data (IEEE BigData), Los Angeles, CA, USA, December 9-12, 2019*, Chaitanya K. Baru, Jun Huan, Latifur Khan, Xiaohua Hu, Ronay Ak, Yuanyuan Tian, Roger S. Barga, Carlo Zaniolo, Kisung Lee, and Yanfang (Fanny) Ye (Eds.). IEEE, 2577–2586. https://doi.org/10.1109/BigData47090.2019.9006327

[46] Mohammad Amin Tajeddini, Hamed Kebriaei, and Luigi Glielmo. 2020. Decentralized Hierarchical Planning of PEVs Based on Mean-Field Reverse Stackelberg Game. *IEEE Trans Autom. Sci. Eng.* 17, 4 (2020), 2014–2024. https://doi.org/10.1109/TASE.2020.2986374

[47] Heinrich Von Stackelberg. 2010. *Market structure and equilibrium*. Springer Science & Business Media.

[48] Omar Abdel Wahab, Jamal Bentahar, Hadi Otrok, and Azzam Mourad. 2021. Resource-Aware Detection and Defense System against Multi-Type Attacks in the Cloud: Repeated Bayesian Stackelberg Game. *IEEE Trans. Dependable Secur. Comput.* 18, 2 (2021), 605–622. https://doi.org/10.1109/TDSC.2019.2907946

[49] Kaidi Wang, Zhiguo Ding, Daniel K. C. So, and George K. Karagiannidis. 2021. Stackelberg Game of Energy Consumption and Latency in MEC Systems With NOMA. *IEEE Trans. Commun.* 69, 4 (2021), 2191–2206. https://doi.org/10.1109/TCOMM.2021.3049356

[50] Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. 2020. A principled approach to data valuation for federated learning. In *Federated Learning*. Springer, 153–167.

[51] Detlof Von Winterfeldt and Gregory W Fischer. 1975. Multi-attribute utility theory: models and assessment procedures. *Utility, probability, and human decision making* (1975), 47–85.

[52] Hui Yin, Ye-Hwa Chen, and Dejie Yu. 2020. Stackelberg-Theoretic Approach for Performance Improvement in Fuzzy Systems. *IEEE Trans. Cybern.* 50, 5 (2020), 2223–2236. https://doi.org/10.1109/TCYB.2018.2883729

[53] Jinsung Yoon, Sercan Arik, and Tomas Pfister. 2020. Data valuation using reinforcement learning. In *International Conference on Machine Learning*. PMLR, 10842–10851.

[54] Jin Zhang and Qian Zhang. 2009. Stackelberg game for utility-based cooperative cognitiveradio networks. In *Proceedings of the 10th ACM Interational Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2009, New Orleans, LA, USA, May 18-21, 2009*, Edward W. Knightly, Carla-Fabiana Chiasserini, and Xiaojun Lin (Eds.). ACM, 23–32. https://doi.org/10.1145/1530748.1530753

[55] Shuyuan Zheng, Yang Cao, and Masatoshi Yoshikawa. 2020. Money Cannot Buy Everything: Trading Mobile Data with Controllable Privacy Loss. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*. 29–38. https://doi.org/10.1109/MDM48529.2020.00024