COTRA: A Data Trading Framework for Multi-Source Data Cooperation

Jin Cheng, Ningning Ding, John C. S. Lui, *Fellow, IEEE*, and Jianwei Huang, *Fellow, IEEE*

***Abstract*—Data trading significantly enhances data-driven technologies by enabling efficient data sharing across mobile devices and communication systems. Despite the clear advantages of incorporating multiple sources often experienced by data users, the topic of multi-source data trading remains largely underexplored. This paper designs a data trading framework that enables multi-source data trading through structured and adaptive cooperation among data sources. The proposed framework aims to enhance data usage efficiency and seller revenue. Key technical challenges addressed include the complex interactions between data sources and buyers, the coupling among diverse data products, and the inherent non-convexity in optimization problems. We employ a Nash bargaining framework to model cooperative seller decisions and a two-stage Stackelberg game for dynamic seller-buyer interactions. By systematically addressing the coupling among data products, we derive closed-form solutions despite the non-convex nature of the optimization problem. Our results reveal that seller revenue remains stable with increasing product coupling until it reaches a threshold, beyond which further coupling increases revenue due to the substitute effect. Adaptive cooperation exhibits greater scalability and resilience, particularly in environments with numerous data sources and complex product interdependencies, while structured cooperation provides more predictability and efficiency in stable conditions. Experimental results demonstrate that this framework can improve seller profits by up to 46.32% compared to existing data trading methods in the current market.**

***Index Terms*—Data trading, multi-source cooperation, substitution effect, Nash bargaining, Stackelberg game.**

Received 19 December 2024; revised 21 June 2025; accepted 17 July 2025. Date of publication 30 July 2025; date of current version 3 December 2025. The work of John C. S. Lui was supported in part by the RGC under Grant GRF-14202923. This work was supported in part by the National Natural Science Foundation of China under Grant 62271434, in part by Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network under Grant ZDSYS20220606100601002, in part by the Shenzhen Stability Science Program 2023, in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society, and in part by Longgang District Shenzhen's "Ten Action Plan" for Supporting Innovation Projects under Grant LGKCSDPT2024002. An earlier version of this paper was presented in part at the 2024 IEEE Global Communications Conference (GLOBECOM) [DOI: 10.1109/GLOBECOM52923.2024.10900935]. Recommended for acceptance by E. E. Tsiropoulou. *(Corresponding author: Jianwei Huang.)*

Jin Cheng is with the Shenzhen Institute of Artificial Intelligence and Robotics for Society and the School of Science and Engineering, Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Chinese University of Hong Kong, Sha Tin Hong Kong (e-mail: jincheng2@link.cuhk.edu.cn).

Ningning Ding is with the Data Science and Analytics Thrust and the Internet of Things Thrust, Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, Guangdong 511453, China (e-mail: ningningding@hkust-gz.edu.cn).

John C. S. Lui is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Sha Tin Hong Kong (e-mail: cslui@cse.cuhk.edu.hk).

Jianwei Huang is with the School of Science and Engineering, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen Key Laboratory of Crowd Intelligence Empowered Low-Carbon Energy Network, and CSIJRI Joint Research Centre on Smart Energy Storage, Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: jianweihuang@cuhk.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMC.2025.3593986, provided by the authors.

Digital Object Identifier 10.1109/TMC.2025.3593986

I. Introduction

A. Background and Motivations

DATA trading has significantly promoted the development of data-driven technologies, notably artificial intelligence (AI), by broadening access to diverse datasets and facilitating comprehensive model training. With the rapid expansion of mobile technologies, such as smartphones, wearable devices, and Internet of Things (IoT) systems, the importance of data trading has become increasingly critical. It enhances access to diverse datasets generated by these mobile devices, thereby enabling real-time analytics and improved model training for mobile applications. In 2022, the global data trading market reached a valuation of $968 million, with a projected compound annual growth rate of 25% from 2023 to 2030 [2]. This burgeoning landscape has witnessed the emergence of numerous data trading platforms in the industry, such as Windows Azure [3], Info Chimps [4], and Xignite [5], alongside extensive theoretical research on data trading over the past decade [6].

However, existing research on data trading neglects multi-source data cooperation, thereby limiting the potential for data utilization. Current studies often either assume a single data source or treat multiple sources as independent, primarily focusing on data pricing properties [7], such as arbitrage-free mechanisms and pricing flexibility. This oversight leads to insufficient data utilization, hindering the advancement of data-driven technologies.

Consider a mobile commerce example in Table I, which involves three data sources from different owners: (1) a mobile customer source (Data A) with customer-centric attributes gathered from mobile devices, such as location; (2) a product source (Data B) containing product-specific details from manufacturing plants, like material type; and (3) a transaction source (Data C) with transaction records, including Customer ID (CID), Product ID (PID), and Transaction ID (TID). As illustrated in Table II, with sensitive information (i.e., IDs) removed, three view types can be marketed as data products. If an investment entity seeks insights into the customer preferences for product *type*s in a given *region*, one needs to integrate these sources into View C (i.e., Product C). However, without cooperation among these data sources, only Views A and B can be offered separately by

TABLE I
DATA SOURCES

Data A	
CID	**Region**
938	California
...	...

Data B	
PID	**Type**
AOJ	A
...	...

Data C		
TID	**CID**	**PID**
1	938	AOJ
...	...	...

TABLE II
DATA PRODUCTS

View A
Region
California
...

View B
Type
A
...

View C	
Region	**Type**
California	A
...	...

their respective owners. Even if the owner of Data C purchases Views A and B, the synthesis of View C remains impossible due to the lack of key and sensitive information, resulting in a significant underutilization of data.

B. Contributions

To address the gap in multi-source data trading, we design the *COTRA* framework, facilitating cooperation among multiple data sources and enhancing data utilization. This framework enables sellers, including data aggregators and multiple data owners, to produce a wide range of product types. The aggregator synthesizes data products through cooperation with various owners and markets these products efficiently. This motivates the first key question of this paper:

- *Key Question 1: How can sellers cooperatively determine the pricing strategy for coupled data products and distribute revenue?*

We model the interactions between sellers and buyers using a two-stage Stackelberg game. In Stage I, data sellers collaboratively determine pricing and revenue distribution strategies based on the Nash bargaining theory [8], ensuring a fair and Pareto-optimal distribution of benefits. In Stage II, each data buyer makes purchase decisions, with the aggregator coordinating product provisioning from multiple sources.

We introduce two cooperation types within this framework: *structured* and *adaptive*. Structured cooperation refers to a scenario in which data source cooperation follows a predefined structure. Specifically, each cooperation instance involves two endogenous sources and one relational source to synthesize data products. This structured pattern facilitates a focused and well-defined analysis of the cooperative dynamics, ensuring clarity and predictability. In contrast, adaptive cooperation is more flexible, integrating any number of data sources to synthesize products from various combinations. This adaptability enables the framework to handle varying market conditions and diverse data demands. While adaptive cooperation is well-suited for scalability and managing complex interdependencies, structured cooperation offers greater predictability and efficiency in stable environments.

In addition to addressing seller cooperation, it is equally important to examine buyer behavior, as understanding this dynamic is crucial for optimizing data utilization and maximizing seller revenue. This leads us to the second key question of this paper:

- *Key Question 2: How do buyers assess the utility of coupled products, and what is the impact on their purchasing decisions?*

A key modeling contribution of this work is the consideration of *coupling* among diverse data products, a critical yet often neglected factor in previous studies. This *coupling* is exemplified by View C's partial overlap with Views A and B, significantly influencing buyer purchase decisions, as demonstrated in our later analysis and experiments. However, this important aspect has been neglected in existing works [6]. Moreover, the low replication cost of data necessitates a distinct pricing model from other products. To address these problems, we systematically categorize utilities and data sources into endogenous and relational types to disentangle product coupling, thereby enhancing our ability to characterize buyer payoffs and pricing strategies.

The technical challenge lies in the *non-convexity* of the optimization problem due to ambiguous numerical relationships among parameters. To tackle this, we categorize our analysis into distinct cases based on system parameters, which may exhibit either convex or non-convex characteristics. For both locally convex and non-convex cases, we systematically identify the optimal solutions through a series of analyses and prove their optimality. Furthermore, we conduct a comprehensive examination of the overall non-convex problem, integrating the insights from both cases and finally deriving a unified closed-form solution. This closed-form solution offers computational efficiency, making it highly practical for real-world applications. Additionally, it enhances understanding of the trading framework, providing insights into how different parameters influence outcomes.

We summarize our key contributions as follows:

- *Multi-source Data Cooperation Framework:* To the best of our knowledge, this paper is the first work that develops a data trading framework with multi-source data cooperation. We introduce and analyze two cooperation mechanisms within the framework—structured cooperation, which offers predictability and efficiency in stable environments, and adaptive cooperation, which provides flexibility and scalability for dynamic adjustments in complex settings. These mechanisms optimize data utilization and enhance revenue generation across varying market conditions.
- *Modeling of Coupling Among Data Products:* We characterize the buyers' utility concerning coupled data products, recognizing their potential to function as substitutes for one another. We systematically categorize the products into endogenous and relational views, accordingly classifying the utility to enhance the ability to characterize buyer payoffs.
- *Closed-form Solution for the Non-convex Pricing Problem:* The technical challenge arises from the inherent non-convexity of the problem due to ambiguous numerical relationships among parameters. Despite the technical challenges, we derive closed-form optimal solutions by partitioning our analysis into several convex and non-convex subproblems and then obtaining optimal solutions for each.
- *Insights of Multi-source Data Cooperation:* We reveal that within the *COTRA* framework, sellers' revenue initially

remains steady as product coupling increases due to the substitute effect. However, revenue begins to rise significantly once the coupling exceeds a specific threshold. Additionally, adaptive cooperation exhibits greater scalability and resilience, particularly when managing numerous data sources and complex product interdependencies. In contrast, structured cooperation provides more predictability and efficiency in stable environments. Experimental results show that *COTRA* can boost sellers' profit by up to 46.32% compared to current trading methods in the data market.

C. Related Works

Related works fall into two categories: single-source data trading and multi-source data evaluation.

1) Single-Source Data Trading: Previous studies in this field primarily focus on desirable properties of data pricing functions, such as being arbitrage-free [6], [9], maximizing revenue [10], [11], [12], preserving privacy [13], [13], [14], and ensuring flexibility [7]. For instance, Chen et al. [9] developed a data trading framework that achieves equitable pricing for graph statistics by adjusting noise levels in query answers to prevent arbitrage. Likewise, He et al. [13] proposed a framework for IoT data trading that enhances privacy and efficiency by combining sampling-based methods with differential privacy, thereby ensuring secure and accurate data aggregation for trading. However, these studies often focus on the properties of pricing functions without providing a detailed characterization of utility or deriving closed-form solutions. Furthermore, they are limited to single-source scenarios, overlooking the complexities and benefits of multi-source data trading.

2) Multi-Source Data Evaluation: Most works along this line evaluate data instances within machine learning contexts. These studies usually employ Shapley Value [15] and develop efficient algorithms to reduce the computation complexity [16], [17], [18], [19], [20]. Wang et al. [21] proposed TKNN-Shapley, an efficient and privacy-preserving data valuation method that integrates differential privacy into the Shapley value framework for fair assessment of individual data contributions. Although these studies evaluate data instances that may potentially originate from different sources, they generally treat the sources as independent, overlooking cooperative dynamics such as information sharing. This limits the potential for enhanced data utilization. Additionally, these approaches do not yield closed-form solutions due to high computational complexity. Recent advances in decentralized optimization [22] address locally non-convex problems over dynamic networks using perturbation-based gradient tracking. In contrast, our work differs in three key aspects: problem formulation, optimization methodology, and application domain, which we elaborate on in the following discussion.

To the best of our knowledge, this is the first work proposing a data trading framework with multi-source data cooperation.

We organize the rest of this paper as follows: Section II outlines the system model. Sections III and IV detail the optimal solutions for buyer and seller decisions under structured and adaptive cooperation, respectively. Section V presents the extensive experiments conducted to evaluate the performance of our trading framework. Finally, Section VI concludes the paper.

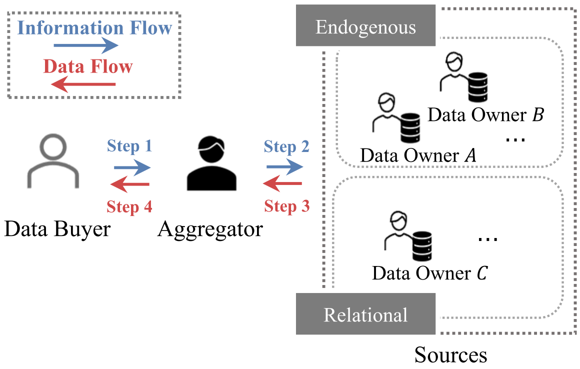


Fig. 1. System model under structured cooperation.

II. System Model

In this section, we introduce a cooperative multi-source data trading framework under structured cooperation, as shown in Fig. 1, focusing on a scenario where data sources cooperate based on predefined and fixed structures. For adaptive cooperation, we will introduce the necessary modifications to the model with a few adjustments in Section IV-A.

For each transaction, a buyer purchases views from the view set $\mathcal{V}$, which includes offerings from sellers (i.e., aggregators and data owners). The framework leverages data virtualization technology [23] to enable real-time access to data stored across multiple sources, a widely used capability in contemporary data management systems. Unlike traditional methods, this technology allows aggregators to operate without a central database or cache to store data products. Instead, the aggregator maintains a catalog of available data products and delivers them on demand through coordination among the data owners, thereby enhancing the flexibility of data trading.

To mitigate the risk of privacy leakage during cross-source aggregation, we assume a privacy-preserving data virtualization environment. Under this setting, sensitive identifiers are not directly stored or exposed. Instead, the aggregator accesses data through privacy-preserving join mechanisms, such as token-based matching [24] or federated join execution [25], thereby minimizing access to raw sensitive attributes and enhancing compliance with data protection requirements.

We categorize products within this framework into two types: endogenous and relational. Endogenous products (e.g., View A) derive their utility from inherent data, while relational products (e.g., View C) rely on both inherent data and their interrelationships. Data sources align with these distinctions. Each synthesis of data products in structured cooperation involves exactly two endogenous sources and one relational source, focusing on two-dimensional relational views. This case is a practical approximation of the scenario when multiple heterogeneous relational sources utilize disjoint endogenous sources, thereby streamlining the complexity involved in multi-source data trading.

A. Buyer Modeling

Each data transaction serves one buyer. The one-buyer setting approximates cases where multiple heterogeneous buyers access the platform one at a time, with sellers maintaining unlimited capacities. This subsection outlines the buyer's decision, data usage, utility, and payoff.

Decision: Let $\mathcal{V} = \mathcal{V}_e \cup \mathcal{V}_r$ be the view sets, where $\mathcal{V}_e$ and $\mathcal{V}_r$ represent the sets of endogenous and relational views, respectively. Then, the buyer's purchase decision is $\boldsymbol{x} = (x_j : x_j \in [0, 1] \text{ and } j \in \mathcal{V})$.[1]

Data Usage: Based on the purchase decision, we can calculate the data usage.[2] We initially assume a uniform distribution of endogenous products within the relational product.[3] We denote the coupling coefficient for relational view $j \in \mathcal{V}_r$ as $\boldsymbol{\alpha}_j = (\alpha_j^i : \forall i \in \mathcal{V}_e)$. If the purchase decision for view j is x_j, then a portion $\alpha_j^i x_j$ of the endogenous view i is covered by view j. When providing endogenous products, sellers prioritize the portions that remain uncovered by any other relational views. Thus, the usage of an endogenous view i becomes $y_i(\boldsymbol{x}) = \min(1, x_i + \sum_{j \in \mathcal{V}_r} \alpha_j^i x_j)$; while for a relational view j, it is $y_j(\boldsymbol{x}) = x_j$.

Utility: We differentiate utility into endogenous and relational categories.[4] In machine learning, utility is often represented as task accuracy [7]. The accuracy function typically shows concavity concerning training data size in feature-based and link-based tasks, which rely on endogenous and relational utility, respectively. Following prior works [28], we denote the basic utility function of usage by:

$$f(y) = \mu \ln(1 + y), \tag{1}$$

where the logarithmic function captures the widely considered diminishing marginal utility, μ is the buyer's preference coefficient, and y is the actual data usage.[5] We incorporate utility coefficients μ_e to signify the preference for endogenous utility and μ_r to signify the preference for relational utility.[6] Therefore, the total utility of the buyer is:

$$g(\boldsymbol{x}) = \sum_{i \in \mathcal{V}_e} \mu_e \ln\left(1 + y_i(\boldsymbol{x})\right) + \sum_{j \in \mathcal{V}_r} \mu_r \ln\left(1 + y_j(\boldsymbol{x})\right), \tag{2}$$

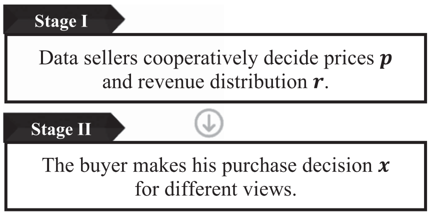


Fig. 2. The two-stage Stackelberg game.

where $y_i(\boldsymbol{x})$, $i \in \mathcal{V}$, is the data usage.

Payoff: The buyer's payoff is the difference between their utility and payment to the sellers. We denote product prices by $\boldsymbol{p} = (p_i : \forall i \in \mathcal{V})$ and the buyer's payoff by:

$$U_b(\boldsymbol{x}, \boldsymbol{p}) = g(\boldsymbol{x}) - \sum_{i \in \mathcal{V}} x_i p_i. \tag{3}$$

B. Seller Modeling

The seller set $\mathcal{S}$ includes both data owner set $\mathcal{S}_o$ and the aggregator set $\mathcal{S}_a = \{a\}$. We introduce the sellers' revenue, cost, and payoff as follows.

Revenue: Sellers cooperatively set the price $\boldsymbol{p} = (p_i : \forall i \in \mathcal{V})$ and determine the revenue distribution ratio $\boldsymbol{r} = (r_j : \forall j \in \mathcal{S})$. The total revenue is given by $R(\boldsymbol{p}) = \sum_{i \in \mathcal{V}} p_i x_i$ and each seller $j \in \mathcal{S}$ receives a portion $r_j R(\boldsymbol{p})$.

Cost: For data owners, the cost corresponds to the transmission expense of data, while aggregators bear costs for both data and view transmission.[7] A common practice [28], [35] of the cost function is a linear with coefficient β. Besides, logical trading implies that the utility of an entire view can cover its own cost, i.e., $\mu_e \ln 2 \geq 2\beta$ and $\mu_r \ln 2 \geq 2\beta$.

Payoff: Each seller calculates his payoff by subtracting his cost from the distributed revenue. Thus, for a data owner $j \in S_o$, the payoff is:

$$U_j(\boldsymbol{p}, \boldsymbol{r}) = r_j R(\boldsymbol{p}) - \beta y_j(\boldsymbol{x}), \ \forall j \in \mathcal{S}_o. \tag{4}$$

For the aggregator, the payoff is:

$$U_a(\boldsymbol{p}, \boldsymbol{r}) = r_a R(\boldsymbol{p}) - \beta \sum_{i \in \mathcal{V}} x_i. \tag{5}$$

C. Two-Stage Stackelberg Game

We model the interactions among sellers and buyers as a two-stage Stackelberg game, as illustrated in Fig. 2.

[1]The proposed model is most suitable for structured and semantically aligned data sources, such as mobile user logs, product metadata, and transactional databases, which are commonly managed within enterprise-level database systems. These sources typically support view-level integration through joinable attributes and standardized formats.

[2]The relational view can partially substitute for the endogenous view due to their overlap, a consideration frequently neglected in existing studies.

[3]There are two reasons for this assumption. First, a uniform distribution of endogenous views within relational views can be achieved by reordering the data items of relational views in the database. Second, this assumption simplifies the inherently complex problem, making the analysis more feasible and manageable without compromising its integrity.

[4]To ensure analytical tractability, our model adopts structural simplifications such as logarithmic utility, linear transmission costs, and view-based product aggregation. These abstractions support closed-form solutions while retaining key features of real-world data trading, including substitution effects, cooperative pricing, and source heterogeneity. Similar simplifications are common in data economics and pricing models [26], [27].

[5]In our framework, adopting log-scale quantization, as studied by Doostmohammadian et al. [29], aligns well with the concave logarithmic modeling of data utility, preserving decision accuracy and reducing transmission costs. This suggests a promising direction for future work.

[6]For clarity, we assume homogeneity within product types, focusing on the key problem. We consider heterogeneity in the adaptive cooperation model, which is a generalization of this model, as detailed in Section IV.

[7]We do not consider processing cost because they are assumed to be either negligible compared to transmission costs or consistent across different scenarios, as works [30], [31] did. This simplification allows the focus on the more variable and impactful costs associated with data and view transmission. While we follow prior works [30], [31] in assuming negligible or uniform processing costs across data views, we acknowledge that this assumption may not hold in all practical deployments. In particular, large-scale systems with real-time or complex queries may incur heterogeneous processing overheads. If such costs scale approximately linearly with data volume, they can be absorbed into the transmission cost term with appropriate adjustments to cost coefficients. Otherwise, when processing costs are nonlinear or product-specific (e.g., due to secure computation [32], graph traversal [33], or adaptive analytics [34]), they can be explicitly modeled as additive terms in seller payoffs, potentially influencing equilibrium outcomes.

- *Stage I:* Sellers cooperatively determine price $\boldsymbol{p}$ and revenue distribution $\boldsymbol{r}$. We model sellers' interactions by Nash bargaining theory [15] to ensure fair and optimal allocation among the decision-makers within this cooperative framework. We formulate the sellers' decision problem as follows:

Problem $\mathbb{P}_1$: Nash Bargaining for Sellers.

$$\max_{\boldsymbol{r},\boldsymbol{p}} \; G(\boldsymbol{r},\boldsymbol{p}) = \prod_{j\in\mathcal{S}} U_j(\boldsymbol{p},\boldsymbol{r}) \tag{6a}$$

$$\text{s.t.} \; \sum_{j\in\mathcal{S}} r_j = 1, \; r_j \in [0,1], \; \forall j \in \mathcal{S}, \tag{6b}$$

$$p_i \geq 0, \; \forall i \in \mathcal{V}. \tag{6c}$$

- *Stage II:* The buyer makes the purchase decision $\boldsymbol{x}$. Thus, we formulate the buyer's decision problem as follows:

Problem $\mathbb{P}_2$: Payoff Maximization for the Buyer.

$$\max_{\boldsymbol{x}} \; U_b(\boldsymbol{x},\boldsymbol{p}) \tag{7a}$$

$$\text{s.t.} \; x_i \in [0,1], \; \forall i \in \mathcal{V}. \tag{7b}$$

In this section, we have introduced the system model for multi-source data trading, defined the buyer's and the sellers' decision process, and outlined the two-stage Stackelberg game between sellers and buyers. To derive closed-form solutions and gain valuable insights, we first focus on a scenario involving two endogenous sources (Data A and B) and one relational source (Data C), as detailed in Tables I and II. Here, the coupling coefficient for the relational product across the two endogenous products is α.[8] This case is a plausible approximation for situations involving multiple heterogeneous relational sources with disjoint endogenous sources and sellers with unlimited capacities. As shown in Sections III-A and III-B, despite its simplicity, this case retains the core aspects of the problem, which presents several challenges due to non-convexity and ambiguous parameter relationships. The following section will examine the Nash equilibrium through backward induction under structured cooperation.

III. Data Trading Under Structured Cooperation

In this section, we study the decisions of buyers and sellers under structured cooperation by backward induction. Section III-A delves into the buyer's optimal purchase decisions, considering the influence of coupling among data products. Following this, Section III-B analyzes sellers' optimal pricing and revenue distribution strategies.

A. Stage II: Buyer's Optimal Decisions

In this subsection, we derive the buyer's optimal decisions in Stage II, given the prices set by sellers in Stage I. Due to the

[8] We also consider that sellers possess complete information about the buyer. Consider a concrete example: the sellers know the buyer's information through user profiling, a widely employed method in market research that makes the assumption reasonable.

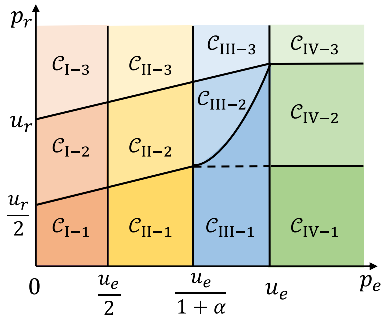


Fig. 3. Theorem 1: Solution of Stage II.

homogeneity of products, decisions regarding the same product type are identical. Consequently, the buyer's decisions fall into two parts: x_e for endogenous products and x_r for relational products. The following lemma describes the relationship between the optimal decisions x_e^* and x_r^*, which is determined by the coupling among the products:

Lemma 1: The buyer's optimal decisions satisfy $x_e^* \leq 1 - \alpha x_r^*$.

Proof of Lemma 1 is provided in Appendix A, available online. This lemma indicates that the purchase level of the endogenous products will not exceed the uncovered portion of the purchased relational products. Applying Lemma 1 and employing variable substitution, we can formally reformulate Problem $\mathbb{P}_2$ as follows:

Problem $\mathbb{P}_2'$: Reformulation of Problem $\mathbb{P}_2$.

$$\max_{y_e,y_r} \; 2\mu_e \ln(1+y_e) + \mu_r \ln(1+y_r)$$
$$- 2(y_e - \alpha y_r)p_e - p_r y_r \tag{8a}$$

$$\text{s.t.} \; y_e, y_r, y_e - \alpha y_r \in [0,1], \tag{8b}$$

where y_e and y_r are exactly the usage of different views determined by $\boldsymbol{x}$, as illustrated in Section II-A. After determining the optimal values y_e^* and y_r^*, we can compute the corresponding $\boldsymbol{x}^*$. We can verify that the Hessian matrix of the objective function in (9) is negative semi-definite, which shows that the reformulated problem is convex.

$$H = \begin{pmatrix} -\frac{2\mu_e}{(1+y_e)^2} & 0 \\ 0 & -\frac{\mu_r}{(1+y_r)^2} \end{pmatrix} \tag{9}$$

We derive the buyer's optimal decisions, as presented in Theorem 1, with a summary of these decisions in Table III and a visual representation in Fig. 3.

Theorem 1. (Optimal Buyer's Decisions): Given the sellers' pricing strategies, p_e and p_r, the buyer's optimal purchasing decisions x_e^* and x_r^* are shown in Table III and Fig. 3.

Proof of Theorem 1 is given in Appendix B, available online. Based on the given optimal buyer's decisions, we have the following insights:

Observation 1. (Buyer's Behaviours): The buyer's optimal decisions fall into the following four categories based on p_e:

TABLE III
THEOREM 1 (BUYER'S OPTIMAL DECISION): SOLUTION OF STAGE II

p_e	p_r	$y_e^*(\boldsymbol{p})$	$y_r^*(\boldsymbol{p})$	$x_e^*(\boldsymbol{p})$	$x_r^*(\boldsymbol{p})$
$(\mathcal{C}_{\text{I}})\ p_e \leq \frac{\mu_e}{2}$	$(\mathcal{C}_{\text{I-1}})\ p_r - 2\alpha p_e \leq \frac{\mu_r}{2}$	1	1	$1-\alpha$	1
	$(\mathcal{C}_{\text{I-2}})\ \frac{\mu_r}{2} < p_r - 2\alpha p_e \leq \mu_r$	1	$\frac{\mu_r - p_r + 2\alpha p_e}{p_r - 2\alpha p_e}$	$1-\alpha y_r^*$	y_r^*
	$(\mathcal{C}_{\text{I-3}})\ \mu_r \leq p_r - 2\alpha p_e$	1	0	1	0
$(\mathcal{C}_{\text{II}})\ \frac{\mu_e}{2} < p_e \leq \frac{\mu_e}{1+\alpha}$	$(\mathcal{C}_{\text{II-1}})\ p_r - 2\alpha p_e \leq \frac{\mu_r}{2}$	$\frac{\mu_e - p_e}{p_e}$	1	$x_e^* - \alpha$	1
	$(\mathcal{C}_{\text{II-2}})\ \frac{\mu_r}{2} < p_r - 2\alpha p_e \leq \mu_r$	$\frac{\mu_e - p_e}{p_e}$	$\frac{\mu_r - p_r + 2\alpha p_e}{p_r - 2\alpha p_e}$	$x_e^* - \alpha y_r^*$	y_r^*
	$(\mathcal{C}_{\text{II-3}})\ \mu_r \leq p_r - 2\alpha p_e$	$\frac{\mu_e - p_e}{p_e}$	0	x_e^*	0
$(\mathcal{C}_{\text{III}})\ \frac{\mu_e}{1+\alpha} < p_e \leq \mu_e$	$(\mathcal{C}_{\text{III-1}})\ p_r - 2\alpha p_e \leq \frac{\alpha\mu_r p_e}{\mu_e - (1-\alpha)p_e}$	α	1	0	1
	$(\mathcal{C}_{\text{III-2}})\ \frac{\alpha\mu_r p_e}{\mu_e - (1-\alpha)p_e} < p_r - 2\alpha p_e \leq \mu_r$	$\frac{\mu_e - p_e}{p_e}$	$\frac{\mu_r - p_r + 2\alpha p_e}{p_r - 2\alpha p_e}$	$x_e^* - \alpha y_r^*$	y_r^*
	$(\mathcal{C}_{\text{III-3}})\ \mu_r \leq p_r - 2\alpha p_e$	$\frac{\mu_e - p_e}{p_e}$	0	x_e^*	0
$(\mathcal{C}_{\text{IV}})\ p_e > \mu_e$	$(\mathcal{C}_{\text{IV-1}})\ p_r \leq \frac{4\alpha\mu_e + \mu_r + \alpha\mu_r}{2(1+\alpha)}$	α	1	0	1
	$(\mathcal{C}_{\text{IV-2}})\ \frac{4\alpha\mu_e + \mu_r + \alpha\mu_r}{2(1+\alpha)} < p_r \leq \mu_r + 2\alpha\mu_e$	αy_0	y_0	0	y_0
	$(\mathcal{C}_{\text{IV-3}})\ \mu_r + 2\alpha\mu_e \leq p_r$	0	0	0	0
where $y_0 = \frac{-(1+\alpha)p_r + 2\alpha\mu_e + \alpha\mu_r + \sqrt{(-1+\alpha)^2 p_r^2 + 2(\alpha-1)\alpha p_r(2\mu_e - \mu_r) + \alpha^2(2\mu_e + \mu_r)^2}}{2\alpha p_r}$.					

- *Case I ("All-In"):* When p_e is sufficiently low (i.e., $\mathcal{C}_I$), the buyer purchases part of the relational view and supplements their usage y_e to 1 with enough endogenous views.
- *Case II ("Partial-Purchase"):* With moderately low p_e (i.e., $\mathcal{C}_{\text{II}}$), x_e decreases due to the elevated price. The final usage of the endogenous views is $(\mu_e - p_e)/p_e \in [\alpha, 1)$.
- *Case III ("Mixed-Purchase"):* For moderately high p_e (i.e., $\mathcal{C}_{\text{III}}$), an interesting scenario emerges. With sufficiently low p_r, the buyer purchases the entire relational view to meet the demand of the endogenous view due to the coupling, highlighting the substitute effects of relational views.
- *Case IV ("No-Purchase"):* High p_e discourages endogenous product purchases, leading buyers to substitute products (i.e., $y_e = \alpha y_r$). If the relational product price (p_r) exceeds the total utility ($\mu_r + 2\alpha\mu_e$), no purchase occurs.

Observation 2. (Substitution Effect): As shown in Case III-1 and IV-1, when the price of endogenous products p_e is sufficiently high and the price of relational ones p_r is sufficiently low, the buyer opts to purchase the entire relational views to satisfy the needs of the endogenous views.

In this subsection, we have determined the buyer's optimal decision $x_e^*(\boldsymbol{p})$ and $x_r^*(\boldsymbol{p})$ in Stage II. Next, we will derive the sellers' optimal decision in Stage I through backward induction.

B. Stage I: Sellers' Optimal Decisions

In this subsection, we analyze the sellers' optimal pricing and revenue distribution strategies in Stage I by backward induction, considering the buyer's optimal decisions in Stage II. We begin by employing the Nash bargaining framework [8] to guide our analysis of the optimal strategies.

Theorem 2 describes the sellers' optimal revenue distribution strategy, derived from the Nash bargaining framework [8], which guides the optimal pricing strategy. For clarity, we define the total cost as $C(\boldsymbol{p}) = \beta\sum_{i\in\mathcal{V}}(y_i^*(\boldsymbol{p}) + x_i^*(\boldsymbol{p}))$.

Theorem 2. (Optimal Revenue Distribution): The sellers' optimal revenue distribution strategy in Stage I is:

$$r_j^*(\boldsymbol{p}) = \frac{\sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p}) - C(\boldsymbol{p}) + 4\beta y_j^*(\boldsymbol{p})}{4\sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p})}, \forall j \in \mathcal{S}_o \quad (10a)$$

$$r_a^*(\boldsymbol{p}) = \frac{\sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p}) - C(\boldsymbol{p}) + 4\beta\sum_{i\in\mathcal{V}} x_i^*(\boldsymbol{p})}{4\sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p})}, \quad (10b)$$

where $y_j^*(\boldsymbol{p})$ is the data usage defined in Section II-A.

Proof of Theorem 2 is given in Appendix C, available online. According to Theorem 2, each seller $j \in \mathcal{S}$ receive the same payoff, which is given by:

$$\left(\sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p}) - C(\boldsymbol{p})\right) \Big/ 4. \quad (11)$$

This indicates that Problem $\mathbb{P}_1$ in Stage I aims to maximize sellers' surplus and can be reformulated as follows:

$$\max_{\mathbf{p}} \sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p}) - C(\boldsymbol{p}) \quad (12a)$$

$$\text{s.t. } p_i \geq 0, \forall i \in \mathcal{V}. \quad (12b)$$

The primary challenge in solving Problem $\mathbb{P}_3$ lies in its inherent non-convexity, stemming from complex parameter relationships, as detailed in Table III. To address this, we divide the analysis into several cases based on system parameters, which exhibit either convexity or non-convexity. Specifically, we identify that Cases $\mathcal{C}_I$ and $\mathcal{C}_{II}$ exhibit convexity, allowing us to decouple decisions into two convex subproblems using variable substitutions. For non-convex Cases $\mathcal{C}_{III}$ and $\mathcal{C}_{IV}$, we determine a unique optimal pricing strategy through detailed analyses, ultimately revealing two potential optimal pricing strategies.

Theorem 3. (Optimal Pricing Strategy): The optimal pricing strategy for the sellers is one of the following:

- *Bundling Strategy:* $\boldsymbol{p}_b^* = (p_{eb}^*, p_{rb}^*)$, where

$$p_{eb}^* = \begin{cases} \mu_e/(1+\alpha^2), & \text{if } \mu_e \leq 2(1+\alpha)^2\beta; \\ \sqrt{2\beta\mu_e}, & \text{if } 2(1+\alpha)^2\beta < \mu_e \leq 8\beta; \\ \mu_e/2, & \text{if } \mu_e > 8\beta; \end{cases} \quad (13a)$$

$$p_{rb}^* = \begin{cases} \sqrt{2(1-\alpha)\beta\mu_r + 2\alpha p_{eb}^*}, & \text{if } \mu_r \leq 8(1-\alpha)\beta; \\ \mu_r/2 + 2\alpha p_{eb}^*, & \text{if } \mu_r > 8(1-\alpha)\beta. \end{cases} \quad (13b)$$

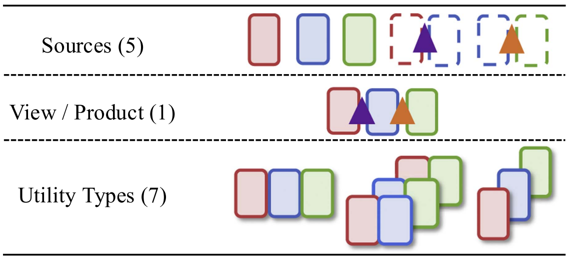


Fig. 4. Example of a multi-dimensional relational product involving multiple sources and utility types. Rectangles represent endogenous sources, and triangles represent relational sources. Data from relational sources can join with data from endogenous sources to form views.

- *Separated Strategy:* $\boldsymbol{p}_s^* = (p_{es}^*, p_{rs}^*)$, where $p_{es}^* = \mu_e$ and $p_{rs}^* = \mu_r + 2\alpha\mu_e$.

The proof of Theorem 3 is given in Appendix D, available online. Theorem 3 shows that, depending on the values of μ_e, μ_r, α, and β, the optimal pricing strategies can be determined as either $\boldsymbol{p}_b^*$ or $\boldsymbol{p}_s^*$, based on which strategy yields higher profits. The bundling strategy $\boldsymbol{p}_b^*$ simultaneously adjusts p_e and p_r to maximize total profits. In contrast, the separated strategy sets a baseline with a high p_e and increases the p_r to ensure profitability from the relational product.

In this section, we have explored how buyers and sellers make optimal decisions within a structured cooperation framework. In the following section, we will explore the interactions and behaviors of buyers and sellers under adaptive cooperation, addressing more dynamic scenarios with flexible data relationships.

IV. Data Trading Under Adaptive Cooperation

In this section, we extend our analysis to the adaptive cooperation framework for multi-source data trading, which provides greater flexibility and scalability compared to structured cooperation. A key distinction is that under adaptive cooperation, relational views can be multi-dimensional, incorporating combinatorial utilities from multiple data sources. We begin by introducing the necessary modifications to the system model from Section II to reformulate the problem in Section IV-A. Following this, we provide detailed analyses of the optimal decisions for both buyers and sellers under adaptive cooperation in Sections IV-B and IV-C.

A. Problem Formulation

In this subsection, we introduce the model modifications and formulate the optimization problems under adaptive cooperation. In this case, sellers gain the ability to synthesize multi-dimensional relational data products, enabling cooperation among multiple sources. This capability allows a single data product to incorporate combinatorial utility types, effectively acting as a partial substitute for any combination of its contributing sources. As shown in Fig. 4, the adaptive cooperation model, which involves three endogenous sources and two relational sources, enables the creation of a three-dimensional relational product. This product is considered three-dimensional because it contains three columns in the final view, each representing a distinct dimension of information. Together, these columns incorporate seven unique utility types, enhancing the product's richness and value.

Following this, we introduce the necessary modifications to the buyer's and sellers' modeling in Sections IV-A1 and IV-A2. Despite these changes, the problems in Stages I and II under adaptive cooperation retain the similar structure as defined in Problems $\mathbb{P}_1$ and $\mathbb{P}_2$ in Section II-C.

1) Buyer Modeling Under Adaptive Cooperation: In the adaptive cooperation framework, the buyer's decision-making process becomes more complex due to the combinatorial utility derived from multi-dimensional data products. Let $\mathcal{V}$ represent the set of data product types. The buyer's decision are denoted as $\boldsymbol{x} = (x_j : x_j \in [0, 1] \text{ and } j \in \mathcal{V})$. Assuming that the data distributions across different sources are independent and uniform,[9] the data usage is denoted by $\hat{\boldsymbol{y}}(\boldsymbol{x}) = (\hat{y}_i(\boldsymbol{x}) : i \in \mathcal{V})$, where each $\hat{y}_i(\boldsymbol{x})$ is:

$$\hat{y}_i(\boldsymbol{x}) = 1 - \prod_{j\in\mathcal{V}}(1 - \alpha_{ji}x_j), \ \forall i \in \mathcal{V}. \tag{14}$$

Here, α_{ji} is the proportion of Product i that is covered by Product j. For tractability, we approximate $\hat{y}_i$ with $y_i(\boldsymbol{x}) = \sum_{j\in\mathcal{V}} \alpha_{ji}x_j$, ignoring multi-layer coupling. To examine this approximation, we define the error between the actual usage $\hat{\boldsymbol{y}}$ and the approximate usage $\boldsymbol{y}$ as

$$\begin{aligned}\epsilon_i(\boldsymbol{x}) &= \hat{y_i(\boldsymbol{x})} - y_i(\boldsymbol{x}) \\ &= -\sum_{j=1}^{M-1}\sum_{k=j+1}^{M} a_{ji}a_{ki}x_jx_k + \cdots + (-1)^{M+1}\Pi_{j=1}^{M} a_{ji}x_j\end{aligned} \tag{15}$$

For the detailed derivation process, Appendix J, available online, provides comprehensive information. We base this approximation on two reasons. First, all α and x values fall within the interval $[0, 1]$, and each error term remains relatively small when the coupling coefficients are small. Second, we consider only single-layer coupling, focusing on the direct impact of one product covering another. We exclude cascading or multi-layer effects that might result from interactions across multiple layers.

Larger coupling coefficients may introduce more errors due to the prominence of multi-layer effects. Effectively managing these effects demands a sophisticated model to capture interdependencies across multiple coupling layers, resulting in a bi-level and non-convex optimization problem [36], [37]. These problems are inherently complex and require advanced techniques for obtaining approximate solutions [38]. As an initial step toward multi-source data cooperation, this work simplifies the model to focus on the most significant and direct interactions, thereby preserving key insights. Future research should address the challenges of strong coupling and multi-layer interactions by developing efficient algorithms to handle these complexities within data trading frameworks.

[9] The reasons for this assumption are discussed in Section II.

The corresponding utility $g(\cdot)$ and payoff $U_b(\cdot)$ for the buyer are:

$$g(\boldsymbol{x}) = \sum_{i\in\mathcal{V}} \mu_i \ln\left(1 + y_i(\boldsymbol{x})\right) \tag{16}$$

and

$$U_b(\boldsymbol{x}) = g(\boldsymbol{x}) - \sum_{i\in\mathcal{V}} p_i x_i, \tag{17}$$

where μ_i is the utility coefficient of Product i.

2) Seller Modeling Under Adaptive Cooperation: We generalize the sellers' modeling presented in Section II-B by incorporating cost heterogeneity across different data and product types. This adjustment ensures that the model more accurately reflects practical scenarios.

In the generalized model, the sellers' payoffs are given by:

$$U_j(\boldsymbol{p}, \boldsymbol{r}) = r_j R(\boldsymbol{p}) - \beta_j^D y_j\left(\boldsymbol{x}(\boldsymbol{p})\right), \quad \forall j \in \mathcal{S}_o, \tag{18}$$

$$U_a(\boldsymbol{p}, \boldsymbol{r}) = r_a R(\boldsymbol{p}) - \sum_{i\in\mathcal{V}} \beta_i^V x_i(\boldsymbol{p}), \tag{19}$$

where β_j^D and β_i^V represent the cost coefficients for transmitting Data j and Product i, respectively. It is important to note that, under the structured cooperation framework, the number of data sources (or owners) equals the number of data products. However, this relationship does not apply in the adaptive cooperation scenario, where any combination of sources can synthesize a data product, as illustrated in Fig. 4.

In this subsection, we have introduced some modifications to the buyer's and sellers' modeling. Although there are some differences in the modeling, the problems in Stages I and II under adaptive cooperation follow the same form as Problems $\mathbb{P}_1$ and $\mathbb{P}_2$ defined in Section II. In the following subsections, we will derive the optimal decisions for buyers and sellers based on backward induction.

B. Stage II: Buyer's Optimal Decisions

In this subsection, we derive the buyer's optimal decision $\boldsymbol{x}$ under adaptive cooperation in Stage II, given the sellers' decisions from Stage I. We begin by formulating the buyer's decision problem in Stage II as Problem $\mathbb{P}_3$, based on the analysis in Section IV-A. In Section IV-B1, we proceed to derive the optimal solution to Problem $\mathbb{P}_3$, followed by practical insights drawn from the homogeneous and non-coupling scenarios, which are discussed in Section IV-B2.

Problem $\mathbb{P}_3$: Payoff Maximization for the Buyer under Adaptive Cooperation in Stage II.

$$\max_{\boldsymbol{x}} \sum_{i=1}^{M} \left(u_i \ln(1 + \sum_{j=1}^{M} a_{ji} x_j) - x_i p_i \right) \tag{20a}$$

$$\text{s.t.} \quad x_i \in [0,1], \quad \forall i \in \mathcal{V}. \tag{20b}$$

Here, M is the number of views.

1) Optimal Solution to the Problem: We first apply the KKT conditions [39] to Problem $\mathbb{P}_3$ and derive the following lemma for the optimal buyer's decisions:

Lemma 2. (Buyer's Optimal Solution in Stage II under Adaptive Cooperation): The buyer's optimal decision $\boldsymbol{x}^*$ under adaptive cooperation satisfies the following conditions:

$$x_j^*(x_j^* - 1)\left(\sum_{i=1}^{M} \frac{u_i a_{ji}}{1 + \sum_{k=1}^{M} a_{ki} x_k^*} - p_j\right) = 0, \quad \forall j \in \mathcal{V}. \tag{21}$$

The proof of Lemma 2 is provided in Appendix E, available online. Based on Lemma 2, we know that the optimal decision x_j^* is either 0, 1, or satisfies $\sum_{i=1}^{M} \frac{u_i a_{ji}}{1+\sum_{k=1}^{M} a_{ki} x_k^*} = p_j$ for all $j \in \mathcal{V}$. The expression $\sum_{i=1}^{M} \frac{u_i a_{ji}}{1+\sum_{k=1}^{M} a_{ki} x_k^*}$ represents the marginal utility of Product j, which depends on the coupling level of each product and the decisions made for each product. The denominator $1 + \sum_{k=1}^{M} a_{ki} x_k^*$ serves as a normalizing factor, adjusting for the cumulative effects of all products based on the decision variables, thereby emphasizing the interconnected nature of their utilities. Additionally, we can derive the following corollary from the lemma:

Corollary 1: For the optimal decision variable x_j^* corresponding to Product j, $x_j^* = 1$ holds for $p_j \leq \sum_{i=1}^{M} \frac{u_i a_{ji}}{2}$, and $x_j^* = 0$ holds for $p_j \geq \sum_{i=1}^{M} u_i a_{ji}$.

Corollary 1 defines the thresholds for guaranteed purchase ($x_j^* = 1$) and non-purchase ($x_j^* = 0$), accounting for product coupling. This result highlights how pricing and product interactions influence buyer behavior. A lower price p_j ensures full purchase of Product j, while a higher price discourages purchase. The coupling between products adjusts these thresholds, emphasizing the importance of interdependencies in optimizing pricing strategies. Since the solution involves only element-wise operations such as inverse or max/min evaluation over the product set, the computational complexity is $\mathcal{O}(M)$, where M is the number of data products.

To gain further insights, we consider a special case with only two types of products. This scenario represents a practical situation where only two products exhibit positive utility, with arbitrary coupling between them. Such cases are common in markets with limited products [40], [41]. For this case, we get the following corollary:

Corollary 2: For the case $M = 2$, the buyer's optimal decision $\boldsymbol{x}^* = (x_1^*, x_2^*)$ is as follows:

$$x_1^* = \frac{\alpha_{21} - \alpha_{22}}{-\alpha_{12}\alpha_{21} + \alpha_{11}\alpha_{22}} + \frac{\alpha_{22} u_1}{\alpha_{22} p_1 - \alpha_{12} p_2} + \frac{\alpha_{21} u_2}{\alpha_{21} p_1 - \alpha_{11} p_2}, \tag{22}$$

$$x_2^* = \frac{\alpha_{11} - \alpha_{12}}{\alpha_{12}\alpha_{21} - \alpha_{11}\alpha_{22}} + \frac{\alpha_{12} u_1}{-\alpha_{22} p_1 + \alpha_{12} p_2} + \frac{\alpha_{11} u_2}{-\alpha_{21} p_1 + \alpha_{11} p_2}. \tag{23}$$

The proof of Corollary 2 is provided in Appendix F, available online. This corollary reveals the *frozen demand effect*. At first glance, one might expect that the demand for Product 1 x_1^* would increase with a larger coupling coefficient α_{12}, since a larger α_{12} implies that Product 1 covers more of Product 2. However, the demand x_1^* remains constant irrespective of α_{12}, particularly when the price p_1 is either below $\check{p}_{1,\alpha_{12}}$ or above $\hat{p}_{1,\alpha_{12}}$. These price thresholds are detailed in (A28) and (A29) in the appendix, available online. This occurs because if p_1 is low enough, consumers purchase the full quantity of Product 1, resulting in $x_1^* = 1$, making α_{12} irrelevant. On the other hand, if the price p_1 is too high, consumers refrain from buying Product 1, leading to $x_1^* = 0$, again making the coupling coefficient α_{12} irrelevant. The substitution and frozen demand effects observed in this case are also applicable in broader scenarios.

2) Insights From the Homogeneous and Non-Coupling Scenarios: Following the optimal solution, we explore two practical scenarios for adaptive cooperation: homogeneous and non-coupling. These cases offer insights into how the system behaves under different product couplings. Homogeneous coupling occurs in markets where data products are standardized [42], with different sources offering similar products that have overlapping functionalities. In contrast, no coupling is observed in specialized fields where data products are tailored to meet specific and non-overlapping needs [43]. Analyzing these cases is crucial for understanding buyer behavior in different data trading contexts and for developing effective pricing strategies that reflect the degree of data coupling.

For the homogeneous case, where the coupling levels across products are uniform, we identify a threshold structure in the buyer's decision-making process, as described below:

Lemma 3. (Threshold Structure of Buyer's Decision in the Homogeneous Case): When the coupling levels across all products are uniform, i.e., $\alpha_{ji} = \bar{\alpha}_i$ for all i and j, there exists a threshold price $\bar{p}$ such that for any Product k:

$$x_k^* = \begin{cases} 1, & \text{if } p_k < \bar{p}, \\ 0, & \text{if } p_k > \bar{p}. \end{cases} \tag{24}$$

Moreover, if $p_j = \bar{p}$, the following condition holds:

$$\sum_{i=1}^{M} \frac{u_i}{\frac{1}{\bar{\alpha}_i} + \sum_{k=1}^{M} x_k^*} = \bar{p}. \tag{25}$$

The proof of Lemma 3 is provided in Appendix G, available online. This lemma demonstrates a *substitution effect:* when the coupling levels across products are uniform, buyers consistently prefer to purchase those products with prices lower than a specified threshold, thereby avoiding more expensive alternatives. Specifically, (25) establishes a threshold price, $\bar{p}$, such that for any product k, the buyer's decision x_k^* is binary—either 1 (purchased) if $p_k < \bar{p}$ or 0 (not purchased) if $p_k > \bar{p}$. The absence of an index j in (25) reflects that the threshold structure applies uniformly across all products k. This behavior is crucial for understanding consumer choices in markets where products have similar interdependencies, emphasizing the role of pricing in competitive market dynamics.

In the non-coupling case, for each product $j \in \mathcal{V}$, the coupling coefficient α_{ji} is non-zero only when $i \neq j, \forall i, j \in \mathcal{V}$.[10] In this scenario, we also derive the buyer's optimal decisions:

[10]This analysis considers the more general case where $\alpha_{ii} \in (0, 1)$, corresponding to a situation where data sellers possess only a partial view of Product i, covering only a fraction α_{ii} of the entire product.

Lemma 4. (Buyer's Decision in the Non-coupling Case): For the non-coupling case, the buyer's optimal decision $\boldsymbol{x}$ for each Product j is characterized as follows:

$$x_j^* = \begin{cases} 0, & \text{if } \frac{u_j}{p_j} - \frac{1}{\alpha_{jj}} \leq 0, \\ \frac{u_j}{p_j} - \frac{1}{\alpha_{jj}}, & \text{if } 0 < \frac{u_j}{p_j} - \frac{1}{\alpha_{jj}} \leq 1, \\ 1, & \text{if } \frac{u_j}{p_j} - \frac{1}{\alpha_{jj}} > 1. \end{cases} \tag{26}$$

Lemma 4 illustrates the *utility threshold effect*, highlighting a scenario in which a product can only substitute for one other product, thereby demonstrating the impact of the absence of coupling. When the utility $u_j \geq p_j$ and the coefficient $\alpha_{jj} \leq p_j/(u_j - p_j)$, the user demand x_j increases with α_{jj} because the intrinsic value of Product j rises. However, when $\alpha_{jj} > p_j/(u_j - p_j)$, the user demand no longer changes with α_{jj}. This is because the buyer has decided to purchase the entire Product j, as it already offers sufficient utility.

In this subsection, we have analyzed the buyer's optimal decision $\boldsymbol{x}$. We now proceed to derive the sellers' optimal pricing decisions, $\boldsymbol{p}$ and $\boldsymbol{r}$.

C. Stage I. Sellers' Optimal Decision

In this subsection, we use backward induction to derive the sellers' optimal decision in Stage I under adaptive cooperation, considering the buyer's optimal decisions in Stage II. We begin by analyzing the optimal revenue distribution strategy in Section IV-C1, followed by an exploration of the optimal pricing strategy in Section IV-C2.

1) Optimal Revenue Distribution: Based on the Nash bargaining solution, we derive the following theorem concerning the sellers' optimal revenue distribution strategy. First, we define the total cost of sellers as

$$C(\boldsymbol{p}) = \sum_{i=1}^{M} \beta_i^D y_i^*(\boldsymbol{p}) + \sum_{i=1}^{M} \beta_i^V x_i^*(\boldsymbol{p}), \tag{27}$$

where $y_i^*(\boldsymbol{p})$ and $x_i^*(\boldsymbol{p})$ represent the buyer's optimal decisions in Stage II.[11] The sellers' optimal revenue distribution strategy under adaptive cooperation is as follows.

Theorem 4. (Sellers' Optimal Revenue Distribution under Adaptive Cooperation): Given price $\boldsymbol{p}$, the sellers' optimal revenue distribution $\boldsymbol{r}^*(\boldsymbol{p})$ under adaptive cooperation is:

$$r_j^*(\boldsymbol{p}) = \frac{\sum_{i=1}^{M} p_i x_i^*(\boldsymbol{p}) - C(\boldsymbol{p}) + (M+1)\beta_j^D y_j^*(\boldsymbol{p})}{(M+1)\sum_{i=1}^{M} p_i x_i^*(\boldsymbol{p})}, \quad j \in \mathcal{S}_o, \tag{28}$$

$$r_a^*(\boldsymbol{p}) = \frac{\sum_{i=1}^{M} p_i x_i^*(\boldsymbol{p}) - \sum_{i=1}^{M} \beta_i^D y_i^*(\boldsymbol{p}) + M \sum_{i=1}^{M} \beta_i^V x_i^*(\boldsymbol{p})}{(M+1)\sum_{i=1}^{M} p_i x_i^*(\boldsymbol{p})}, \tag{29}$$

where $y_i^*(\boldsymbol{p})$ denotes the usage.

The proof of Theorem 4 is provided in Appendix H, available online. The solution requires only basic arithmetic operations such as summation and normalization across cooperative sellers, and thus has time complexity $\mathcal{O}(MN)$, given buyer's decisions.

[11]For simplicity, we express $\boldsymbol{y}(\boldsymbol{x}(\boldsymbol{p}))$ as $\boldsymbol{y}(\boldsymbol{p})$.

According to Theorem 4, the sellers' payoffs for data owners $j \in \mathcal{S}_o$ and the aggregator a are equal, expresses as:

$$U_j(\boldsymbol{p}, \boldsymbol{r}^*(\boldsymbol{p})) = U_a(\boldsymbol{p}, \boldsymbol{r}^*(\boldsymbol{p})) = \frac{\sum_{i=1}^{M} p_j x_j^*(\boldsymbol{p}) - C(\boldsymbol{p})}{M+1}. \tag{30}$$

This result shows that the objective function of the optimization problem in Stage I corresponds to the sellers' surplus, i.e., $\sum_{i=1}^{M} p_j x_j^*(\boldsymbol{p}) - C(\boldsymbol{p})$. By substituting the total cost function $C(\boldsymbol{p})$ from (27), we express the sellers' surplus, which serves as the pricing optimization objective function, as follows:

$$\sum_{i=1}^{M} \left(x_i^*(\boldsymbol{p})(p_i - \beta_i^V) - \beta_i^D \sum_{j=1}^{M} a_{ji} x_j^*(\boldsymbol{p}) \right). \tag{31}$$

2) Optimal Pricing Strategy: The sellers' pricing problem in Stage I, with the objective function in (31), presents significant challenges due to the inherent non-convexity of $\boldsymbol{x}^*(\boldsymbol{p})$ as outlined in Theorem 2. This non-convexity primarily arises from the coupling between different data products and the interaction terms within the utility functions. To tackle this complexity, we initially apply a single-level reduction to reformulate the problem, as detailed in Section IV-C2-a. Following this, we offer a closed-form solution for a simplified case in Section IV-C2-b. Finally, we present a general solution approach for the whole problem in Section IV-C2-c.

a) Single-Level Reduction for the Pricing Problem: Since Problem $\mathbb{P}_3$ in Stage II is convex, we can leverage single-level reduction [44] to simplify the bilevel optimization framework. This approach replaces the lower-level optimization problem with its corresponding KKT conditions, allowing us to incorporate the buyer's optimal decisions directly into the objective function in Stage I, as shown in (31). By reducing the bilevel problem to a single level, we streamline the analysis and enable more efficient solution methods. The resulting formulation for the sellers' pricing problem is as follows:

Problem $\mathbb{P}_4$: Pricing Optimization for Sellers under Adaptive Cooperation in Stage I.

$$\max_{\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\mu}} \sum_{i=1}^{M} \sum_{m=1}^{M} \left(\frac{u_m a_{im} x_i}{1 + \sum_{k=1}^{M} a_{km} x_k} - \beta_i^D \alpha_{mi} x_m \right) - \sum_{i=1}^{M} \lambda_i - \sum_{i=1}^{M} x_i \beta_i^V \tag{32a}$$

$$\text{s.t.} \quad \lambda_j (1 - x_j) = 0, \quad \forall j \in \mathcal{V}, \tag{32b}$$

$$\mu_j x_j = 0, \quad \forall j \in \mathcal{V}, \tag{32c}$$

$$p_j \geq 0, \quad \forall j \in \mathcal{V}, \tag{32d}$$

$$0 \leq x_j \leq 1, \quad \forall j \in \mathcal{V}, \tag{32e}$$

$$\lambda_j \geq 0, \quad \forall j \in \mathcal{V}, \tag{32f}$$

where λ_j is the dual variable associated with the constraint $x_j \leq 1$, and u_j is the dual variable associated with the constraint $x_j \geq 0$, for all $j \in \mathcal{V}$.

b) Insights from the Simple Case with $M = 2$: Through the single-level reduction approach, we initially derive a closed-form solution for the simple case where $M = 2$ to gain intrinsic insights in Corollary 3. This analysis directly relates to Corollary 2 from Section IV-B, which highlights the fundamental principles and interactions involved. In this way, we establish a foundation for comprehending more complex scenarios.

Corollary 3: For the case where $M = 2$,[12] the optimal pricing strategy $\boldsymbol{p}^* = (p_1^*, p_2^*)$ for the seller is:

$$p_1^* = \frac{u_1 \alpha_{11}}{1 + \alpha_{11} x_1^* + \alpha_{21} x_2^*} + \frac{u_2 \alpha_{12}}{1 + \alpha_{12} x_1^* + \alpha_{22} x_2^*}, \tag{35}$$

$$p_2^* = \frac{u_1 \alpha_{21}}{1 + \alpha_{11} x_1^* + \alpha_{21} x_2^*} + \frac{u_2 \alpha_{22}}{1 + \alpha_{12} x_1^* + \alpha_{22} x_2^*}. \tag{36}$$

The corresponding optimal solutions for the buyers are $x_1^* = \max(0, \min(1, \hat{x}_1))$ and $x_2^* = \max(0, \min(1, \hat{x}_2))$, where $\hat{x}_1$ and $\hat{x}_2$ are defined in (33) and (34) are shown at the bottom of the next page.

Appendix I, available online, provides the proof of Corollary 3, which uncovers a *utility-cost paradox*. Intuitively, one might expect the revenue R^* to increase with the utility coefficient u_1. However, the revenue R^* decreases with the utility u_1 when $\beta_1^D > u_1 + (\alpha_{12}\beta_2^V - \alpha_{22}\beta_1^V)/(\alpha_{22}\alpha_{11} - \alpha_{12}\alpha_{21})$. This paradox arises because when the data transmission cost β_1^D is sufficiently high, any increase in utility u_1 leads to a disproportionately larger increase in costs, particularly β_1^D. As a result, the increased costs outweigh the benefits from the higher utility, causing the overall revenue R^* to decrease. In other words, the transmission cost grows faster than the added value provided by the increased utility, leading to a net reduction in revenue.

c) General Approach to Solve Problem $\mathbb{P}_4$: To effectively solve Problem $\mathbb{P}_4$, we begin by reformulating the original optimization problem into a mixed-integer optimization problem with linear constraints. This reformulation simplifies the computational complexity, making the problem more tractable within existing optimization frameworks.

To manage the nonlinear constraints $\lambda_j(1 - x_j) = 0$ and $\lambda_j x_j = 0$, we convert them into linear conditions suitable for mixed-integer programming. This reformulation includes:

For $\lambda_j(1 - x_j) = 0$:

$$\begin{cases} \lambda_j \leq U(1 - z_j^1), \\ x_j \geq 1 - U(1 - z_j^2), \\ z_j^1 + z_j^2 \geq 1, \\ z_j^1,\ z_j^2 \in \{0, 1\}, \end{cases} \tag{37}$$

For $\mu_j x_j = 0$:

$$\begin{cases} -U(1 - z_j^3) \leq \mu_j \leq U(1 - z_j^3), \\ x_j \leq U(1 - z_j^4), \\ z_j^3 + z_j^4 \geq 1, \\ z_j^3,\ z_j^4 \in \{0, 1\}, \end{cases} \tag{38}$$

where U is a sufficiently large number.

[12] This corresponds to the scenario where only two products have positive utility, with arbitrary coupling between them. This case is in alignment with the analysis in Section IV-B

These reformulated constraints align with mixed-integer nonlinear programming (MINLP) techniques, effectively handling discrete decision variables and reducing non-convexity. This enables the use of robust optimization methods like the Branch and Bound [45], ensuring global optimality and faster convergence. In addition to exact methods, several approximation techniques can be employed to improve scalability. For example, convex relaxations (with a piecewise linear approximation of non-convex terms) reduce the problem to a polynomial-time solvable convex program, typically with time complexity $\mathcal{O}(M^3)$. Meta-heuristic methods such as genetic algorithms (GA) and simulated annealing (SA) can further reduce the practical runtime to $\mathcal{O}(M \cdot N_{\text{iter}})$, where N_{iter} is the iteration count. When the coupling structure is sparse, decomposition techniques allow distributed computation with per-iteration complexity $\mathcal{O}(M) \sim \mathcal{O}(M^2)$. These methods offer flexible trade-offs between accuracy and runtime, making them well-suited for real-time or large-scale data trading applications.

In this section, we have extended our analysis to data trading under adaptive cooperation, which provides greater flexibility and scalability than structured cooperation. We have introduced the necessary modifications to reformulate the problem and analyzed optimal strategies for both buyers and sellers under adaptive cooperation. Next, we will conduct a series of experiments to validate the framework's performance.

V. Experimental Results

In this section, we evaluate the performance of the multi-source data trading framework *COTRA* using a real-world dataset under two distinct cooperation types. Section V-A provides a detailed analysis of structured cooperation, focusing on its performance and key implications. In Section V-B, we extend the evaluation to adaptive cooperation, examining the framework's adaptability to more dynamic and flexible data trading scenarios. We highlight key insights in *italics* for easy identification throughout the section.

A. Structured Cooperation

This subsection evaluates the framework under structured cooperation using a real-world dataset from Taobao [46] to showcase its practical effectiveness. The dataset comprises 23 million data points, capturing the purchasing behaviors of 20,000 customers over a specific period.

We begin by calibrating the model parameters with this dataset to capture market dynamics accurately. Our primary focus is on the data utility for model training tasks, as detailed in Section II-A. We define "endogenous utility" as the model's ability to reflect the actual data distribution, quantified through the Kullback-Leibler (KL) divergence. Specifically, we evaluate each data segment's distribution against the overall dataset's distribution by calculating their KL divergence d_{KL} and define each segment's endogenous utility as $1 - d_{KL}$.

Additionally, we define "relational utility" as the model's ability to predict relationships between different data elements. This is assessed through its performance in recommendation tasks, where utility is measured by the accuracy of forecasting the types of products consumers are likely to purchase. We conduct the fitting process using an XGBoost model [47], a standard approach for such analyses. Unless stated otherwise, the coupling coefficient α is set to 0.5, and the cost coefficient β is set to 0.05.

Next, we evaluate how changes in the coupling coefficient α and the cost coefficient β affect the performance of COTRA under structured cooperation. We examine these effects across various market scenarios, comparing them to two commonly used benchmarks in the current data trading market: the uniform pricing framework ("UNIFORM") and the non-cooperation framework ("NOCO") [6], [7]. The former assigns a uniform price to all products without distinguishing different utilities, whereas the latter restricts product availability.

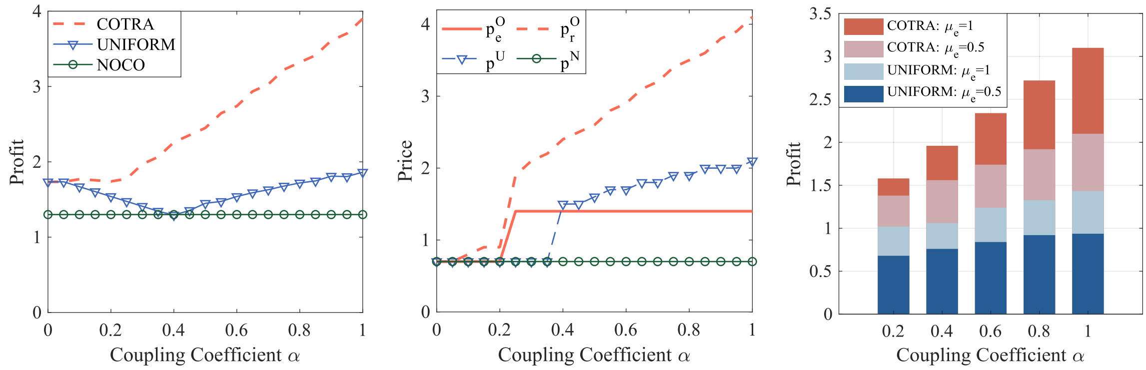
1) Coupling Coefficient: Fig. 5 illustrates the impact of varying the coupling coefficient α. Fig. 5(a) shows that *COTRA* outperforms the existing methods [6], [7] by averages of 37.48% and 46.32%, respectively. At lower values of α, products are considered independent, leading sellers to focus on maximizing sales without considering substitution effects. *As α increases, there is a notable shift towards relational products, allowing sellers to adjust baseline prices (i.e., p_e) and relational prices (i.e., p_r) to enhance profitability*, as illustrated in Fig. 5(b). This strategy creates a market dynamic where buyers gravitate towards relational products that better meet their specific needs, thereby maximizing revenue. In contrast, the uniform pricing and non-cooperation frameworks exhibit less flexibility and adaptability, resulting in lower revenues, as shown in Fig. 5(a) and (c).

2) Cost Coefficient: Fig. 6 illustrates the impact of varying the cost coefficient β. As shown in Fig. 6(a), *COTRA* generates higher revenues, surpassing the benchmark methods by averages of 43.29% and 38.45%, respectively. This comparison highlights that, unlike the uniform and non-cooperation strategies, which respond similarly to the changes, the *COTRA* framework adapts distinctly, demonstrating resilience to cost variations.

Specifically, Fig. 6(b) shows that in *COTRA*, *as costs increase, the demand for endogenous products x_e diminishes to zero and then stabilizes*. Conversely, Fig. 6(c) illustrates *initial stability*

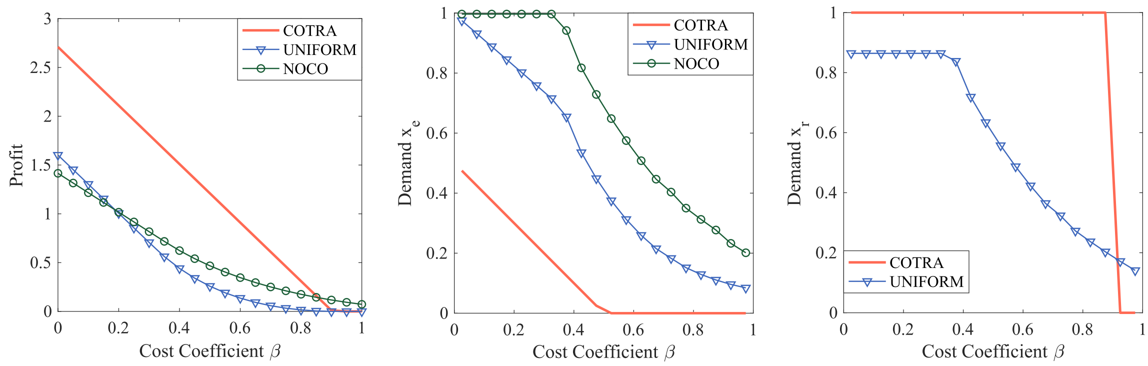
$$\hat{x_1} = \frac{\alpha_{21} - \alpha_{22} + \alpha_{22}\sqrt{\frac{\alpha_{12}\alpha_{21}u_1 - \alpha_{11}\alpha_{22}u_1}{\alpha_{12}\alpha_{21}\beta_1^D - \alpha_{11}\alpha_{22}\beta_1^D - \alpha_{22}\beta_1^V + \alpha_{12}\beta_2^V}} - \alpha_{21}\sqrt{\frac{\alpha_{12}\alpha_{21}u_2 - \alpha_{11}\alpha_{22}u_2}{\alpha_{12}\alpha_{21}\beta_2^D - \alpha_{11}\alpha_{22}\beta_2^D + \alpha_{21}\beta_1^V - \alpha_{11}\beta_2^V}}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}} \tag{33}$$

$$\hat{x_2} = \frac{-\alpha_{11} + \alpha_{12} - \alpha_{12}\sqrt{\frac{\alpha_{12}\alpha_{21}u_1 - \alpha_{11}\alpha_{22}u_1}{\alpha_{12}\alpha_{21}\beta_1^D - \alpha_{11}\alpha_{22}\beta_1^D - \alpha_{22}\beta_1^V + \alpha_{12}\beta_2^V}} + \alpha_{11}\sqrt{\frac{\alpha_{12}\alpha_{21}u_2 - \alpha_{11}\alpha_{22}u_2}{\alpha_{12}\alpha_{21}\beta_2^D - \alpha_{11}\alpha_{22}\beta_2^D + \alpha_{21}\beta_1^V - \alpha_{11}\beta_2^V}}}{\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}} \tag{34}$$



(a) Profit (b) Price (c) Profit

Fig. 5. Impact of varying the coupling coefficient α.



(a) Profit (b) Demand x_e (c) Demand x_r

Fig. 6. Impact of varying the cost coefficient β.

in the demand for relational products x_r, followed by a sharp decline. This trend occurs because sellers must raise prices to cope with higher costs. Moreover, the substitution effect helps lower the price of endogenous products while maintaining the prices of relational products to sustain sales. However, when costs become excessively high, all market transactions cease.

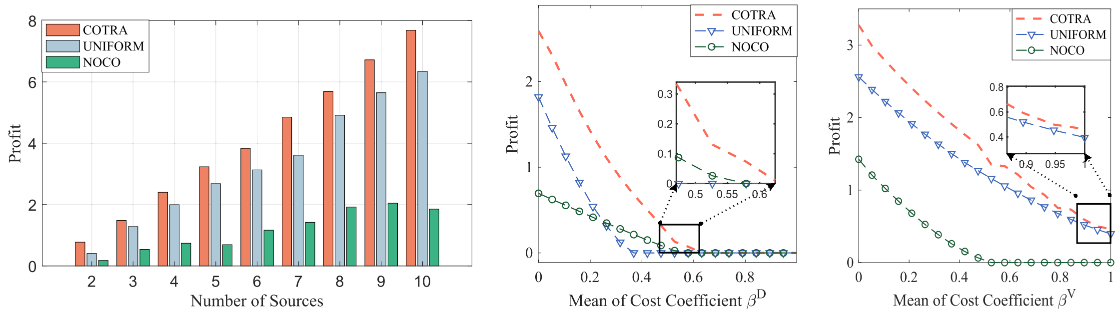
The experiments in this subsection demonstrate the efficiency of *COTRA* under structured cooperation, laying a foundation for exploring more dynamic data trading environments. In the following subsection, we extend the analysis to adaptive cooperation, where we evaluate how *COTRA* performs in more flexible and dynamic scenarios.

B. Adaptive Cooperation

This subsection analyzes the experimental results obtained under adaptive cooperation. Specifically, we examine how system performance is influenced by the number of data sources (Section V-B1), the coupling coefficients (Section V-B2), and the cost coefficients (Section V-B3). The structured cooperation experiments use real-world data from the Taobao platform, involving user behavior logs, product metadata, and transactional records. These sources are joined to form structured relational views. In contrast, the adaptive cooperation experiments use refined and extended views derived from the original dataset by reorganizing and combining existing features (e.g., geolocation, category hierarchies, behavioral patterns). This allows controlled evaluation of the framework's performance in flexible multi-source aggregation scenarios while preserving the structural properties of the real data.

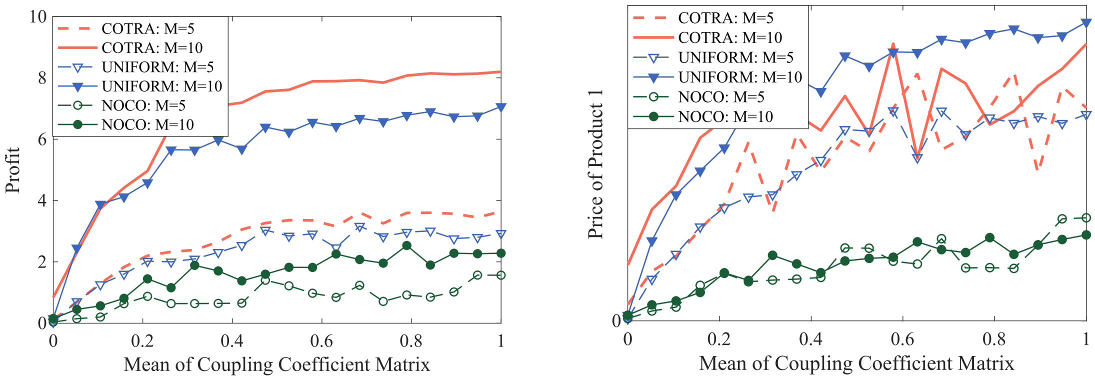
1) Number of Data Sources: We evaluate the scalability of the framework by examining its performance with varying numbers of data sources. Fig. 7(a) illustrates a*positive correlation between the number of data sources and the profit generated by the framework*. For example, with 10 data sources, the *COTRA* framework generates a profit of approximately 7.72 units, compared to 6.30 units for UNIFORM and 1.84 units for NOCO. This increase in profit suggests that a greater number of data sources enhances the richness of data products, thereby increasing their value to buyers. Across all scenarios, *COTRA* consistently outperforms both the UNIFORM and NOCO methods. Specifically, *COTRA* shows a profit improvement of up to 22.54% over UNIFORM and generates more than four times the profit compared to NOCO when there are 10 data sources. The superior performance of *COTRA* can be attributed to its efficient management and integration of multiple data sources, leveraging cooperative dynamics among sellers to optimize pricing and maximize revenue.

2) Coupling Coefficients: Fig. 8(a) and (b) illustrate the impact of varying the coupling coefficient matrix on system



(a) Number of Sources (b) Cost Coefficient β^D (c) Cost Coefficient β^V

Fig. 7. Impact of number of data sources, cost coefficient.



(a) Profit (b) Pricing p_1

Fig. 8. Impact of coupling coefficient.

performance, with Fig. 8(a) showing the profit and Fig. 8(b) detailing the pricing of Product 1. As the mean of the coupling coefficient matrix increases, the profit generated by the *COTRA* framework exhibits a significant upward trend, particularly when the number of data sources M is larger, as indicated by the distinct curves for $M = 5$ and $M = 10$. This suggests that *stronger interdependencies between data products enhance the value of the combined offerings, leading to higher profits*. Specifically, for $M = 10$, *COTRA* outperforms the other two methods by averages of 18.07% and 73.54%, effectively leveraging these interdependencies to maximize revenue.

Additionally, Fig. 8(b) shows that the pricing of Product 1 under the *COTRA* framework rises with the coupling coefficient, reflecting *the added value from stronger relational dependencies between products*. This pricing is also influenced by the number of data sources, with a more substantial increase observed when $M = 10$. In contrast, the UNIFORM and NOCO show relatively stable and lower pricing, indicating their limited ability to capitalize on the coupling effect.

3) Cost Coefficients: We analyze the impact of the cost coefficients on framework performance in Fig. 7(b) and (c). They examine the effects of varying means of β_D (related to data transmission costs) and β_V (associated with view transmission costs), respectively. Fig. 7(b) indicates that as the mean of β_D increases, the profit generated by the *COTRA* framework generally declines, reflecting the expected outcome that higher transmission costs reduce overall profitability. However, *COTRA's profit curve shows a more gradual decline compared to the steeper decreases observed in the UNIFORM and NOCO methods*, suggesting *COTRA*'s ability to adjust pricing strategies and sustain profitability despite rising costs. On average, *COTRA* outperforms the UNIFORM and NOCO methods by 47.68% and 71.94%, respectively.

Similarly, Fig. 7(c) shows that as the mean of β_V increases, profits decrease across all frameworks. On average, *COTRA* outperforms the UNIFORM and NOCO methods by 20.78% and 78.83%, respectively. *COTRA appears to maintain relatively higher profits at moderate levels of β_V, with significant declines only at higher cost levels*. This trend implies that *COTRA* may be more effective at managing higher costs through strategic pricing adjustments and optimizing the selection of profitable views.

To improve clarity and provide direct data support for our experimental analysis, we have added two summary tables as shown in Tables IV and V.

These experimental results reflect practical scenarios, where integrating data from multiple sources is a common strategy to improve model accuracy and decision quality. The consistent performance gains observed across tasks demonstrate that our framework not only supports theoretical analysis, but also delivers measurable benefits in real-world data applications.

TABLE IV
PERFORMANCE COMPARISON UNDER STRUCTURED COOPERATION

Parameter	Setting	COTRA	NOCO	UNIFORM	Gain over NOCO (%)	Gain over UNIFORM (%)
Coefficient α	0.0	1.75	1.30	1.70	34.62	2.94
	0.2	1.82	1.30	1.60	40.00	13.75
	0.4	2.15	1.30	1.50	65.38	43.33
	0.6	2.60	1.30	1.70	100.00	52.94
	0.8	3.25	1.30	1.90	150.00	71.05
	1.0	3.90	1.30	1.95	200.00	100.00
Coefficient β	0.0	2.80	1.45	1.55	93.10	80.65
	0.2	2.40	1.10	1.25	118.18	92.00
	0.4	1.85	0.80	1.00	131.25	85.00
	0.6	1.30	0.55	0.70	136.36	85.71
	0.8	0.65	0.30	0.35	116.67	85.71
	1.0	0.15	0.08	0.10	87.50	50.00

TABLE V
PERFORMANCE COMPARISON UNDER ADAPTIVE COOPERATION

Parameter	Setting	COTRA	NOCO	UNIFORM	Gain over NOCO (%)	Gain over UNIFORM (%)
# Sources	2	1.35	0.54	0.88	150.00	53.41
	3	2.08	0.88	1.28	136.36	62.50
	4	2.83	1.21	1.95	133.88	45.13
	5	3.78	1.89	2.37	100.00	59.49
	6	4.56	2.29	3.01	99.13	51.50
	7	5.74	2.73	3.48	110.26	64.94
	8	6.38	2.88	4.03	121.53	58.32
	9	6.97	3.12	4.49	123.40	55.23
	10	7.89	3.56	5.16	121.62	52.71
Coefficient β^D	0.0	3.18	1.98	2.15	60.61	47.91
	0.2	2.84	2.05	2.11	38.54	34.60
	0.4	2.33	1.58	1.66	47.47	40.36
	0.6	1.88	1.29	1.42	45.74	32.39
	0.8	1.49	1.12	1.17	33.04	27.35
	1.0	1.12	0.90	0.93	24.44	20.43
Coefficient β^V	0.0	2.95	1.91	2.09	54.45	41.15
	0.2	2.62	1.84	1.91	42.39	37.17
	0.4	2.21	1.52	1.64	45.39	34.76
	0.6	1.78	1.21	1.37	47.11	29.93
	0.8	1.31	1.03	1.08	27.18	21.30
	1.0	0.96	0.80	0.85	20.00	12.94
Coupling Matrix	0.0	0.88	0.38	0.52	131.58	69.23
	0.2	2.93	1.28	1.88	128.91	55.85
	0.4	5.17	2.31	3.29	123.81	57.10
	0.6	6.89	3.12	4.61	120.83	49.46
	0.8	8.61	3.94	5.73	118.53	50.26
	1.0	9.15	4.18	6.28	118.89	45.70

VI. CONCLUSION

In this paper, we introduced *COTRA*, a data trading framework designed to facilitate multi-source data cooperation. Our key contribution lies in addressing the critical yet often overlooked issue of coupling among data products. By supporting both structured and adaptive cooperation patterns, *COTRA* significantly enhances data utilization and optimizes seller revenues. We developed closed-form solutions for the inherently non-convex optimization problems associated with multi-source data trading.

Under *structured cooperation*, sellers operate within a predefined framework, offering predictability and efficiency in stable environments where data relationships are clear and well-defined. In contrast, *adaptive cooperation* provides greater scalability and flexibility, which is particularly advantageous in scenarios with numerous data sources and intricate product interdependencies. Our extensive experiments demonstrate that *COTRA* can increase seller profits by up to 46.32% compared to traditional data trading methods. Moreover, results indicate that when the coupling level exceeds a certain threshold, the benefits of adaptive cooperation become increasingly significant, boosting profitability and resilience in varied market conditions.

For future work, we plan to explore more complex dynamics and diverse market structures, such as competitive environments with dynamic pricing. This will expand *COTRA*'s applicability and enhance its robustness across various data trading contexts. Future work could also explore incorporating cryptographic privacy-enhancing techniques such as secure multiparty computation [48], homomorphic encryption [49], or differential privacy [50] into the data aggregation process, enabling stronger end-to-end protection of sensitive attributes in multi-source cooperative environments.

REFERENCES

[1] J. Cheng, N. Ding, J. C. Lui, and J. Huang, "Cooperative multi-source data trading," in *Proc. 2024 IEEE Glob. Commun. Conf.*, 2024, pp. 1–6.

[2] Grand View Research, "Data marketplace market report," 2023. [Online]. Available: https://www.grandviewresearch.com

[3] Microsoft Azure, 2025. [Online]. Available: https://azuremarketplace.microsoft.com

[4] Info Chimps, 2025. [Online]. Available: http://infochimps.org

[5] Xignite, 2025. [Online]. Available: https://www.xignite.com

[6] M. Zhang, F. Beltrán, and J. Liu, "A survey of data pricing for data marketplaces," *IEEE Trans. Big Data*, vol. 9, no. 4, pp. 1038–1056, Aug. 2023.

[7] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang, "Data pricing in machine learning pipelines," *Knowl. Inf. Syst.*, vol. 64, pp. 1417–1455, 2022.

[8] K. Binmore, A. Rubinstein, and A. Wolinsky, "The Nash bargaining solution in economic modelling," *RAND J. Econ.*, vol. 17, pp. 176–188, 1986.

[9] C. Chen, Y. Yuan, Z. Wen, Y.-P. Wang, and G. Wang, "GShop: Towards flexible pricing for graph statistics," in *Proc. IEEE 40th Int. Conf. Data Eng.*, 2024, pp. 2612–2624.

[10] X. Zhang, W. T. Yue, Y. Yu, and X. Zhang, "How to monetize data: An economic analysis of data monetization strategies under competition," *Decis. Support Syst.*, vol. 173, 2023, Art. no. 114012.

[11] J. Chen, J. Farooq, and Q. Zhu, "QoS-based contract design for profit maximization in IoT-enabled data markets," *IEEE Internet Things J.*, vol. 10, no. 11, pp. 10080–10094, Jun. 2023.

[12] H. Xiao, L. Cai, J. Feng, Q. Pei, and W. Shi, "Resource optimization of MAB-based reputation management for data trading in vehicular edge computing," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5278–5290, Aug. 2023.

[13] Z. He and Z. Cai, "Trading aggregate statistics over private Internet of Things data," *IEEE Trans. Comput.*, vol. 73, no. 2, pp. 394–407, Feb. 2024.

[14] L. Xue, J. Ni, D. Liu, X. Lin, and X. Shen, "Blockchain-based fair and fine-grained data trading with privacy preservation," *IEEE Trans. Comput.*, vol. 72, no. 9, pp. 2440–2453, Sep. 2023.

[15] A. E. Roth, *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[16] J. Zhang et al., "Dynamic shapley value computation," in *Proc. IEEE Int. Conf. Data Eng.*, 2023, pp. 639–652.

[17] J. Zhang, Q. Sun, J. Liu, L. Xiong, J. Pei, and K. Ren, "Efficient sampling approaches to shapley value approximation," in *Proc. ACM Manage. Data*, vol. 1, no. 1, pp. 1–24, 2023.

[18] H. Xia et al., "Equitable data valuation meets the right to be forgotten in model markets," in *Proc. VLDB Endowment*, vol. 16, no. 11, pp. 3349–3362, 2023.

[19] X. Luo and J. Pei, "Applications and computation of the shapley value in databases and machine learning," in *Proc. 2024 Int. Conf. Manage. Data*, 2024, pp. 630–635.

[20] Z. Liu, H. A. Just, X. Chang, X. Chen, and R. Jia, "2D-Shapley: A framework for fragmented data valuation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 21730–21755.

[21] J. T. Wang, Y. Zhu, Y.-X. Wang, R. Jia, and P. Mittal, "A privacy-friendly approach to data valuation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 60429–60467.

[22] M. Doostmohammadian, Z. R. Gabidullina, and H. R. Rabiee, "Nonlinear perturbation-based non-convex optimization over time-varying networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 6, pp. 6461–6469, Nov./Dec. 2024.

[23] Denodo, 2025. [Online]. Available: https://www.denodo.com/en/data-management/data-virtualization

[24] A. Gkoulalas-Divanis, D. Vatsalan, D. Karapiperis, and M. Kantarcioglu, "Modern privacy-preserving record linkage techniques: An overview," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4966–4987, 2021.

[25] A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt, "FedX: Optimization techniques for federated query processing on linked data," in *Proc. 10th Int. Semantic Web Conf.*, Springer, 2011, pp. 601–616.

[26] B. Brooks and S. Du, "Optimal auction design with common values: An informationally robust approach," *Econometrica*, vol. 89, no. 3, pp. 1313–1360, 2021.

[27] Y. Wang, B. Chen, and D. Simchi-Levi, "Multimodal dynamic pricing," *Manage. Sci.*, vol. 67, no. 10, pp. 6136–6152, 2021.

[28] M. A. Alsheikh, D. T. Hoang, D. Niyato, D. Leong, P. Wang, and Z. Han, "Optimal pricing of Internet of Things: A machine learning approach," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 4, pp. 669–684, Apr. 2020.

[29] M. Doostmohammadian, M. I. Qureshi, M. H. Khalesi, H. R. Rabiee, and U. A. Khan, "Log-scale quantization in distributed first-order methods: Gradient-based learning from distributed data," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 10948–10959, 2025.

[30] C. Ying, X. Hua, M. Zhuo, C. Xin, and H. Jiwei, "Cost-efficient edge caching for noma-enabled IoT services," *China Commun.*, vol. 21, no. 8, pp. 182–191, 2024.

[31] X. Cao et al., "Communication-efficient distributed learning: An overview," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 851–873, Apr. 2023.

[32] Q. Xu, J. Wang, S.-F. Sun, Z. Liu, and X. Chen, "Practical Equi-join over encrypted database with reduced leakage," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 5, pp. 2846–2860, May 2025.

[33] S. Anuyah, V. Bolade, and O. Agbaakin, "Understanding graph databases: A comprehensive tutorial and survey," 2024, *arXiv:2411.09999*.

[34] X. Xiong, J. Yu, and Z. He, "AutoQuo: An adaptive plan optimizer with reinforcement learning for query plan selection," *Knowl.-Based Syst.*, vol. 306, 2024, Art. no. 112664.

[35] N. Ding, L. Gao, and J. Huang, "Joint participation incentive and network pricing design for federated learning," in *Proc. IEEE Conf. Comput. Commun.*, 2023, pp. 1–10.

[36] S. Lin, D. Sow, K. Ji, Y. Liang, and N. Shroff, "Non-convex bilevel optimization with time-varying objective functions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2024, Art. no. 1293.

[37] G. Stomberg, A. Engelmann, and T. Faulwasser, "Decentralized non-convex optimization via bi-level SQP and ADMM," in *Proc. IEEE 61st Conf. Decis. Control*, 2022, pp. 273–278.

[38] P. Guo and X. Zhu, "Focus programming: A bi-level programming approach to static stochastic optimization problems," *Int. Trans. Oper. Res.*, vol. 30, no. 6, pp. 3833–3862, 2023.

[39] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[40] W. Wu, J. Zhou, A. He, S. Han, J. Chen, and B. Zheng, "MetaSplit: Meta-split network for limited-stock product recommendation," in *Proc. ACM Web Conf.*, 2024, pp. 216–225.

[41] R. W. Hamilton and A. S. Hosany, "On the strategic use of product scarcity in marketing," *J. Acad. Marketing Sci.*, vol. 51, no. 6, pp. 1203–1213, 2023.

[42] X. Liu, Y. Zhang, W. Qi, and J. Wang, "Optimal pricing of online products based on customer anchoring-adjustment psychology," *Int. Trans. Oper. Res.*, vol. 31, no. 1, pp. 448–477, 2024.

[43] K. Toften and T. Hammervoll, "Niche marketing research: Status and challenges," *Marketing Intell. Plan.*, vol. 31, no. 3, pp. 272–285, 2013.

[44] V. V. Kalashnikov, S. Dempe, G. A. Pérez-Valdés, N. I. Kalashnykova, and J.-F. Camacho-Vallejo, "Bilevel programming and applications," *Math. Problems Eng.*, vol. 2015, no. 1, 2015, Art. no. 310301.

[45] E. L. Lawler and D. E. Wood, "Branch-and-bound methods: A survey," *Operations Res.*, vol. 14, no. 4, pp. 699–719, 1966.

[46] T. Tianchi, "Alibaba product recommendation dataset," 2018. [Online]. Available: https://tianchi.aliyun.com/dataset/46

[47] W. Wang et al., "A user purchase behavior prediction method based on XGBoost," *Electronics*, vol. 12, no. 9, 2023, Art. no. 2047.

[48] A. C. Yao, "Protocols for secure computations," in *Proc. 23rd Annu. Symp. Found. Comput. Sci.*, 1982, pp. 160–164.

[49] C. Gentry, "Fully homomorphic encryption using ideal lattices," in *Proc. 41st Annu. ACM Symp. Theory Comput.*, 2009, pp. 169–178.

[50] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata Lang. Program.*, Springer, 2006, pp. 1–12.

Jin Cheng received the BS degree from Xiamen University, and the MS degree from the University of Chinese Academy of Sciences. She is currently working toward the PhD degree with the joint doctoral program between the Chinese University of Hong Kong, Shenzhen and the Chinese University of Hong Kong. Her research interests include data trading, network economics, online optimization, and database management systems.

Ningning Ding received the BS degree in information science and engineering from Southeast University, in 2018, and the PhD degree in information engineering from the Chinese University of Hong Kong, in 2022. She is a tenure-track assistant professor with the Data Science and Analytics Thrust and the Internet of Things Thrust within the Information Hub, Hong Kong University of Science and Technology (Guangzhou). Before that, she was a postdoctoral scholar with the Department of Electrical and Computer Engineering, Northwestern University, USA. Her research focuses on interdisciplinary areas of artificial intelligence, network systems, and network economics, with a current emphasis on federated learning, machine unlearning, and data trading.

John C. S. Lui (Fellow, IEEE) received the PhD degree in computer science from UCLA. He is currently the Choh-Ming Li chair professor with the Department of Computer Science & Engineering (CSE), Chinese University of Hong Kong (CUHK). After his graduation, he joined the IBM Laboratory. He later joined the CSE Department, CUHK. His current research interests include quantum networks, online learning algorithms and applications (e.g., multi-armed bandits, reinforcement learning), machine learning on network sciences and networking systems, large-scale data analytics, network economics, large-scale storage systems, and performance evaluation theory. He has served with the IEEE Fellow Review Committees. He is an elected member of the IFIP WG 7.3, fellow of ACM, senior research fellow of the Croucher Foundation, fellow of the Hong Kong Academy of Engineering Sciences (HKAES).

Jianwei Huang (Fellow, IEEE) is a presidential chair professor and associate vice president with the Chinese University of Hong Kong, Shenzhen. He is also the associate director with the Shenzhen Institute of Artificial Intelligence and Robotics for Society. He has served as the editor-in-chief of the *IEEE Transactions on Network Science and Engineering* and the associate editor-in-chief of the *IEEE Open Journal of the Communications Society*. He is an IEEE ComSoc distinguished lecturer, a Clarivate Web of Science highly cited researcher, and an Elsevier Most Cited Chinese researcher.