2024 IEEE Global Communications Conference: Selected Areas in Communications: Big Data

Cooperative Multi-source Data Trading

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

GLOBECOM 2024 - 2024 IEEE Global Communications Conference | 979-8-3503-5125-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/GLOBECOM52923.2024.10900935

***Abstract*—In the era of big data, data trading significantly enhances data-driven technologies by facilitating data sharing. Despite the clear advantages often experienced by data users when incorporating multiple sources, the topic of multi-source data trading remains largely unexplored. This paper designs a novel data trading framework, which enables multi-source data trading through multi-source cooperation. The proposed framework aims to improve data usage efficiency and increase seller revenue. In particular, we model data sellers' cooperative decisions through the Nash bargaining framework and systematically outline the interactions between sellers and buyers as a two-stage Stackelberg game. A key contribution of this work is the consideration of coupling among diverse data products, which is essential but often overlooked in prior studies. We properly classify data's utility into endogenous and relational categories to disentangle the coupling. Despite the inherent non-convex nature of the optimization problem, we methodically derive the closed-form optimal solutions by decomposing the problem into several subproblems. Interestingly, we reveal that, under our proposed framework, sellers' revenue initially remains steady with the increase of product coupling level, but begins to rise once the level exceeds a certain threshold due to the substitute effect. Finally, experimental results show that our proposed framework can improve the seller's profit by up to 46.32% compared to traditional data trading methods in the current data market.**

I. Introduction

A. Background and Motivation

Data trading has significantly promoted the development of data-driven technologies, such as artificial intelligence (AI), through expanding access to diverse datasets and facilitating model training. In 2022, the global data trading market attained a valuation of $968 million, with a projected compound annual growth rate of 25% from 2023 to 2030 [1]. This burgeoning landscape has witnessed the emergence of numerous data trading platforms in the industry, such as Windows Azure [2], Info Chimps [3], and Xignite [4], and extensive research on data trading theoretical analysis over the past decade [5].

However, existing research on data trading has largely ignored the benefits of multi-source data cooperation, consequently restricting data utilization. Current works assumed either a single data source or multiple independent ones and mainly focused on the properties that the data pricing should exhibit [6] (e.g., arbitrage-free and flexible). This oversight leads to insufficient data utilization, impeding the improvement of data-driven technologies. Consider the example in Table I, which involves three data sources from different owners: (1) customer source (Data A), which contains customer-centric attributes (e.g., customer's region); (2) product source (Data B), which stores product-specific details from the manufacturing plant (e.g., material type); and (3) transaction source (Data C), which holds transaction records with Customer ID (CID), Product ID (PID), and Transaction ID (TID). Logically, with sensitive information (i.e., IDs) removed, three view types can be marketed as data products, as shown in Table II. 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. However, without cooperation among different data sources, only Views A and B can be offered separately by different owners. Even if the owner of Data C purchases Views A and B, View C cannot be synthesized due to the lack of key and sensitive information, resulting in a significant underutilization of data.

TABLE I: Data sources.

Data A		Data B		Data C		
CID	**Region**	**PID**	**Type**	**TID**	**CID**	**PID**
938	California	AOJ	A	1	938	AOJ
...	...	...	...	...	...	...

TABLE II: Data products.

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

B. Contributions

To fill this gap, we design a novel data trading framework, *COTRA*, to enable cooperation among multiple data sources. *COTRA* facilitates a wide range of product types, consequently enhancing data utilization. In the proposed framework, sellers include an aggregator and multiple data owners. The aggregator synthesizes data products by cooperating with different owners and subsequently markets these products. We use a two-stage Stackelberg game to model the interactions among sellers and buyers. In Stage I, data sellers cooperatively set

Jin Cheng is with the School of Science and Engineering, Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen (Email: jincheng2@link.cuhk.edu.cn). Ningning Ding is with the Data Science and Analytics Thrust, Information Hub, Hong Kong University of Science and Technology (Guangzhou) (Email: ningningding@hkust-gz.edu.cn). John C.S. Lui is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong (Email: 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, The Chinese University of Hong Kong, Shenzhen (Corresponding Author, Email: jianweihuang@cuhk.edu.cn).

This work is supported by the National Natural Science Foundation of China (Project 62271434), Shenzhen Science and Technology Innovation Program (Project JCYJ20210324120011032), Guangdong Basic and Applied Basic Research Foundation (Project 2021B1515120008), Shenzhen Key Lab of Crowd Intelligence Empowered Low-Carbon Energy Network (No. ZDSYS20220606100601002), Shenzhen Stability Science Program 2023, and the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

prices and revenue distribution strategies based on the Nash bargaining theory [7], ensuring fair and Pareto optimal benefit sharing among sellers. In Stage II, each data buyer makes purchase decisions, and the aggregator coordinates product provisioning with multiple sources.

A key modeling contribution of this work is the consideration of *coupling* among diverse data products, which is important but often overlooked in the past literature. For example, we consider the inherent coupling exemplified by View C's partial coverage of Views A and B. This aspect notably influences buyer purchase decisions, as we will show in the later analysis and experiments, but is often overlooked in existing works [5]. Additionally, the low replication cost of data necessitates a distinct pricing model from other products. To address the coupling problem, we systematically categorize the utility into endogenous and relational categories to disentangle the coupling, enhancing our ability to characterize buyer payoffs and pricing strategies. Hence, two key questions arise from cooperative multi-source data trading:

- **Key Question 1**: For sellers, *how to cooperatively decide the pricing strategy for coupled data products and distribute the revenue?*
- **Key Question 2**: For buyers, *how to characterize the utility derived from coupled products, and what is the impact of the coupling regarding their purchasing decisions?*

The technical challenge arises due to the inherent *non-convexity of the optimization problem*, stemming from ambiguous numerical relationships among parameters. Addressing this is challenging, given its non-convex and non-monotonic nature. To tackle this, we categorize our analysis into distinct cases based on system parameters, which may exhibit either convex or non-convex characteristics. In both locally convex and non-convex cases, we identify the unique optimal solution through a series of analyses and prove the optimality. Finally, we conduct a comprehensive analysis of the non-convex problem, deriving the closed-form solution.

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. This holds significant implications for enhancing data utilization and further expanding the data market.
- ***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, under the *COTRA* framework, sellers' revenue initially remains steady with the increase of product coupling level due to the substitute effect but begins to rise once the level exceeds a certain threshold. Experimental results show that *COTRA* can improve the sellers' profit by up to 46.32% compared to existing trading methods.

II. Related Works

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

Single-source Data Trading: Prior research in this strand predominantly revolves around desired properties for data pricing function, such as arbitrage-free [5], revenue maximization [8], and flexibility [6]. These studies usually focus on designing the properties that the pricing function should hold but do not give a concrete characterization of the utility and the closed-form solution. Furthermore, this line of work only considers a single data source, neglecting the scenarios involving multiple sources.

Multi-source Data Evaluation: Most works along this line evaluate data instances in machine learning. These studies usually utilize Shapley Value [7] and develop efficient algorithms to reduce the computation complexity [9]. Although these works evaluate different data instances potentially from different sources, the sources remain independent, with a lack of cooperative dynamics(e.g., sharing sensitive information), which limits data utilization. Besides, due to the high computation complexity, there is no closed-form solution.

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

III. System Model

In this section, we present a cooperative multi-source data trading framework, *COTRA*, as shown in Fig. 1. For each transaction, one buyer will purchase views in the view set $\mathcal{V}$ from sellers $\mathcal{S}$, comprising an aggregator and data owners. We integrate data virtualization technology [10], which can achieve real-time access to data stored across multiple data sources and is widely used in current data management systems. This technology enables the aggregator to operate efficiently without relying on a central database or cache to store data products. Instead, it maintains a catalog of available data products and delivers products on-demand through multi-party coordination, enhancing the flexibility of data trading.

We categorize products into endogenous (e.g., View A) and relational (e.g., View C) types. Endogenous products derive utility from inherent data, while relational products rely on both inherent data and data relationships. We consider two-dimension relational views, noting that the multi-dimension case can be captured by multiple binary joins [11]. Data sources are classified in line with these distinctions.

A. Buyer Modeling

Each data transaction serves one buyer. The one-buyer scenario is a reasonable approximation of the scenario when

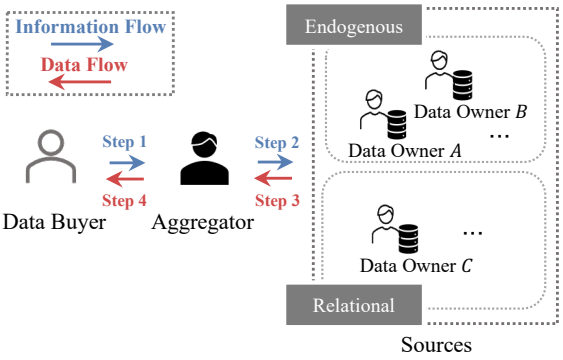

Fig. 1: The data trading framework, *COTRA*.

multiple heterogeneous buyers come to the platform one at a time, and sellers have unlimited capacities. Next, we will introduce 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 $\boldsymbol{x} = (x_j : x_j \in [0,1] \text{ and } j \in \mathcal{V})$.

Data Usage: Given the purchase decision, we can calculate the data usage[1]. We initially assume a uniform distribution of endogenous products within the relational product[2]. 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 of the views j is x_j, then $\alpha_j^i x_j$ of the endogenous views i will be covered. To ensure equitable trading, sellers prioritize the uncovered portions when selling endogenous products, which is also easily implemented in practice. As such, the actual usage of an endogenous view i is $y_i(\boldsymbol{x}) = \min(1, x_i + \sum_{j \in \mathcal{V}_r} \alpha_j^i x_j)$, whereas that of a relational view j is $y_j(\boldsymbol{x}) = x_j$.

Utility: We differentiate utility into endogenous and relational categories. Within the domain of machine learning, utility is often represented as task accuracy [6]. 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 [12], we denote the basic utility function of usage by:

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

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. We incorporate utility coefficients μ_e to signify the preference for endogenous utility and μ_r to signify the preference for relational utility[3]. Therefore, the total utility of the buyer is expressed as follows:

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

[1]Notably, the relational view partially overlaps with the endogenous view, allowing it to act as a substitute, a fact often overlooked in existing studies.

[2]This assumption is not restrictive but strategic for two reasons. First, we can practically implement this through data item reordering. Second, this simplification is essential, considering the inherent challenge of the problem.

[3] For clarity, we assume homogeneity within product types, keeping our focus on the core problem. Consideration of heterogeneity is reserved for future work, as detailed in Section VII.

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, denoted as $\mathcal{S}$, encompasses both data owner set $\mathcal{S}_o$ and the aggregator set $\mathcal{S}_a$. Next, we introduce the sellers' revenue, cost, and payoff.

Revenue: The sellers cooperatively determine the price $\boldsymbol{p} = (p_i : \forall i \in \mathcal{V})$ and the revenue distribution ratio $\boldsymbol{r} = (r_j : \forall j \in \mathcal{S})$. This implies that, for revenue $R(\boldsymbol{p}) = \sum_{i \in \mathcal{V}} p_i x_i$, seller $j \in \mathcal{S}$ will get distributed revenue $r_j R(\boldsymbol{p})$.

Cost: For data owners, the cost corresponds to the transmission expense of data, while aggregators bear costs for both data and view transmission. A common practice [12, 13] 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., $\mu_e \log 2 \geq 2\beta$ and $\mu_r \log 2 \geq 2\beta$.

Payoff: Each seller's payoff is calculated by deducting the cost from the distributed revenue. Therefore, the payoff of data owner $j \in \mathcal{S}_o$ is:

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

The payoff of the aggregator is:

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

C. Two-stage Stackelberg Game

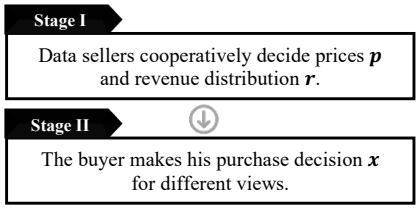

Fig. 2: The two-stage Stackelberg game of *COTRA*.

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

- In Stage I, sellers cooperatively determine price $\boldsymbol{p}$ and revenue distribution $\boldsymbol{r}$. We model sellers' interactions through Nash bargaining theory [7], ensuring fair and optimal allocation among decision-makers in a cooperative framework. We formulate the decision problem for sellers as follows:

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

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

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

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

TABLE III: Theorem 1 (Buyer's Optimal Decision): solution of Stage II.

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

where $y_0 = \frac{-(1+\alpha)p_r + 2\alpha\mu_e + \alpha\mu_r + \sqrt{(-1+\alpha)^2 p_r^2 + 2(\alpha-1)\alpha p_r(2\mu_e - \mu_r) + \alpha^2(2\mu_e + \mu_r)^2}}{2\alpha p_r}$.

- In Stage II, the buyer determines the purchase decision $\boldsymbol{x}$. Thus, we formulate the buyer's decision problem as follows:
Problem $\mathbb{P}_2$: *Payoff Maximization for the Buyer.*

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

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

To derive the closed-form solutions and valuable insights into the problem, we investigate the case involving two endogenous sources (i.e., Data A and B) and one relational source (i.e., Data C) as detailed in Tables I and II, where the coupling coefficient for the relational product across the two endogenous products is α [4]. This case is a plausible approximation of the scenario when multiple heterogeneous relational sources with disjoint endogenous sources and sellers have unlimited capacities. We defer the analysis of the general case with joint endogenous sources to our forthcoming journal paper. As observed in Sections IV and V, even though this case is straightforward, we retain the key part of the problem for focus, and the problem presents several challenges due to non-convexity and ambiguous parameter relationships. Next, we will analyze the Nash equilibrium by backward inductions.

IV. Stage II: Buyer's Optimal Decisions

In this section, we derive the optimal decisions of the buyer in Stage II given the sellers' prices in Stage I. Decisions associated with the same kind of products are identical due to the homogeneity. Consequently, for endogenous and relational products, the buyer's decisions fall into two parts, x_e and x_r. Next, we reformulate the original problem into an equivalent convex problem, utilizing Lemma 1.

Lemma 1: *Buyer's optimal decisions satisfy:* $x_e^* \le 1 - \alpha x_r^*$.

Proof: Suppose that $x_e^* > 1 - \alpha x_r^*$, then the buyer's payoff is:

$$U_b(\boldsymbol{x}^*, \boldsymbol{p}) = 2\mu_e \ln 2 + \mu_r \ln(1 + x_r^*) - 2x_e^* p_e - x_r^* p_r. \tag{8}$$

[4]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.

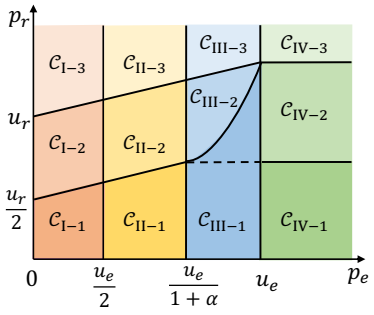

Fig. 3: Theorem 1: solution of Stage II.

For a given small positive value ϵ, we have $x_e' = x_e^* - \epsilon \ge 1 - \alpha x_r^*$, and $x_r' = x_r^*$. We can observe that:

$$U_b(\boldsymbol{x}', \boldsymbol{p}) > U_b(\boldsymbol{x}^*, \boldsymbol{p}), \tag{9}$$

which contradicts the initial assumption.

This lemma states 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:

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

$$s.t. \;\; y_e, y_r, y_e - \alpha y_r \in [0,1],$$

where y_e and y_r are exactly the usage of different views determined by $\boldsymbol{x}$, as illustrated in Section III-A. After determining y_e and y_r, we can get corresponding $\boldsymbol{x}$. We can verify that the Hessian matrix is negative semi-definite. Hence, the reformulated problem is convex.

Applying the Karush-Kuhn-Tucker (KKT) conditions [14], we derive the optimal buyer's decisions, summarized in Table III and visually presented in Fig. 3 for clarity.

Theorem 1 (Optimal Buyer's Decisions): *There are 12 cases of optimal buyer's decisions decided by the sellers' pricing strategy p_e and p_r, as shown in Table III and Fig. 3.*

Proof of Theorem 1 is given in Appendix B in the technical report [15]. Based on the given optimal buyer's decisions, the following insights emerge:

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

- ***Case I ("All-In")**: When p_e is sufficiently low (i.e., $\mathcal{C}_I$), the buyer purchases part of the relational view and buys enough endogenous views to supplement their usage y_e to 1.*
- ***Case II ("Partial-Purchase")**: With moderately low p_e (i.e., $\mathcal{C}_{II}$), 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., $\mathcal{C}_{III}$), an interesting scenario emerges. With sufficiently low p_r, the buyer purchases the entire relational view to fulfill the demand of the endogenous view due to the coupling, which shows 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.*

In this section, we get 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 by backward induction, considering the buyer's optimal decisions.

V. Stage I: Sellers' Optimal Decisions

In this section, we analyze sellers' optimal pricing strategy and revenue distribution in Stage I by backward induction, considering buyers' optimal decisions in Stage II. We first delve into the Nash bargaining framework to guide our analysis of revenue distribution and pricing strategy. We present the solutions to Problem $\mathbb{P}_1$ in Theorems 2 and 3. For clarity, we define 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): *Sellers' optimal revenue distribution 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 III-A.

Proof of Theorem 2 is given in Appendix C in the technical report [15]. According to Theorem 2, the payoff for seller $j \in \mathcal{S}_o$ is determined as follows:

$$U_j(\boldsymbol{p}, r_j^*(\boldsymbol{p})) = U_a(\boldsymbol{p}, r_a^*(\boldsymbol{p})) = (\sum_{i\in\mathcal{V}} p_i x_i^*(\boldsymbol{p}) - C(\boldsymbol{p}))/4, \quad (11)$$

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

Problem $\mathbb{P}_3$: *Pricing Optimization for Sellers.*

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

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

The challenge in solving Problem $\mathbb{P}_3$ lies in the inherent non-convexity of the problem, due to non-convexity parameter relationships as detailed in Table III. To tackle the challenge, we categorize our analysis into several cases based on system parameters, which exhibit either convexity or non-convexity. We identify that Cases $\mathcal{C}_I$ and $\mathcal{C}_{II}$ are convex, and by using variable substitutions, we split the problem in these cases into two convex subproblems to decouple the decisions. In non-convex Cases $\mathcal{C}_{III}$ and $\mathcal{C}_{IV}$, we identify a unique optimal pricing strategy through a series of analyses. Accordingly, there are two potential optimal pricing strategies.

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

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

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

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

- ***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 in the online report [15]. Theorem 3 shows that, given the values of μ_e, μ_r, α, and β, the optimal pricing strategies can be determined as either $\boldsymbol{p}_b^*$ or $\boldsymbol{p}_s^*$, depending on which yields higher profits. The bundling strategy $\boldsymbol{p}_b^*$ simultaneously controls p_e and p_r to increase total profits. In contrast, the separated strategy sets a high p_e as the baseline and increases the p_r to ensure the profits from the relational product.

VI. Experimental Results

This section evaluates *COTRA* through a real-world dataset from Taobao [16], showcasing its practical performance. The dataset contains 23 million data points detailing the purchasing behaviors of 20,000 customers over a specified period.

We begin by calibrating model parameters using the dataset to reflect market dynamics accurately. We focus on the data utility for model training tasks, as illustrated in Section III-A. We define "endogenous utility" as the model's ability to reflect the actual data distribution, measured by the Kullback-Leibler (KL) divergence. We compare each data segment's distribution to the overall dataset's distribution by their KL divergence KL_d, defining each segment's endogenous utility as $1 - KL_d$. We define "relational utility" as its capacity to predict relationships between different data elements, as assessed by its performance in recommendation tasks. We measure this utility as the accuracy in forecasting the types of products that consumers are likely to purchase. We conduct the fitting based on an XGBoost model [17], a standard approach for such analyses. Additionally, we set the coupling coefficient (α) to 0.5 and the cost coefficient (β) to 0.05 by default.

Next, we assess the effects of varying the coupling coefficient (α) and the cost coefficient (β) on the performance of the

COTRA framework. We analyze these impacts across diverse market scenarios, using two benchmarks commonly used in the current data trading market for comparison: the uniform pricing (i.e., "UNIFORM") and the non-cooperation frameworks (i.e., "NOCO") [5, 6]. The former framework assigns a uniform price to all products without distinguishing different utilities, while the latter has limited product availability.

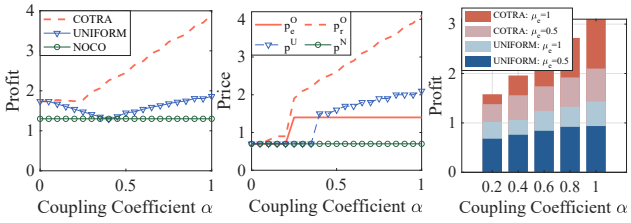

Fig. 4: Impact of varying the coupling coefficient α.

- ***Coupling Coefficient*** $\boldsymbol{\alpha}$: Fig. 4 demonstrates the impact of varying the coupling coefficient α. Fig. 4a reveals that *COTRA* outperforms the existing methods in [5, 6] 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 the substitution effect. As α increases, there is a notable transition towards relational products, permitting sellers to modify baseline prices p_e and relational prices p_r to enhance profitability, as illustrated in Fig. 4b. This strategy creates a market dynamic where buyers gravitate towards relational products that satisfy their specific needs, thereby maximizing revenue. In contrast, uniform pricing and non-cooperation frameworks exhibit less flexibility and adaptability, leading to lower revenues, as depicted in Figs. 4a and 4c.

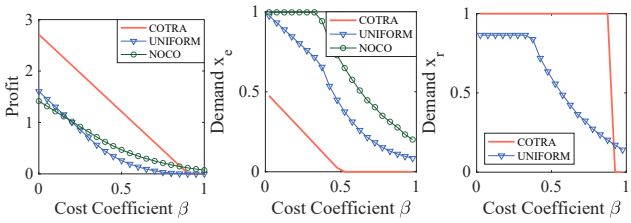

Fig. 5: Impact of varying the cost coefficient β.

- ***Cost Coefficient*** $\boldsymbol{\beta}$: Fig. 5 illustrates the impact of varying the cost coefficient β. As shown in Fig. 5a, *COTRA* yields higher revenues, surpassing the benchmark methods by averages of 43.29% and 38.45%, respectively. This comparison underscores that, unlike the uniform and non-cooperation strategies, which respond similarly to changes in β, our model adapts distinctly, reflecting its resilience to cost variations. Specifically, Fig. 5b shows that in *COTRA*, as costs increase, the demand for endogenous products x_e diminishes to zero and then levels off. Conversely, Fig. 5c exhibits shows initial stability in the demand for relational views x_r, followed by a marked decline. This trend happens because sellers have to raise prices to deal with higher costs. Besides, the substitution effect helps lower the price of endogenous products while keeping relational product prices steady to keep sales going. However, when costs get too high, all market transactions stop.

These experiments show the adaptability and efficiency of *COTRA* under various market conditions, setting a foundation for more dynamic data trading environments.

VII. Conclusion

This paper provides a data trading framework, *COTRA*, for cooperative data provisioning from multiple sources. We consider an important yet overlooked aspect: the coupling among data products. Despite the challenges from non-convex optimization, we derived close-formed solutions. We revealed that with *COTRA*, sellers' revenue initially remains steady but rises once the coupling level surpasses a certain threshold. Our experiments indicate that *COTRA* can increase the seller profit by up to 46.32% compared to current data trading methods. Future research works will focus on enhancing the versatility of the proposed framework by investigating its applicability to a broader range of real-world scenarios, encompassing a diversity of data product types.

References

[1] Grand View Research, "Data marketplace market report, 2023," https://www.grandviewresearch.com.
[2] Microsoft Azure, https://azuremarketplace.microsoft.com.
[3] Info Chimps, http://infochimps.org.
[4] Xignite, https://www.xignite.com.
[5] M. Zhang, F. Beltrán, and J. Liu, "A survey of data pricing for data marketplaces," *IEEE Transactions on Big Data*, 2023.
[6] Z. Cong, X. Luo, J. Pei, F. Zhu, and Y. Zhang, "Data pricing in machine learning pipelines," *Knowledge and Information Systems*, vol. 64, pp. 1417–1455, 2022.
[7] A. E. Roth, *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988.
[8] S. Chawla, S. Deep, P. Koutris, and Y. Teng, "Revenue maximization for query pricing," *arXiv:1909.00845*, 2019.
[9] J. Zhang, H. Xia, Q. Sun *et al.*, "Dynamic shapley value computation," in *IEEE ICDE*, 2023.
[10] "Denodo," https://www.denodo.com/en/data-management/data-virtualization.
[11] Z. Chang, D. Xie, S. Wang, F. Li, and Y. Shen, "Towards practical oblivious join processing," *IEEE TKDE*, 2023.
[12] M. A. Alsheikh, D. T. Hoang, D. Niyato *et al.*, "Optimal pricing of internet of things: A machine learning approach," *IEEE JSAC*, vol. 38, no. 4, pp. 669–684, 2020.
[13] N. Ding, L. Gao, and J. Huang, "Joint participation incentive and network pricing design for federated learning," in *IEEE INFOCOM*, 2023.
[14] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[15] "Online technical report," https://www.dropbox.com/scl/fi/36have3x3udrzs0a26jiu/Proof.pdf?rlkey=npjq1ft9lvynx17d6847vbgen&dl=0.
[16] Taobao Tianchi, "Alibaba product recommendation dataset," https://tianchi.aliyun.com/dataset/46.
[17] W. Wang, W. Xiong, J. Wang, L. Tao, S. Li, Y. Yi, X. Zou, and C. Li, "A user purchase behavior prediction method based on xgboost," *Electronics*, vol. 12, no. 9, p. 2047, 2023.