Fair Revenue Distribution in Data Markets

Extended Abstract

Nathan Joe Joseph
University of Illinois at Urbana-Champaign
Urbana, United States
njoseph4@illinois.edu

Jiaxin Song
University of Illinois at Urbana-Champaign
Urbana, United States
jiaxins8@illinois.edu

Bhaskar Ray Chaudhury
University of Illinois at Urbana-Champaign
Urbana, United States
braycha@illinois.edu

ABSTRACT

The immense success of ML systems relies heavily on large-scale, high-quality data. The high demand for data has led to several paradigms that involve selling and exchanging data. A key model in this landscape is a data market– a two-sided marketplace that (i) receives ML task requests from buyers, (ii) addresses these requests using the datasets hosted by the sellers, (iii) collects revenue from the buyers, and (iv) compensates the sellers from the earned revenue. Typically, a data market compensates the sellers based on their contributions to the total earned revenue (this is usually measured by standard credit sharing rules, e.g., Shapley shares). However, we observe that multiple data allocations can yield the same optimal revenue while resulting in vastly different compensation outcomes. For example, when multiple sellers offer equally valuable, *non-complimentary* datasets, a revenue-maximizing allocation may select only one, thereby excluding others from compensation despite their comparable data quality. Such discrepancies highlight the need for fairness in revenue distribution. In this paper, we develop a revenue maximization framework for data markets that incorporates fairness constraints for seller compensation. We show that while this problem is NP-hard, we can still obtain a $O(\log n)$-bi-criteria approximation (approximating revenue and fairness) in polynomial time.

CCS CONCEPTS

• **Theory of computation → Algorithmic game theory**.

KEYWORDS

Data marketplace; fairness

ACM Reference Format:
Nathan Joe Joseph, Jiaxin Song, and Bhaskar Ray Chaudhury. 2026. Fair Revenue Distribution in Data Markets: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/VHME3099

1 INTRODUCTION

The success of ML systems is highly dependent on the availability of high-quality heterogeneous data. This demand has led to several paradigms that involve selling (data markets [4]), exchanging (data-exchange [11]) and sharing (federated learning (FL) [37])[1] data. In response to this pressing need for principled economic frameworks handling data, a long line of pioneering work has emerged on data markets [2–4, 8, 10]: a two-sided real-time marketplace that facilitates the buying and selling of data. Most of the foregoing work involves (i) the data marketplace receiving ML task requests from buyers, (ii) the marketplace addressing these requests using the data hosted by the sellers, and (iii) compensating the sellers from the earned revenue.

In this paper, we concern ourselves with *fair revenue distribution* for the sellers in the data market: [4] adopt the policy that the data market divides the revenue among the sellers based on their contribution in generating the total revenue (measured using standard credit sharing rules like the Shapley share). However, there are cases where multiple distinct data allocations[2] yield the same optimal total revenue, but produce significantly different compensation profiles across sellers. Disparities in revenue compensation for the sellers can lead to reduced seller loyalty to the platform and may raise concerns under fair competition regulations, such as the Digital Markets Act (Official Journal of the European Union, 2023).

Furthermore, research indicates that a lack of fairness and transparency in the compensation of sellers negatively affects not only sellers, but also the platform and buyers: undercompensated sellers tend to invest less in curating and annotating their datasets [36], which impairs the platform's ability to effectively match data with buyer needs [36]. This, in turn, reduces service quality and overall revenue.

Given these challenges, it can be argued that fair revenue distribution should be a core principle in data markets. In this work, we develop models and methodologies to support revenue maximization while ensuring fair compensation for data sellers, and we identify fundamental limitations inherent to this goal.

1.1 Our Contributions

Fair Revenue Compensation Problem. We introduce a formal model of revenue maximization where a data marketplace $\mathcal{D}$ hosts datasets from heterogeneous sellers. Given prediction tasks from each buyer, a data marketplace $\mathcal{D}$ charges each buyer based on the marginal improvement in the buyer's prediction achieved by using the datasets hosted in $\mathcal{D}$. Thereafter, $\mathcal{D}$ distributes the revenue among the sellers based on their contribution (usually measured by standard credit-sharing functions like the Shapley Share ([33]).

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/VHME3099

[1] We remark that while FL does not strictly involve agents sharing data, there are works [26] that abstract FL as a data sharing incentivization problem

[2] A data allocation refers to the assignment of datasets to meet the prediction tasks of individual buyers.

We incorporate fairness through a *fair-thresholding constraint* for each seller: that is, seller j must receive a payment of at least τ_j from the data market. Here, τ_j represents seller j's fair compensation threshold. Our fairness criterion aligns with the notion of *share-based fairness* studied in computational social choice [6, 7, 13], which is particularly well-suited for markets involving heterogeneous datasets[3]. In such settings, enforcing equal revenue across sellers may be impractical or undesirable, as datasets can vary significantly in their value to different buyers and sellers may incur differing costs for annotation and curation.

Mathematically, the core problem faced by the data market $\mathcal{D}$ is to maximize its total revenue while ensuring that each seller receives at least their fair compensation threshold τ_j.

Computational Results. We formulate the revenue maximization in data markets, albeit with fairness constraints, as a linear optimization problem with exponentially many variables, call this program Fair-Revenue-Max. The large number of variables is primarily attributed to the complex *combinatorial value of the dataset*; research [15, 21] shows that utilities for a combination of datasets are more complex than sum of utilities for individual datasets. We first show intractability results in solving Fair-Revenue-Max exactly.

Theorem 1. *The Fair-Revenue-Max problem is NP-hard to solve even when there is only one type of buyer and all sellers have the same thresholds $\tau_j = \tau$ for all j.*

We then look at approximation algorithms for solving Fair-Revenue-Max. When the buyer-utilities for datasets are *submodular*, we come up with a polynomial time $\tilde{O}(\log m)$ [4] bi-criteria approximation algorithm for the problem, where m is the number of data-sellers in the data market. In particular, given a data market $\mathcal{D}$ with n types of buyers and m sellers, and thresholds τ_j for each seller j, we outline an algorithm that determines a data allocation to buyers, and a revenue compensation for sellers, such that each seller j receives at least $\tilde{\Omega}(\tau_j/\log m)$ revenue, and the data market receives $\tilde{\Omega}(r^*/\log m)$ total revenue, where r^* is the optimal revenue generated by solving Fair-Revenue-Max exactly. This algorithm is developed by solving the LP corresponding to Fair-Revenue-Max through the algorithmic framework of Plotkin-Shmoys-Tardos [32], and designing the necessary oracles for the foregoing framework.

Theorem 2. *For any Fair-Revenue-Max instance with submodular buyer utilities, in polynomial time, one can construct a solution* $\mathbf{x}$ *such that the data market receives a revenue of* OPT/η*, and each seller j receives a compensation of τ_j/η, where $\eta \in \tilde{O}(\log m)$, and* OPT *refers to the optimal solution to Fair-Revenue-Max.*

In summary, we initiate the study of fair revenue compensation for sellers in data markets, drawing a natural parallel to similar efforts (incorporating fairness) in two-sided optimization settings such as assortment planning [19] and other related online two-sided marketplaces [18]. We introduce a formal mathematical framework, establish intractability results, and develop approximation algorithms under mild and realistic assumptions. We hope our work serves as a starting point for future research on mechanisms for fair competition in data economies.

[3]Datasets that are valued differently by different buyers.

[4]$\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ hides a logarithm of the maximum money. a buyer is willing to pay for per-unit increase of accuracy in her prediction task.

1.2 Related Work

Data economics has been a vibrant area of research, and a full survey of all concepts is well beyond the scope of this paper. We only mention some work that is most relevant to our paper. A long line of work [2, 3, 8, 10, 14, 17] investigates revenue-maximizing strategies of a monopolist data seller. The framework of the data marketplace discussed in our paper has also been explored in the prior literature, e.g., three-layer data market [22] and real-time data marketplace [4]. Pricing policies from other first principles have also been explored in data markets [9, 16, 28, 31]. There are also competitive pricing and allocation rules studied in the context of digital goods that behave similarly to data [25]. Data markets have also been studied in the presence of externalities [1, 20, 24]. [5, 11, 23, 34] consider the data exchange economies, where a group of agents in possession of datasets exchange datasets amongst themselves fairly and voluntarily for mutual benefit without monetary compensation.

The concept of fairness has also been widely explored in the data exchange markets, where the data owners also act as data users. The concept of reciprocal fairness has been proposed by [5, 11, 29], whereby an agent receives payment proportional to its contribution to others' prediction task. The Shapley share [33] and proportional value [35] from cooperative game theory are commonly used to measure the contribution of one's data [5, 11, 12, 29]. Economic concepts, including pricing [38], fairness [29], incentives [29, 30, 39], and steady-states [27, 30] have also been explored in decentralized machine learning frameworks like Federated Learning (FL), which inherently involve principles of data sharing.

2 SUMMARY AND FUTURE DIRECTIONS

This paper introduces a formal model of fair revenue compensation in data markets. We investigate algorithmic aspects of the model and prove that while the general problem is intractable, we can obtain an efficient approximation algorithm using the PST framework.

We see several future directions that stem out from our work. Obviously, the main direction is to investigate our guarantees beyond submodular accuracies, e.g., when datasets differ by features, then a combination of datasets is more useful than summing the benefits from individual datasets (*superadditive accuracies*). Secondly, all of our computational results assume oracle access to only agent utilities for given sets. One could consider *stronger oracle* accesses, e.g., access to $\psi_j^i(S)$ given S. While Shapley shares are hard to compute exactly, good approximations are achievable through efficient sampling. We suspect that we may be able to provide better approximation guarantees for the exact oracle required for the PST framework used in this paper. Another interesting direction is to account for the *strategic behavior* of the buyers and sellers. In particular, the buyer payments and seller compensations are enforced through contracts. Designing mechanisms that incentivize buyers and sellers to truthfully report the μ_i and τ_j seems a natural question. Lastly, we believe that it would be interesting to study the *price of fairness*, i.e., the ratio between the maximum achievable revenue without fairness constraints and that with such constraints. Establishing non-trivial bounds on this quantity in meaningful or practical instances would offer deeper insights into the trade-offs between fairness and efficiency.

ACKNOWLEDGMENTS

The research of Bhaskar Ray Chaudhury and Jiaxin Song was supported by the NSF Career Award CCF-2441580. We also thank the attendees from the workshop on Incentives for Collaborative Learning and Data Sharing at TTIC for their insightful suggestions.

REFERENCES

[1] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. 2022. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics* 14, 4 (2022), 218–256.
[2] Anat R Admati and Paul Pfleiderer. 1986. A monopolistic market for information. *Journal of Economic Theory* 39, 2 (1986), 400–438.
[3] Anat R Admati and Paul Pfleiderer. 1990. Direct and indirect sale of information. *Econometrica: Journal of the Econometric Society* (1990), 901–928.
[4] Anish Agarwal, Munther A. Dahleh, and Tuhin Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *EC*. ACM, 701–726.
[5] Hannaneh Akrami, Bhaskar Ray Chaudhury, Jugal Garg, and Aniket Murhekar. 2025. On the Theoretical Foundations of Data Exchange Economies. In *Proceedings of the 26th ACM Conference on Economics and Computation* (Stanford University, Stanford, CA, USA) *(EC '25)*. Association for Computing Machinery, New York, NY, USA, 444. https://doi.org/10.1145/3736252.3742566
[6] Moshe Babaioff, Tomer Ezra, and Uriel Feige. 2024. Fair-Share Allocations for Agents with Arbitrary Entitlements. *Math. Oper. Res.* 49, 4 (2024), 2180–2211.
[7] Moshe Babaioff and Uriel Feige. 2022. Fair Shares: Feasibility, Domination and Incentives. In *EC*. ACM, 435.
[8] Moshe Babaioff, Robert Kleinberg, and Renato Paes Leme. 2012. Optimal Mechanisms for Selling Information. *CoRR* abs/1204.5519 (2012).
[9] Dirk Bergemann, Alessandro Bonatti, and Tan Gan. 2022. The economics of social data. *The RAND Journal of Economics* 53, 2 (2022), 263–296.
[10] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. 2018. The design and price of information. *American economic review* 108, 1 (2018), 1–48.
[11] Aditya Bhaskara, Sreenivas Gollapudi, Sungjin Im, Kostas Kollias, Kamesh Munagala, and Govind S. Sankar. 2024. Data Exchange Markets via Utility Balancing. In *WWW*. ACM, 57–65.
[12] S. Brânzei, N. Devanur, and Y. Rabani. 2021. Proportional dynamics in exchange economies. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 180–201.
[13] Eric Budish. 2011. The Combinatorial Assignment Problem: Approximate Competitive Equilibrium from Equal Incomes. *Journal of Political Economy* 119, 6 (2011), 1061–1103. https://doi.org/10.1086/664613
[14] Yang Cai and Grigoris Velegkas. 2020. How to sell information optimally: An algorithmic study. *arXiv preprint arXiv:2011.14570* (2020).
[15] Raul Castro Fernandez. 2023. Data-sharing markets: model, protocol, and algorithms to incentivize the formation of data-sharing consortia. *Proceedings of the ACM on Management of Data* 1, 2 (2023), 1–25.
[16] Bhaskar Ray Chaudhury, Jugal Garg, Aniket Murhekar, and Jiaxin Song. 2026. Data Pricing via Competitive Equilibrium. In *Proceedings of the ACM on Web Conference (WWW)*. http://jugal.ise.illinois.edu/papers/www26.pdf
[17] Bhaskar Ray Chaudhury, Jugal Garg, Eklavya Sharma, and Jiaxin Song. 2025. Data Pricing via Competitive Equilibrium. http://jugal.ise.illinois.edu/papers/data-rev.pdf
[18] Hongyu Chen, Hanwei Li, David Simchi-Levi, Michelle Xiao Wu, and Weiming Zhu. 2021. Assortment display, price competition and fairness in online marketplaces. *Price Competition and Fairness in Online Marketplaces (August 30, 2021)* (2021).
[19] Qinyi Chen, Negin Golrezaei, and Fransisca Susan. 2022. Fair Assortment Planning. *CoRR* abs/2208.07341 (2022).
[20] J. P. Choi, D. S. Jeon, and B. C. Kim. 2019. Privacy and personal data collection with information externalities. *Journal of Public Economics* 173 (2019), 113–124.
[21] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.
[22] Alireza Fallah, Michael I. Jordan, Ali Makhdoumi, and Azarakhsh Malekian. 2024. On Three-Layer Data Markets. arXiv:2402.09697 [econ.TH] https://arxiv.org/abs/2402.09697
[23] Rashida Hakim, Christos Papadimitriou, and Mihalis Yannakakis. 2026. Fair Data-Exchange Mechanisms. *arXiv preprint arXiv:2602.11417* (2026).
[24] Safwan Hossain and Yiling Chen. 2024. Equilibrium of Data Markets with Externality. In *ICML*. OpenReview.net.
[25] Kamal Jain and Vijay Vazirani. 2010. Equilibrium Pricing of Semantically Substitutable Digital Goods. *arXiv preprint arXiv:1007.4586* (2010).
[26] Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I. Jordan. 2022. Mechanisms that Incentivize Data Sharing in Federated Learning. *CoRR* abs/2207.04557 (2022).
[27] Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I Jordan. 2022. Mechanisms that incentivize data sharing in federated learning. *arXiv preprint arXiv:2207.04557* (2022).
[28] Sameer Mehta, Milind Dawande, Ganesh Janakiraman, and Vijay Mookerjee. 2021. How to sell a data set? Pricing policies for data monetization. *Information Systems Research* 32, 4 (2021), 1281–1297.
[29] A. Murhekar, J. Song, P. Shahkar, B. R. Chaudhury, and R. Mehta. 2024. You Get What You Give: Reciprocally Fair Federated Learning. In *Forty-second International Conference on Machine Learning*.
[30] Aniket Murhekar, Zhuowen Yuan, Bhaskar Ray Chaudhury, Bo Li, and Ruta Mehta. 2023. Incentives in Federated Learning: Equilibria, Dynamics, and Mechanisms for Welfare Maximization. In *NeurIPS*.
[31] Jian Pei. 2020. A survey on data pricing: from economics to data science. *IEEE Transactions on knowledge and Data Engineering* 34, 10 (2020), 4586–4608.
[32] Serge A Plotkin, David B Shmoys, and Éva Tardos. 1995. Fast approximation algorithms for fractional packing and covering problems. *Mathematics of Operations Research* 20, 2 (1995), 257–301.
[33] Lloyd S Shapley. 1953. A value for n-person games. *Contribution to the Theory of Games* 2 (1953).
[34] Jiaxin Song, Pooja Kulkarni, Parnian Shahkar, and Bhaskar Ray Chaudhury. 2025. On the Existence and Complexity of Core-Stable Data Exchanges. *arXiv preprint arXiv:2509.16450* (2025).
[35] Hugo Steinhaus. 1949. Sur La Division Pragmatique. *Econometrica* 17 (1949), 315–319.
[36] TechCrunch. 2024. AI Training Data Has a Price Tag That Only Big Tech Can Afford. https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford Available at https://techcrunch.com/2024/06/01/ai-training-data-has-a-price-tag-that-only-big-tech-can-afford.
[37] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai, and Wensheng Zhang. 2023. A survey on federated learning: challenges and applications. *Int. J. Mach. Learn. Cybern.* 14, 2 (2023), 513–535.
[38] Zhenyu Wen, Wanglei Feng, Di Wu, Haozhen Hu, Chang Xu, Bin Qian, Zhen Hong, Cong Wang, and Shouling Ji. 2025. FLMarket: Enabling Privacy-preserved Pre-training Data Pricing for Federated Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 1587–1598. https://doi.org/10.1145/3690624.3709346
[39] Y. Zhan, P. Li, Z. Qu, D. Zeng, and S. Guo. 2020. A learning-based incentive mechanism for federated learning. *IEEE Internet of Things Journal* 7, 7 (2020), 6360–6368.