7-8-2023

Selling the Data Product: Pricing Strategies and Welfare Implications

Jinzhao Wang
Fudan University, wangjinzhaofdu@163.com

Yifan Dou
Fudan University, yfdou@fudan.edu.cn

Lihua Huang
Fudan University, lhhuang@fudan.edu.cn

Xiaoyang Zhang
Fudan University, zhangxy_slade@hotmail.com

Qifeng Tang
Shanghai Data Exchange, keven@chinadep.com

Selling the Data Product: Pricing Strategies and Welfare Implications

Short Paper

Jinzhao Wang
Fudan University
670 Guoshun Road, Shanghai, China
jzwangfdu@gmail.com

Yifan Dou
Fudan University
670 Guoshun Road, Shanghai, China
yfdou@fudan.edu.cn

Lihua Huang
Fudan University
670 Guoshun Road, Shanghai, China
lhhuang@fudan.edu.cn

Xiaoyang Zhang
Fudan University
670 Guoshun Road, Shanghai, China
zhangxy22@m.fudan.edu.cn

Qifeng Tang
Shanghai Data Exchange
999 Guidan Road, Shanghai, China
keven@chinadep.com

Abstract

This paper examines the pricing and welfare implications of data as a factor of production with a stylized economic model. We introduce a generalized framework that specifies two types of data: 1) public data pricing, which maximizes social welfare, and 2) commercial data pricing, which maximizes the profit. The model reveals two takeaways: first, two prices may converge in the data economy. It is due to that data come from citizens and may be used to create value back to them. Therefore, a profit-seeking data seller might find it optimal to extend the user base, which is in line with the interest of the welfare maximizer. Second, the pricing gap between optimal prices does not change monotonically with the improvement of data quality. These findings shed new light on the current and future of data product operations, particularly in the understudied public sectors.

Keywords: data market, data pricing, analytical modeling, social welfare

Introduction

The broad utilization of internet services and digital devices has enabled companies to accumulate and amass massive data in the digital economy era. The rapid growth and accumulation of data have become indispensable driving forces for value creation and constitute new strategic resources. To this end, data that are isolated are deemed ineffective as they cannot be incorporated with other data or production factors. Only when data are shared between entities (often facilitated by marketplaces and pricing tools) can they increase productivity and bolster the economy.

Among various data resources, public data's value is widely considered underrated and underexplored. That being said, governments and private sectors are increasingly aware of the potential of public data. For example, in May 2022, the World Bank and the Saudi Data and Artificial Intelligence Authority co-hosted

a webinar to discuss realizing the value of public data generated and managed by governments and public institutions[1].

It motivates us to examine the research question of this paper: *To unlock the value of data, how are the pricing strategies different between the public and private data providers?* By characterizing this discrepancy, our research is among the first to differentiate the data market operations based on the data type. Public data refer to data resources generated and collected by state departments, institutions, organizations authorized by law to manage public affairs, and organizations providing public services, such as electricity supply, gas supply, water supply, and public transportation, in the course of performing public service responsibilities. Introducing public data into the data markets can foster more possibilities for purchases, generate greater productivity, and release the value of data as the production factor. In this paper, we capture the supplier of the public data as a data product seller aiming to maximize social welfare.

We propose a generalized analytical framework including three parties to facilitate data flow in three interrelated markets. Specifically, the three parties include the data product seller, the data product buyer, and a mass of citizens. The seller gathers the data from citizens through the raw data market and sells the data product to the buyer in the data product market. After purchasing the data product, the data buyer analyzes it and merges it with other production factors to produce end products, which are eventually sold to citizens through the end product market. Compared to similar models in previous literature, our model takes a step forward and captures the characteristic of data as a "factor of production" that flows from one entity to another in the digital economy.

This short paper reveals two important takeaways from comparing the optimal prices of welfare-maximizing and profit-seeking data product sellers. First and interestingly, two pricing strategies may converge in the data economy, which crucially depends on the data product quality of the seller side and the analytics capability on the buyer side; Second, with the improvement of data quality, the gap between optimal prices does not change monotonically. Specifically, under a certain threshold, the gap grows even larger. These findings shed new light on the pricing of data products, particularly in the understudied public sectors.

Literature Review

Our study is related to three streams of literature, including 1) data as a factor of production, 2) the value of data, and 3) data pricing.

As data play an increasingly important role in helping firms with decision-making and improving production efficiency in the digital era, they work in a way similar to the traditional factors of production, such as land, capital, and labor. By integrating with other traditional factors of production, data engage in the production process, exert the multiplier effect, and help the firm achieve value promotion. This study follows the vast literature on the factors of production to consider a baseline paradigm with the Cobb-Douglas production function (Cobb & Douglas, 1928). Besides, we focus on the role of data in enabling product innovation and creating user value (Gregory et al., 2021) rather than using data for precision targeting (which may be privacy-risky) of existing products.

There is a growing body of literature on how data generate measurable value for firms, such as empowering innovation, driving R&D transformation, and enhancing the market-direct capabilities of the firm (Suoniemi et al., 2020), so which brings firms competitive advantages. It is also commonly assumed that "data network effects" can lead to winner-take-all outcomes and become a significant barrier to market entry (Ichihashi, 2021). In addition, data-driven price personalization has been widely discussed in marketing and economics. However, all of the research above does not discuss the perspective of data as a factor of production, which is the central issue in this paper. We propose a generalized model framework to describe the whole life cycle from raw data to end product. It contains 1) the raw data market, 2) the data product market, and 3) the end product market, such that it captures how data are gathered, explored, refined, and eventually encapsulated for use.

[1] https://www.worldbank.org/en/events/2022/05/13/unlocking-the-potential-value-of-data-an-emphasis-on-public-data#1

The last stream of related literature is on data transaction and pricing, receiving increasing attention in economics and management in recent years. Data are nonrival, which means data can be used by any number of agents simultaneously without being diminished (Jones and Tonetti, 2020). Data has externalities because some data might reveal information about others. The novel attributes of data (e.g., nonrivalry, externalities, non-competitiveness, unlimited supply, easy replication, and extremely low marginal cost) are the causes of the pricing challenges in the data market, which is hugely different from the traditional markets. Previous studies attempt to derive mechanisms for buying and selling data (Mehta et al., 2021), especially in specific settings, such as data for machine-learning tasks (Agarwal et al., 2019) or business-to-business context (Ray et al., 2020). However, data property rights, information asymmetry, and transaction uncertainty make it difficult for buyers and sellers to reach a consensus on the data price. The pricing and transaction challenges in the data market root in the immaturity of the data lifecycle and low data asset specificity (Huang et al., 2021). Our study concentrates on the differences in data pricing strategies between different data suppliers- a social planner whose objective is to maximize the overall social welfare and a firm to maximize its profit (Bergemann et al., 2022; Tirole, 2021). We do not consider the legal and technical issues associated with data transactions and pricing. We assume that data are in a tradable state. Our findings reveal that the differences in data pricing between the welfare maximizers and profit maximizers depend on the supply side's data quality and the demand side's data utilization capability (Gurkan and Vericourt, 2022).

The Model

Consider a simplified data economy that consists of three parties who are referred to as the "seller", the "buyer", and the "citizens", respectively. The transactions among three parties can be categorized into three interrelated markets: the raw data market (data from citizens to the seller), the data product market (data from the seller to the buyer), and the end product market (data from the buyer to citizens).

In the raw data market, the data seller provides products or services to citizens, and simultaneously amasses original records (i.e., the raw data) from citizens. For example, the State Grid Corporation and Public Transportation Corporation obtain citizens' detailed records of electricity usage and public transportation routes when citizens use their services. A more straightforward example is the internet marketplaces such as Taobao and JD.com. These Internet giants accumulate a massive amount of data from users' browsing history and content contribution. In the data product market, the data seller offers data products that are cleaned from the raw data and encapsulated into structural and legally binding forms. Lastly, in the end product market, the data buyer sells the end product to citizens, which is produced through analyzing the data product and merging it with other production factors.

In this short paper, we set the second market – the data market – as our starting point of analysis due to the page limit and our research interest. In other words, the cost of collecting raw data is sunk to the data seller and thus is no longer considered when pricing the data products. The timeline of the model proceeds as follows. In the first stage, the data seller determines the price of the data product. In the second stage, the data buyer decides the amount of data product to purchase. The data buyer then cleans, processes the data, and integrates the data product into the production of the end product. In the meantime, the data buyer also chooses the end product's price. Finally, each citizen decides whether to purchase the end product. Without loss of generality, we assume that a citizen demands at most 1 unit of the end product.

Among these participating parties, we are particularly interested in the role of the data seller. We differentiate two types of data sellers: profit maximizer (who optimizes the revenue from the data product, such as commercial data sellers) and welfare maximizer (who optimizes the overall welfare of all parties, such as public data providers). Both types of data sellers face pricing issues, and we label them with subscripts R (profit-maximizing) and W (welfare-maximizing), respectively. It is worth noting that we are interested in examining the differences in pricing strategies between profit maximizer and welfare maximizer, and specifically, the potential distortion that may exist in the data economy. We consider the public data pricing as the benchmark. Next, we follow the backward induction to explain the decision-making process.

Stage 2: The end product market

The data buyer first decides the amount of data, $d \in [0,1]$. Note that we use d_R and d_W to denote the amount of data purchased from profit-maximizing and welfare-maximizing sellers, respectively.

It takes two steps for the data buyer to turn the data into use. The first step is to analyze, prepare, and process data, preparing them for subsequent integration with other production factors. It is often challenging financially because it incurs significant expenses before any tangible return on investments is realized. We assume the following form for the cost of analyzing data (called the internal cost):

$$c(d) = \frac{d^2}{m}, \tag{1}$$

in which $m \geq 0$ represents the data analysis and processing capability. The larger m, the more skilled and experienced the data buyer is in analyzing and utilizing data. In other words, for a same amount of data to be put into use, the associated analyzing cost is smaller the data buyer is more proficient in data processing (i.e., better infrastructure, larger data team). Note that the internal cost is a convex function since the larger size of the data imposes a greater technical barrier to the resources and the tools for analytics.

The data buyer then merges the data with other production factors to produce the end product. We follow the classic Cobb-Douglas functional form to assume the following add-on value created from data:

$$v = \beta dK, \tag{2}$$

where K represents the amount of traditional production factors, and $\beta \in [0,1]$ represents the quality of the data product (i.e., the data purchased from the data seller).

We consider a continuum of consumers who are heterogeneous with their valuation of the basic product functionality, which is captured by the consumer type θ. A greater θ indicates a higher willingness-to-pay for the end product, and θ follows a uniform distribution in [0,1]. A type- θ consumer derives the following willingness-to-pay from purchasing the end product:

$$u(\theta) = \theta + v. \tag{3}$$

Given the end product price p, the market equilibrium is characterized by the marginal consumer type $\hat{\theta}$ which satisfies:

$$\hat{\theta} = p - \beta dK. \tag{4}$$

Therefore, the data buyer's profit is calculated by the difference between the end product market revenue and the costs:

$$\pi(p, d) = p\left(1 - \hat{\theta}\right) - \frac{d^2}{m} - rd, \tag{5}$$

where r is the unit price of data and chosen by the data seller.

Stage 1: The Data Product Market

In the data product market, a profit-maximizing seller's optimization problem is:

$$\max_{r_R \geq 0} r_R d_R - f(\beta), \tag{6}$$

in which $f(\beta)$ represents the cost of raw data collection and preparation. This short paper assumes exogenous β, which can be treated as a constant and omitted in the following decision process (this term is omitted in the following welfare-maximizing scenario likewise). We then denote $b = \beta K$ to simplify the formulation because both β and K are exogenous. The economic interpretation of b is the integration efficiency between data product and other production factors, which is later examined as an important dimension.

Similarly, for the welfare-maximizing seller, the objective is given by

$$\max_{r_W \geq 0} \int_{\hat{\theta}}^{1} (\theta + bd_W) d\theta - \frac{{d_W}^2}{m}. \tag{7}$$

This equation is the difference between the willingness-to-pay of the citizens (i.e., consumers) for the end product, as captured in the first term, and the data analysis costs incurred by the data buyers, as represented in the second term. Other financial transfers within the transaction processes, such as payment and receipt between parties in the end product market or between participants in the data product market, do not affect the overall social welfare.

The model and variables described above can be summarized in the following Figure 1.

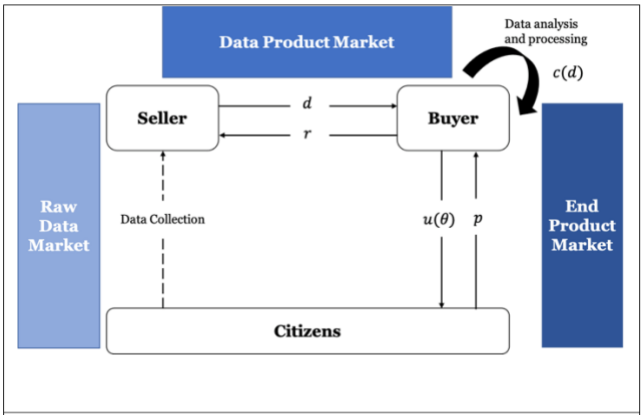

Figure 1. The Model

Analysis and Results

This section presents the optimal pricing strategies for two types of data sellers. We then compare the pricing regions between them to obtain the welfare implications. The detailed proof, including the solution to the data buyer's optimization (as the stage-2 sub-problem to the stage-1 data product seller's pricing problem), is omitted here due to the page limit and available upon request.

We start with the welfare-maximizing seller. The optimal price is given by following Proposition 1.

Proposition 1. *The profit-maximizing data product seller's optimal price r_R^* satisfies:*

a) *For $m \geq 4/b^2$, then $r_R^* = (b^2m + 2bm - 4)/(4m)$;*
b) *For $8/[b(2b+1)] \leq m < 4/b^2$, then $r_R^* = (b^2m + bm - 4)/(2m)$;*
c) *For $m < 8/[b(2b+1)]$, then $r_R^* = b/4$.*

Proposition 1 suggests that the profit-seeking seller's optimal price is determined jointly by the data product b and the data seller analytic capability m. Either is large enough will drive to the extreme case where the data buyer wishes to acquire as much data as possible. On the contrary, with a low analytic capability, the data product buyer's interest in data is limited (i.e., $d_R^* < 1$ & $r_R^* = b/4$) because the purchased data are unlikely to be fully explored and integrated with other production factors.

Besides, Proposition 1 also offers the implication that, for the data product market to flourish, both the availability of superior data products and the expertise of data buyers are needed. Besides, it is also intuitive to see that the profit-maximizing data product seller always charges a non-zero price, but the price might be independent of the buyer's internal cost (i.e., m) when the threshold is on the data quality (i.e., $m < 8/[b(2b+1)]$).

Next, we move on to the welfare-maximizing seller. The optimal price is given by following Proposition 2.

Proposition 2. *The welfare-maximizing data product seller's optimal price r_W^* satisfies:*

a) *For $m \geq 4/b^2$, then any $r_W^* \in [0, (b^2m + 2bm - 4)/(4m))$ is satisfied;*
b) *For $4/[b(b+1)] \leq m < 4/b^2$, then any $r_W^* \in [0, (b^2m + bm - 4)/(2m))$ is satisfied;*

c) For $m < 4/[b(b+1)]$, then $r_W^* = 0$.

It can be inferred from Proposition 2 that the optimal price for welfare-maximizing data product seller is not unique when it is substantially large, which is in sharp contrast to the profit-maximizing seller's strategy. Additionally, Proposition 2 shows that a zero price is necessary (case c) either when the data utilization capability of the data buyer is weak, or the quality of the data product is low, causing $m < 4/[b(b+1)]$.

It should be pointed out that the lower bound of prices in Proposition 2 is the same as the price strategies in Proposition 1, which means that the profit-maximizing price can also be a welfare-maximizing price (or at least close to it). We conduct comparisons in Proposition 3 to explore this interesting direction.

Proposition 3. *The optimal prices in two cases, r_R^* and r_W^*, satisfy:*

a) For $m \geq 8/[b(2b+1)]$, the profit-maximizing price r_R^* also achieves welfare maximization;
b) For $m < 8/[b(2b+1)]$, the profit-maximizing price r_R^* cannot achieve social welfare maximization. Specifically, the gap between the two prices $\Delta r^* = r_R^* - r_W^*$ first increases and then decreases with b, suggesting a non-monotonic relationship with the data product quality.

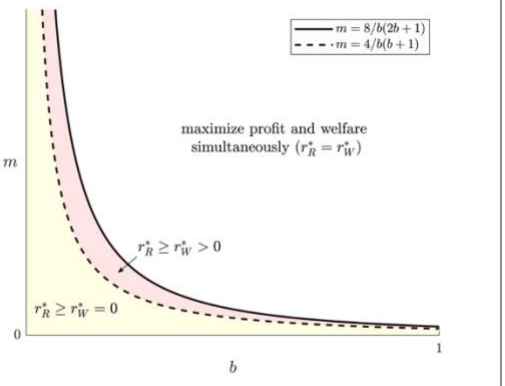

Figure 2. The Comparison between Pricing Regions

We visualize the pricing regions under two optimal prices in Figure 2, in which the horizontal axis represents the data quality b, and the vertical axis represents the data buyer's analytical capability m. In the upper-right corner (white), where both m and b are sufficiently large, two pricing strategies converge because the data buyer wishes to purchase the maximum number of data products to enhance the end product, which maximizes the social welfare.

Moving from right to left into red and yellow regions in Figure 2, we can see that this convergence no longer exists, and the discrepancy between the two strategies shows up. Therefore, we further portray the difference of $\Delta r^* = r_R^* - r_W^*$ in Figure 3. Given m, we can calculate the solutions of b satisfying $m = 8/[b(2b+1)]$ and $m = 4/[b(b+1)]$ respectively (i.e., $\widehat{b_1}$ and $\widehat{b_2}$). As shown in Figure 3, we consider two cases: a) $m = 5$ and b) $m = 10$. For clarity, we carry over the same set of background colors (yellow and red) from Figure 2 to highlight the same intervals.

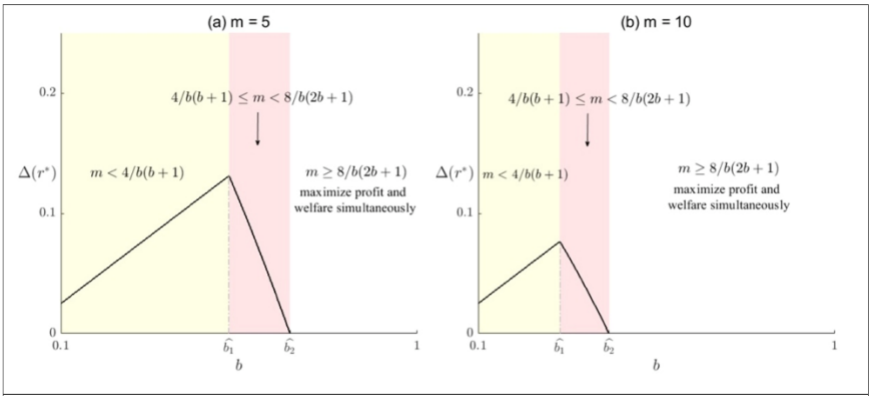

Figure 3. The Difference between Optimal Data Prices ($\Delta r^* = r_R^* - r_W^*$)

As suggested by Proposition 3c, we find an interesting, non-monotonic pattern between two prices. Specifically, when either m or b is small enough (i.e., yellow area), the profit-seeking seller charges a positive price ($r_R^* = b/4$), while the profit-maximizer provides the data for free ($r_W^* = 0$). The discrepancy between them is further enlarged with b, which implies that when the quality of the data product improves, the welfare-maximizing seller has to bear a higher opportunity cost for not charging for it. Interestingly, as b further increases and moves into the red region, the profit gap vanishes gradually because the profit-seeking optimal price (i.e., $r_R^* = b/4$) is no longer affected by m. In this case, a greater b always incentivizes the data product buyer to spend more on the data product, which is aligned with the interest of the welfare maximizer.

Conclusion

As people become increasingly cognizant of the advantages of data from both a social and commercial standpoint, the demand to unlock the value of data, ranging from the public to private sectors, is also growing stronger. However, the value exploration of data cannot be achieved overnight. It often requires data providers to process the data from its original raw form, and the data users must also be equipped with the necessary processing and analysis tools and skills. These factors increase the layers of complexities. This short paper proposes a general framework that incorporates both sides of data (i.e., value and costs, as explained above) while examining the pricing strategies based on the different roles of data product sellers.

Our results give two important implications for unlocking the value of data. First, data have the unique feature that they are created by users, and they can be used to create value for users. Consequently, a profit-seeking seller may prefer a larger number of users, which aligns with the goal from an optimal social viewpoint. It is rare in a traditional economy. However, the data-associated costs, such as cleaning, preparation, and analysis, cannot be easily internalized. Accordingly, the combination of social and economic success hinges on the excellence and proficiency of data products and analytics, as demonstrated in our paper.

Second, our study highlights that the pricing difference changes with data quality non-monotonically. The important implication is that, in the first stage of the development of the digital economy, it is necessary to invest in additional subsidies to unleash the value of data, particularly for the public sector. The good news is that the welfare-maximizing price might eventually be incentive-compatible with the profit-seeking seller, which also appears new in academic literature.

Due to space limitations, we omit the discussion of the raw data market in this paper and skip the formation logic and means of data quality improvement. Subsequent studies can extend the perspective to the framework. In addition, we only consider the most typical type of data charge-pay-per-volume, while other forms of data price charge (e.g., monthly subscriptions) could be explored for future research. We do not consider the issue of data timeliness, which is our limitation. We believe our work can serve as a foundation for future work on data pricing mechanisms.

Acknowledgements

This research is supported by National Natural Science Foundation of China (72241424), the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission, and WU Jiapei Award for Information Economics in 2022 (M22106023).

References

Agarwal, A., Dahleh, M., & Sarkar, T. (2019). A marketplace for data: An algorithmic solution. *Proceedings of the 2019 ACM Conference on Economics and Computation*, 701-726.

Bergemann, D., Bonatti, A., & Gan, T. (2022). The economics of social data. *The RAND Journal of Economics,* 53(2), 263-296.

Cobb, C. W., & Douglas, P. H. (1928). A theory of production. *American Economic Review*, 18(1), 139–165.

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. (2021) The role of artificial intelligence and data network effects for creating user value. *Academy of Management Review*, 46(3): 534-551.

Gurkan, H., & de Véricourt, F. (2022). Contracting, pricing, and data collection under the AI flywheel effect. *Management Science*, 68(12), 8791-8808.

Huang, L., Dou, Y., Liu, Y., Wang, J., Chen, G., Zhang, X., & Wang, R. (2021). Toward a research framework to conceptualize data as a factor of production: the data marketplace perspective. *Fundamental Research*, 1(5), 586-594.

Ichihashi, S. (2021). Competing data intermediaries. *The RAND Journal of Economics*, 52(3), 515-537.

Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the Economics of Data. *American Economic Review*, 110(9), 2819-58.

Mehta, S., Dawande, M., Janakiraman, G., & Mookerjee, V. (2021). How to sell a data set? Pricing policies for data monetization. *Information Systems Research*, 32(4), 1281-1297.

Ray, J., Menon, S., & Mookerjee, V. (2020). Bargaining over data: When does making the buyer more informed help?. *Information Systems Research*, 31(1), 1-15.

Suoniemi, S., Meyer-Waarden, L., Munzel, A., Zablah, A. R., & Straub, D. (2020). Big data and firm performance: The roles of market-directed capabilities and business strategy. *Information & Management*, 57(7): 103365.

Tirole, J. (2021). Digital dystopia. *American Economic Review*, 111(6), 2007-2048.