Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)
© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230933

A Survey on Data Pricing: Methods, Challenges, and Prospects

Zangbo CHI[a1], Hua ZHAO[a], Zhi FANG[a], Hong ZHANG[a],
Liangliang JI[a] and Zengwen YU[a,b2]
[a] *Beijing Institute of Computer Technology and Application, Beijing, China*
[b] *School of Computer Science and Technology, Xidian University, Xi'an 710071, China*

Abstract. As organizations increasingly rely on data-driven decision-making, understanding the true value and potential of data becomes crucial. Data pricing, which aims to determine the financial value of information, plays a pivotal role in data circulation and transactions. However, the existing surveys on data pricing have limitations that need to be addressed. This paper presents a comprehensive survey of big data pricing methods from a data science perspective. It begins by providing an overview of the fundamental concepts underlying data pricing and the data market. Subsequently, it delves into the general principles and challenges associated with data pricing. The survey categorizes and summarizes various approaches and methods employed in data pricing, assessing their respective advantages and disadvantages. In conclusion, this paper identifies potential research directions to enhance our understanding of data pricing. By rectifying the deficiencies in existing surveys, this comprehensive study aims to contribute to the development of effective data pricing strategies and foster advancements in the field of data science.

Keywords. Big data, data pricing, data market, data science, pricing model, privacy

1. Introduction

Data has emerged as the "new oil" of the modern era [1], possessing immense potential to unlock valuable insights, uncover market trends, and enhance service quality. Just as oil requires pricing to facilitate its utilization, data pricing aims to assign a monetary value to data, enabling its effective utilization and exchange. Data pricing plays a crucial role within data markets, serving as a fundamental component of data transactions.

However, in the practical process of data trading, the parties involved, including consumers, data owners, and platforms (or data brokers), often have different perspectives and priorities. Consumers expect data prices to accurately reflect the value of data for their specific tasks. They seek high-quality and reliable data to support their business decisions and drive innovation. On the other hand, data owners and platforms typically price data based on the costs associated with data collection and management.

[1] Zangbo CHI, Beijing Institute of Computer Technology and Application; E-mail: shidaide2019@gmail.com
[2] Corresponding Author: Zengwen YU, Beijing Institute of Computer Technology and Application; School of Computer Science and Technology, Xidian University; E-mail: 22033110404@stu.xidian.edu.cn.

They aim to obtain returns from data transactions that can compensate for the investments made in data acquisition, storage, and maintenance. These divergent interests among transaction parties create challenges and hinder the development of a consensus, impeding the smooth progress of data trading.

The main challenge lies in developing a fair data pricing strategy that takes into account the systematic study of data demand, supply, and the realization of data value under different circumstances. Such a strategy should encourage companies and organizations possessing valuable data to willingly participate in data trading, effectively balancing the economic interests of data owners and the task requirements of consumers. To overcome these challenges, an interdisciplinary approach is needed, combining principles from both data science and economics. By considering factors such as data demand, market competition, and pricing mechanisms, researchers and practitioners can develop robust data pricing strategies that align the interests of all transaction parties. This will foster a conducive environment for data trading, enabling fair and efficient exchanges that benefit both data owners and consumers.

In light of the challenges posed by data pricing and transactions, prior studies have made significant contributions in tackling these issues. Some researchers have conducted relevant surveys on data pricing, but there are still many shortcomings [2] - [5]. Pei et al. [2] provides a comprehensive description of data pricing from an economic perspective, summarizing the essential considerations and guidelines in data pricing, along with corresponding methods. However, they do not present a comprehensive classification of existing pricing methods. Similarly, Zhang et al. [3] review the theories and methods applicable to data pricing within economics, yet they lack a comprehensive summary of pricing strategies. Liu et al. [4] introduce big data pricing methods from a social science perspective, focusing more on institutional and framework-based narratives rather than delving into the specific details of pricing methods. Cai et al. [5] overlook the integral aspect of data transactions while categorizing data pricing approaches in detail.

Given this background, this paper offers a comprehensive summary of recent research on data pricing. It introduces the fundamental concepts and relevant properties of data pricing, and analyzes existing methods based on query-based pricing and privacy-based pricing, highlighting their respective advantages and disadvantages. Finally, the paper outlines the challenges and future directions in the field of data pricing.

2. Data Pricing

2.1. Data Pricing

Data pricing involves the process of assigning a monetary value to data, encompassing the considerations and transactions that occur within data markets. To develop effective data pricing strategies, it is essential to have a comprehensive understanding of data demand, supply, and how data value is quantified in different contexts. In data pricing transactions, the primary focus is often on the records contained within a dataset. Although individual data units, such as a single data point, may have limited information after technical processing, the dataset as a whole possesses significant aggregation. Customers typically aggregate these fundamental data units along various dimensions to reveal the underlying value of the data. For example, a retailer may find little utility in a customer's individual purchase record after appropriate anonymization. However, when

anonymized and aggregated with the purchase records of all customers within a specific region, these data can provide valuable insights to the retailer, enabling them to make informed business decisions and ultimately generate higher economic returns.

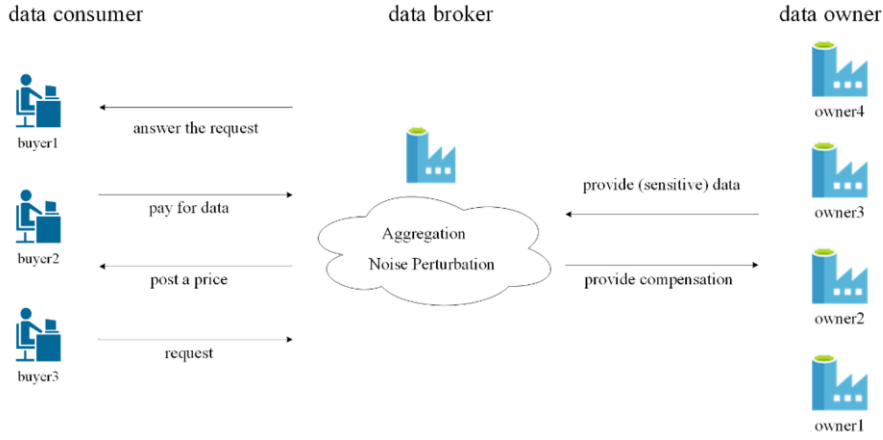

Figure 1. An example of a data market.

2.2. Data Market

A data market is a centralized environment or platform that facilitates data transactions, requiring appropriate pricing for the data involved. As depicted in Figure 1, the participants in a data trading marketplace typically consist of data owners, data consumers, and platforms (data brokers). Data owners are individuals or organizations that provide data, and the quality and value of the data they offer determine the transaction price and the level of demand for the data. Data consumers are individuals or organizations that have specific data needs. They select relevant data based on their requirements and pay the corresponding price. Platforms refer to third-party intermediaries that offer data trading platform services. They provide technical and service support to facilitate transactions between data owners and consumers. Data owners can set prices for their data products or services and sell them in the market, while data consumers can purchase or subscribe to data according to their needs. Data marketplaces serve as transparent, efficient, and trustworthy platforms for data circulation and exchange.

2.3. Principles in Data Pricing

When conducting data pricing and data transactions, to ensure that the participants receive higher finical returns, the participants involved in the data transaction need to follow some common guidelines.

- Authenticity [2].

Authenticity refers to the fact that, during data trading, every participant only offers prices that maximize their own benefits. This prevents fraudulent activities in data trading. Regardless of how others behave, no supplier or consumer can increase their profits by falsely reporting the true value of the data.

- Revenue Maximization [2].

Revenue maximization is an important metric in data pricing models, aiming to achieve long-term economic gains. There are significant differences in calculating revenue maximization between traditional products and data products. For traditional products, sellers can achieve revenue maximization when marginal cost equals marginal revenue. However, data products have nearly zero marginal cost, rendering this rule inapplicable. Furthermore, in modern data trading models, intermediaries often interact with buyers, and the platforms acquiring data may not necessarily be aware of the specific use of the data. This makes it challenging to determine a price that maximizes revenue for data products. Therefore, in data pricing models, it is necessary to consider the unique characteristics of data products and employ specialized pricing strategies to maximize revenue.

- Fairness [2].

Fairness ensures that all data contributors receive income according to the level of their contribution. There are some requirements for achieving fairness [2], such as symmetry (the allocation of cooperative gains should not depend on the labels or ordering of individuals within the cooperation), efficiency (the sum of individual benefits should be equal to the overall coalition value; otherwise, it would be considered inefficient), redundancy (if a member does not contribute to any cooperation alliance they participate in, they should not receive any gains from the overall cooperation), additivity (when there are multiple collaborations, the distribution of benefits for each collaboration should be independent of the outcomes of other collaborations. If the returns for two tasks, T1 and T2, are v1 and v2 respectively, then completing both tasks, T1+T2, should yield a return of v1+v2).

A widely applicable way to fairly distribute income is based on Shapley value in game theory [2], the core idea of the Shapley value is to measure the contribution of each participant to the outcome of a cooperative game. It calculates the average contribution of each participant to the cooperative gains by considering all possible permutations and combinations of participants. Specifically, it determines the contribution value by adding each participant to the game process and observing the impact of their inclusion on the final outcome.

- Arbitrage-free [2].

Arbitrage is one of the most important issues in big data pricing, which refers to the behavior of buyers obtaining data products by some means at a price lower than the seller's specifications, and the existence of arbitrage opportunities will lead to inconsistent data pricing and greatly increase the risk of information leakage. For a pricing function π and a service S that can be disassembled into m($m \geq 1$) subservices, if the pricing function is arbitrage-free, then

$$\pi(S) \leq \Sigma\pi(Si), i = 1, 2, \cdots, m \quad (1)$$

- Privacy protection [2].

Privacy and information exchange are closely intertwined, and the issue of privacy in information products is increasingly garnering attention. Due to the relatively low cost of tracking informational goods, collecting user privacy data has become relatively easy. In the market for information goods, safeguarding privacy becomes particularly crucial. Under normal circumstances, transactions within the marketplace can potentially

expose the privacy of all parties involved through various means, such as data leaks, unauthorized access and dissemination of third-party data, and data linkage.

Protecting privacy is highly desirable in the data market. In general, transactions within the market can potentially compromise the privacy of all parties involved [2]. Firstly, buyers' privacy is highly vulnerable. Details such as their identities, purchase locations and times, specific products purchased, prices, and total amounts spent can all reveal their privacy. Secondly, the privacy of information providers can also be jeopardized. For instance, medical information held by hospitals is highly valuable to commercial entities like pharmacies and medical device companies. Imagine a scenario where hospitals appropriately collect and anonymize medical data and offer corresponding data products in the market, ensuring that individual patients cannot be re-identified. However, buyers may be able to infer the success rate of a specific treatment from the data, which could be considered a breach of the hospital's privacy. Lastly, transactions within the market can also expose the privacy of third parties involved. For example, AI technology companies may offer machine learning model building services to buyers of data products. However, if the machine learning models are stolen, it can be viewed as a violation of privacy for the AI technology companies.

To protect privacy in the market for information products, various approaches are being explored. These approaches include concealing information about buyers' purchases, the timing of their purchases, and the amount they spend. Efforts are being made to establish decentralized and trusted privacy-preserving data markets. Trade-offs between privacy, payment, and accuracy are being investigated in the context of privacy considerations. Additionally, there is ongoing exploration of aggregating unverifiable information from privacy-sensitive populations. Differential privacy [6], as a method for measuring the similarity between datasets, holds significant applications in quantifying privacy loss, thereby enabling fair compensation for data providers.

2.4. Challenges

The pricing of data has gained significant attention, but it still faces many challenges.

Data consumers require diverse types of data. Analyzing and processing data from multiple heterogeneous sources, integrating massive business data for storage and sale, and determining the fundamental value of complex data are among the current challenges.

Data pricing involves a multi-party balance between data providers, data users, and data intermediaries. Data consumers expect data prices to reflect the value of the data for their tasks, while data owners and platforms often base data pricing on factors such as privacy loss, data collection, and management costs. Establishing systematic, efficient, and accurate value assessment principles for all parties in the data market, encouraging companies and organizations with data to have a higher willingness to sell data while ensuring consumers' economic interests and meeting their task requirements, poses a highly challenging problem.

The value of data, data products, and data services may exhibit temporal characteristics. Exploring the underlying changes in supply-demand relationships and developing corresponding dynamic pricing mechanisms remains a hot research topic.

Designing transaction mechanisms and building trustworthy trading platforms to maximize the benefits for participants, ensuring fair and efficient data transactions, is also a topic worthy of in-depth research.

3. Data Pricing Models

Data pricing is one of the most crucial tasks in data transactions, and the value of data is typically reflected in the price during these transactions. In this section, we will explore various mechanisms used by different parties to measure data prices from the perspective of data science. Specifically, we will discuss two distinct approaches: query-based pricing and privacy-based pricing. As shown in Table 1, Query-based pricing models focus on determining the value of data based on the queries made by data consumers, taking into account factors such as data relevance and accuracy. On the other hand, privacy-based pricing models take into consideration the privacy risks associated with the data being exchanged, with higher prices often associated with data that carries a higher privacy risk. By understanding and considering these different mechanisms, data market participants can make informed decisions regarding data pricing, taking into account both the utility of the data and the privacy concerns involved.

Table 1. Data Pricing Methods

Pricing Models	Ideas
Query-based	Determine the price based on the value that the data product can generate for data consumers in performing a specific task [7 - 17].
Privacy-based	Determine the price of the data product based on its intrinsic value, such as the level of privacy inclusion and the quality of the data [18 - 26].

3.1. Query-Based Pricing Models

To enable buyers to initiate arbitrary query requests and support complex query operations in data market transactions, Koutris et al. [7] proposed a query-based pricing framework for pricing internet data. It allows sellers to set explicit prices on a small number of views (or sets of views). When buyers submit query requests, the prices are automatically derived from the explicit prices of the views, rather than being pre-defined by the sellers. Buyers with different needs can freely choose which queries to purchase based on the associated value of the data, without sellers having to explicitly set prices on an exhaustive catalog of all possible queries. The proposed framework not only satisfies the arbitrage-free axiom but also needs to fulfill the discount-free theorem, which states that the pricing function should not compute a query price using a view's price that is lower than the seller's pre-defined price point. Additionally, the paper demonstrates that if a query cannot be reduced to a chain or loop query, it becomes an NP-hard pricing problem. Polynomial-time algorithms for chain and loop queries are also presented. However, the methods proposed in the paper only support simple query statements and cannot meet the demand for complex queries in the data market.

Based on the aforementioned approach, Koutris et al. [8] introduced the QueryMarket system, which transforms the arbitrage-free problem into an Integer Linear Programming (ILP) problem, enabling pricing for complex operations such as join queries and binding queries. Furthermore, as data consumers may perform multiple queries on the same information when purchasing data, the paper introduces a method to address the issue of duplicate charges by recording query history to achieve dynamic pricing. Calculating the price of an associated query on a relation of around 1000 tuples using QueryMarket takes approximately 1 minute.

Similarly, Qirana [9] is a query-based pricing system that allows data sellers to choose from a set of arbitrage-free pricing functions. The core idea is to view queries as a mechanism to reduce uncertainty and achieve real-time pricing for large-scale SQL queries, such as aggregate operations. Inspired by Qirana, Chawla et al. [10] studied the challenges faced by brokers selling data access rights and explored the problem of revenue maximization with monopolistic buyers and unlimited supply while maintaining the non-arbitrage assumption.

Wang et al. [11] proposed a pricing mechanism for approximate queries and demonstrated its effectiveness through theoretical analysis. They used sampling techniques to obtain approximate results with bounded error on specific data queries. They also introduced a transformation function that converts the original pricing function into a pricing function that supports approximate aggregate queries while preserving the arbitrage-free property.

In order to tackle the challenge of trading correlated queries, Niu et al. [12] conducted an analysis of data transactions and presented the Erato [13] framework. Their focus was on trading noisy aggregate statistical data from the viewpoint of data brokers in the data market. They observed that in many instances, transactions primarily involve aggregated outcomes rather than raw data. As an illustration, they derived three types of aggregate statistics from the raw data, namely weighted sums, probability distribution fitting, and degree distribution.

Cai et al. [14] focused on the challenge of trading and pricing multiple correlated queries involving privacy-preserving web browsing history data. They devised an innovative online data commercialization framework that utilized an enhanced matrix mechanism to perturb query results. Their approach introduced a query pricing mechanism based on ellipsoids, guided by a given linear market value model. The objective of this mechanism was to identify and leverage approximate optimal dynamic prices in each round, striking a delicate balance between data utility and privacy protection, while ensuring efficient runtime performance.

Additionally, Cai et al. [15] investigated the trading of correlated data queries on high-dimensional privacy data. They constructed a model to capture the correlation between user attributes in high-dimensional settings and developed an initial attribute clustering scheme. By addressing the Optimal Attribute Clustering (OAC) problem, they devised a novel data perturbation mechanism that enhanced the data utility of traded data and generated high-dimensional privacy-preserving datasets closely aligned with the original data distribution. Additionally, they quantified the privacy loss resulting from the NP-hardness of the OAC problem and introduced an auction mechanism to compensate data owners.

In addition, Miao et al. [16] tackled the pricing challenge associated with incomplete data using iDBPricer. They introduced the concept of a lineage set and proposed two pricing functions: usage and completeness-aware price function (UCA price) and quality, usage, and completeness-aware price function (QUCA price). Moreover, they introduced the notion of historical awareness to assess the reliability of decisions made on incomplete data, understand the impact of missing information on the final decision, and identify factors influencing potentially unreliable query results.

Chen et al. [17] conducted in-depth research on trading graph data in the data market. Based on graph simulation and subgraph isomorphism theory, they demonstrated the arbitrage-free property of queries in graph data pricing. They proposed effective algorithms for precise, approximate, and dynamic pricing problems on graph data, achieving good performance on multiple datasets.

Query-based pricing methods are essentially task-oriented operations. They first set tuple-based or item-based base prices and generate prices for any view that data consumers wish to purchase based on these prices. Query-based pricing methods are suitable for easily queryable data stored in structured or unstructured databases, offering advantages such as flexible pricing and minimal maintenance after setting the base prices. However, since the data sold through query-based pricing methods is a collection of multiple entries without individual item value, the interpretability of generated prices is relatively low. Moreover, the time complexity of price generation is generally high. Additionally, due to the strong temporal nature of big data, offline pricing algorithms suffer from the inability to update prices in real time. Therefore, addressing these issues is an important consideration for researchers in this field.

3.2. Privacy-Based Pricing Models

Data transactions often involve the exchange of personal data, and the privacy contained therein can serve as a crucial indicator for measuring the value of the data. To address the issue of privacy loss experienced by sellers in data transactions and to incentivize more individuals to sell their personal data, many studies argue that a certain level of privacy compensation should be provided to data owners. This section aims to investigate the transactions between data owners and data consumers (potentially involving data intermediaries), starting from the data itself. It considers multiple factors such as the sensitivity and privacy risks associated with personal data, data scarcity, and relevant legal regulations. The objective is to calculate the appropriate amount of privacy compensation that should be given to data owners for their privacy loss.

Ghosh et al. [18] investigated the problem of truthful auctioning of private data from the perspective of differential privacy. They proposed personal privacy data transactions and provided privacy compensation to sellers, suggesting that this mechanism can be viewed as a simple setting of multi-unit procurement auctions without loss of generality. The paper argues that privacy transactions can be conducted in a similar manner to the trading of other commodities such as stocks and bonds, and data intermediaries can purchase arbitrary amounts of privacy from each data owner by providing sufficient incentives. To reveal the privacy attitudes of data owners regarding the sale of their data, data platforms or brokers employ an auction-based approach, where each data owner submits bids that reflect their privacy preferences. Based on the received bids, the privacy level to be purchased from the data owner is determined, and a noisy query output is generated to ensure the preservation of that privacy level.

However, the issue with this privacy compensation mechanism is that even if data owners from the same batch of data for sale have different valuations of privacy, the mechanism calculates the same ε value for all owners whose data is being used and provides privacy compensation accordingly. This inevitably results in excessive privacy protection for certain data owners and lacks the ability to customize the privacy compensation mechanism for individualized privacy data.

In order to tackle the issue of "pseudo personalization" in the privacy compensation discussed earlier, Zhang et al. [19] introduced a pricing mechanism that guarantees personalized privacy requirements for sellers. They utilized differential privacy as a measure of privacy loss and ensured that the customized differential privacy parameters set by data owners are met, while also supporting queries with high precision outputs.

For data intermediaries interested in purchasing data from data owners, they proposed a mechanism that employs reverse auctions to select data owners and determine rewards. The underlying principle was to maximize the expected purchase privacy, resulting in accurate outputs in common query scenarios such as counting, median, and linear prediction.

Similarly, Li et al. [20] adopted a differential privacy-based approach to quantify privacy loss. Their primary focus was on assessing the relationship between the accuracy of query results and their associated costs. They formalized the concept of noise query arbitrage and established a formal relationship between privacy loss and payment to data owners. The paper introduced a micro-payment function to strike a balance between query pricing and privacy compensation within the framework designed for a given query. However, due to the utilization of a linear privacy measurement mechanism and the flexibility for users to define their privacy loss coefficients, there is a tendency for users to set excessively high privacy coefficients under the same ε value, resulting in unjustifiably high profits. Ensuring trustworthy and arbitrage-free transactions in the pricing of privacy data remains a crucial area for future research.

Nget et al. [21] introduced a personal data pricing framework that supports aggregated queries on noisy data. They proposed the concept of privacy compensation, which ensures that each data seller receives fair compensation while maintaining a balance between price and privacy. Additionally, they defined non-decreasing pricing functions for scenarios involving low-risk and low-return, as well as high-risk and high-return. Building upon this personal data transaction framework, they developed a balanced pricing mechanism that calculates query prices and perturbed results for data buyers while providing data owners with compensation based on their privacy loss.

In addition, Shen et al. [22] presented a pricing method for personal big data based on differential privacy. They devised both forward pricing and reverse pricing approaches. The forward pricing method involves calculating privacy compensation for each user based on their privacy loss and then determining the price based on the total privacy compensation. On the other hand, the reverse pricing method first computes the payment price for buyers and then distributes compensation among users based on the degree of their privacy loss. These mechanisms were designed to achieve reasonable pricing for personal big data, taking into account the privacy concerns of users.

Xiao et al. [23] aimed to maximize profits in online data markets by adequately compensating data owners and determining reasonable pricing for data collectors. The researchers examined a scenario involving untrustworthy data collectors, sensitive data costs and transaction behaviors, data collectors' preference for market attributes (specifically, profit maximization), and the requirement of non-negative payments for data owners. In order to address the challenges posed by compensating data owners and unknown data costs, they introduced an enhanced online learning algorithm called Modified Stochastic Gradient Descent (MSGD). MSGD leverages interactions between data owners and collectors to infer the cost model and employs auxiliary parameters to correct biased gradients resulting from noise. To safeguard the privacy of data owners during transactions, a Local Differential Privacy (LDP) framework was adopted, enabling owners to perturb their actual data and transaction behavior.

Fallah et al. [24] considered building a platform that collects data from privacy-sensitive users and formulated the problem as a Bayesian optimal mechanism design. Individuals could share their (verifiable) data in exchange for monetary rewards or

services. Differential privacy (Central Differential Privacy and Local Differential Privacy) was used to quantify the privacy cost. Lower bounds for estimating errors were established, and optimal estimators (close to) achieving a given user's privacy loss level were derived.

Feng et al. [25] proposed a personalized privacy-aware data trading approach based on contract theory. This method offered self-interested data owners an optimal contract that specified different levels of privacy protection and corresponding data trading prices. Data owners uploaded perturbed data based on the negotiated privacy protection level, and the data was ultimately aggregated using a group-weighted maximum likelihood estimation method.

In a similar vein, Jiang et al. [26] explored methods to enhance privacy-preserving data transactions by employing a transaction model rather than directly trading raw data. The objective was to provide satisfactory privacy compensation and query pricing. Initially, data agents utilized a GAN-based model to train a generator, which augmented the data to alleviate data scarcity while introducing noise during the training process to protect the owners' privacy. Rényi differential privacy was then employed to quantify the privacy loss at the data item level during GAN training, and each owner was compensated based on their respective privacy policies. Subsequently, data agents charged fees for each data consumer's queries, with prices set lower than the total privacy compensation. Although this approach ensured data security and avoided direct data trading, its lack of generality made it challenging to train a universal model.

Privacy-based pricing is fundamentally a pricing approach based on the intrinsic value of data, considering the data owner's capacity to bear privacy loss risks and their desire for returns to determine compensation prices. At the same time, it is important to consider the issue of excessively low privacy exposure, which may reduce the utility for data consumers.

4. Future Directions

Data pricing is playing an increasingly important role in the digital economy era. However, there are still many research directions worth exploring in this field:

4.1. How to Choose Suitable Datasets and Evaluation Metrics?

Current research on data pricing lacks standardized datasets and evaluation methods designed specifically for pricing tasks. Due to significant differences in data types and scales across different domains, the datasets and evaluation methods used in research also vary. This lack of a unified standard makes it difficult for researchers to compare the effectiveness and performance of different algorithms and to generalize and apply these algorithms to different domains. Although metrics such as Mean Square Error (MSE) and Root Mean Square Error (RMSE) are commonly used to evaluate the impact of adding noise in pricing problems, the evaluation of these metrics often involves subjectivity and lacks universal baselines and quantitative analytical explanations. This limits the comparability of algorithm effectiveness and the scope of their application. Therefore, in the future, it is necessary to establish public datasets and evaluation standards to promote further development in data pricing research.

4.2. How to Price Different Types of Data Appropriately?

Data pricing needs to adopt different strategies based on types. For structured data, pay-as-you-go is a simple and effective pricing model. For unstructured data, a pricing model based on data content can better reflect the value of the data. Pricing personal and sensitive data requires more caution and privacy protection. For public and open data, using free or open pricing models can promote data sharing and collaboration. Data intermediaries need to determine appropriate prices based on the characteristics and usage scenarios of the data to ensure their interests and protect user privacy. At the same time, consumers also need to choose suitable data and pricing models based on their needs and budgets.

4.3. How to Establish an Effective Dynamic Pricing Theory Framework?

In the current data product market, different participants have different expectations and evaluations of data products. Existing single-indicator data pricing methods are unable to fully meet the needs of all parties. Additionally, most existing pricing methods are static, while data has strong timeliness and consumer demand for data changes over time. Therefore, data prices should also be adjusted accordingly. To make data prices more practical, future research can focus on dynamic data pricing problems. This involves capturing the relationship between data prices and time through function models, predicting changes in data content and prices, developing dynamic pricing algorithms, and achieving fair and efficient dynamic data pricing mechanisms.

4.4. How to Improve Existing Data Trading Mechanisms?

Data pricing focuses on the monetary value of data, while data trading considers the impact of market types, mechanism designs, and participant behaviors on data prices when data circulates in the market. The design of data trading mechanisms directly affects the willingness of data owners to sell and data consumers to purchase, requiring appropriate pricing and trading mechanisms to address the issues present in the current data market. Improvements in data trading mechanisms can be made in several aspects.

- Privacy and copyright protection.

Data inherently contains varying degrees of privacy, and due to data's replicability, sold data can spread at a low cost, jeopardizing the data owner's rights. Excessive data trading can not only lower data prices but also reduce the desire of data owners to trade, impacting the development of the data market. Therefore, privacy and copyright protection mechanisms must be studied in data trading, addressing the protection aspects from institutional and technical perspectives.

- Creating a fair and truthful trading environment.

Although some research has explored the fairness and truthfulness of transactions, there are still limitations and trade-offs. In the future, it is necessary to analyze the market environment during each transaction and design corresponding trading mechanisms to ensure fair prices for participants. Additionally, the behavior of transaction participants should be regulated to ensure accurate reporting of their costs or income during transactions, ensuring the integrity of the trading environment.

- Establishing appropriate feedback mechanisms.

In general, the feedback obtained from transaction data is very limited, making it difficult to accurately estimate the market value of the data and apply efficient online learning algorithms effectively. Establishing a record mechanism for each transaction or

transaction failure, analyzing the reasons for the transaction results, and setting up comprehensive feedback channels can guide subsequent transactions effectively, promoting the healthy development of the data market.

5. Conclusions

This paper provides a comprehensive overview of data pricing issues from the perspective of data science. We have commenced by reviewing existing surveys in the field of data pricing, highlighting the significance of this topic. Subsequently, we have summarized the various stakeholders involved in data transactions, the criteria employed for data pricing, and the challenges associated with pricing data. We have outlined different data pricing models, along with their respective strengths and limitations. Moreover, we have emphasized the existing issues and potential research directions for the future. The primary objective of this paper is to offer guidance and insights for further research in the domain of data pricing, fostering the healthy development of data markets, and providing valuable references for scholars and practitioners in related fields.

References

[1] The Economists. The world's most valuable resource is no longer oil, but data. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, 2017.
[2] Pei, Jian. "A survey on data pricing: from economics to data science." *IEEE Transactions on knowledge and Data Engineering* **34.10** (2020): 4586-4608.
[3] Zhang, X. W., D. Jiang, and Y. Yuan. "A survey of game theory and auction-based data pricing." *Big Data Research* **7.4** (2021): 61-79.
[4] Liu, N., et al. "A review and comparative analysis of domestic and foreign research on big data pricing methods." *Big Data Research* (2021).
[5] Li, C. A. I., et al. "Survey of Data Pricing." *Journal of Frontiers of Computer Science & Technology* **15.9** (2021): 1595.
[6] Jiang, Honglu, et al. "Differential privacy and its applications in social network analysis: A survey." *arXiv preprint* arXiv:2010.02973 (2020).
[7] Koutris, Paraschos, et al. "Query-based data pricing." *Journal of the ACM (JACM)* **62.5** (2015): 1-44.
[8] Koutris, Paraschos, et al. "Toward practical query pricing with querymarket." *proceedings of the 2013 ACM SIGMOD international conference on management of data*. 2013.
[9] Deep, Shaleen, and Paraschos Koutris. "QIRANA: A framework for scalable query pricing." *Proceedings of the 2017 ACM International Conference on Management of Data*. 2017.
[10] Chawla, Shuchi, et al. "Revenue maximization for query pricing." *Proceedings of the VLDB Endowment 13.1 (2019)*: 1-14.
[11] Wang, Xingwang, et al. "On pricing approximate queries." *Information Sciences* **453** (2018): 198-215.
[12] Niu, Chaoyue, et al. "Unlocking the value of privacy: Trading aggregate statistics over private correlated data." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018.
[13] Niu, Chaoyue, et al. "ERATO: trading noisy aggregate statistics over private correlated data." *IEEE Transactions on Knowledge and Data Engineering* **33.3** (2019): 975-990.
[14] Cai, Hui, et al. "Online pricing and trading of private data in correlated queries." *IEEE Transactions on Parallel and Distributed Systems* **33.3** (2021): 569-585.
[15] Cai, Hui, et al. "Towards Correlated Data Trading for High-Dimensional Private Data." *IEEE Transactions on Parallel and Distributed Systems* (2023).
[16] Miao, Xiaoye, et al. "Towards query pricing on incomplete data." *IEEE Transactions on Knowledge and Data Engineering* **34.8** (2020): 4024-4036.
[17] Chen, Chen, et al. "GQP: A Framework for Scalable and Effective Graph Query-based Pricing." 2022 *IEEE 38th International Conference on Data Engineering (ICDE). IEEE*, 2022.

[18] Ghosh A, Roth A. Selling privacy at auction[C]//*Proceedings of the 12th ACM conference on Electronic commerce*. 2011: 199-208.
[19] Zhang, Mengxiao, Fernando Beltran, and Jiamou Liu. "Selling data at an auction under privacy constraints." *Conference on Uncertainty in Artificial Intelligence. PMLR*, 2020.
[20] Li, Chao, et al. "A theory of pricing private data." *ACM Transactions on Database Systems (TODS)* **39.4** (2014): 1-28.
[21] Nget, Rachana, Yang Cao, and Masatoshi Yoshikawa. "How to balance privacy and money through pricing mechanism in personal data market." *arXiv preprint* arXiv:1705.02982 (2017).
[22] Shen, Yuncheng, et al. "Personal big data pricing method based on differential privacy." *Computers & Security* **113** (2022): 102529.
[23] Xiao, Mingyan, Ming Li, and Jennifer Jie Zhang. "Locally Differentially Private Personal Data Markets Using Contextual Dynamic Pricing Mechanism." *IEEE Transactions on Dependable and Secure Computing* (2023).
[24] Fallah, Alireza, et al. "Optimal and differentially private data acquisition: Central and local mechanisms." *arXiv preprin*t arXiv:2201.03968 (2022).
[25] Feng, Zhenni et al. “Towards personalized privacy preference aware data trading: A contract theory based approach.” *Computer Networks* **224** (2023): 109637.
[26] Jiang, Xikun, et al. "Pricing GAN-based data generators under Rényi differential privacy." *Information Sciences: An International Journal* **602**-(2022):602.