A Survey on Big Data Market: Pricing, Trading and Protection

Fan Liang*, Wei Yu*, Dou An†, Qingyu Yang‡, Xinwen Fu§ and Wei Zhao¶

*Department of Computer and Information Sciences, Towson University, USA

Emails: fliang1@students.towson.edu, wyu@towson.edu

† MOE Key Lab for Intelligent Network and Network Security, Xi'an Jiaotong University, China

Email: adkaka.an@gmail.com

‡ SKLMSE lab, Xi'an Jiaotong University, China. Email: yangqingyu@mail.xjtu.edu.cn

§ Department of Computer Science, University of Central Florida, USA. Email: xinwenfu@ucf.edu

¶ American University of Sharjah, Sharjah, UAE. Email: zhao8686@gmail.com

Abstract—**Big data is considered to be the key to unlocking the next great waves of growth in productivity. The amount of collected data in our world has been exploding due to a number of new applications and technologies that permeate our daily lives, including mobile and social networking applications, and Internet of Thing (IoT)-based smart-world systems (smart grid, smart transportation, smart cities, etc.). With the exponential growth of data, how to efficiently utilize the data becomes a critical issue. This calls for the development of a big data market that enables efficient data trading. Via pushing data as a kind of commodity into a digital market, the data owners and consumers are able to connect with each other, sharing and further increasing the utility of data. Nonetheless, to enable such an effective market for data trading, several challenges need to be addressed, such as determining proper pricing for the data to be sold or purchased, designing a trading platform and schemes to enable the maximization of social welfare of trading participants with efficiency and privacy preservation, and protecting the traded data from being resold to maintain the value of the data. In this paper, we conduct a comprehensive survey on the lifecycle of data and data trading. To be specific, we first study a variety of data pricing models, categorize them into different groups, and conduct a comprehensive comparison of the pros and cons of these models. Then, we focus on the design of data trading platforms and schemes, supporting efficient, secure, and privacy-preserving data trading. Finally, we review digital copyright protection mechanisms, including Digital Copyright Identifier (DCI), Digital Rights Management (DRM), Digital Encryption, Watermarking, and others, and outline challenges in data protection in the data trading lifecycle.**

Keywords—**Big Data, Data Pricing, Privacy and Digital Copyright Protection, Data Trading, Data Utilization, Internet of Things.**

I. INTRODUCTION

With a number of new technologies integrated into our daily lives, such as mobile and social networking applications, and Internet of Thing (IoT)-based smart-world systems (smart grid, smart transportation, smart city, and others), massive amounts of data will be collected [1], [2], [3], [4], [5], [6], [7]. The different kinds of sensors and smart devices generate large datasets continuously from all aspects and domains. Thus, unprecedented, comprehensive, and complex data, namely big data, becomes more valuable. Furthermore, with the advancement of data analytics provided by machine learning and data mining techniques, and the computing capabilities supported by cloud and edge computing infrastructures, the potential values of the generated big data become more impressive [8], [9], [10], [11], [12], [13], [14]. Thus, big data is the impetus of the next waves of productivity growth. Nonetheless, there are a number of significant challenges, including data collection, storage, analysis, sharing, updating, and others. To maximize the utility of the data collected, one viable solution is to design an effective big data trading market that allows data owners and consumers (i.e., buyers) to carry out data trading effectively and securely.

In the past few decades, businesses have moved from pricing Internet service at a fixed hourly rate, to data plan-based flat-rate pricing models. Indeed, in 1996, the largest U.S. Internet Service Provider, AOL, switched to a monthly data plan [15]. This was the first time that a business endowed the data itself with a commercial value. Since the emergence of big data, datasets have become a type of "new money" of the digital world [16]. Thus, big data trading has become a growing and promising area in research and industry communities. Unlike traditional commodities, data is a virtual item, the basic characteristics of which are variability, variety, volume, velocity, and complexity. To realize the real value and utility of big data, traditional pricing models and strategies must be reevaluated and further improved.

With growing attention on the economic value of big data in improving the efficiency and decision making of utilities, customer experience, and others, several third-party big data trading markets have been designed [17], [18]. For instance, Global Big Data Exchange (GBDEX) owns 150 PB of authorized tradable data collected from thousands of companies and organizations. Nonetheless, due to the shortage of feasible protocols, existing big data trading markets are still in the initial stages. To enable an effective market for data trading, several challenges need to be addressed. The first issue is related to how to determine the proper price for the data to be traded, a problem which must consider the market structure in the design of corresponding data pricing models. Via proper price, the economic benefits of both data owners and consumers can be ensured. The second issue is related to the data trading platform and schemes. In this regard, feasible

trading platforms and schemes must be designed to ensure profit, fairness, truthfulness and privacy of the participants in the market. For example, a trusted third-party platform needs to be created to ensure that the data stored in different locations can be circulated to provide reliable service to heterogeneous users. In addition, to prevent privacy leakage or other attacks, the data trading process demands high level security and privacy. The third issue is related to data copyright protection, as digital products can be easily counterfeited or duplicated. Specifically, if the purchased data is resold by buyers, the value of data from the original data owners as sellers will be significantly affected, leading to the unwillingness of data owners to participate in the market. Thus, data copyright protection schemes must be designed to ensure the owners' legal rights.

To address the aforementioned issues, in this paper we conduct a comprehensive survey of big data trading to assist newcomers and provide a general understanding of this complex discipline and emergent research area. Our contributions are listed as follows:

- We review existing research related to big data, and identify the big data lifecycle for data trading, including data collection, data analytics, data pricing, data trading, and data protection. It is worth noting that, because a significant volume of research has been devoted to data collection and data analytics, our survey focuses on data pricing, data trading, and data protection, which have not been well explored.
- We review existing research related to big data pricing. We first illustrate the principles of data pricing and explain the reasons why this process is important. We then categorize the popular market structures, data pricing strategies, and data pricing models, and list the advantages and limitations of each category.
- We investigate the data trading process, and summarize data trading issues and the solutions for handling those issues. We further systematically investigate the auction, as one popular trading strategy, and detail different auction schemes, related platforms, and issues with respect to efficiency, security, and privacy protection.
- We study the final piece of the big data lifecycle: data protection. We summarize the existing copyright protection schemes and illustrate the advantages and disadvantages of those methods, as well as outline the challenges of copyright protection for big data.

The remainder of this paper is organized as follows: In Section II, we briefly discuss the principles and basic concepts of big data, as well as list the challenges and potential value of big data. In Section III, we identify the big data lifecycle and outline challenges related to data pricing, data trading, and data protection. In Section IV, we review the existing data pricing models, categorize these pricing models, and discuss their advantages and disadvantages. In Section V, we focus on data trading, reviewing data trading platforms and schemes, and discussing related issues. In Section VI, we discuss data copyright protection schemes and outline challenges for data copyright protection. Finally, we conclude

the paper in Section VII.

II. Basic Concept of Big Data

In this section, we introduce the basic concept of big data, including the definition, challenges, and applications.

A. Definition of Big Data

The total amount of data in the world is exploding with an estimated 2.5 quintillion bytes of data generated every day. Indeed, almost 90 % of the data in the world was created in the last two years alone [19]. The data sources are diverse, especially as IoT is ever more involved in our daily lives, supporting numerous smart-world systems [3], [1], [20], [21], [22]. Such diverse data sources result in the expansive volume of data, likewise creating massive potential commercial value. We refer to those kinds of data as big data.

While there is no consensus definition of big data, as shown in Figure 1, the *three V's* is the most-used definition of big data: (i) *Volume:* Huge data size is the first characteristic of big data. The size of the dataset can range from terabyte to zettabytes, or greater. For instance, Facebook stored roughly 100 petabytes of media (photos and videos) as of 2012, which was uploaded by 845 million users [23]. (ii) *Velocity:* Velocity is the characteristic of how rapidly the data stream is changing and being generated. Multiple data sources constantly generate data such that big data has an unbelievably high refresh rate. It also has only a short time frame to process the data. Even though the total amount of data is roughly 100 petabytes in Facebook, there are still 1.13 billion daily active users uploading 900 million photos each day [24]. (iii) *Variety:* The data can have a wide array of different and complementary formats, such as log data from various devices and applications, database files, and XML files, among others. In addition, the data can have unstructured data types (images, video and audio streams, etc.). Thus, big data is massive, continuous, and comprehensive, and has a high potential commercial value thanks to advances in data analytic techniques, such as machine learning and data mining [9], [14].
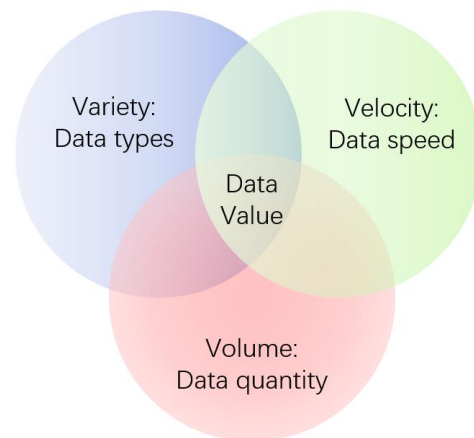

Fig. 1: Three V of Big Data

Notice that the terms data mining and Business Intelligence (BI) [25] are often used interchangeably to describe the

processing of big data. Obviously, those concepts are related to data analysis. Thus, the goal of big data is not only to collect data, but also to conduct data analysis to extract business value. As an extension of the traditional definition of big data, another V has been considered, namely Value. Particularly relevant to data trading, the relationship between the three Vs and the value of data, and how they affect each other, are important and challenging aspects of big data research.

B. Benefits and Challenges of Big Data

Compared with traditional data sources, big data has both advantages and disadvantages. In [26], the following differences are categorized:

- *Comprehensiveness:* Big data not only captures major activities, but also captures the related data with details for future analysis. For instance, as smartphones grew in popularity, so too did the use of social networking to connect people and distribute pictures and video. Traditional data source may capture only the contact list, whereas big data can involve a lot of sensors and data in the smartphones, recording as much information as possible (location, facial information, voice information, etc.). Such additional information could provide comprehensive details to describe the person and help big data applications to carry out the future analysis and provide tailored services.
- *Constancy:* Big data captures information constantly. As an example, most people experience annual or biannual physical health checkups. The hospital or doctor records the basic health index for each patient, including blood pressure, body temperature, height, weight, and more. Nowadays, new systems like the Apple Watch and Sports Bracelets with sensors are able to record these metrics continuously, anytime and anywhere. Such technology has the potential to obtain highly frequent data of large populations for in-depth big data analysis.
- *Multiplicity:* In big data, there are more and more semi-structured and unstructured data, as opposed to structured data [27]. Most traditional datasets are arranged as structured datasets, because the designers already know the type and the structure of the traditional data source, and the data is destined for traditional databases. For instance, a receipt from a market, a salary payroll, and an inventory list are typical business applications with traditional structured types of data, and are easy to use and manage. In contrast, unstructured data sources are difficult to control or manage. Video streams, audio files, and text data are the examples of this category, which have largely varying size, encoding, and context. Analyzing and managing the unstructured data is difficult as the data bits are not predefined.

C. Big Data Applications

To make big data useful, big data analysis software tools can extract useful information. From a big data user's perspective, big data applications can be used to analyze and mine value from big data source.

1) The Purpose of Big Data Applications: In last several decades, every level of economic entity in the world has turned toward using data intensive technologies. This widespread adoption depends to some degree on economic development and education level, which promotes data growth. Thus, Oracle, IBM, Microsoft, Dell, and many other companies have invested heavily in applications development for big data management and analytics. In addition, the big data application industry is growing at around 10 % every year, almost twice as fast as the traditional software field [28]. Thus, big data management and analytics applications are the keys to data value creation.

Many specific fields, such as government, manufacturing, health-care, education, Internet, social media, and IoT-driven smart-world systems all need big data applications to mine the value of their own collected datasets to better support applications. For instance, most of the data-intensive business-based companies like Facebook, Google, and Tencent extract value from their own datasets generated by their user platforms. The main purpose of this process is to sell those valuable datasets to potential advertisers, other third parties, or presenting them to investors to generate further investment. Thus, it is important that an efficient big data management and analytics application for mining commercial value from the collected data must be in place. The big data application becomes an important reference for data pricing as well.

2) The Challenges of Big Data Applications: One of the challenges for big data applications is that there is no direct and simple way to quantify the value of the datasets. As we discussed before, by increasing a big data application's performance, this should also increase commercial value for the resulting datasets [29]. Following this rule, to pursue the maximum value of the datasets, one efficient way is to increase the application's performance in the process of generating value from the datasets. To increase such a performance, one obviously needs to increase computation capability and running efficiency, and reduce computation resource requirements and data storage costs. Nonetheless, the issue remains how to quantify the improvement, and it must be noted that there is no guarantee that the commercial value of the datasets will increase by carrying out these simple improvements. Thus, it is necessary to design a comprehensive performance evaluation model. By modeling the application's performance, technicians and managers are able to make informed decisions, and the experimental results can serve as a reference to design future improvements for value generation.

Extending the discussion above, the next challenge for big data applications is the design and development of an appropriate model for the evaluation of the value generation process. There are many interrelated and complex scenarios and parameters used to measure the performance of big data applications in such a process. For instance, each computation task may involve a number of discreet computation nodes for big data applications. Furthermore, during a certain computation task, the involved computation nodes can be changed via scheduling strategies [29]. Considering the complexity of structure and interaction activities for big data, the modeling and performance evaluation of big data applications require specialized

knowledge. For instance, in [30], Structured Infrastructure for Multi-formalism modeling and Testing of Heterogeneous formalisms and Extensions for SYStems (SIMTHESys) was defined as a new framework for big data modeling. In addition, SIMTHESys [31] is the modeling framework that is designed to adapt the rapid and randomly changing system models [32]. Other modeling frameworks were proposed as well, including AToMe [33], OsMoSys [34], and Mobius [35].

3) The Convergence of Big Data and Other Technologies: Big data is the fundamental source/input for Artificial Intelligence (AI) and Machine Learning. In the big data era, the vast number of datasets feed those technologies to obtain meaningful results. Nonetheless, the ability to randomly access vast amounts of data momentarily and with agility is a challenging issue for designing effective big data applications [36]. In addition, instead of working with limited sample sets of data in statistics fields and data analysis sciences, as in the past, big data allows scientists to access and analyze unlimited datasets. The result is obviously improved analyses, due to massively increased sample sizes of big datasets, as well as more variety and detail in sources and sensors. This is the reason why a number of organizations have transitioned from experiential-based analytics strategy to a big data-based strategy. Organizations are able to develop their own applications to fit their unique requirements. Furthermore, during the analysis processing, the redundant or unnecessary data can be filtered out. This refines the source data, and the datasets become consolidated. Running refine loop constantly, datasets can be analyzed via "analytical sandboxes" and big data "centers of excellence", and can also improve the flexibility of data management [36].

Machine learning techniques, such as deep learning, are viable approaches to exploiting the value of big data [37]. Machine learning is driven by big data sources, is suitable for large, complex datasets that change rapidly, and can be further improved through the assistance of cloud and edge computing infrastructures [38]. Unlike traditional analysis techniques, machine learning is capable of thriving on growing datasets. In this way, the more data is fed into a machine learning system, the more it can learn, leading to results with higher quality. Thus, merging big data and machine learning can help organizations improve the extraction of business value from their own datasets and expand their big data application analytics capacity.

D. Value From Big Data

Big data is now the most important resource of the data technology era. To trade or share data resources, how to evaluate the commercial value for those datasets is a fundamental issue. Furthermore, capturing and mining value from datasets can further increase the value of data. To determine commercial value from big data, we need to define what is the commercial value of a dataset. As we discussed previously, the most commonly cited definition of big data is proposed by Gartner (2012): "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [39]." Although this is a usable

characterization of big data, this definition is not clear enough to give us an explicit distinction and scale from high to low value. Using this definition, we cannot measure the value of a dataset. Thus, an evaluation-based definition is necessary for capturing data value.

Obviously, collecting and storing large amounts of data is not the goal for all companies and organizations. Yet, they are all interested in analyzing the data to extract and create actual commercial value [40]. In [41], Davenport listed a number of real and anecdotal examples of how organizations design strategies for using collected datasets and mine value from those datasets. Furthermore, a comprehensive research from [42] indicated that data-driven decision making could lead to better performance over other decision-making methods in terms of productivity and profitability. There are a number of studies [43], [44], [45] on identifying the issues of how big data analytics creates commercial value, and where commercial value from big data can be obtained. Based on the systematic study, there are two main aspects of big data, from which commercial value for organizations can be created. The first is in the ability of big data to be used to improve and optimize current business processes, services and practices. The second is in the development of new business models, products, and practices that can be developed and innovated from the analysis of big data. Thus, capturing value from big data needs to identify the relationships between business models and the big data analyzed.

Data mining is one of common methods to capture value from datasets. Nonetheless, there are some challenges involved with the application of data mining for big data. The first challenge focuses on data accessing and computing procedures. Due to the distributed storage systems and the volume of continuously growing data, the computing platform must have the ability to handle the distributed and large-scale data storage. Most data mining algorithms require loading of all the necessary data into main memory, which is obviously a technical challenge in the case of big data, since moving data from the distributed storage system is expensive [46]. The second challenge is the various big data applications. More specifically, applications exist in different domains, with different data privacy and data sharing schemes between the data owners and consumers. The third challenge is designing effective machine learning and data mining algorithms. The learning and mining algorithms have to handle the difficulty of large volumes, and distributed, complex and dynamic data characteristics [46].

III. BIG DATA LIFECYCLE

We now define the big data lifecycle, separated into five stages. Based on this lifecycle, we outline challenges related to key stages, such as data pricing, data trading, and data protection. Figure 2 illustrates the detailed stages of the big data lifecycle.

- *Stage 1 – Data Collection.* Data collection is the first stage of the big data lifecycle. With the development of smart devices and IoT, it becomes easier to collect useful data everywhere. There are three steps for data collection:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2806881, IEEE Access

5

(i) *Gather data*, different types of data are collected via different collection methods and all raw data is stored by the data owners. (ii) *Clean data*, after collection, the data owner needs to pre-process the raw data, remove the noise, and sort the different types of data into reasonable groups. For instance, the unstructured data and structured data will be separated for further processing. (iii) *Verify data*, to make sure that the original data is usable and makes sense, data verification is necessary. In addition, sample data will be randomly selected and to check the usability.

- *Stage 2 – Data Analytics.* Data analytics is the second stage of the big data lifecycle. After the collection and pre-processing of the raw datasets, data analytics supported by machine learning and data mining techniques is the most important stage to extract commercial value from the datasets. In addition, approximately 74 % of information technology organizations use at least one of the advanced analytics methods mentioned [47]. There are numerous benefits of data analytics, including social influence marketing (61 %), customer-based marketing (41 %), and opportunities of sales marketing (38 %) [48].
- *Stage 3 – Data Pricing.* As the datasets have clear commercial value, data pricing models and methods are chosen after data analysis has been applied. In this stage, the data owners give each dataset a reasonable price in order to push those datasets into digital markets. The factors that affect the price include the data size and customers demands, among others. The owners can use various data pricing models to evaluate the datasets and obtain the best profit.
- *Stage 4 – Data Trading:* Data trading is the fourth stage of the big data lifecycle, and must be considered as distinct from traditional goods trading. Data as a kind of digital commodity, needs an appropriate market and trading methods. To trade data safely, fairly, and obtain the best profit, the design of effective data trading schemes, such as auctions, is important.
- *Stage 5 – Data Protection:* The last stage of the big data lifecycle is data protection. After data is traded, it is necessary to protect the copyright of the data and the data owner's legal rights. This is also an indispensable stage for the closed loop of the big data lifecycle.

There have been a number of research efforts devoted to Data Collection and Data Analytics. Thus, in this paper, we instead focus on the remaining stages. Thus, in the following sections, we will further study Data Pricing, Data Trading, and Data Protection. To be specific, for data pricing, we review existing data pricing models, categorize these pricing models, discuss the pros and cons of the models, and outline challenges and future research in Section IV. Then, with respect to data trading, we present the challenges of data trading platforms and schemes, discuss related issues, and outline challenges and future research in Section V. Finally, in regard to data protection, we study existing copyright protection technologies, illustrate the benefits and drawbacks of some existing schemes, and outline the challenges and future research for
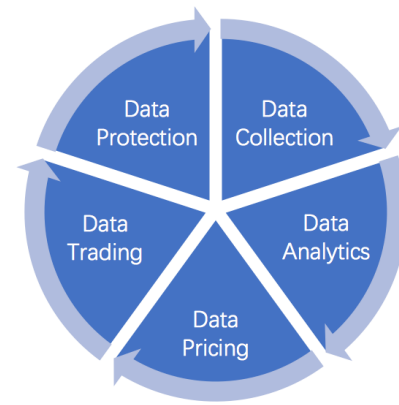

Fig. 2: Big Data Lifecycle

data copyright protection for traded big data in Section VI.

IV. DATA PRICING MODEL

Big data continues to grow exponentially. Accompanied by the growth of big data and the development of big data-driven applications, the data itself becomes more valuable. As we mentioned above, data mining and machine learning processes can generate commercial value from the datasets, based on sufficient and comprehensive data samples. Thus, big data becomes a new kind of data asset, which consequently needs an effective and fair method for evaluation and pricing. In this section, we first review existing data pricing models. We then categorize the outlined pricing models and discuss the pros and cons of each.

A. Principle of Digital Commodity Pricing

Data can be considered as a kind of digital commodity to be bought and sold in the market. Early research of physical goods trading in economics shows that the differentiation of price point for physical goods is dominated by the feature differentiation of the product line. This model was proposed by Mussa and Rosen in 1978 and named Vertical Segmentation or Quality Segmentation [49]. In this model, the consumers obviously prefer a higher quality commodity to a lower quality commodity, since it has the same price. In order to satisfy different consumers, the producers usually provide various product lines with different quality levels of products. Generally speaking, the producers have to consider both incremental cost and quality dependent cost to make decisions for the differentiation of the products. For example, in information technology product firms, different levels of product with different prices are commonly offered, and the available prices cover most price points from high to low. Like the research in physical commodity trading, people pay more attention to the quality differentiation for digital and information goods. Thus, the version-based strategy is one common way to determine the price of a digital commodity.

Furthermore, for physical commodity production, the re-production cost is part of the prime cost that needs to be considered. On other hand, for digital commodity productions, there is almost zero cost for re-production (version control,

integrity check, maintenance, and others). Thus, the factors of commercial price measurement for a digital commodity is developing cost, collocation or analytics cost, and maintenance cost. Meanwhile, to satisfy different consumers, the price of a digital commodity also needs differentiation. For instance, raw or pre-processed climate record datasets can be re-packaged into several levels of products by using different precision, time frequency, and others. These datasets with different prices and features can satisfy various consumer needs. Based on the study of current big data pricing strategies, we categorize the existing pricing strategies into different groups.

B. Data Market Structures

As we know, the market structures determine the price in the trading of physical goods, and the same holds true for the data market, in which those structures likewise affect data price. Thus, summarizing data market structures is the first step for data pricing. We organize the market structures in Table I.

In monopoly structures, the same dataset is future analyzed or produced by different monopolists. In other words, the various qualities and levels of commodities are produced based on the same original dataset, we call a data commodity. To do this, the monopolists have enough power to increase the profits for data commodities. Commonly, instead of setting a single price for all data commodities, those monopolists set different price points depending on the qualities and details of the data commodities in order to satisfy the demands for different levels from consumers. This strategy is defined as price discrimination [50]. Nonetheless, the monopolists usually conceal those tricks first, and try to investigate the preferences of consumers. In the data market, monopolists often set a price as a reference and monitor the reactions from consumers. Depending on this reference, and slight increases and decreases to price, this method enables monopolists to subdivide the demand models and price functions. Conversely, this also pushes the profits of the data commodity to maximum.

In competition, most monopolists lose their market positions, and only a few winners survive. Those winners mostly control the market resources, and this results in an oligopoly structure. Particularly, in the data market, original datasets only belong to a few owners. To this end, the data owners have strong abilities to control the machine learning and data mining process, market price, competition, and opportunities. Thus, the data owners achieve maximum profits in the data market. Nonetheless, this data market structure is diseased, and it is impossible to make the data market flourish. Oligopolies critically affect consumer demands and the services of providers. In addition, lack of competition makes the data market sluggish.

In the case of the strong competition structure, the selling price should approach the marginal cost, which increases the market transparency. It brings numerous benefits to consumers, such as lowering prices and generating better services. Nonetheless, in the long term, this structure can cause problems. Because all competitions lead to less profit, and in order to sell more commodities in a strong competition structure,

owners have to reduce the selling price as much as they can. This definitely reduces the owner's benefit and hence decreases the number of competitors. Especially relevant to the data market, the strong competition structure usually appears in emerging markets. Those new fields have a lower threshold for entering the market. Thus, many owners swarm the field with homogeneity and inferior quality products, and there is, in reality, no competitiveness. Thus, only a decreased selling price provides an efficient method to provide competitiveness to the data commodities. The competitive price-cutting will lead to cut-throat competition and shrink the market.

C. Data Pricing Strategies

Commonly, considering costs is the only rule for pricing a commodity, especially for digital commodities. In fact, only considering cost is a common defect, and should be just one factor of reasonable pricing. A mature pricing strategy will be the primary factor for maximizing profit rather than reducing cost. Thus, selecting an appropriate pricing strategy is also important. There are a number of ways to identify the pricing strategies utilized by different producers or companies. For example, the authors in [50] organized the different types of data pricing strategies into the following six main categories for the data market.

- **Free Data Strategy** is the publishing data online or to share in public storage, and trading is not the goal for free data. For instance, data samples, low-accuracy data, and public databases are examples of free data. Free data can attract some potential customers who hesitate to purchase the complete datasets and stimulate consumption. Meanwhile, free data pricing model has the flexibility. Based on the requirements, the owners are able to adjust the free data strategy to other pricing strategies in order to maximize the profits for the owners.
- **Usage-Based Pricing Strategy** refers to a measurement that counts the data stream usage and service time. This strategy is involved for some primary market actions. For instance, initially, mobile phone carriers sold data service based on usage for each user. The companies count data usage and calculate the price each month. Similarly, network providers offered Internet service to users, counting the service time and calculating the price. In recent years, service providers have merged data usage and service time together, which changes the price dynamically. They consider both peak time and usage together, which is more reasonable for pricing data and service.
- **Package Pricing Strategy** is an enhanced usage-based pricing strategy. Some vendors such as T-mobile, Verizon, etc., provide a data package plan with a fixed price [51], [52]. The package pricing strategy is established based on a number of research efforts and data collected on the usage-based pricing strategy. To maximize the profit for vendors, additional efforts need to be performed, including user usage analytics, peak time monitoring, network traffic control, and others. Depending on the research results, the vendors can create a reasonable

TABLE I: Characteristics of Data Market Structures

Structure	Market share rate	Price determination	Competition	Profit	Fairness
Monopoly	Some vendors share the market	Have ability to control the price	Weak competition	High	Relatively fair
Oligopoly	Very few vendors share the market	Strong ability to control the price	No competition	Very high	Not fair
Strong competition	Low market share rate	Have no ability to control the price	Malignant competition	Low	Fair

pricing model for their digital commodities and services. Package pricing optimization is currently a highly active research topic.

- **Flat Pricing Strategy** is the simplest pricing strategy. In this strategy, time is the only parameter, and the vendor simply considers selling each digital commodity once. This pricing strategy is commonly used in software licenses and hosting. In addition, using flat pricing strategy is convenient for vendors in predicting expected profits, and developing future plans and activities. Nonetheless, the flat pricing lacks flexibility and diversity for consumers.

- **Two-Part Tariff Strategy** is a combination of package pricing and flat pricing strategies. In this scenario, consumers need to pay two parts of the total price. The first is the flat fee for software licenses, and the second is the constant service and data support. This strategy is broadly used by network service providers, mobile phone carriers, software companies, and others. These companies sell their digital products with a fixed price at first, and in addition, the second part includes the service fee, update fee, or data usage beyond the fixed packages.

- **Freemium Strategy** is a new strategy heavily utilized by many vendors in recent years. The main idea of this pricing strategy is providing the basic products or limited service to consumers for free. Meanwhile, the vendors also provide additional value of service (premium service) to the consumers at a cost. The pricing strategy for premium service can be any of the strategies listed above. This strategy is often adopted by small companies, such as small developers on the Apple and Google Play stores. They upload their products to the store as free to download. Nonetheless, the full functional version requires additional fees to unlock.

Given the strategies outlined above, designing a pricing model for data commodities requires consideration of both data market structures and data pricing strategies. Some existing data pricing models consider only market structures, such as auction, cost-based pricing, and others, while others involve a distributed data storage structure (cloud computing, edge computing, etc.) and IoT for assistance. We discuss these in further detail below.

D. Data Pricing Models

The primary factors of data pricing are the cost of data collection, the cost of data analysis, the cost of data management, and the demand of consumers. After the market structures and pricing strategies illustrated above, Figure 3 categorizes the pricing models into two main groups: (i) *Economic-Based Pricing Model*, and (ii) *Game Theory-Based Pricing Model*.

In the following, we first present key factors and challenges of data pricing, and then describe data price models in detail.
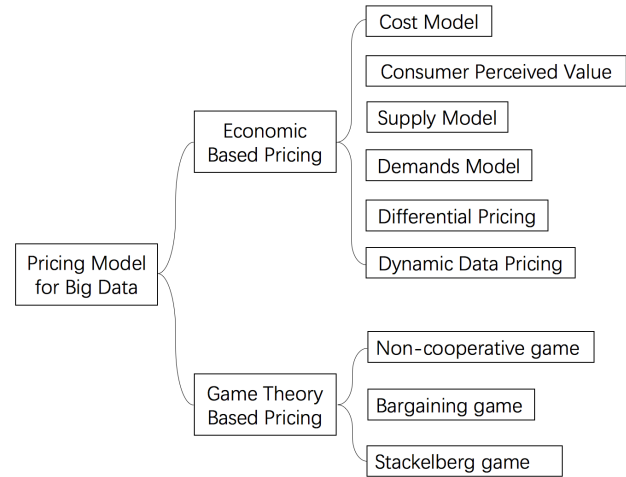

Fig. 3: Data Pricing Models

1) Main Factors and Challenges of Data Pricing: Data, as a unique type of commodity, has a number of characteristics not found in common physical commodities. Thus, the following are challenges for pricing a digital dataset.

- **Diverse Data Sources:** Along with billions of smart personal devices and sensors, IoT-driven smart systems have become major infrastructures that contribute data to the collection process. The diverse devices and associated deployment costs can pose significant challenges to evaluating collection costs. Meanwhile, the collected data has various types, and is difficult to classify and evaluate. Furthermore, how to motivate owners of those devices to contribute and share the collected data is an added challenge.

- **Complexity of Data Management:** Big data creates a huge data volume that is constantly increasing. Thus, how to manage (analyze, store, update, etc.) the data is another challenge for data pricing. In reality, there is a large cost for maintaining big data. From a technical perspective, most big data is stored in cloud or edge storage, and maintaining the storage and data usability, and securing the data incurs high costs. Those processes are also difficult to evaluate and price. Meanwhile, raw data needs to be analyzed before it is usable. Developing efficient applications to analyze datasets is also the factor for the evaluation of data pricing.

- **Diversity of Data:** In order to sell datasets, vendors usually process the raw data to satisfy various demands. This approach raises a number of complex issues for pricing evaluation. For instance, an original dataset needs

to be re-produced and divided into different levels of various volumes, precisions, and types. Then, how to evaluate the price of those different commodities remains a challenging issue.

There are a number of studies on how to handle these challenges. For instance, IoT provides the most important network infrastructure for data collection, as billions of devices run automatically and constantly collect data across various domains. Thus, to quantify the cost for data collection, and handle diverse data sources, we need to understand how IoT works, and encourage all components to provide better performance in IoT [53], [1], [54], [55]. To encourage sensors to upload data and achieve better profits for the sensor owners, an appropriate pricing model has become more critical. Introducing pricing mechanisms is one viable approach to encourage attendees to contribute their own data. Pricing mechanisms adjust price and payment schedule to guarantee the enough scale of participants and improve the data service, data accuracy, and data coverage.

There are several different pricing strategies for evaluating cost according to the different scenarios [56]. The most common strategy is economic-based pricing, which establishes the price model based on economic principles. The second strategy is game theory-based pricing. In such a strategy, the model considers the price to be affected via competition, and is dynamic.

2) Economic-Based Pricing Model: Economic-based pricing models are based on economic principles. In the following, we present the details of classical economic concepts for data pricing.

- *Cost Model:* It considers the total cost for any commodities and sets a ratio of the total cost as the profit [57]. We assume I as the desired income, C as the total cost, and p as the percentage of profit. The Equation $I = C(1 + p)$ represents the relationship between cost and income. The cost typically includes fixed and variable costs for the commodities. Generally speaking, fixed costs are resource costs, equipment costs, energy consumption, and others. The variable costs include labor costs, development costs, and others. The advantage of this pricing model is its simplicity, as it only considers internal factors to determine the selling price [58]. On the other hand, no external factors are involved, such as competition and demand, which are disadvantageous to this pricing model [59].

- *Consumer Perceived Model:* Since the cost-based pricing model is easy to imitate and copy by competitors, for a long term view, the vendors need to consider the feedback from consumers. Especially for digital commodities, there are almost zero re-produce costs. Thus, using the perceived pricing model is more reliable. The consumer perceived price is determined by the price, which all the consumers are willing to pay. In [60], the authors proposed five main factors to affect the data pricing, represented by $P_v = (v_p, v_c, v_m, v_s, v_e)$. Here, v_p represents the performance based on the feedback data from consumers. The key factor is utility, and the utility is the satisfaction measurements for the consumers, who purchase the commodities or services. Thus, in the digital market field, this is the most important factor for

the vendors to determine the price accuracy that satisfies the consumers. In addition, v_c is the market environment factors that could affect the behaviors of the consumers, v_m is the motivations of a consumer who is willing to purchase digital commodities, v_s is the supplier value, which represents the credit of the vendors and the main feedback from consumers, and v_e represents economic value, which depends on the demands of consumers, and the perception of price from consumers.

- *Supply and Demand Model:* The relationship between supplier and consumer is part of the commercial model. Depending on the relationship, the markets can determine the price for commodities. In the economics field, the supply and demand function is used to represent this relationship. In a market, denote P as the price of a commodity and Q as the quantity of a commodity. Thus, we have two linear functions for documenting this relationship. Equation $P = a - b \cdot Q_s$ is the supply function, and $P = c - d \cdot S$ is the demand function. Here, a, b, c, and d are the coefficients, and $b > d$ [61]. Based on those two equations, we could establish a relationship between supply and demand, shown in Figure 4. As shown in the figure, the orange line represents supply function and green line represents demand function. The actions between vendors and consumers are balanced conditions. As we can see from the figure, the orange line has a higher slope than the green line, and therefore there must be an intersection. This intersection is the balanced condition for vendors and consumers. At this point, we obtain P from Equation: $\frac{P-a}{b} = \frac{P-c}{d}$, and Q from Equation: $a - b \cdot Q_d = c - d \cdot Q_s$. There are two basic characteristics for this model [62]: (i) this is a consistent action between vendors and consumers, since the commodity is flowed into the market, and (ii) the vendors and consumers cannot change this process, and the decision making process is determined by the market. Thus, this model guarantees fairness in the market.
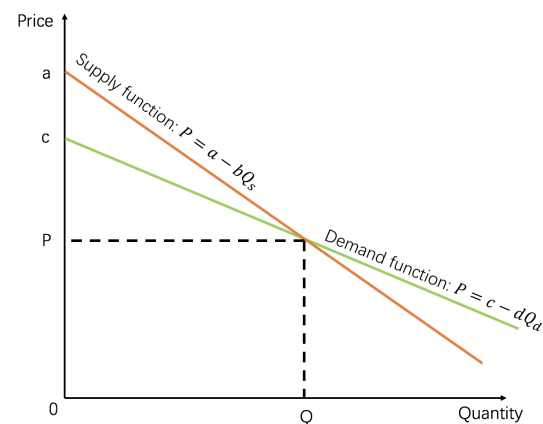

Fig. 4: Supply and Demand Functions

- *Differential Pricing Model:* In order to satisfy the various demands of commodities, the vendors should provide various commodities with different characteristics. The differential pricing model considers the difference be-

tween those commodities and provides different prices. For instance, the high accuracy data package should have a higher price than a low accuracy data package, and a full function digital application should also have a higher price than a demo version application.

- *Dynamic Data Pricing (Smart Data Pricing) Model:* This is a special case for the differential pricing model to avoid peaks in demand and data flow. It is also called the smart data pricing model (SDP). The dynamic data pricing model monitors the market and evaluates whether the system is busy or idle. Depending on the evaluation, the price of the digital commodities can be dynamically adjusted so that the resources of vendors and money of consumers can be saved. Furthermore, there are two main mechanisms to achieve this goal. The first is time-based pricing, and the second is usage-based pricing. For the time-based pricing mechanism, the price will be changed over time. For instance, the network provider usually offers a lower data price at night to encourage consumers to use network service during off-peak hours, and discourage use during peak demand. In the same way, the usage-based pricing mechanism will vary the pricing according to data usage. For instance, a supplemental usage-based data plan provided by a network provider is implemented once the usage exceeds the original flat data plan. Thus, it will encourage the demand of purchasers to reside within a fixed or planned window and discourage consumers from exceeding the predetermined limit.

3) Game Theory-Based Pricing Model: Game theory is a useful method used in the fields of pricing and markets, especially in pricing data commodities. In the following, we first introduce three different game theory schemes that are used for data pricing models: (i) *Non-cooperative game*, (ii) *Stackelberg game*, and (iii) *Bargaining game*. Then, we discuss how to use those game theory schemes in pricing digital data commodities.

Non-cooperative Game: In a non-cooperative game, all participants are assumed to not cooperate with each other. To illustrate the detail of a non-cooperative game, there are some terminologies: (i) *Player* is the individual who participates and makes decisions in a game. (ii) *Payoff* is the real profit or utility, and represents the expected result for a player. (iii) *Rationality* describes a condition in which all players want to maintain their individual maximum profit during the game process. (iv) *Strategy* is the finite actions that are operated by the player, and each player's strategy can be different. The payoff result will be affected not only by one play, but also by others [63].

In [56] and [64], the authors designed a pricing model to evaluate IoT sensing data. In the model, all vendors sell their data in a competitive manner, and they define this model as non-cooperative game. The vendors act in the role of players, and they determine the pricing strategy. Denote (V, π) as a n players (sellers) game, where V_i represents the pricing strategies used by player i. The V is the Cartesian product of each strategy sets: $V = V_1 \cdot V_2 \cdot V_3 \cdots V_n$, and π_i is the vector that represents the payoff for seller i. We set v_i as the pricing strategy for player i. Then, we obtain the vector of

strategies $v = (v_1, v_2, v_3, \ldots, v_n)$ for the number of n players. Meanwhile, we also obtain a vector $\bar{v}_i$, that represents all the chosen pricing strategies without the pricing strategy, which is chosen by player i. Thus, $\bar{v}_i = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots v_n)$. The relationship between those three factors is that player i uses a given pricing strategy v in order to achieve the payoff π. The Nash equilibrium of V represents the condition, in which none of the players can improve his/her profit by only changing his/her own pricing strategy without other players changing their pricing strategy. The inequality was given by [65] $\forall i, v_i \in V_i : \pi_i(v_i^*, \bar{v}_i^*) \geq \pi_i(v_i, \bar{v}_i^*)$.

From this inequality, we know that the players in the Nash equilibrium have no motivation to change their pricing strategy, because it would result in a worse payoff. Thus, achieving the Nash equilibrium is the only way to solve the problem. Notice that there is no Nash equilibrium in some conditions, and sometimes there is more than one Nash equilibrium in the opposite. Thus, to find the Nash equilibrium and the game only has one Nash equilibrium is the necessary and sufficient condition for using the non-cooperative game to price the datasets.

Stackelberg Game: In the non-cooperative game model, all players must publish a pricing strategy, and the pricing strategy is transparent. Nonetheless, this is not always true in the real digital market, as the players cannot compute the Nash equilibrium. Thus, they cannot price the digital commodities. Instead, the players need to wait until some other players announce their pricing strategy. For instance, when the mobile carriers sell them the data plan, they obviously need to consider the price from their competitors. Thus, the later a mobile carrier announces its pricing strategy, the better performance is. This is a typical Stackelberg game model [66].

In [67], the authors proposed a Stackelberg game model to protect the players who announce their pricing strategy. It defines two positions, the leader and follower. Assume there are two players in a market, and $V_1 and V_2$ are the pricing strategy sets for player 1 and 2. If player 2 announces V_2 first, then player 2 will be the leader and player 1 will be the follower. Using this proposed Stackelberg game, they proved that the leader could obtain a better payoff than in using other models. Notice that involving the Stackelberg game into the digital market can maximize the payoff for all the players, and especially for the leaders [68]. In addition, existing studies showed the use of the Stackelberg game model in spectrum trading and resource allocation [69], and could improve network performance and robustness in IoT systems [70].

Bargaining Game: The last scheme is the Bargaining game. This game theory refers to a condition in which the vendors and consumers reach an agreement, and to achieve this agreement, the vendors and consumers need to negotiate. Considering this process in a simple digital market, the trade only occurs when the vendors and consumers agree on the selling price for a certain commodity.

In a pricing model that was proposed by [71], denote r_v as the reserved price that ensures an acceptable payoff/profit for the vendor. Similarly, the consumer also provides r_c as the reserved price for the consumer, who is willing to purchase.

Meanwhile, similar to the other game theories, both vendors and consumers submit their pricing strategy P_v and P_c. Then, the vendors are able to determine the selling price strategy P_v^* from the expected profit $\pi_v(P_v, r_v)$ when $\pi_v^*(P_v^*, r_v) \geqslant \pi_v(P_v, r_v)$. Just like the consumers, they are able to determine the consumption price strategy P_c^* from the expected profit $\pi_c(P_c, r_c)$ when $\pi_c^*(P_c^*, r_c) \geqslant \pi_c(P_c, r_c)$ as well. Then, the vendors and consumers will compare P_v^* and P_c^*. If $P_v^* > P_c^*$, the negotiations need to continue. Otherwise, if $P_v^* \leqslant P_c^*$, a Bargaining game is enacted and the selling pricing P will be $P = kP_c^* + (1-k) \cdot P_v^*$, where $0 \leqslant k \leqslant 1$. Finally, we obtain the Nash bargain equilibrium at the price set (P_v^*, P_c^*).

Since the Bargaining game is a proper scheme to a complex negotiation condition, it is often used to improve the performance of the data auction process [72], network resource auction and sharing [73], and energy efficiency management [74].

Finally, based on our study, we organize the existing data pricing models in Table II.

To summarize, in this section we have provided a comprehensive overview of data pricing and reviewed major concepts related to big data pricing such as the principle of digital commodity pricing, data market structures and data pricing strategies. We have also reviewed economic-based pricing models and game theory-based pricing models in detail. Nonetheless, the models for big data are relatively few, and most state-of-the-art pricing models have been investigated for traditional goods. As future research directions, more efforts should be conducted in the big data science for understanding different types of data and the design of proper models to realize the exact values for different kinds of data users.

V. BIG DATA TRADING

Data pricing and data trading are complementary processes. Since data has a commercial price, data markets and data trading schemes become effective ways to assist the data pricing and sharing process. In addition, the value of data motivates a number of studies such as the design of data trading technologies to ensure that the data trading process is fair, secure, and efficient. In this section, we systematically study data trading schemes and platforms, and related issues.

A. The Main Purpose of Big Data Trading

Since the volume of data is growing immensely, and IoT technology is progressing at a similar pace, massive datasets with comprehensive content and detail become increasingly valuable. The main purpose or benefits of trading big data can be separated into two aspects. On one hand, the data trading process shall maximize profits for data owners. On the other hand, this process shall also satisfy the demands of consumers for massive data. The consumers can further utilize those datasets to improve their products or services. This is definitely a beneficial process for both owners and consumers.

- **For Data Owners:** Big data is the foundation for the next wave of productivity resolution: Data Technology (DT). Data owners, such as Facebook, Google, Amazon, Tencent, and Alibaba, collect massive data via the services that they provide [84]. Obviously, via the advancements in big data analytics supported by machine learning and data mining techniques, those datasets produce huge value for those companies. For instance, with the assistance of machine learning and data mining techniques, e-commerce companies are able to push commodities on consumers' wish lists or browsing history. The location-based service providers are able to distinguish the home or work locations for a customer, and provide the best route at the appropriate time. Nonetheless, not all the companies have the ability to collect the demanding data, since collecting huge and comprehensive datasets requires a significant infrastructure investment and long-term efforts. In terms of providing services, stimulating productivity, and maximizing the value of data, the data owners have strong aspirations to trade their own datasets with others.

- **For Data Consumers:** In high-competition environments, information is the key for a company to discover new business opportunities, values, and customers. Nonetheless, a big challenge is where the consumers can obtain the necessary datasets, since they have no ability to collect the data by themselves. To this end, the data consumers have a strong desire to purchase data from the market, and use those valuable datasets to improve their services or products. As an example, based on sufficient information, manufacturers are able to maximally match the requirements for many different consumers with product differentiation, and service providers are able to refine their service plans to improve and target their services to their customers [85]. Thus, data trading is one viable approach to satisfy those needs.

Without data trading, the data remains static, and forms individual information islands. Thus, data trading pushes the data as a dynamic flow, realizing the commercial value of the data, and establishing a win-win market. Indeed, data trading is the general trend for managing big data and a key to the expansion of the big data era. In addition, data trading can stimulate data analytics supported by machine learning, data mining, and other technologies, and provide benefits for both owners and consumers. In the following, we first outline key issues of big data trading, and then present big data markets with supported platforms and trading techniques.

B. The Issues of Big Data Trading

Big data trading involves resource trading and allocation via information communication technology. There are abundant research investigations that focus on resource trading and allocation, and leverage various algorithms or game theory schemes to optimize the trading process [86], [87], [88]. Nonetheless, some issues remain unsolved, including how to ensure the maximum profit for multiple vendors, how to ensure the trading is truthful, how to protect the privacy for both vendors and consumers, and how to establish a trusted trading platform. In the following, we discuss these issues in detail.

Multiple Owners Data Trading: Most of the research related to data trading has the limitation of only considering a

TABLE II: Data pricing models

Ref.	Market Structures	Pricing Models	Algorithms/Schemes	Goals	Approaches
[16]	Monopoly	Economic-Based Pricing	Information Entropy	Reduce uncertainty by reducing the information entropy to avoid the interference from uncertain data content	Define five tuples parameters for data and reduce the uncertainty information entropy
[75]	Strong competition	Economic-Based Pricing	Gaussian Process	Capture the uncertain contents of source data	Create a trading market, and simulate the trading process. Then, use statistics approach and Gaussian process to analyze the data trading and maximum profit process
[76]	Monopoly	Economic-Based Pricing	Max-Flow problem	Improve the transparency of data price	Propose a Generalized Chain Queries and obtain the price by computing the time complexity of GChQ query
[77]	Strong competition	Game Theory-Based Pricing	Stackelberg Game-Based Pricing	Use pricing incentives method to improve the Quality of Service (QoS) and encourage upload data for the sensors	Set a bundle with data and service; use Stackelberg game to encourage the service and negotiate the data price with consumers
[52]	Monopoly	Economic-Based Pricing	Alternate optimal multi-step pricing scheme	Compute the most closest pricing in case of fallacies	Propose a general framework to compute the profit for a licensed datasets pricing and consider with Willingness to Pay (WTP) of consumer side, and then obtain the maximizing profit
[78]	Monopoly	Economic-Based Pricing	Economic-Based schemes	Create the differential digital product	Calculate the differentiation of digital product and create the "differential delay", then, pricing the digital product depends on the differentiations
[79]	Monopoly	Game Theory-Based Pricing	Stackelberg Game-Based Pricing	Use concave increasing function and Stackelberg game to maximize the profit of datasets	Create a digital market model with sensors, datasets, and consumers and use Stackelberg game model to obtain the maximum profit
[80]	Strong competition	Economic-Based Pricing and Auction-Based Pricing	Economic-based schemes and Bargaining game	Calculate the cost and build economic model	Build the economic model to analyze and calculate the cost of a certain datasets, and then, use Bargaining game to confirm price of the datasets
[81]	Oligopoly	Economic-Based Pricing	Statistics and Machine learning (Time-Dependent Pricing based on K-Nearest Neighbors (TDP-KNN) algorithm pricing algorithm and Time-Dependent Pricing based on Transition Rules (TDP-TR) pricing algorithm)	Collect the usage of a certain dataset	Use machine learning TDP-KNN algorithm) to collect and analyze the usage of some datasets and calculate the future usage expectation, and then, price the data
[82]	Monopoly	Economic-Based Pricing	Genetic algorithm	Design an assessment methods for exist data product and obtain its multi-dimensionality and dimensions	Create a data-pricing tow level programming model depends on the quality and the profit, and using genetic algorithm solves the model
[83]	Strong competition	Economic-Based Pricing	Snowballing Algorithm	Design the market rule and the pricing models	Design a market model to create the selling database, and analyze the database

single data owner [18]. Nonetheless, in real-world practice, there are many owners of a commodity in a data market. The challenge, then, is how to quantitatively analyze the ownership for each owner. When there are multiple owners, they are in competition [89]. For instance, if a certain dataset has two owners, both of the owners want to sell the dataset through their own market. Although, the demand is constant, the competition appears. Thus, it is difficult to design the mathematic model to describe these complex requirements. In addition, as we mentioned, the maintenance cost of a data commodity is another important component of the total cost. Big data is usually uploaded and stored in the cloud, and the responsibility for updating, maintenance, and modification is difficult to quantify [90], [91]. Thus, determining and considering the maintenance cost for each owner is critical.

Trading Fairness and Truthfulness: Much like traditional commodities trading, the most important concern is fairness and truthfulness, which are the fundamental requirements for all trading processes. There are two primary aspects of fairness and truthfulness. The first is between vendors and consumers, and the other is between vendors, consumers and trading organizations [86]. Both of these aspects are challenging for data and digital commodities trading, since all the traded commodities are virtual goods, and all the trading processes are occurring via network, which are "blind" for all the vendors and consumers [92]. To handle this issue, there has been some research focused on establishing a fair trading platform [93], [94], [95], [96], [97], while others have focused on data commodity based on cryptography-based techniques [98], [99], [100], [101], [102]. Nonetheless, these proposed schemes all have some limitations. For instance, the authors in [94] proposed a fair trading market with a fair protocol, and the trading process can be finished or terminated at any time to ensure that there is no loss for both vendors and consumers. Nonetheless, the platform cannot discriminate false information, and only considers one trading process at a time.

Privacy Protection: Privacy is an important factor for both vendors and consumers. In the data trading process, some personal information of consumers should be hidden to protect privacy. Similarly, for the data commodities, privacy is obviously important as well. Generally speaking, people use both legal supervision and technical protection, such as copyright law, watermarks, encrypted licenses, and others. Nonetheless, copyright laws only focus on protecting the legal rights of the owner, and cannot protect the privacy of the data directly; they are not preventative. Also, watermarking technologies can only be used as evidence in an investigation to determine misuse. Both of these protection schemes are reactive. Regarding data encryption technologies, ever greater expenditure of computation resources is required to ensure higher privacy demands [103], [104], [105]. Some investigations have focused on privacy protection. For instance, a minimized design strategy was proposed in [106]. The principle of the proposed strategy is to reduce the risk of privacy leakage by providing the minimum amount of data at each time interval, and further increase the price for larger data packages. Cryptography-based technologies are often implemented for privacy protection. For example, a hiding design strategy was proposed to encrypt and hide a part of the data from the original source [107]. The encryption process can use different efficient encryption technologies, and encrypt data while the owners upload the data to the cloud/edge storage nodes.

Third-Party Trading Platform: With the increase in data trading demand, it is difficult or not effective/scalable for data owners establish their own trading platform. Thus, the third-party trading platform becomes a viable way to accomplish this. The data owners entrust the third-party trading platform to sell the data commodities to consumers, similar to trading traditional goods in an online market. Nonetheless, the trustworthiness for the platform is a big risk, since the data commodities incur almost zero cost for duplication. A number of studies proposed schemes to avoid the third-party trading platform from stealing the data commodities or leaking the information by selling the licenses and content separately. More typically, the owner encrypts data commodities and uploads them to the trading platform, and sells the key to the consumer. Thus, only the consumer who purchases the license can decrypt the data commodities.

C. Big Data Markets

Similar to the traditional markets that are important for trading traditional goods, data trading also needs data markets to support data trading. Notice that data is a virtual item/digital commodity, and has its own characteristics. Thus, to trade data in the market fairly and securely, establishing data markets is essential. There are a number of research works related to data market platforms and the supporting mechanisms. In the following, we discuss data markets in detail.

1) Market Platforms: A successful data market, which is necessary to offer both vendors and consumers an optimal experience in selling and purchasing, also needs to protect the privacy of both data commodities and personal information. To meet these requirements, we review some existing schemes.

Trading Query: Before the customers decide to purchase the datasets, there are many query processes for searching. Nonetheless, the query operations are not free. For instance, the Worldwide Historical Weather in Microsoft Azure Marketplace is $12 for every 100 "transactions". For this reason, the markets should have an efficient query system to minimize the cost for consumers. For optimizing such queries, a big data learning scheme was proposed by [108]. Nonetheless, the proposed scheme requires rich data statistics. Unfortunately, as data commodities are distinct from traditional goods, there are fewer statistical records in data markets (i.e., no purchase history, no value distribution) and only basic information such as the size and attributes of datasets may be available, which are obviously not enough.

To find an optimal solution to this problem, an optimized learning-based optimizer was proposed in [109]. This optimization scheme could reduce the number of queries in the purchasing process by designing an effective algorithm to reduce the amount of intermediate data. This scheme includes a parser, optimizer, and execution engine. In detail, the parser first obtains the local table information when the consumer registers with data market. Then, the optimizer optimizes the

query by loading reference data from the local data table and statistics of the data market information. Finally, the result is sent to the execution engine. After optimization, the scheme can avoid some accesses to the data market to reduce the cost for the consumer.

Dynamic Trading: Existing data markets often have two limitations. The first is that data markets usually only sell whole datasets, instead of requirement-oriented subsets, and do not support arbitrary querying, as we mentioned above. The second is that data markets typically do not support data update and maintenance, since the original datasets are uploaded by owners, and the data commodities are static. Nonetheless, the data commodities need frequent update, as data is dynamic. In [110], the authors proposed a dynamic data market framework to solve this problem. In this framework, an online sharing plan selection algorithm was used to ensure the efficiency for maintaining the data commodities' views. Then, through maintaining the view of data commodities, the commodities were kept updated.

In another study [111], the authors proposed a distributed algorithm with notions from matching game theory in terms of selling the data by demands. The scheme compares the formulated preference functions of vendors and consumers, captures the requirements of consumers and finds the matching part of a data commodity, and then sells the matching part to the consumers. This scheme supports the self-organization feature for all the participants into a matching table, and ensures the matching process and results dynamically adapt to the demands of consumers. Via the simulation, the results show that, with the use of their proposed scheme, the average utility for every consumer increased by 25 % to 50 %.

Privacy Protection: In the data trading process, a critical problem is how to trust a trading platform for both vendors and consumers. Neither the vendors, nor the consumers want to expose sensitive personal information to each other. Generally speaking, cryptography is an efficient way to protect sensitive information, and a number of research has focused on trading data using cryptography-based schemes [86], [102], [112]. For example, the authors in [86] proposed a Truthfulness and Privacy preservation in Data Markets (TPDM) mechanism. Particularly, TPDM adopted homomorphic encryption with signatures (identity recognition). It protects the privacy and data confidentiality, while improving batch verification and the data trading process. In contrast to traditional encryption schemes, the identity-based signature component processes the data in cipher-text space. In addition, all the signatures from data owners and consumers are their real identifications, and it prevents against all malicious vendors or adversaries.

2) Data Auction: One of the most popular data trading mechanisms is via the auction process. Generally speaking, an auction is an economically-driven scheme, which aims to allocate commodities and establish corresponding prices through buyers' and sellers' bidding process [113]. Auction theory has been well explored in several areas (economic, electricity market, mobile market, and others) [114], [115], [116], [117]. Due to the capabilities of ensuring fairness and efficiency, auction mechanisms show great potential to address big data trading problems. Before a detailed review of related

works on auction theory in the big data market, we introduce basic concepts of the auction mechanism as follows:

- *Bidder:* In the auction process, a bidder is the one who submits the bids and aims to buy commodities in the market. In the big data market, the bidder is typically data consumers (start-up companies who want to investigate an application using a particular dataset, researchers, and others).
- *Auctioneer:* The auctioneer plays the role of an agent who runs the auction process, enacts winner determination, and conducts payments and allocations. In the big data market, an auctioneer can be an agent in the cloud.
- *Seller:* The seller is the owner of the commodities to be bid upon and sold. In a big data market, this includes the utilities (Google, Facebook, etc.) that generate, collect and store large-scale data from different platforms and devices for further sale.
- *Valuations:* In the auction process, buyers and sellers both put valuations on each unit of the commodities that they request or sell. Moreover, the valuations can be greater or lower than the final clearing price, which is determined by the auctioneer in the auction process.
- *Clearing price:* In the auction process, the seller and buyer submit asks and bids. The asks indicate the asking price on the commodity to be sold while bids indicate the bidding price for the requested commodity. A clearing price will be determined by the auctioneer according to the optimization goal, such as social welfare maximization. In other words, the clearing price is the price at which the buyer and seller will make a deal.

Data Auction Models: In the recent past, a vast number of research efforts have been conducted toward auction mechanisms, and their applications tested. It follows, then, that many of these have been applied toward use in big data trading, which has gained in popularity. We now present some typical auction types that have been used in big data trading or have the potential to tackle the trading issues inherent to big data markets. Figure 5 illustrates a typical framework of the auction-based big data trading process [118].

- *One-side Auction:* One-side auctions include forward and reverse auctions [119], [120]. Forward auctions are also denoted as seller-side auctions, in which buyers compete for the commodities of the seller. For example, to enable effective data circulation among data producers and data users, An *et al.* in [118] proposed a Multi-rounds False name Proof forward Auction (MFPA) mechanism, which aims to maximize the social welfare of the data owner and consumer. To defend against false name bidding attacks, the volume of the data is traded in bundle sizes in MFPA. The authors conducted theoretical analysis to prove that the bidders can achieve maximum utility if and only if their bids and asks are truthfully submitted. In the case of reverse auctions, the sellers compete to sell commodities to buyers. Generally speaking, in big data markets, the reverse auction mechanism is suitable for the situation, in which multiple data owners tend to sell data to one data consumer or data collector.
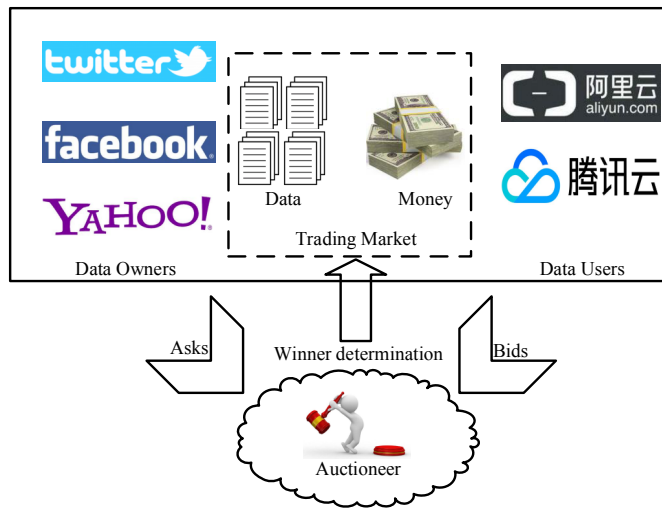
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2806881, IEEE Access

14

Fig. 5: A Framework of Auction-Based Big Data Trading Process [118]

- *Double Auction:* Double auction [121], [122] is one of the most commonly used auctions in the real-world practice, and has been widely used in the New York Stock Exchange [117], the smart grid [123], [116], and in the mobile market [115]. In the double auction process, multiple buyers and multiple sellers submit bids and asks to the auctioneer. Figure 6 illustrates typical curves of bids and asks from buyers and sellers [124]. Here, the black and red curves indicate the ascending order of asks from sellers and descending order of bids from buyers, respectively. After collecting the bidders' profiles, the auctioneer matches these bids and asks by the clearing price, as well as payments from the buyers to the sellers. Several related works have been explored to design double auction mechanisms in big data trading markets [87], [125].

For example, to prevent the low trading efficiency that is caused by selfish action, the authors in [87] proposed an iterative auction mechanism. This auction mechanism can avoid selfish actions and prevent direct access to private information. The procedure of the iterative auction involves four steps. In the first step, the auctioneer announces the allocations, pricing, and auction rules for the data commodities to all consumers. In the second step, each consumer computes the bidding price in order to maximize the utilities. In the third step, the auctioneer receives the bidding prices and, according to the rule and price, announces the result. Those three steps also exist in common auction processes. The unique aspect of their proposed auction mechanism is the fourth step, based on the prior auction process. In this step, the auctioneer can adjust and re-announce a new starting price and auction rule to start a brand new auction. This iterative auction process encourages consumers to list a reasonable price during the auction process. In addition, in the secondary mobile market, Susanto *et al.* in [125] proposed a McAfee-based double auction mechanism

to enable mobile data trading in a heterogeneous and dynamic environment. Their theoretical analysis proved that the proposed double auction scheme is capable of achieving Nash equilibrium and truthfulness.
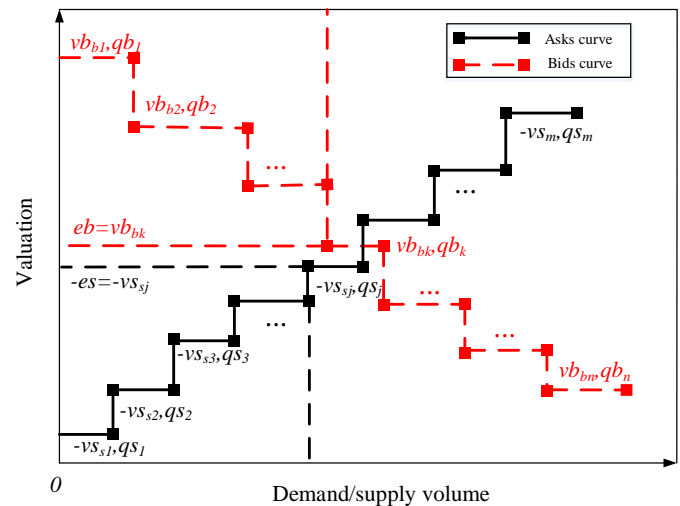

Fig. 6: The Bids and Asks Curves in Double Auction [124]

- *Seal-bid Auction:* In Seal-bid auctions, the buyers privately submit their bids to the auctioneer without knowing the bidding information of other buyers. Unlike conventional auctions, the seal-bid auction is a one-time auction, and leads to non-open competition for the buyers. Seal-bid Auctions have been well explored, and typical examples include the kth-price auction [126], [127], VCG auction [128], [129], and McAfee auction [113]. Recall that the kth-price auction can be divided into the first price auction and the second price auction. In the first price auction, the winner is the bidder who submits the highest bidding price and would thus pay the highest price to win the auction. In the second price auction, which is also known as Vickrey auction, the winner is the bidder who submits the highest bidding price, while the winner would pay the second highest price to win the auction. Notice that the first price auction ensures the maximum profit of the seller, while the second price auction induces the buyers to report truthfully, ensuring the fairness of the auction scheme. Vickrey-Clarke-Groves (VCG) auction seems to be a generalized form of Vickrey auction. Regarding the McAfee auction, it is an extension of Vickrey auction. Specifically, the buyers and sellers submit private bids to the auctioneer, of the buyers (sellers) whose bidding prices are larger (smaller) than a threshold price, and the winners would pay the highest price that does not win the auction. In the big data market, a few seal-bid auction schemes have been investigated [88], [118]. For example, the authors in [88] proposed an optimal price seal-bid auction market model based on the Bayesian optimal mechanism. First, the data sources are divided into three groups: Crowdsensing data, Social data, and Sensing data. Then, the cost function, satisfaction rate function, and data utility function are defined. Based on those

functions, the starting prices of the data commodities are identified. During the Bayesian profit maximization auction process, the valuation distribution function was computed. Based on this function, the most optimal price point and secondary optimal price point were identified. Meanwhile, the optimal size of data that takes from those collectors is identified. Nonetheless, this auction scheme only considers one round of auctions.

- *Combinatorial Auction:* In the big data trading market, the buyers' demands for data and the sellers' supply of data are always manifold. Thus, when applying the above-mentioned auction schemes, both buyers and sellers cannot be satisfied by simply rubbing the data together to trade. As a result, the combinatorial auction [130], [131] was designed for such a situation. In a combinatorial auction, the bidders in the market are allowed to bid on combinations and bundles of the commodities. Particularly, the bidders submit bids that contain the combination of a variety of commodities and the price of the combination. The auctioneer then makes the optimal allocation for the bidders according to their bids and asks.

Privacy Protection in Data Auction Models: An efficient auction scheme tends to induce the bidders to truthfully submit bidding profiles to ensure fairness, as well as achieving social welfare maximization, which is the property of strategy-proofness. In addition, as virtual commodities, data can be transacted only through the Internet during the auction process. Thus, a bidder's truthful behavior will put them at the risk of releasing private information. This private information is related to the bidder's preferences on the types of data, the bidder's active time, their economic situation, and even their geographic location. The release of such information would not only cause economic loss of the bidders, but also threaten the their personal safety. For instance, if the type of data that the user is interested is released, the sellers may increase their valuation, the bidders would suffer from these malicious biddings in the future to the effect that their profit would be damaged. In addition, if a bidder's active time or their location was released, their personal safety would be seriously threatened by competitors or other malicious actors. In summary, privacy preservation remains a critical issue in big data auction markets. Unfortunately, few research efforts have been carried out in this area.

Nonetheless, a large and extensive body of work has focused on designing privacy preserving auction schemes in other types of auction markets, such as spectrum markets [132], [133], [134], mobile crowd sensing [135], [136], cloud computing markets [137], and Electrical Vehicle (EV) charging markets [138]. Generally speaking, privacy preserving methods in auction schemes can be divided into three aspects: anonymity [139], [104], [140], cryptosystems [112], and perturbation [141], which have the potential to be extended for the preservation of privacy in big data trading markets.

Specifically, anonymity provides an efficient method to protect bidder privacy from the public, such as with auction results. This method simply anonymizes sensitive parts of the public information. Nonetheless, the privacy will be released by attacks (the linkage attack [138], etc.) when applying the anonymity method. Cryptosystems are able to prevent adversaries from invading the auction system to obtain privacy information. One of most common methods in cryptosystems is the homomorphic encryption system, which adds an agent to the auction system to help with the auction process, and ensures that each part of the auction system cannot hold all of a bidder's private information [112]. Finally, the perturbation method, which includes differential privacy, can be applied when the adversary seeks to infer the bidders' profiles by comparing the auction results generated by several similar bids [104]. The differential privacy scheme adds random noise to the results of the auction, and ensures that the same bidders' profiles will not generate the same result of the auction. Therefore, the adversary cannot infer the exact bidders' profiles.

Third-party Auction Platforms: Based on the growth of data auctions, the data owners will find it hard to build their own auction platforms. Thus, the third-party auction platform is emerging as the primary contender in the data auction field. Security and truthfulness are especially important for third party auction platforms. There are only a few research efforts that focus on auction platform strategies. One privacy-preserving big data auction scheme using homomorphic encryption has been designed. Particularly, the auction platform was designed based on the concept of homomorphic encryption [142] to meet the needs of privacy preservation. In this work, the entire system consists of two entities that are independent of each other: The Auctioneer (AC) and the Intermediate Platform (IP). All sensitive bids are encrypted using a Paillier cryptosystem [143] assisted with a one-time pad. Under this structure, all bids are first received by the Intermediate Platform in the form of ciphertext encrypted using Paillier. These bids will be disguised with a pad before being sent to the Auctioneer. In addition, this design enables the target auction data to be accessible only by the winner of the auction. Finally, a digital signature feature of the Paillier cryptosystem is applied to ensure that the data has not been manipulated, either in transmission, or by a compromised Auctioneer or Platform. This design addresses the issue of privacy protection for data auctions with untrusted third-party auctioneers. While the winner of the auction can be determined by using encrypted biddings, both the seller and the bidders do not need to worry about the leakage of sensitive information. The processes and algorithms are well designed with an overall time-complexity of $O(n \log n)$, which allows for large-scale deployment. Meanwhile, the structure is proven to be secure against different types of attacks that the participants are concerned about, including fake bids and the situation where the platform is compromised.

To summarize, in this section we have first discussed the main purpose of big data trading from the perspective of data owners and data consumers. We have then outlined the issues of big data trading with respect to multiple owner data trading, trading fairness and truthfulness, privacy protection, and third-party trading platform. Additionally, we have reviewed the big data market platforms and data auction models comprehensively. Nonetheless, designing effective trading platforms and auction models for big data trading remains a challenging issue. Further research efforts are needed to support big

data trading, including design of secure third-party trading platforms, creation of effective auction models that ensure truthful trading among multiple data owners and consumers, and development of privacy protection mechanisms that ensure sensitive information cannot be inferred by the adversary, among others.

VI. DATA PROTECTION

With the digitization of traditional media growing daily, content is increasingly stored in digital volumes instead of in traditional goods or analog contents (films, newspaper, design drawings, customers information, office documents, etc.). In other words, commodities are changing from practical items to virtual items. In this way, the contents are easily distributed and copied. Thus, data protection emerges as the key provision for securing the ownership of the data. Data pricing, data trading, and data protection comprise a three-dimensional closed loops, which impact each other. To reach the maximum profit for data owners and maximum value of data, data protection is an indispensable part in the loop. In the following, we discuss the last important stage in the big data lifecycle, which is data protection.

A. Digital Rights Management

The digital rights management (DRM) has been established for the prevention of digital content from being deliberately copied, shared, and stolen, acting more importantly as a guideline in the development of digital copyright protection. In early 2001, W3C established the first DRM group as the standard organization to participate in digital rights management worldwide [144]. There have been various solutions to realize DRM, including XrML Rights Expression Language [145], Microsoft DRM [146], Apple HLS DRM [147], Adobe Flash access DRM [148], RealNetworks Helix DRM [149], and the OMA DRM Specification [150].

All of these DRM solutions require five key components [151]: (i) *Security.* It focuses on encryption of the content and the creation of hashes, watermarks, and digital signatures for digital content. (ii) *Access control.* It is responsible for identity and access management, and the provision of credentials for users who need to access the protected digital contents. In addition, this component monitors the behaviors of authorized users, and sets different access rights for different users. (iii) *Usage control.* It monitors the usage for each authorized user, and records the usage as history. (iv) *License management.* It releases license (keys, XrML files, authentication code) to authorized users, and controls and checks the lifetime (validity period) for a license. (v) *Payment management.* This component works with usage control, and calculates the fee that users need to pay. This is the main goal for digital business.

We would like to use Microsoft DRM as an example to explain how DRM works. As shown in Figure 7, an anonymous user tries to access to the content server to play or download some content, which is protected by a DRM server. He/she sends the request to the individualization server first. The server then checks the individualization application on client side devices. If there is an application running, the application will send the license requirement to the DRM server. The individualization application is a client-side DRM software, called individualized black box (IBX). Without this software, the DRM server cannot release the license for decrypting the content. To meet the requirement of IBX, the DRM server releases the encrypted license. In addition, the IBX protects the sensitive information when the user tries to decrypt the license, this kind of process is called individualization. After releasing the license, the DRM server checks the user status. If this is the first time accessing the server, the DRM will ask the user to join the domain. Different domains distinguish the contents and rights of a particular user. This is how DRM realizes access control. Finally, the user is allowed to access the content server, which sends back the content [152].
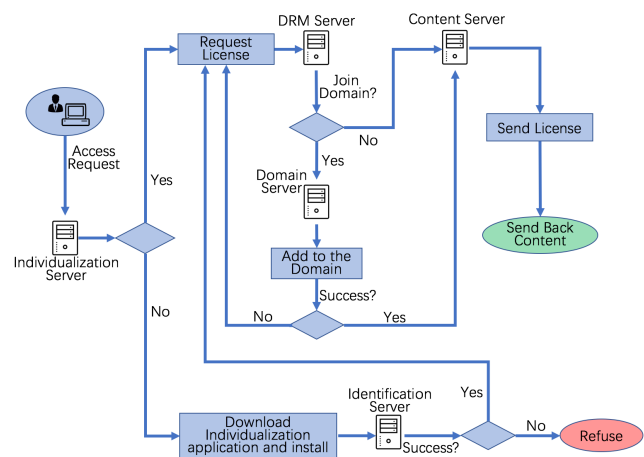

Fig. 7: Microsoft Digital Rights Management Workflow

B. Digital Rights Management Models

Based on different digital contents, we categorize the DRM models into the following three groups: (i) software-based DRM, (ii) multimedia-based DRM, and (iii) unstructured data-based DRM.

1) Software-Based DRM: The most common DRM is software-based DRM, since software is the most widely used application on a computer. Belonging to digital commodities, software is easy to copy and re-produce with zero cost. Thus, software development companies usually design mechanisms to protect copyright and prevent piracy incursion. An optimal DRM mechanism can record installation times and PC identification information, and support multiple installations and hosts.

There are two main approaches that are involved, online authentication and offline authentication. For online authentication, the software checks the Internet connection first while the user starts the installation process. If there is an Internet connection, the software sends an authentication request to the DRM server, as in the common DRM strategy we discussed above. Otherwise, the installation will be stopped when there is no Internet connection, or will only install a software demo. Offline authentication is more critical than online authentication. Without offline DRM support, the local

license file is weak and easy to decrypt. A number of research efforts have focused on offline authentication. For example, the authors in [153] proposed a cost function to measure the complexity of decryption. If the decryption cost is larger than the price that is determined by this function, the software is secure. The authors in [154] proposed a structure to increase the complexity of decrypting the license file. The model uses code injection and Software Copyrights Protection (SCP) technology with both static and dynamic code to encrypt the license file to protect the original software. The protection Dynamic-Link Library (DLL) layer was introduced between the software layer and the license layer (license file, hardware token management file, library file, etc.). Thus, the user cannot access the authentication information directly.

2) Multimedia-Based DRM: Multimedia is the most important component of digital commodities. More than 80 % of Internet traffic is dedicated to video content [155]. Thus, a big challenge is how to properly protect the copyrights for multimedia content. Generally speaking, encryption and watermarking technologies are used in this direction. The big difference between software and multimedia (video and audio) is online streaming. Online video and audio supports Real Time Protocol (RTP)/Real Time Streaming Protocol (RTSP) protocols in order to realize online streaming, and in some situations, needs to support group domain authentications (family members, enterprise users, etc.). Commonly, adversaries run malicious clients on hosts to interrupt and monitor the stream to analyze the encryption key. To approach this issue, the authors in [156] proposed a scheme using selective video decryption to ensure the security of the content, while reducing the encryption time. In addition, the selective decryption is a variation of efficient video encryption [157], and the proposed algorithm only operates on the sign bits of the transform parameters. Thus, it does not need additional space and the stream encrypted by the algorithm results in a H.264 bit stream. Meanwhile, the scheme pushes the encryption process into a secure environment by confining the access states. For instance, a user is either at the encryption state or the decrypting state, but not in both states, which is forbidden.

In addition, watermarking technology has been widely applied in video and audio DRM [158], [159], [160]. Embedding a watermark into video content requires the complete decoding of the video content. This is a critical issue, as this process requests a lot of computation resources and reduces the quality of the video [161], [162]. To avoid the increasing complexity of embedding a watermark by an increasing video bit-rate, a blind watermarking algorithm based on the H.264 codec standard was proposed by [163]. Notice that H.264 is the popular high-quality codec standard and is based on motion compensation. The H.264 standard uses few macroblocks to represent the frame, along with each macroblock luma and chroma (Cb and Cr). The watermark algorithm scans the macroblock and selects the optimal prediction model. Thus, according to the characteristics of H.264, the blind watermarking algorithm embeds a watermark right in the selected macroblock, preventing collusion attacks and maintaining the quality of video during the decoding process.

Another watermarking algorithm was proposed in [164],

which applies to Depth Image Based Rendering (DIBR)-based 3D video content. While traditional watermark systems will either damage the 3D video, cause irreversible deformation, or are easy to attack, the proposed synthesized-unseen watermarking algorithm overcomes these issues. The designed algorithm embeds the watermark into depth maps based on the pseudo-3D-discrete cosine transform (3D-DCT) and quantization index modulation (QIM), and it increases the robustness of the watermark and avoids the damaging of video content. It is worth noting that images are also considered multimedia content, similar to video and audio content, and watermarking technologies are the most common approach to protect copyrights. For image-based watermarking systems, usually Discrete Wavelet Transform (DWT), Least Significant Bit (LSB), and Discrete Cosine Transform (DCT) algorithms are used to embed the watermark into a secure key. Furthermore, multiple watermarks are can be imbedded into one image. In addition, watermarking schemes have been used to trace anonymous Internet malicious traffic flows for identifying the malicious sources for the purpose of forensics [140], [165], [166].

3) Unstructured Data-Based DRM: Unstructured data, such as Microsoft Word documents, PDF documents, various databases, source code, and others, are digitalized data, which can be conveniently spread and stored. Nonetheless, it is fragile, and it is very difficult to prevent deliberate replication and tampering with unstructured data. In addition, unstructured data usually has very high commercial value and contains sensitive information, the leaking of which will lead to critical business loss for the data owners. Thus, unstructured data protection is an active topic today, otherwise known as Data Leakage Protection (DLP) [167]. Unstructured data DRM is completely different than other types of DRM, because the data is easy to manipulate and damage. Thus, encryption, as the most secure method, is usually involved to protect unstructured data. Nonetheless, with data size constantly expanding, the encrypting process will continue to cost more and more. For instance, the authors in [168] proposed a probabilistic data structure (Bloom Filter)-based protection scheme. This scheme records the status into a Matrix Bloom Filter with a positive or genitive tag. The scheme includes an analyzer, which analyzes and scans the content. In comparison with encryption schemes, this scheme demonstrates better performance.

To summarize, in this section we have reviewed three models of digital rights management, and have discussed relevant existing approaches for each model. The different types of digital content management, namely software-based DRM, multimedia-based DRM, and unstructured data-based DRM, have been well explored. As we can see, digital management techniques serve as the key method to protecting big data from being stolen and copied. Nonetheless, with the rapid increase of digital content and the trade properties of big data, the feasibility of existing data protection schemes and more advanced techniques should be further investigated.

VII. Conclusion

In this paper, we have addressed the issue of big data trading. To be specific, we first reviewed existing research

relevant to big data, and identified the big data lifecycle for data trading, including data collection, data analytics, data pricing, data trading, and data protection. Then, we reviewed existing works related to big data pricing. With regard to data pricing, we clarified its importance, categorized different market structures, data pricing strategies, and data pricing models, and then listed the advantages and limitations of each category. For the data trading process, we outlined key issues associated with data trading and their possible solutions. We further investigated auction strategies and detailed different schemes, trading platforms, and related issues. Finally, we investigated data protection as the last stage of the big data lifecycle. We categorized existing copyright protection schemes and outlined the challenges of big data copyright protection. Notice that the main purpose of this survey is to provide a clear and deep understanding of big data trading. We outlined the breadth of topics related to data pricing, data trading and data protection, and highlighted areas that remain unresolved, in an effort to further promote the research and development of big data.

REFERENCES

[1] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: architecture, enabling technologies, security and privacy, and applications," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1125–1142, Oct 2017.

[2] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016.

[3] J. A. Stankovic, "Research directions for the Internet of Things," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 3–9, Feb 2014.

[4] J. Wu and W. Zhao, "Design and realization of WInternet: From Net of Things to Internet of Things," *ACM Trans. Cyber-Phys. Syst.*, vol. 1, no. 1, pp. 2:1–2:12, Nov. 2016. [Online]. Available: http://doi.acm.org/10.1145/2872332

[5] X. Yang, X. Ren, J. Lin, and W. Yu, "On binary decomposition based privacy-preserving aggregation schemes in real-time monitoring systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 10, pp. 2967–2983, Oct 2016.

[6] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," *IEEE Internet of Things journal*, vol. 1, no. 1, pp. 22–32, 2014.

[7] S. Mallapuram, N. Ngwum, F. Yuan, C. Lu, and W. Yu, "Smart city: The state of the art, datasets, and evaluation platforms," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, May 2017, pp. 447–452.

[8] F. Chen, T. Xiang, X. Fu, and W. Yu, "User differentiated verifiable file search on the cloud," *IEEE Transactions on Services Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[9] X. W. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[10] Z. Chen, G. Xu, V. Mahalingam, L. Ge, J. Nguyen, W. Yu, and C. Lu, "A cloud computing based network monitoring and threat detection system for critical infrastructures," *Big Data Research*, vol. 3, pp. 10–23, 2016.

[11] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang, "A survey on the edge computing for the Internet of Things," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2017.

[12] W. Yu, G. Xu, Z. Chen, and P. Moulema, "A cloud computing based architecture for cyber security situation awareness," in *2013 IEEE Conference on Communications and Network Security (CNS)*, Oct 2013, pp. 488–492.

[13] N. D. Nguyen, T. Nguyen, and S. Nahavandi, "System design perspective for human-level agents using deep reinforcement learning: A survey," *IEEE Access*, vol. 5, pp. 27 091–27 102, 2017.

[14] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sept 2009.

[15] P. Lewis, "An "all you can eat" price is clogging internet access," *New York Times. December*, vol. 17, p. A1, 1996.

[16] Y. Shen, B. Guo, Y. Shen, X. Duan, X. Dong, and H. Zhang, "A pricing model for big personal data," *Tsinghua Science and Technology*, vol. 21, no. 5, pp. 482–490, 2016.

[17] M. Khan, X. Wu, X. Xu, and W. Dou, "Big data challenges and opportunities in the hype of industry 4.0," pp. 1–6, May 2017.

[18] X. Cao, Y. Chen, and K. R. Liu, "Data trading with multiple owners, collectors, and users: An iterative auction mechanism," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 2, pp. 268–281, 2017.

[19] G.-H. Kim, S. Trimi, and J.-H. Chung, "Big-data applications in the government sector," *Communications of the ACM*, vol. 57, no. 3, pp. 78–85, 2014.

[20] G. Xu, W. Yu, D. Griffith, N. Golmie, and P. Moulema, "Toward integrating distributed energy resources and storage devices in smart grid," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 192–204, Feb 2017.

[21] J. Lin, W. Yu, and X. Yang, "Towards multistep electricity prices in smart grid electricity markets," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 1, pp. 286–302, Jan 2016.

[22] J. Lin, W. Yu, X. Yang, Q. Yang, X. Fu, and W. Zhao, "A real-time en-route route guidance decision scheme for transportation-based cyberphysical systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2551–2566, March 2017.

[23] , "Visualizing Facebook?s Media Storage: How Big Is 100 Petabytes?" 2012, https://techcrunch.com/2012/02/02/visualizing-facebooks-media-store-how-big-is-100-petabytes/.

[24] ——, "Volume, velocity, and variety: Understanding the three V's of big data," 2016, http://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/.

[25] Xiaomeng Su et al. [Particle Data Group], "Introduction to Big Data," 2012, https://www.ntnu.no/iie/fag/big/.

[26] B. Franks, "Taming the big data tidal wave," *Finding opportunities in huge data streams*, 2012.

[27] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health information science and systems*, vol. 2, no. 1, p. 3, 2014.

[28] , "Data, data everywhere," 2010, http://www.economist.com/node/15557443.

[29] E. Barbierato, M. Gribaudo, and M. Iacono, "Performance evaluation of NoSQL big-data applications using multi-formalism models," *Future Generation Computer Systems*, vol. 37, pp. 345–353, 2014.

[30] M. Iacono, E. Barbierato, and M. Gribaudo, "The SIMTHESys multiformalism modeling framework," *Computers & Mathematics with Applications*, vol. 64, no. 12, pp. 3828–3839, 2012.

[31] E. Barbierato, M. Gribaudo, and M. Iacono, "Defining formalisms for performance evaluation with SIMTHESys," *Electronic Notes in Theoretical Computer Science*, vol. 275, pp. 37–51, 2011.

[32] M. Iacono and M. Gribaudo, "Element based semantics in multi formalism performance models," in *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2010 IEEE International Symposium on*. IEEE, 2010, pp. 413–416.

[33] J. De Lara and H. Vangheluwe, "AToMˆ3: A tool for multi-formalism and meta-modelling," in *FASE*, vol. 2. Springer, 2002, pp. 174–188.

[34] V. Vittorini, M. Iacono, N. Mazzocca, and G. Franceschinis, "The OsMoSys approach to multi-formalism modeling of systems," *Software & Systems Modeling*, vol. 3, no. 1, pp. 68–81, 2004.

[35] G. Clark, T. Courtney, D. Daly, D. Deavours, S. Derisavi, J. M. Doyle, W. H. Sanders, and P. Webster, "The Mobius modeling tool," in *Petri Nets and Performance Models, 2001. Proceedings. 9th International Workshop on*. IEEE, 2001, pp. 241–250.

[36] , "Big Data is powerful on its own. So is artificial intelligence. What happens when the two are merged?" 2016, http://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale/.

[37] ——, "Why do Machine Learning on Big Data?" 2016, http://www.skytree.net/machine-learning/why-do-machine-learning-big-data/.

[38] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[39] J. S. Ward and A. Barker, "Undefined by data: a survey of big data definitions," *arXiv preprint arXiv:1309.5821*, 2013.

[40] S. F. Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *International Journal of Production Economics*, vol. 165, pp. 234–246, 2015.

[41] T. H. Davenport, "Competing on analytics," *harvard business review*, vol. 84, no. 1, p. 98, 2006.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2806881, IEEE Access

19

[42] A. McAfee, E. Brynjolfsson, T. H. Davenport et al., "Big data: the management revolution," *Harvard business review*, vol. 90, no. 10, pp. 60–68, 2012.

[43] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[44] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: The real-world use of big data," *IBM Global Business Services*, pp. 1–20, 2012.

[45] P. M. Hartmann, M. Zaki, N. Feldmann, and A. Neely, "Big data for big business? A taxonomy of data-driven business models used by start-up firms," *A Taxonomy of Data-Driven Business Models Used by Start-Up Firms (March 27, 2014)*, 2014.

[46] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[47] P. Russom et al., "Big data analytics," *TDWI best practices report, fourth quarter*, vol. 19, p. 40, 2011.

[48] S. Sagiroglu and D. Sinanc, "Big data: A review," in *Collaboration Technologies and Systems (CTS), 2013 International Conference on*. IEEE, 2013, pp. 42–47.

[49] M. Mussa and S. Rosen, "Monopoly and product quality," *Journal of Economic theory*, vol. 18, no. 2, pp. 301–317, 1978.

[50] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, "Pricing approaches for data markets," in *International Workshop on Business Intelligence for the Real-Time Enterprise*. Springer, 2012, pp. 129–144.

[51] D. Dash, G. Gratsias, A. Ailamaki et al., "Predicting cost amortization for query services," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM, 2011, pp. 325–336.

[52] A. Kushal, S. Moorthy, and V. Kumar, "Pricing for data markets," technical report, Tech. Rep., 2012.

[53] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[54] S. Mallapuram, W. Yu, P. Moulema, D. W. Griffith, N. T. Golmie, and F. Liang, "An integrated simulation study on reliable and effective distributed energy resources in smart grid," in *Proccedings of 2017 ACM International Conference on Research in Adaptive and Convergent Systems (RACS)*, 2017.

[55] C. Perera, "Sensing as a Service (S2aaS): Buying and selling IoT data," *arXiv preprint arXiv:1702.02380*, 2017.

[56] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Data collection and wireless communication in Internet of Things (IoT) using economic analysis and pricing models: a survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2546–2590, 2016.

[57] A. Roncoroni, "Commodity price models," *Encyclopedia of Quantitative Finance*, 2010.

[58] T. T. Nagle, J. Hogan, and J. Zale, *The Strategy and Tactics of Pricing: New International Edition*. Routledge, 2016.

[59] E. F. Fama and K. R. French, "Commodity futures prices: Some evidence on forecast power, premiums, and the theory of storage," in *The World Scientific Handbook of Futures Markets*. World Scientific, 2016, pp. 79–102.

[60] R. Harmon, H. Demirkan, B. Hefley, and N. Auseklis, "Pricing strategies for information technology services: A value-based approach," in *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 2009, pp. 1–10.

[61] T. F. Bresnahan, "The oligopoly solution concept is identified," *Economics Letters*, vol. 10, no. 1-2, pp. 87–92, 1982.

[62] A. K. Sen, "Rational fools: A critique of the behavioral foundations of economic theory," *Philosophy & Public Affairs*, pp. 317–344, 1977.

[63] J. Nash, "Non-cooperative games," *Annals of mathematics*, pp. 286–295, 1951.

[64] H. Yaïche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for bandwidth allocation and pricing in broadband networks," *IEEE/ACM Transactions On Networking*, vol. 8, no. 5, pp. 667–678, 2000.

[65] J. W. Friedman, "A non-cooperative equilibrium for supergames," *The Review of Economic Studies*, vol. 38, no. 1, pp. 1–12, 1971.

[66] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A stackelberg game approach," *IEEE Journal on Selected areas in Communications*, vol. 30, no. 3, pp. 538–549, 2012.

[67] S. Haddadi and A. Ghasemi, "Pricing-based stackelberg game for spectrum trading in self-organised heterogeneous networks," *IET Communications*, vol. 10, no. 11, pp. 1374–1383, 2016.

[68] X. Lv, R. Zhang, and J. Yue, "Competition and cooperation between participants of the Internet of Things industry value chain," *Advances in Information Sciences and Service Sciences*, vol. 4, no. 11, pp. 406–412, 2012.

[69] D. B. Rawat, S. Shetty, and C. Xin, "Stackelberg-game-based dynamic spectrum access in heterogeneous wireless systems," *IEEE Systems Journal*, vol. 10, no. 4, pp. 1494–1504, 2016.

[70] A. Danak, A. R. Kian, and B. Moshiri, "Inner supervision in multisensor data fusion using the concepts of stackelberg games," in *Multisensor Fusion and Integration for Intelligent Systems, 2006 IEEE International Conference on*. IEEE, 2006, pp. 65–70.

[71] Y. Mao, T. Cheng, H. Zhao, and N. Shen, "A strategic bargaining game for a spectrum sharing scheme in cognitive radio-based heterogeneous wireless sensor networks," *Sensors*, vol. 17, no. 12, p. 2737, 2017.

[72] G. Berz, *Game Theory Bargaining and Auction Strategies: Practical Examples from Internet Auctions to Investment Banking*. Springer, 2016.

[73] S. M. Azimi, M. H. Manshaei, and F. Hendessi, "Cooperative primary–secondary dynamic spectrum leasing game via decentralized bargaining," *Wireless Networks*, vol. 22, no. 3, pp. 755–764, 2016.

[74] C. Yang, J. Li, A. Anpalagan, and M. Guizani, "Joint power coordination for spectral-and-energy efficiency in heterogeneous small cell networks: A bargaining game-theoretic perspective," *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 1364–1376, 2016.

[75] Z. Zheng, Y. Peng, F. Wu, S. Tang, and G. Chen, "An online pricing mechanism for mobile crowdsensing data markets," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 2017, p. 26.

[76] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-based data pricing," *Journal of the ACM (JACM)*, vol. 62, no. 5, p. 43, 2015.

[77] D. Niyato, D. T. Hoang, N. C. Luong, P. Wang, D. I. Kim, and Z. Han, "Smart data pricing models for the Internet of Things: A bundling strategy approach," *IEEE Network*, vol. 30, no. 2, pp. 18–25, 2016.

[78] R. K. Chellappa and A. Mehra, "Versioning 2.0: A product line and pricing model for information goods under usage constraints and with R&D costs," in *Conf. Inform. Systems Technol., Minneapolis, MN, Oct*, 2013, pp. 5–6.

[79] D. Niyato, M. A. Alsheikh, P. Wang, D. I. Kim, and Z. Han, "Market model and optimal pricing scheme of big data and Internet of Things (IoT)," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[80] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart Data pricing (SDP): Economic solutions to network congestion," *Recent Advances in Networking*, vol. 1, pp. 221–274, 2013.

[81] Y.-C. Tsai, Y.-D. Cheng, C.-W. Wu, Y.-T. Lai, W.-H. Hu, J.-Y. Jeng, and Y.-C. Tseng, "Time-dependent smart data pricing based on machine learning," in *Canadian Conference on Artificial Intelligence*. Springer, 2017, pp. 103–108.

[82] H. Yu and M. Zhang, "Data pricing strategy based on data quality," *Computers & Industrial Engineering*, vol. 112, pp. 1–10, 2017.

[83] S. A. Fricker and Y. V. Maksimov, "Pricing of data products in data marketplaces," in *International Conference of Software Business*. Springer, 2017, pp. 49–66.

[84] W.-T. Wang, Y.-S. Wang, and E.-R. Liu, "The stickiness intention of group-buying websites: The integration of the commitment–trust theory and e-commerce success model," *Information & Management*, vol. 53, no. 5, pp. 625–642, 2016.

[85] P. Groves, B. Kayyali, D. Knott, and S. V. Kuiken, "The 'big data' revolution in healthcare: Accelerating value and innovation," 2016.

[86] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen, "Trading data in good faith: Integrating truthfulness and privacy preservation in data markets," in *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*. IEEE, 2017, pp. 223–226.

[87] X. Cao, Y. Chen, and K. R. Liu, "An iterative auction mechanism for data trading," Technical Report, Tech. Rep., 2016.

[88] Y. Jiao, P. Wang, D. Niyato, M. A. Alsheikh, and S. Feng, "Profit maximization auction and data management in big data markets," in *Wireless Communications and Networking Conference (WCNC), 2017 IEEE*. IEEE, 2017, pp. 1–6.

[89] H. M. Getson, S. Vallie, A. Peterson, and K. Rodriguez, "Systems and methods for allocating capital to trading strategies for big data trading in financial markets," Mar. 9 2015, uS Patent App. 14/642,577.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2806881, IEEE Access

20

[90] W. Zhang, S. Xiao, Y. Lin, T. Zhou, and S. Zhou, "Secure ranked multi-keyword search for multiple data owners in cloud computing," in *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on*. IEEE, 2014, pp. 276–286.

[91] W. Zhang, Y. Lin, S. Xiao, J. Wu, and S. Zhou, "Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing," *IEEE Transactions on Computers*, vol. 65, no. 5, pp. 1566–1577, 2016.

[92] X. Ding, H. Wang, D. Zhang, J. Li, and H. Gao, "A fair data market system with data quality evaluation and repairing recommendation," in *Asia-Pacific Web Conference*. Springer, 2015, pp. 855–858.

[93] J. Yu, M. H. Cheung, and J. Huang, "Economics of mobile data trading market," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2017 15th International Symposium on*. IEEE, 2017, pp. 1–8.

[94] S. Delgado-Segura, C. Pérez-Sola, G. Navarro-Arribas, and J. Herrera-Joancomartí, "A fair protocol for data trading based on bitcoin transactions," *Future Generation Computer Systems*, 2017.

[95] H. Susanto, H. Zhang, S.-Y. Ho, and B. Liu, "Effective mobile data trading in secondary ad-hoc market with heterogeneous and dynamic environment," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 645–655.

[96] X. Cao, "User behavior analysis and data trading in multi-agent systems," Ph.D. dissertation, University of Maryland, College Park, 2017.

[97] X. Wang, L. Duan, and R. Zhang, "User-initiated data plan trading via a personal hotspot market," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7885–7898, 2016.

[98] P. Paillier *et al.*, "Public-key cryptosystems based on composite degree residuosity classes," in *Eurocrypt*, vol. 99. Springer, 1999, pp. 223–238.

[99] C. Gentry, A. Sahai, and B. Waters, "Homomorphic encryption from learning with errors: Conceptually-simpler, asymptotically-faster, attribute-based," in *Advances in Cryptology–CRYPTO 2013*. Springer, 2013, pp. 75–92.

[100] C. Gentry *et al.*, "Fully homomorphic encryption using ideal lattices." in *STOC*, vol. 9, no. 2009, 2009, pp. 169–178.

[101] A. López-Alt, E. Tromer, and V. Vaikuntanathan, "On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012, pp. 1219–1234.

[102] W. Gao, W. Yu, F. Liang, W. G. Hatcher, and C. Lu, "Privacy-preserving big data auction using homomorphic encryption," in *Technical Report and Submitted to IEEE International Conference on Communication (ICC) 2018*, 2018.

[103] C. Perera, R. Ranjan, and L. Wang, "End-to-end privacy for open big data markets," *IEEE Cloud Computing*, vol. 2, no. 4, pp. 44–53, 2015.

[104] X. Yang, T. Wang, X. Ren, and W. Yu, "Survey on improving data utility in differentially private sequential data publishing," *IEEE Transactions on Big Data*, vol. PP, no. 99, pp. 1–1, 2017.

[105] X. Yang, X. Ren, J. Lin, and W. Yu, "On binary decomposition based privacy-preserving aggregation schemes in real-time monitoring systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 10, pp. 2967–2983, Oct 2016.

[106] S. Gürses, C. Troncoso, and C. Diaz, "Engineering privacy by design," 2011.

[107] D. Goldschlag, M. Reed, and P. Syverson, "Onion routing," *Communications of the ACM*, vol. 42, no. 2, pp. 39–41, 1999.

[108] M. Stillger, G. M. Lohman, V. Markl, and M. Kandil, "LEO-DB2's learning optimizer," in *VLDB*, vol. 1, 2001, pp. 19–28.

[109] Y. Li, E. Lo, M. L. Yiu, and W. Xu, "Query optimization over cloud data market." in *EDBT*, 2015, pp. 229–240.

[110] Z. Liu and H. Hacigümüs, "Online optimization and fair costing for dynamic data sharing in a cloud data market," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. ACM, 2014, pp. 1359–1370.

[111] B. Lorenzo and F. J. Gonzalez-Castano, "A matching game for data trading in operator-supervised user-provided networks," in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–7.

[112] D. Li, Q. Yang, W. Yu, D. An, X. Yang, and W. Zhao, "A strategy-proof privacy-preserving double auction mechanism for electrical vehicles demand response in microgrids," in *Proc. of 36th IEEE International Performance Computing and Communications Conference (IPCCC)*, 2017.

[113] R. P. McAfee, "A dominant strategy double auction," *Journal of economic Theory*, vol. 56, no. 2, pp. 434–450, 1992.

[114] D. An, Q. Yang, W. Yu, X. Yang, X. Fu, and W. Zhao, "Sto2auc: A stochastic optimal bidding strategy for microgrids," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2260–2274, Dec 2017.

[115] G. Iosifidis, L. Gao, J. Huang, and L. Tassiulas, "An iterative double auction for mobile data offloading," in *International Symposium on Modeling & Optimization in Mobile, Ad Hoc & Wireless Networks*, 2013, pp. 154–161.

[116] D. An, Q. Yang, W. Yu, X. Yang, X. Fu, and W. Zhao, "SODA: strategy-proof online double auction scheme for multimicrogrids bidding," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, no. 99, pp. 1–14, 2017.

[117] J. Yang, *The Efficiency of an Artificial Double Auction Stock Market with Neural Learning Agents*. Physica-Verlag HD, 2002.

[118] D. An, Q. Yang, W. Yu, D. Li, and Y. Zhang, "Towards truthful auction for big data trading," in *Proc. of 36th IEEE International Performance Computing and Communications Conference (IPCCC)*, 2017.

[119] Q. Wu, M. C. Zhou, Q. Zhu, and Y. Xia, "Vcg auction-based dynamic pricing for multigranularity service composition," *IEEE Transactions on Automation Science & Engineering*, vol. PP, no. 99, pp. 1–10, 2017.

[120] L. Li, X. Liu, and Z. Hu, *A Bid Evaluation Method for Multi-attribute Online Reverse Auction*. Springer Singapore, 2017.

[121] X. Zhou and H. Li, "Buying on margin and short selling in an artificial double auction market," *Computational Economics*, no. 8, pp. 1–17, 2017.

[122] T. Zhou, B. Chen, C. Zhu, and X. Zhai, "TPAHS: A truthful and profit maximizing double auction for heterogeneous spectrums," in *Trustcom/bigdatase/i?spa*, 2017, pp. 27–33.

[123] D. Li, Q. Yang, W. Yu, D. An, and X. Yang, "Towards double auction for assisting electric vehicles demand response in smart grid," in *Proc. of 13th IEEE Conference on Automation Science and Engineering (CASE)*, 2017.

[124] ——, "Towards double auction for assisting electric vehicles demand response in smart grid," in *Proc. of 13th IEEE Conference on Automation Science and Engineering (CASE)*, 2017.

[125] H. Susanto, H. Zhang, S. Ho, and B. Liu, "Effective mobile data trading in secondary ad-hoc market with heterogeneous and dynamic environment," in *ICDCS*. IEEE Computer Society, 2017, pp. 645–655.

[126] O. Kirchkamp, E. Poen, and J. P. Rei?, "Outside options: Another reason to choose the first-price auction," *European Economic Review*, vol. 53, no. 2, pp. 153–169, 2009.

[127] B. Edelman and M. Schwarz, "Internet advertising and the generalized second price auction: Selling billions of dollars worth of keywords," 2007, pp. 242–259.

[128] Nisan, Noam, Ronen, and Amir, "Computationally feasible VCG mechanisms," *Journal of Artificial Intelligence Research*, vol. 29, no. 6, pp. 242–252, 2000.

[129] L. M. Ausubel and P. Milgrom, "The lovely but lonely vickrey auction," in *Combinatorial Auctions, Chapter*, 2006.

[130] S. D. Vries and R. Vohra, "Combinatorial auctions: A survey," *Discussion Papers*, vol. 15, no. 3, pp. 284–309, 2000.

[131] P. Cramton, Y. Shoham, and R. Steinberg, *Combinatorial Auctions*. MIT Press,, 2006.

[132] Q. Huang, Y. Gui, F. Wu, and G. Chen, "A general privacy-preserving auction mechanism for secondary spectrum markets," *IEEE/ACM Transactions on Networking*, pp. 1–1, 2015.

[133] X. Jin and Y. Zhang, "Privacy-preserving crowdsourced spectrum sensing," in *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*.

[134] R. Zhu and K. G. Shin, "Differentially private and strategy-proof spectrum auction with approximate revenue maximization," in *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*, 2015.

[135] Y. Zhang, Y. Mao, and S. Zhong, *Joint Differentially Private GaleC-Shapley Mechanisms for Location Privacy Protection in Mobile Traffic Offloading Systems*. IEEE Press, 2016.

[136] H. Jin, L. Su, B. Ding, K. Nahrstedt, and N. Borisov, "Enabling privacy-preserving incentives for mobile crowd sensing systems," in *IEEE International Conference on Distributed Computing Systems*, 2016, pp. 344–353.

[137] M. Tebaa and S. E. Hajji, "Secure cloud computing through homomorphic encryption," *Eprint Arxiv*, 2014.

[138] Q. Xiang, L. Kong, X. Liu, J. Xu, and W. Wang, "Auc2Reserve: a differentially private auction for electric vehicle fast charging reservation (invited paper)," in *Proc. of IEEE International Conference on Embedded and Real-Time Computing Systems and Applications*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2806881, IEEE Access

21

[139] S. Li, X. Li, M. X. He, and S. K. Zeng, "Sealed-BID electronic auction without the third party," in *International Computer Conference on Wavelet Active Media Technology and Information Processing*, 2014, pp. 336–339.

[140] W. Yu, X. Fu, S. Graham, D. Xuan, and W. Zhao, "DSSS-based flow marking technique for invisible traceback," in *2007 IEEE Symposium on Security and Privacy (SP '07)*, May 2007, pp. 18–32.

[141] A. Pingley, W. Yu, N. Zhang, X. Fu, and W. Zhao, "CAP: a context-aware privacy protection system for location-based services," in *2009 29th IEEE International Conference on Distributed Computing Systems*, June 2009, pp. 49–57.

[142] Z. Brakerski and V. Vaikuntanathan, "Efficient fully homomorphic encryption from (standard) LWE," *SIAM Journal on Computing*, vol. 43, no. 2, pp. 831–871, 2014.

[143] N. Fazio, R. Gennaro, T. Jafarikhah, and W. E. Skeith, "Homomorphic secret sharing from paillier encryption," in *International Conference on Provable Security*. Springer, 2017, pp. 381–399.

[144] J. Van Tassel, *Digital rights management: Protecting and monetizing content*. Focal Press, 2016.

[145] A. Hohmann, "Rights expression languages in libraries: Development of an application profile," 2016.

[146] C. Skipper, E. D. Lewis, D. Konetski, C. Burchett, R. W. Schuckle, J. M. Burke, W. W. Robbins, and C. E. Gates, "Systems and methods for providing secure data," Aug. 5 2015, uS Patent App. 14/819,322.

[147] C. D'Orazio and K.-K. R. Choo, "An adversary model to evaluate DRM protection of video contents on iOS devices," *Computers & Security*, vol. 56, pp. 94–110, 2016.

[148] V. Swaminathan and K. Y. Kishore, "Methods and apparatus for integrating digital rights management (DRM) systems with native HTTP live streaming," Aug. 12 2014, uS Patent 8,806,193.

[149] T. Gaber, "Digital rights management: Open issues to support E-commerce," *E-Marketing in Developed and Developing Countries, IGI Global*, pp. 69–87, 2013.

[150] J. Choi, W. Aiken, J. Ryoo, and H. Kim, "Bypassing the integrity checking of rights objects in OMA DRM: a case study with the MelOn music service," in *Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication*. ACM, 2016, p. 62.

[151] M. R. Lambert, "Digital rights management," Feb. 19 2013, uS Patent 8,380,849.

[152] Microsoft, "Digital Rights Management (DRM)," 2016, https://msdn.microsoft.com/es-es/library/cc838192(v=vs.95).aspx.

[153] K. Reavis Conner and R. P. Rumelt, "Software piracy: An analysis of protection strategies," *Management Science*, vol. 37, no. 2, pp. 125–139, 1991.

[154] H. BARAPATRE, P. NIMJE, and A. NIMBALKAR, "Software piracy protection," *Imperial Journal of Interdisciplinary Research*, vol. 3, no. 4, 2017.

[155] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2016?2021," 2017, https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html.

[156] A. David and N. Zaidenberg, "Maintaining streaming video DRM," in *Proceedings of The International Conference on Cloud Security Management ICCSM-2014*, 2014, p. 36.

[157] C. Shi, S.-Y. Wang, and B. Bhargava, "Mpeg video encryption in real-time using secret key cryptography," in *in Proc. Int. Conf. Parallel and Distributed Processing Techniques and Applications*. Citeseer, 1999.

[158] F. Arab, S. M. Abdullah, S. Z. M. Hashim, A. A. Manaf, and M. Zamani, "A robust video watermarking technique for the tamper detection of surveillance systems," *Multimedia Tools and Applications*, vol. 75, no. 18, pp. 10 855–10 885, 2016.

[159] J. Hao, X. Yao, J. Huang, Y. Qian, and J. Jagannathan, "Video content protection," Feb. 9 2016, uS Patent 9,258,584.

[160] K. Jain and U. Raju, "A digital video watermarking algorithm based on LSB and DCT," 2015.

[161] N. Azeem, I. Ahmad, S. R. Jan, M. Tahir, F. Ullah, and F. Khan, "A new robust video watermarking technique using H. 264/AAC Codec luma components based on DCT," *International Journal of Advance Research and Innovative Ideas in Education, Online ISSN-2395-4396*, 2016.

[162] S. Wang, D. Zheng, J. Zhao, W. J. Tam, and F. Speranza, "Adaptive watermarking and tree structure based image quality estimation," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 311–325, 2014.

[163] Z. Ma, J. Huang, M. Jiang, and X. Niu, "A video watermarking DRM method based on h. 264 compressed domain with low bit-rate increasement," *Chinese Journal of Electronics*, vol. 25, no. 4, pp. 641–647, 2016.

[164] X. Liu, F. Li, J. Du, Y. Guan, Y. Zhu, and B. Zou, "A robust and synthesized-unseen watermarking for the DRM of DIBR-based 3D video," *Neurocomputing*, vol. 222, pp. 155–169, 2017.

[165] Z. Ling, X. Fu, W. Jia, W. Yu, D. Xuan, and J. Luo, "Novel packet size-based covert channel attacks against anonymizer," *IEEE Transactions on Computers*, vol. 62, no. 12, pp. 2411–2426, Dec 2013.

[166] Z. Ling, J. Luo, W. Yu, X. Fu, D. Xuan, and W. Jia, "A new cell-counting-based attack against tor," *IEEE/ACM Transactions on Networking*, vol. 20, no. 4, pp. 1245–1261, Aug 2012.

[167] N. I. Readshaw, J. Ramanathan, and G. G. Bray, "Method and apparatus for associating data loss protection (dlp) policies with endpoints," Apr. 12 2016, uS Patent 9,311,495.

[168] L. Shi, S. Butakov, D. Lindskog, R. Ruhl, and E. Storozhenko, "Applicability of probablistic data structures for filtering tasks in data loss prevention systems," in *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*. IEEE, 2015, pp. 582–586.