Computers & Industrial Engineering 112 (2017) 1–10

Contents lists available at ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie

Data pricing strategy based on data quality ☆

Haifei Yu, Mengxiao Zhang *

School of Business Administration, Northeastern University, Shenyang 110819, PR China

ARTICLE INFO

Article history:
Received 31 December 2016
Received in revised form 2 June 2017
Accepted 7 August 2017
Available online 9 August 2017

Keywords:
Big data
Data marketplace
Data pricing
Production management
Bi-level programming model

ABSTRACT

This paper presents a bi-level mathematical programming model for the data-pricing problem that considers both data quality and data versioning strategies. Data products and data-related services differ from information products or services in terms of quality assessment methods. For this problem, we consider two aspects of data quality: (1) its multidimensionality and (2) the interaction between the dimensions. We designed a multi-version data strategy and propose a data-pricing bi-level programming model based on the data quality to maximize the profit by the owner of the data platform and the utility to consumers. A genetic algorithm was used to solve the model. The numerical solutions for the data-pricing model indicate that the multi-version strategy achieves a better market segmentation and is more profitable and feasible when the multiple dimensions of data quality are considered. These results also provide managerial guidance on data provision and data pricing for platform owners.

1. Introduction

The advent and ubiquity of Web 2.0, social networks, cloud computing, and "Software-as-a-Service" has expanded the volume of personal, business, and public data at an alarming rate. Big data volumes, and the diversity of such data, are a defining feature of the modern world, with significant financial and commercial implications. Enterprises rely not only on the acquisition of data in itself, but also on professional third-party platforms that collect data from various sources (Mohanty, Jagadeesh, & Srivatsa, 2013). Increasingly, data providers appreciate the gradual commercialization of data, and have established network platforms for data trading (Schomm, Stahl, & Vossen, 2013), thereby giving rise to data marketplaces.

Armstrong and Durfee (1998) introduced the term 'data marketplace' to denote the ensemble of agents involved in commercial transactions. A typical data market comprises three main roles: data providers, data consumers, and a data-market owner. Data providers supply data to the data market and set the corresponding prices. Data consumers buy the data that they need. Acting as the intermediary between providers and consumers, the owner negotiates the pricing mechanism with those providers and manages the data transactions (Tang, Amarilli, Senellart, & Bressan, 2014). Currently emerging data platforms include Factual,[1] Infochimps,[2] Xignite,[3] and the Windows Azure Data Marketplace[4] (Stahl, 2013). The latter, for example, encompasses more than one hundred data sources for sale, Infochimps contains 15,000 data collections, and Xignite focuses on financial data.

The emergence of data markets has prompted the design of a new kind of business model in which information and analysis tools effectively become tradable electronic goods (Muschalle, Stahl, Löser, & Vossen, 2012). In data markets, data products are processed and sold like information products at appropriately defined prices to data consumers. The present study defines data products as datasets in the form of tradable data goods after crawling, reformatting, cleaning, encrypting, and other processes. This includes government data, medical data, financial data, e-commerce data, and traffic data.

The pricing of data products is an important issue. Most data-product transactions are completed through offline negotiations between data sellers and buyers, a small proportion of which is done online. The main pricing models for data markets are as follows: (1) Free models are those where data services can be used for free. (2) Freemium models combine free services and value-added services. In the pricing model, consumers have limited access to data for free and pay for the premium services. (3) In packaging models, consumers buy a certain amount of data at a

☆ This work is supported by Fundamental Research Funds for the Central Universities (N150604003).

* Corresponding author.
E-mail address: mengxiaozhang47@hotmail.com (M. Zhang).

[1] https://www.factual.com.
[2] http://www.infochimps.com.
[3] http://www.xignite.com.
[4] https://datamarket.azure.com.

http://dx.doi.org/10.1016/j.cie.2017.08.008
0360-8352/

fixed price. (4) In pay-per-use models, consumers pay for data services based on their usage. (5) Flat-fee models involve data consumers paying a monthly subscription fee in return for unfettered access to data services. (6) In two-part-tariff models, consumers pay a fixed basic fee that becomes supplemented by an additional fee when their usage exceeds some pre-defined quota (Muschalle et al., 2012; Schomm et al., 2013).

Common weaknesses in existing data-pricing mechanisms include (1) the lack of a standardized pricing model. On data platforms such as Aggdata[5] and CustomLists,[6] data are traded mainly through private agreements between data providers and consumers, whereas Infochimps and Azure DataMarket charge their members a monthly subscription fee. (2) Issues relating to data quality tend to be neglected. Few data-pricing models for data markets consider data quality, despite the availability of relevant tools and technologies for assessing and improving data quality. (3) Opaqueness. Pricing strategies are mainly seller-driven, with the cost of data acquisition, cleaning, and packaging being invisible to consumers (Balazinska, Howe, & Suciu, 2011). These shortcomings call for the development of a rigorous and reasonable pricing model for data marketplaces.

The proper assessment of data value is the basis of a rigorous and reasonable data-pricing model. Heckman, Boehmer, Peters, Davaloo, and Kurup (2015) suggest focusing on the intrinsic value and quality of data, instead of the value of the information that underlies the data, in the interest of transparency and fairness. However, data value is determined by many, rather than one, attribute. We therefore consider the multiple dimensions of data quality and establish a linear method of multi-dimensional quality assessment. Data value is also determined by the complex interaction of multiple factors (Heckman et al., 2015). For example, an increase in the timeliness of a particular dataset may occur at the expense of its completeness. Additional costs would therefore be incurred by the data provider to increase the timeliness while simultaneously preserving completeness. By considering the interactions of multiple elements, we establish a nonlinear method for evaluating the integrated value of data.

The present study aims at realizing an effective data valuation on a data-market platform by extending it to an integrated and multi-dimensional quality assessment. Furthermore, we examine whether or not the multi-version strategy is suitable for a data-market owner when considering the linear and the integrated assessment model, and provide some guidance to the data-platform owner on how to produce, provide, and price data products.

Based on the linear and the integrated assessment model, we adopt the perspective of the data value and consider both the profit derived from a data platform and the utility to data consumers, in order to propose a fair and reasonable data-pricing model. We establish a bi-level programing model with two kinds of cost functions to analyze the production-decision behavior of the data-platform owner and the purchasing-decision behavior of data consumers. Data-platform owners may have some monopoly power that allows them to personalize pricing through price segmentation, including versioning, segmenting, and negotiating (Pantelis & Aija, 2013). In the model, as a leader, the owner decides the planned number of data-product versions, the data quality, and the prices accordingly; as followers, consumers choose the ideal data product that is provided on the data platform and that maximizes utility. The model determines the actual number of versions, data qualities, and the corresponding prices based on the total revenue of the data-platform provider and utility of each consumer, and assist consumers and providers in making reasonable decisions. The features of multiple versions are analyzed and managerial implications are presented for data-platform owners.

The rest of this paper is organized as follows: we firstly review the existing relevant literatures in Section 2. Section 3 then describes the data-pricing problem, based on data quality, and establishes a bi-level programing model that involves a data-platform owner and data consumers. Numerical applications and managerial implications are discussed in Sections 4 and 5, respectively, followed by conclusions and proposed avenues for future work in Section 6.

2. Literature review

The value assessment of intangibles such as intellectual products is not a new challenge for entrepreneurs and scholars. The pricing of information products and information services has generated a substantial literature. We here review representative works on these methods, before selectively reviewing research on data pricing.

Information-service markets involve three commonly used pricing schemes: "pure flat-fee" pricing, "pure usage-based" pricing, and "two-part tariff" pricing (Wu & Banker, 2010). Wu and Banker (2010) found that marginal and monitoring costs can influence a firm's choice of pricing scheme. Huang, Kauffman, and Ma (2015) argue for the existence of service interruptions in cloud software, to which some consumers are sensitive. In such a market, it is sensible for a vendor of cloud-computing services to adopt a hybrid pricing strategy that mixes fixed-price reserved services with spot-price on-demand services. Mei, Li, and Nie (2013) constructed a pricing model based on the Stackelberg game and advocated adopting a pure-bundling strategy, instead of pure components, when device prices are high and consumers' evaluations vary widely. Balasubramanian, Bhattacharya, and Krishnan (2015) considered differences in the use of frequencies and the psychological costs to consumers that are associated with a pay-per-use model. They concluded that two factors can affect a seller's profit by analyzing two pricing mechanisms for information products, namely the fixed-fee and pay-per-use mechanisms. Sundararajan (2004) argued that administering usage-based pricing incurs transaction costs, which influence the optimal pricing of information goods when the available information is incomplete.

On the other hand, versioning is a widespread differentiation strategy used in information-product markets. Under this scheme, a firm customizes information products according to the customers' need and encourages them to pay the highest possible price for goods to maximize its overall revenue (Shapiro & Varian, 1998). Bhargava and Choudhary (2001) analyzed the optimal strategy for vertically differentiated information products in the context of a monopoly. They showed that the optimal product line of a firm depends on the benefit-to-cost ratio of qualities when the consumer's valuation is a linear function of product quality and consumer type. Li, Feng, Chen, and Kou (2013) defined a nonlinear function to describe the "willingness to pay" and the utility to a consumer who has a specific quality requirement, and developed hybrid steady-state evolutionary algorithms. They observed that a monopoly can achieve more profit by using a multi-version strategy. Chen and Seshadri (2007) considered a two-stage development problem and found that versioning is an optimal strategy for sellers if the consumers have a convex-shaped reservation utility function. Because data and information products have many features in common, the pricing methods used for information products provide insight for our present research. However, these

[5] http://www.aggdata.com/.
[6] http://www.customlists.net/home/.

methods may not be compatible with the intrinsic characteristics of the data.

Various authors have addressed the issues of data provision and data pricing, as summarized in Table 1. Tang et al. (2014) proposed a framework for the pricing of XML documents and devised PTIME algorithms. Heckman et al. (2015) combined qualitative and quantitative methods to determine the data value for buyers and sellers, in order to propose a grand pricing model. Query pricing is a common method used in data markets. Koutris, Upadhyaya, Balazinska, Howe, and Suciu (2015) proposed a 'query-based pricing' that satisfies arbitrary-free and discount-free. This pricing function allows the price of any query to be determined automatically. In 2013, they considered an updated database and overlapping information, and proposed a new pricing system that avoids recharging (Koutris, Upadhyaya, Balazinska, Howe, & Suciu, 2013). Bergemann and Bonatti (2015) proposed a model of data provision and data pricing for a single data provider selling individual consumers' characteristics (web cookies) to individual firms (advertisers).

Some authors have focused on a special category, which is personal data. Valuable sensitive data are sometimes sold to third parties. Li, Li, Miklau, and Suciu (2014) proposed a theoretical framework for selling and pricing noisy private data, and for compensating data owners whose privacy is affected as a consequence. Jaisingh, Barron, Mehta, and Chaturvedi (2008) argue that the valuation of privacy and personal information varies between consumers, which therefore affects research on strategies for acquiring and pricing personal information. Li and Raghunathan (2014) developed an incentive-compatible mechanism that enables a data owner to price and distribute private data when the data users' true preference for the sensitivity level and quantity of data is unclear. The existing literature on data pricing either surveys published data-pricing methods or investigates new methods that focus on relevance and privacy. Data quality, a key factor influencing data valuation, has been conspicuously neglected to date.

The issues surrounding data quality have generated interest since the mid-1990s (Batini, 2003). Wang and Strong (1996) conducted a two-stage survey and a two-phase sorting study to develop a hierarchical framework for determining the quality characteristics of data. They identified fifteen relevant dimensions out of a total of 179 gathered criteria. Batini, Cappiello, Francalanci, and Maurino (2009) analyzed the classification of data-quality dimensions provided by earlier research, and defined a basic set of such dimensions, including accuracy, completeness, consistency, and timeliness. In 2016, Batini and Scannapieco (2016) provided a detailed description of a set that included accuracy, completeness, redundancy, readability, accessibility, consistency, usefulness, and trustworthiness. In a fair data-market system, the most important among these dimensions are completeness, accuracy, consistency, and timeliness or currency (Batini, 2003; Ding, Wang, Zhang, Li, & Gao, 2015). Heckman et al. (2015) listed some attributes that affect data valuation significantly, and proposed a general linear model for value assessment, suitable for all data types. In this model, the estimated value of data is influenced by many characteristics, such as the fixed cost, age, periodicity, volume, and accuracy of the data.

Data quality is an integrated and multifaceted concept, and requires the knowledge of many data characteristics. The present study thus selects key factors for evaluating data quality and hence proposes a quality-based data-pricing model. The model outcomes carry implications for management by data-platform owners.

Table 1
Selected literature relating to data pricing.

Topic	Contribution	Author
Query pricing	Present QueryMarket to price SQL queries automatically.	Koutris et al. (2015)
	Present a framework to price conjunctive queries.	Koutris et al. (2015)
Private data pricing	Propose a theoretical framework for pricing noisy private data and compensating privacy loss.	Li et al. (2014)
	Pricing personal data when customers show different valuations of privacy.	Jaisingh et al. (2008)
	Propose an incentive-compatible mechanism to price private data.	Li and Raghunathan (2014)
Others	Devise PTIME algorithms to price sampling-based XML documents.	Tang et al. (2014)
	Propose a model to price web cookies between customers and advertisers	Bergemann and Bonatti (2015)

3. Problem formulation and solution approach

3.1. Problem description

Within the context of data pricing, the stakeholders are the data providers, market owners, and consumers. On a data-market platform, the providers supply raw data they have derived from various sources, and hence provide a certain number of data samples to their potential consumers. These data consumers search for datasets on the platform according to their needs and preferences, and make purchasing decisions based on perceived value and willingness to pay. Government agencies, corporations, and even individuals can be data providers and consumers. Data-platform owners manipulate the data obtained from multiple data providers, and update dynamic data continually. As the owner of valuable data, the platform owner provides a trading platform and sets the rules that govern the transactions between providers and consumers. As with other (intangible) goods, data transactions are guided by quality, price, and the potential consumers' "willingness to pay" (Pantelis & Aija, 2013). To maximize market coverage and the overall production profit, information-product manufacturers deploy a multi-version strategy for vertically differentiated markets. We therefore approach the pricing problem from the perspective of a data-market owner, considering the data quality and the consumers' willingness to pay, with the aim of helping data-platform owners to make decisions on the version numbers, data quality, and price of each version, and hence to maximize their profit.

By considering the characteristics of the data, we here propose a data-pricing model based on quality, and analyze the effects of multidimensional data-quality evaluation methods and of different cost functions on customer choice, version design, and total profit for the data-platform owner. This method is more transparent and fairer for both data consumers and providers, and easier for the data platform owner to implement.

3.2. Model description

We assume a monopolistic data market, in which a monopolist is entitled to set rules and provide a trading platform for data providers and consumers. The monopolistic data-platform owner supplies M data-product versions that are vertically differentiated in K quality dimensions for N potential data consumers, who are heterogeneous in terms of data-quality preference in each dimension. The data platform makes versioning decision based on the maximum total profit, whereas the data consumers maximize their utilities through self-selection. The key parameters are summarized in Section 3.2.1, and the features of data-platform owner and the data consumers are given in And 3.2.23.2.3, respectively.

3.2.1. Key parameters and their explanations

Table 2 summarizes the key parameters and descriptions used in the model.

3.2.2. Features of the data-platform owner

Data-platform owners initially invested significantly in constructing data infrastructures to allow data processing. Once construction is completed, the invested funding becomes a sunk cost and can then be ignored, as it no longer influences decision-making on versions and prices. In the production process, in order to provide high-quality data products, an owner must acquire, integrate, analyze, and store data continually from a wide range of sources. This, however, results in variable costs. Furthermore, multiple versions of data products can be provided to data consumers who are heterogeneous in terms of data quality. In general, higher-quality data implies more data processing and higher costs. However, the cost of reproducing and distributing a successful version of a given data product is negligible. The characteristics of data production cost raise the need of a special pricing mechanism for data products.

The cost of data production relates directly to data quality, which must therefore be evaluated correctly and effectively. In the general linear model proposed by Heckman et al. (2015), many characteristics, such as age, periodicity, volume, and accuracy, can influence data valuation. In fact, research into product line design has considered multiple attributes explicitly (Kim & Chhajed, 2002; Krishnan & Zhu, 2006; Nair, Thakur, & Wen, 1995). Based on the above analysis, we established a multi-dimensional quality assessment method for data pricing.

Conventional formulations of cost as a linear function $c_k q_{ik}$ or a quadratic function $c_k q_{ik}^2$ (Kim & Chhajed, 2002; Li et al., 2013) fail to capture the complexity of interactions between multiple factors. For instance, the cost of simultaneously increasing the timeliness and the completeness of a given dataset is more than the sum of these individual components.

We here consider two kinds of cost functions to describe the different costs c_i that result when an owner provides different quality levels of data products i at a price p_i. We denote the quality level of each dimension as q_{ik} and assume that the production cost is an increasing function of quality. We first consider a linearly increasing cost function

$$c_i^L = c q_{ik}^L (i = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K) \tag{1}$$

where

$$q_{ik}^L = \frac{1}{K} \sum_{k=1}^{K} q_{ik} (i = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K) \tag{2}$$

This function, commonly used in earlier studies, signifies that the cost of producing a data product i varies linearly with its quality level in each dimension. A second approach uses an integrated cost function

$$c_i^I = c q_{ik}^I (i = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K) \tag{3}$$

where the integrated quality is

$$q_{ik}^I = q_{i(k-1)}^I + q_{ik}(1 - q_{i(k-1)}^I)(i = 1, 2, \ldots, M,\ k = 1, 2, \ldots, K) \tag{4}$$

When $k = 1$, we have $q_{i1}^I = q_{i1}$. The function is then represented as in Fig. 1 (taking a two-dimensional case as an example), where $q_{i1}, q_{i2} \in [0, 1]$. On the one hand, q_i^I increases as q_{ik} increases; on the other hand, when q_{i1} increases for constant q_{i2}, $q_{i2}(1 - q_{i1})$ decreases, indicating that the increase in the quality level along one axis impacts negatively on the other quality dimension. A platform owner who wishes to provide a higher-quality data product must spend more on data processing. Besides, if each quality level takes a value between 0 and 1, the total quality, integrated over multiple dimensions, remains a value between 0 and 1.[7] The results for these two cost functions can be compared easily.

Table 2
Model variables.

Variable	Description
i	The number of data-product versions, $i = 1, 2, \ldots, M$
j	The number of data consumers, $j = 1, 2, \ldots, N$
k	The number of data-quality dimensions, $k = 1, 2, \ldots, K$
c_i^L	The linear cost of data product i
c_i^I	The integrated cost of data product i
p_i	The price of data product i
c	The parameter of the cost function
q_{ik}	The quality level of data product i in quality dimension k
q_{ik}^L	The linear quality of data product i with quality dimension k
q_{ik}^I	The integrated quality of data product i with quality dimension k
q_{jk}^R	The reservation quality of consumer j in quality dimension k
q_{jk}^S	The saturation quality of consumer j in quality dimension k
θ_{jk}	The quality preference of consumer j in quality dimension k
w_{ijk}	The willingness to pay of consumer j for data product i in quality dimension k
w_{ij}	The willingness to pay of consumer j for data product i
w_{ijk}^L	The linear willingness to pay of consumer j for data product i with quality dimension k
w_{ijk}^I	The integrated willingness to pay of consumer j for data product i with quality dimension k
u_{ij}	The utility of consumer j for data product i
x_{ij}	The purchasing decision of consumer j for data product i, $x_{ij} = 0$ or 1
y_i	The production decision of the data platform owner for data product i, $y_i = 0$ or 1

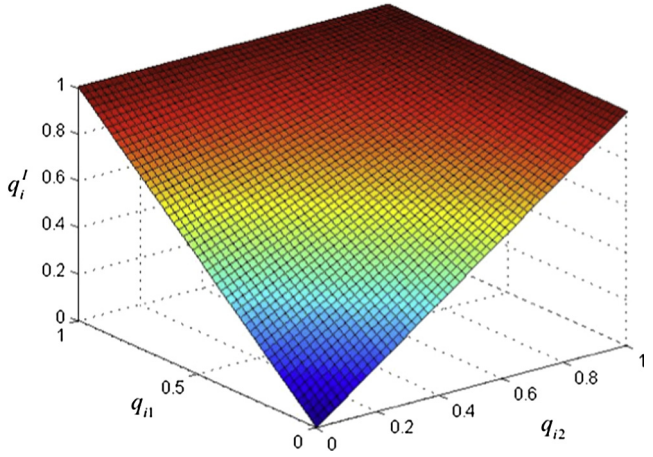

Fig. 1. Integrated quality function.

3.2.3. Features of data consumers

Every consumer in the marketplace has individual preferences and interests. Because of the inherent variability in the valuation of the attributes of a given product, it is often impossible to substitute even similar products. However, a seller can consider the distribution of potential consumers' preferences and demands, instead of identifying individual preferences before purchase. The seller therefore provides some products with different qualities to different markets and assigns different customer types to different varieties of goods. On the other hand, consumers make their own purchasing decisions according to their requirements, prefer-

[7] It is easy to prove that $q_1 + q_2(1 - q_1) \geqslant 0$, since $0 \leqslant q_1 \leqslant 1$, $0 \leqslant q_2 \leqslant 1$, and. $0 \leqslant 1 - q_1 \leqslant 1$. Also, when $0 \leqslant q_2 \leqslant 1$, $q_2(1 - q_1) \leqslant 1 - q_1$ must be true. Additionally, $q_1 + (1 - q_1) = 1$ is constantly tenable, then $q_1 + q_2(1 - q_1) \leqslant 1$ can be derived.

ences, and prices by a process of self-selection (Mussa & Rosen, 1978).

This self-selection is described by a consumer's utility function. For a given consumer, this function characterizes valuation and the willingness to pay: $w(\theta, q) = \theta q$, where θ quantifies the consumer's marginal willingness to pay (characterizing the consumer type θ) and q is the quality of the products that the consumer wants to buy. We have $\theta \in [0, \theta_{\max}]$, $q \in (0, q_{\max}]$, where $\theta_{\max}$ and $q_{\max}$ are the corresponding maximum values. The consumer gains utility at the price p, and so $u(\theta, p, q) = \theta q - p$.The above model assumes that the willingness to pay is a linear function of the quality and reflects the heterogeneity of consumers' quality preferences. On this basis, Krishnan and Zhu (2006) proposed a willingness-to-pay function that incorporated notions of saturation and reservation qualities. They assumed that, for a given consumer, the willingness to pay equals 0 whenever the quality of a product falls below some lower threshold value (signifying that the consumer would not consider buying that product), and remains constant whenever the quality increases beyond some upper threshold. These upper and lower thresholds are called the saturation and reservation quality levels, respectively. In the present study, we adopt this model and assume that consumer $j(1 \leqslant j \leqslant N)$ has a preference θ_{jk} in quality dimension $k(1 \leqslant k \leqslant K)$ and a willingness to pay w_{ijk} for quality dimension k of the product $i(1 \leqslant i \leqslant M)$ (see Fig. 2):

$$w_{ijk} = \begin{cases} 0, & \text{if } q_{ik} < q_{jk}^R \\ \theta_{jk}(q_{ik} - q_{jk}^R), & \text{if } q_{jk}^R \leqslant q_{ik} \leqslant q_{jk}^S \quad i(1 \leqslant i \leqslant M) \\ \theta_{jk} q_{jk}^S, & \text{if } q_{ik} > q_{jk}^S \end{cases} \tag{5}$$

The willingness to pay of consumer j for data product i is written w_{ij}. To include the effect of the integrated quality on a consumer's willingness to pay, we propose a linear willingness to pay expressed as w_{ijk}^L and an integrated willingness to pay denoted as w_{ijk}^I:

$$w_{ijk}^L(\theta_{jk}, q_{ik}) = \begin{cases} 0, & \text{if } q_{ik} < q_{jk}^R \\ \frac{1}{K}\sum_{k=1}^{K} w_{ijk}, & \text{otherwise} \end{cases} \tag{6}$$

$$w_{ijk}^I(\theta_{jk}, q_{ik}) = \begin{cases} 0, & \text{if } q_{ik} < q_{jk}^R \\ w_{ij(k-1)}^I + w_{ijk}(1 - w_{ij(k-1)}^I), & \text{otherwise} \end{cases} \tag{7}$$

When $k = 1$, $w_{ij1}^I = w_{ij1}$. The utility derived from the purchase of data product i is

$$u_{ij}(\theta_{jk}, q_{ik}) = \begin{cases} 0, & \text{if } q_{ik} < q_{jk}^R \\ w_{ij} - p_i, & \text{otherwise} \end{cases} \tag{8}$$

The above model assumes that the marginal willingness to pay (or the consumer type θ) of all potential data consumers on the data platform is uniformly distributed between 0 and $\theta_{\max}$ ($\theta \in [0, \theta_{\max}]$). Consumer j buys a data product i only if $u_{ij}(\theta_{jk}, q_{ik}, p_i) \geqslant 0$ .When more than one version satisfies this condition, consumer j chooses the version with the maximum value, i.e., $q_{ik}^* = \arg\max_i \{u_{ij}(\theta_{jk}, q_{ik}, p_i), i = 1, 2, \ldots, M\}$.

3.2.4. Data-pricing model based on data quality

Considering both the revenue of a data platform and the utilities of data consumers, we established a bi-level programming model involving one leader (a monopolistic data-platform owner) and many followers (all the potential data consumers) to address the issue of the versioning and pricing of data products. At the first level, the owner decides the number of data-product versions, the quality levels of the multiple dimensions involved, and the selling prices, with the aim of maximizing the overall profit to be made by providing these versions. At the second level, the potential data consumers make their own purchasing decisions by the self-selection process. The model described above can be formulated as follows:

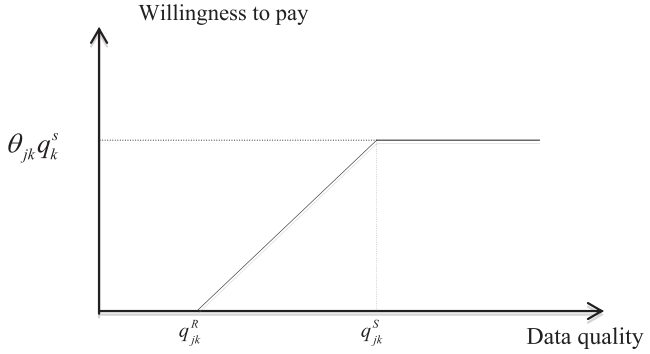

Fig. 2. Willingness to pay w_{ijk} for quality dimension k.

(1) First level: production decision of the monopolist

$$\max \ \pi(q_{ik}, p_i, x_{ij}) = \sum_{j=1}^{N}\sum_{i=1}^{M} p_i x_{ij} - \sum_{i=1}^{M} c_i y_i \tag{9}$$

where

$$y_i \leqslant \sum_{j=1}^{N} x_{ij}, \quad i = 1, 2, \ldots, M \tag{10}$$

$$p_i > 0, \quad i = 1, 2, \ldots, M \tag{11}$$

$$0 < q_{ik} \leqslant 1, \quad i = 1, 2, \ldots, M, \ k = 1, 2, \ldots, K \tag{12}$$

$$y_i = 0 \text{ or } 1, \quad i = 1, 2, \ldots, M \tag{13}$$

M is the maximum number of data-product versions and is constrained by the data-platform owner's resources. The quality levels and price of each version $\{q_{ik}, p_i\}$, $i = 1, 2, \ldots, M$ are decision variables of the monopolist. The function $\pi(q_{ik}, p_i, x_{ij})$ is the total profit of the data-platform owner made with the multi-version strategy. Constraint (10) signifies that the owner does not provide data product i if there is no demand for it. Constraint (11) ensures that the selling price of any version of data product is a positive number, while constraint (12) ensures that the quality in each dimension lies between 0 and 1.

(2) Second level: self-selective decision by every customer j $(j = 1, 2, \ldots, N)$

$$\max \ u_j(x) = \sum_{i=1}^{M} [u_{ij}(\theta_{jk}, q_{ik}, p_i) x_{ij}] \tag{14}$$

where

$$x_{i_1 j} x_{i_2 j} = 0, \ \text{if } i_1 \neq i_2; \ i_1, \ i_2 = 1, 2, \ldots, M \tag{15}$$

$$x_{ij} u_{ij}(\theta_{jk}, q_{ik}, p_i) \geqslant 0, \quad i = 1, 2, \ldots, M; \ j = 1, 2, \ldots, N \tag{16}$$

$$x_{ij} = 0 \text{ or } 1, \quad i = 1, 2, \ldots, M; \ j = 1, 2, \ldots, N \tag{17}$$

with x_{ij} denoting the decision variable of an individual consumer j, and u_j is the utility gained by consumer j through the purchase of a product at a certain quality level. Each consumer chooses only one data-product version that yields a non-negative utility, as expressed by constraints (15) and (16).

The model simultaneously considers the profit made by the data-platform owner and data consumers, reflecting the consumer-driven mode in an e-commerce environment. In addition, we perform an in-depth analysis of the multiple quality dimensions involved, and conclude that production and pricing decisions can reflect the characteristics of the data products.

3.3. Optimal solution for the quality-based data-pricing model

Bi-level programs describe the hierarchical structures that occur naturally in real-world situations, and generated considerable interest in 1980s (Colson, Marcotte, & Savard, 2005). However, bi-level programming problems are difficult to solve. Even the simplest linear bi-level program has been shown to be NP-hard (Ben-Ayed & Blair, 1990; Hansen, Jaumard, & Savard, 1992; Jeroslow, 1985). There are currently five main methods for solving such problems: the extreme-point-search method, the Karush-Kush-Tucker method, the descent method, the direct-search method, and non-numerical optimization methods (which include simulated annealing, genetic algorithms, and "ant colony" algorithm). Because our proposed model, which involves many integer variables and a nonlinear utility function, is difficult to solve analytically, we propose a genetic algorithm (GA) based on the bi-level programing problem (called the GA-BLP) to solve it numerically.

To solve the bi-level programming model, the leader's "decision" vector $z = \{q, p\}$ is reproduced according to the solution of the GA, while the followers' decision vector x is obtained by solving the second-level programming problem. Initially, GA-BLP randomly produces an initial population in which an individual's chromosome is represented by a real-valued vector z. The population size N denotes the number of such vectors z. In the subprocedure BLPF, given $z = \{q, p\}$, the followers' decision vector x is obtained according to the objective at the second level. The following steps are the same as for the pure GA.

procedure GA-BLP

set $t = 0$;
initialize: $P(0)$, which contains N vectors:
$z_j = \{q_j, p_j\}, j = 1, 2, \ldots, N$
calculate the individual fitness:
$\{f_1(0), f_2(0), \ldots, f_N(0)\} = BLPF(P(0))$;
while the stopping criteria is not met, do
produce offspring by performing selection, crossover, and mutation;
$p_{offspring}(t) = S \circ C \circ M(P(t))$;
calculate individual fitness:
$\{f_1(t), f_2(t), \ldots, f_{offspring}(t)\} = BLPF(P_{offspring}(t))$;
insert a new individual: $P(t+1) = insert(P(t), P_{offspring}(t))$;
set $t = t + 1$;
end while
return $\{P(t)\}$;
end procedure
subprocedure BLPF
for i = 1:n do
solve the programming problem at the second level:
$x_j = \arg F_j^j(k); j = 1, 2, \ldots, N$
end for
obtain $y_j(x_j)$;
calculate the value of the objective at the first level:
$f_j(k) = F_j^1(P(0))$;
end subprocedure BLPF

4. Numerical applications

Consider a data platform comprising 1000 potential data consumers, whose quality preference along any of the quality dimensions is uniformly distributed between 0 and 1. With regard to the data products, a dataset has multiple attributes. For example, the "record count", "region included" and "last updated" are displayed on AggData. For simplification and without loss of generality, we focus on two principal data-quality dimensions. However, our results can be generalized relatively easily to more dimensions. The features of consumers and data products are listed in Tables 3 and 4, and the parameter settings for GA-BLP are given in Table 5.

We determined the maximum profit and optimal pricing strategy for two dimensions of quality using, separately, the linear cost function and the integrated cost function. The results serve to illustrate the impact of different forms of the cost function on optimal multiple-version pricing strategies. The two cost functions considered were

(1) $c_i = c(q_{i1} + q_{i2})/2$.
(2) $c_i = c(q_{i1} + q_{i2} - q_{i1}q_{i2})$.

We developed a GA to solve the problem and ran it 30 times in each experiment to obtain more accurate and stable optimal numerical solutions. Solving the problem of the data-pricing strategy based on two-dimensional quality using cost functions (1) and (2), we determined the maximal profit and the market-coverage rate for the data-platform owner with the different numbers of versions set by data-platform owner in advance, as shown in Figs. 3 and 4.

Figs. 3 and 4 show a logarithmic increase in the total profit and market-coverage rate with the maximal number of versions, albeit with a gradual leveling off. The profit corresponding to cost function (2) is always higher than that for cost function (1), even though the market-coverage rate for cost function (2) is not necessarily higher. The reason is that, when consumers evaluate data in the integrated way, their willingness to pay can increase (see Fig. 5), and therefore platform owners can set relatively higher prices for each version to increase profit.

The experimental results also show that owners cannot always provide the planned number of versions, because some versions, even when produced and provided on the market platform, are never chosen as a result of product competition. The numbers of versions provided by the data-platform owner is presented in Fig. 6. As the number of planned versions increases, the number of actual versions associated with cost functions (1) and (2) converges to 5. The optimal quality levels, determined for different numbers of versions is showed in Fig. 7, for both cost functions.

Interestingly, the optimal quality levels with each maximal version number show the similar trends, independently of the choice of cost function. Specifically, the version of the highest quality {1,1} is always provided, regardless of the number of versions. Beyond two versions, the versions with intermediate quality ({0.3,0.3}) are added to the market. As the number of versions continues to grow, two versions with a high level in one dimension and an intermediate level in another (e.g., {1,0.3}, {0.3,1}) are provided in the market. No version with a low quality level close to zero is provided, even in the case of ten versions. This result differs significantly from that of Feng et al. (2013), who considered a linear willingness-to-pay function for whom the three-version scheme ({1,1}, {1,0}, {0,1}) was optimal. In the present study, when the saturation and reservation qualities are considered, the strategy of providing three versions is not optimal. Instead, we

Table 3
Parameter settings for heterogeneous customers.

Customer type	$\theta_{jk} \in [0, \theta_{\max}], \theta_{\max} = 1$
Customer distributions with type	$\theta_{jk} \sim U(0,1)$
Customer willingness-to-pay function	$q^R_{jk} = 0.5v_{jk}$, $q^S_{jk} = 1.1v_{jk}$
Potential market size	$M = 1000$

Table 4
Parameter settings for data product.

Maximum version number	$K = 1, 2, \ldots, 10$
Highest quality version	$q_H = 1.0$
Cost function	$c = 2.5$

Table 5
Parameter settings for GA.

Population	Population size: 20, initial range: [0; 1]
Fitness scaling	Scaling function: rank
Selection	Selection function: stochastic uniform
Reproduction	Elite count: 2, crossover fraction: 0.8
Mutation	Mutation function: constraint dependent
Crossover	Crossover function: scattered
Migration	Direction: forward, fraction: 0.2, interval: 20
Stopping criteria	Generations: 500; function tolerance: 1e−8

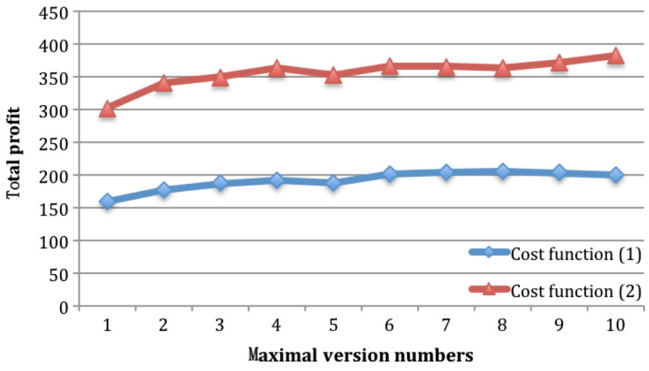

Fig. 3. Total profit with maximal version numbers.

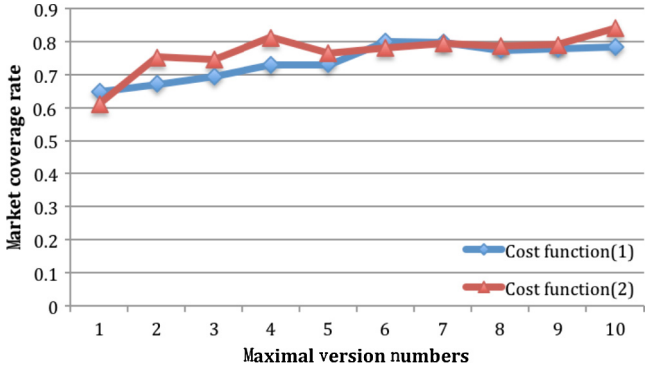

Fig. 4. Market coverage with maximal version numbers.

found that it is most profitable for the platform owners to provide four or five versions ({1,1}, {0.3,1}, {1,0.3} and {0.3,0.3}). This arises from the utility function of the potential data consumers. When a provided data product is of low quality, few customers have their requirements met and derive non-negative utilities;

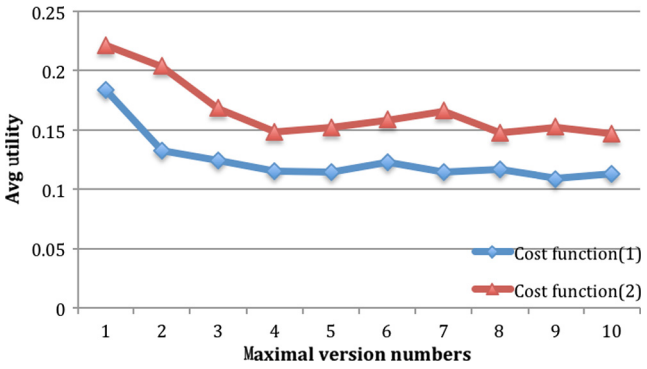

Fig. 5. Average utility with maximal version numbers.

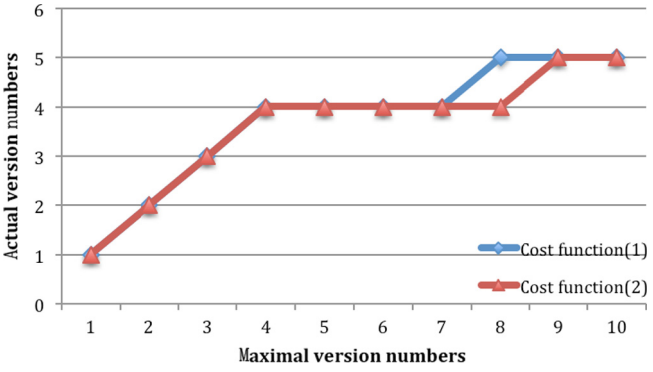

Fig. 6. Actual version numbers.

consequently, this product is not chosen, which makes the owner unwilling to produce a low-quality version despite the low cost.

A multiple-version strategy is profitable because the availability of more versions of data products with various quality levels makes the data market better segmented and the price of each version increases with the number of versions. The performance-to-price ratios of each version are plotted in Fig. 8. The performance-to-price ratio of each version increases with the actual version number, which means that data owners can make additional profit from each version.

Compared to the main pricing models for data markets outlined in Section 1, the model in this paper is more reasonable and transparent since the pricing is based on the intrinsic characteristics of the data. Versioning is widely used for differentiating products and for segmenting consumer market in information-product markets. We applied the versioning strategy to data markets and extended it by considering the multiple dimensions of data quality and the interactions between these dimensions. Our proposed model provides a new approach for data pricing, with the potential to help data-platform owners increase their profit.

5. Managerial implications

The managerial implications of the numerical results are summarized as follows:

(1) To evaluate the value of data reliably and reasonably on a data-market platform, and to provide a fair and transparent pricing scheme for both data providers and consumers, we considered the multiple dimensions of the data quality. For instance, we considered the multiple attributes of a dataset. Based on the values of these attributes, we can assess the

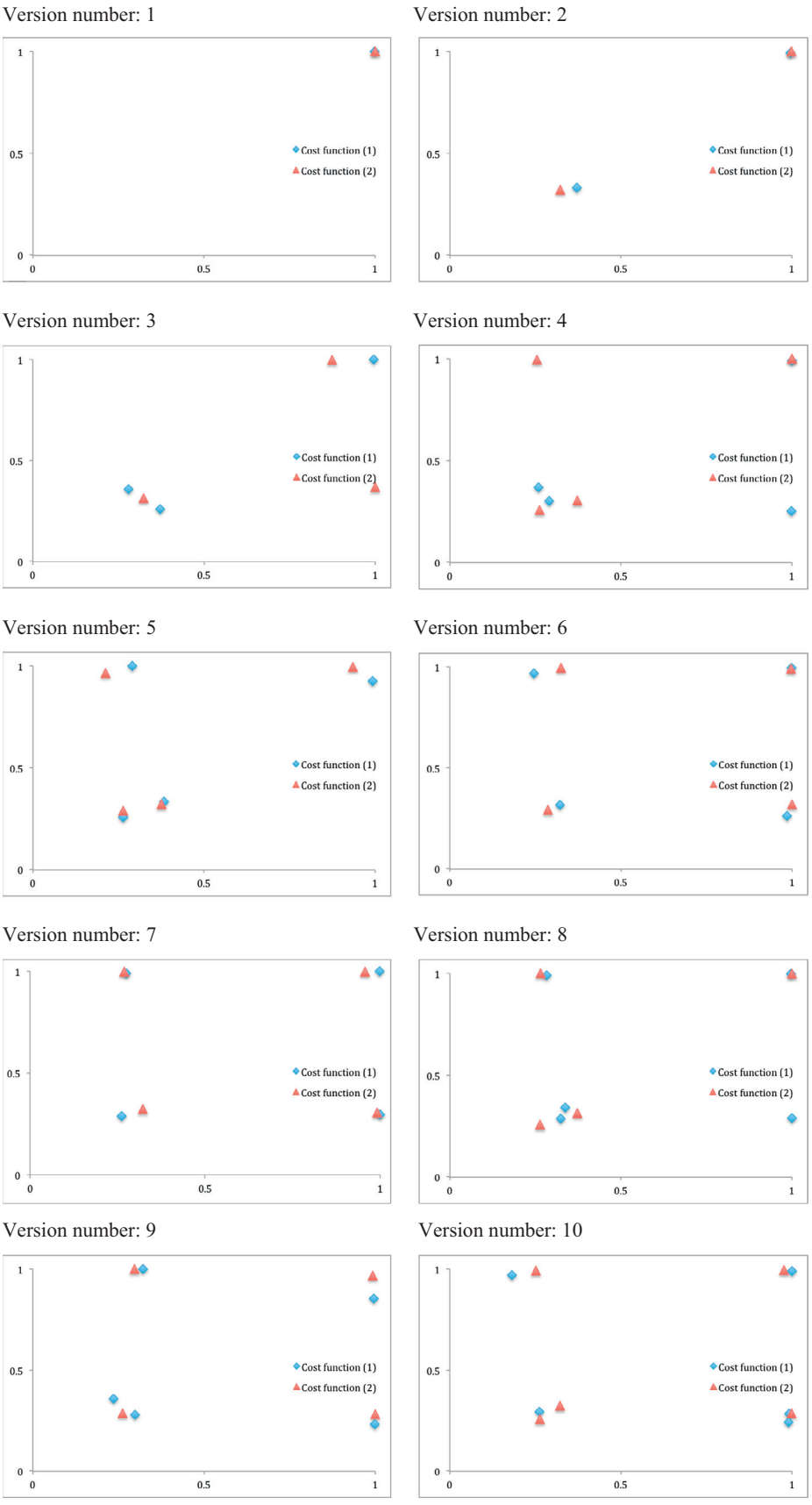

Fig. 7. Optimal quality level.

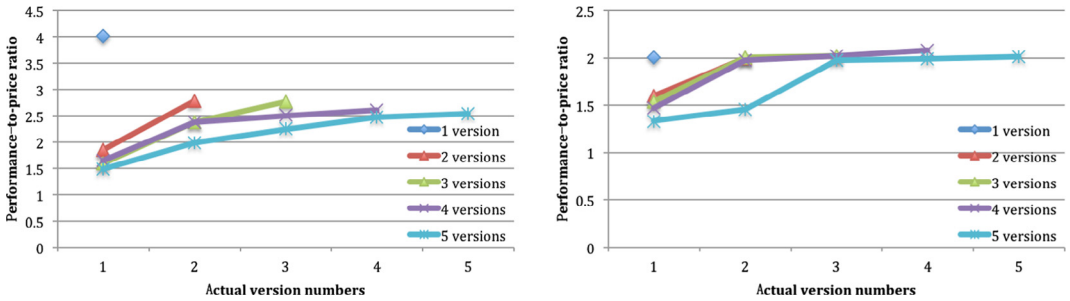

Fig. 8. Performance-to-price ratio as a function of actual version numbers.

value of a dataset clearly. This serves as the basis for a fair and transparent data-pricing scheme. Besides, the consideration of multiple dimensions makes the market segmentation more detailed, since the data-platform owner can differentiate a given data product in each quality dimension for heterogeneous consumers. Thus, a multi-version strategy is more profitable and feasible in a data market where data-value assessment involves many factors than when considering only one dimension.

(2) Traditionally, the value of a dataset is expressed as a linear function of each quality dimension (Heckman et al., 2015), which are assumed to be independent. We proposed an integrated cost function and assumed that the data consumers evaluate the data in an integrated manner. We also compared the pricing decisions in the linear and integrated manners. The experimental results show that data-platform owner can make more profit when considering the interactions.

(3) A consumer's willingness-to-pay function, characterized with saturation and reservation quality levels, significantly influences production decisions. Our results recommend four versions (i.e., {1.0,1.0}, {1.0,0.3}, {0.3,1.0}, and {0.3,0.3}) to be provided in order to avoid cannibalization. This differs from the conclusion of Feng et al. (2013), who considered a linear willingness-to-pay function.

(4) Since the emergence of data transactions as an emerging business in recent years, their development has met many challenges, such as data privacy and security and intellectual property (Lundqvist, 2016). For instance, once a data transaction is done and the data reaches the data consumers, these consumers can become providers themselves. As a first consequence, data-platform owners may lose control of their data without intellectual-property protection, as data consumers can then transfer, share, or sell the data. Secondly, the permission for data consumers to sell, extract, or process the data can generate competition and conflict in the data marketplace. Under such conditions, data platform owners lose their pricing advantages if the consumers can, in turn, directly sell copies of the data at lower prices. If the consumers can sell the data indirectly, transforming the data into new products through extraction, processing, or design, the data-platform owners' revenue will be remarkably damaged. Data licensing[8] may provide a remedy. Specifically, data licensing can clearly address the manner of delivery, maintenance, and control of data. Data-security policies, practices and protocols are also effective, particularly when the data containing personal or sensitive financial, technical, or commercial information is concerned.

6. Conclusions and future works

The emergence of data markets (e.g., Factual, Infochimps, Xignite and the Windows Azure Data Marketplace) has facilitated data transactions between providers and consumers. However, the shortcomings in existing data-pricing mechanisms in such marketplaces and the characteristics of the data products highlight the need for a fair and transparent pricing scheme. We here proposed a new pricing scheme for the data marketplace, aimed at providing a useful decision tool for both data-platform owners and data consumers. This scheme focuses on the value of data itself and considers the multiple attributes of the data. We specifically also consider the interactions between these attributes, which have so far been ignored. Furthermore, we analyzed the effect of a nonlinear utility function on production decision.

Our results and analysis indicate that this multi-version strategy helps the data platform owner to make more profit by segmenting the market along each quality dimension. In addition, when a data-platform owner and data consumers assess the data value in the integrated manner, the platform owner can make additional profit. Our results indicate that four versions ({1.0,1.0}, {1.0,0.3}, {0.3,1.0} and {0.3,0.3}) should be provided when considering a nonlinear willingness-to-pay function with reservation and saturation quality levels. Finally, we also discussed some issues that are relevant to data-market management.

The limitations of this study will form the subject of future research. Firstly, the data quality is not the only factor that affects data pricing. Data volume should also be considered (Heckman et al., 2015). A future study will include both data quality and volume in the data-pricing model.

Pricing strategies must be tailored to data characteristics. Batini and Scannapieco (2016) classified data into stable, long-term-changing, and frequently changing data categories, according to the temporal dimension. These three categories require their own pricing strategy, another subject of future research.

References

Armstrong, A. A., & Durfee, E. H. (1998). Mixing and memory: Emergent cooperation in an information marketplace. In *Paper presented at the proceedings of the international conference on multi agent systems.*

Balasubramanian, S., Bhattacharya, S., & Krishnan, V. V. (2015). Pricing information goods: A strategic analysis of the selling and pay-per-use mechanisms. *Marketing Science, 34*(2), 218–234.

Balazinska, M., Howe, B., & Suciu, D. (2011). Data markets in the cloud: An opportunity for the database community. *Proceedings of the VLDB Endowment, 4*(12), 1482–1485.

Batini, C. (2003). Data quality assessment. *Communications of the ACM, 45*(4ve), 211–218.

[8] http://legalsolutions.thomsonreuters.com/law-products/news-views/corporate-counsel/data-licensing-taking-into-account-data-ownership-and-use.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR), 41*(3), 16.

Batini, C., & Scannapieco, M. (2016). *Erratum to: Data and information quality: Dimensions, principles and techniques data and information quality*. Springer, pp. E1–E1.

Ben-Ayed, O., & Blair, C. E. (1990). Computational difficulties of bilevel linear programming. *Operations Research, 38*(3), 556–560.

Bergemann, D., & Bonatti, A. (2015). Selling cookies. *American Economic Journal: Microeconomics, 7*(3), 259–294.

Bhargava, H. K., & Choudhary, V. (2001). Information goods and vertical differentiation. *Journal of Management Information Systems, 18*(2), 89–106.

Chen, Y.-J., & Seshadri, S. (2007). Product development and pricing strategy for information goods under heterogeneous outside opportunities. *Information Systems Research, 18*(2), 150–172.

Colson, B., Marcotte, P., & Savard, G. (2005). A trust-region method for nonlinear bilevel programming: Algorithm and computational experience. *Computational Optimization and Applications, 30*(3), 211–227.

Ding, X., Wang, H., Zhang, D., Li, J., & Gao, H. (2015). A fair data market system with data quality evaluation and repairing recommendation. In *Paper presented at the Asia-Pacific web conference*.

Feng, H., Li, M., Chen, F., Feng, H., Li, M., & Chen, F. (2013). Optimal versioning in two-dimensional information product differentiation under different customer distributions. *Computers & Industrial Engineering, 66*(4), 962–975.

Hansen, P., Jaumard, B., & Savard, G. (1992). New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing, 13*(5), 1194–1217.

Heckman, J. R., Boehmer, E. L., Peters, E. H., Davaloo, M., & Kurup, N. G. (2015). *A pricing model for data markets*.

Huang, J., Kauffman, R. J., & Ma, D. (2015). Pricing strategy for cloud computing: A damaged services perspective. *Decision Support Systems, 78*, 80–92.

Jaisingh, J., Barron, J., Mehta, S., & Chaturvedi, A. (2008). Privacy and pricing personal information. *European Journal of Operational Research, 187*(3), 857–870.

Jeroslow, R. G. (1985). The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming, 32*(2), 146–164.

Kim, K., & Chhajed, D. (2002). Product design with multiple quality-type attributes. *Management Science, 48*(11), 1502–1511.

Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2013). Toward practical query pricing with QueryMarket. In *Paper presented at the proceedings of the 2013 ACM SIGMOD international conference on management of data*.

Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., & Suciu, D. (2015). Query-based data pricing. *Journal of the ACM (JACM), 62*(5), 43.

Krishnan, V., & Zhu, W. (2006). Designing a family of development-intensive products. *Management Science, 52*(6), 813–825.

Li, M., Feng, H., Chen, F., & Kou, J. (2013). Optimal versioning strategy for information products with behavior-based utility function of heterogeneous customers. *Computers & Operations Research, 40*(10), 2374–2386.

Li, C., Li, D. Y., Miklau, G., & Suciu, D. (2014). A theory of pricing private data. *ACM Transactions on Database Systems (TODS), 39*(4), 34.

Li, X.-B., & Raghunathan, S. (2014). Pricing and disseminating customer data with privacy awareness. *Decision Support Systems, 59*, 63–73.

Lundqvist, B. (2016). Big data, open data, privacy regulations, intellectual property and competition law in an internet of things world.

Mei, L., Li, W., & Nie, K. (2013). *Pricing decision analysis for information services of the internet of things based on Stackelberg game Liss 2012*. Springer.

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *"Big data" in the enterprise big data imperatives*. Springer.

Muschalle, A., Stahl, F., Löser, A., & Vossen, G. (2012). Pricing approaches for data markets. In *Paper presented at the international workshop on business intelligence for the real-time enterprise*.

Mussa, M., & Rosen, S. (1978). Monopoly and product quality. *Journal of Economic theory, 18*(2), 301–317.

Nair, S. K., Thakur, L. S., & Wen, K. W. (1995). Near optimal solutions for product line design and selection: Beam search heuristics. *Management Science, 41*(5), 767–785.

Pantelis, K., & Aija, L. (2013). Understanding the value of (big) data. In *Paper presented at the 2013 IEEE international conference on Big Data*.

Schomm, F., Stahl, F., & Vossen, G. (2013). Marketplaces for data: An initial survey. *ACM SIGMOD Record, 42*(1), 15–26.

Shapiro, C., & Varian, H. R. (1998). Versioning: The smart way to. *Harvard Business Review, 107*(6), 107.

Stahl, F. (2013). High quality information provisioning and data pricing. In *Paper presented at the 2013 IEEE 29th international conference on data engineering workshops (ICDEW)*.

Sundararajan, A. (2004). Nonlinear pricing of information goods. *Management Science, 50*(12), 1660–1673.

Tang, R., Amarilli, A., Senellart, P., & Bressan, S. (2014). Get a sample for a discount. In *Paper presented at the international conference on database and expert systems applications*.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.

Wu, S.-Y., & Banker, R. D. (2010). Best pricing strategy for information services. *Journal of the Association for Information Systems, 11*(6), 339–366.