DiVA

http://www.diva-portal.org

Preprint

This is the submitted version of a paper presented at *18th International Conference on Software Business, (ICSOB), Essen*.

Citation for the original published paper:

Fricker, S., Maksimov, Y. (2017)
Pricing of data products in data marketplaces
In: Werder K.,Ojala A.,Holmstrom Olsson H. (ed.), *Lecture Notes in Business Information Processing* (pp. 49-66). Springer Verlag
Lecture Notes in Business Information Processing
https://doi.org/10.1007/978-3-319-69191-6_4

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:bth-15615

Pricing of Data Products in Data Marketplaces

Samuel A. Fricker[1,2], Yuliyan V. Maksimov[1]

[1]i4Ds Centre for Requirements Engineering,
University of Applied Sciences Northwestern Switzerland (FHNW), Windisch, Switzerland
[samuel.fricker, yuliyan.maksimov]@fhnw.ch

[2]Software Engineering Research Laboratory (SERL-Sweden),
Blekinge Institute of Technology, Karlskrona, Sweden
samuel.fricker@bth.se

Abstract. Mobile computing and the Internet of Things promises massive amounts of data for big data analytic and machine learning. A data sharing economy is needed to make that data available for companies that wish to develop smart systems and services. While digital markets for trading data are emerging, there is no consolidated understanding of how to price data products and thus offer data vendors incentives for sharing data. This paper uses a combined keyword search and snowballing approach to systematically review the literature on the pricing of data products that are to be offered on marketplaces. The results give insights into the maturity and character of data pricing. They enable practitioners to select a pricing approach suitable for their situation and researchers to extend and mature data pricing as a topic.

Keywords: data pricing, data marketplace, systematic literature review.

1 Introduction

With the rise of Mobile Computing and the Internet of Things, massive amounts of data are being produced [1]. Already today, a substantial portion of the population owns a smartphone that is packed with sensors. In the near future, Internet nodes with sensing capabilities are expected to reside in almost any everyday thing. The data, analyzed with big data analytics and machine learning, offers an opportunity to bring about breakthroughs in processing images, video, speech, and audio [2]. Data of importance are generated by industrial vendors, private citizens, or the government [3]. Politics and executive floors of global businesses underline the importance of such data [4].

Marketplaces are enablers for the exchange of data [5]. A *data marketplace* is a platform on which dataset can be offered and accessed [3]. Marketplaces enable trade by offering services for buying and selling data, finding datasets, and obtaining access to vendors. Often cited examples are the Microsoft Azure Marketplace, Xignite, Gnip, AggData, and Cvedia. Data that are being offered may be static archives or online streams of new data. Different modes of access may be offered, e.g. whole repositories, APIs for answering queries, or subscriptions. We call such variants *data products*.

According to an early survey of data vendors, estimating the value of data and setting the right price for a data product offering is a key challenge [6]. For vendors, the *pricing*

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

is part of the value-creation with data. For customers, wrong pricing makes data unattractive. While overviews of the pricing of software products exist [7], there is no consolidated overview of the state-of-the-art for pricing data products.

Given the drastic changes that the software industry is undergoing at this moment with the move towards ‘smart everything everywhere,' it is critical that a better understanding of the business with data is obtained. It is urgent that the so far young and small research area is being developed, especially because it has hardly been discussed in the domain of software business. The lack of consolidation limits the uptake of good practice by practitioners and hinders the planning of research in this area.

This paper offers an overview of the current research in the pricing of data for data marketplaces. It utilizes a systematic approach to identifying, screening, analyzing, and synthesizing the research literature. The paper describes the research on data pricing, the contexts in which data pricing was investigated, and the maturity of the area. For owners of data products, the results offer guidance of how to do pricing. For researchers, the results offer insights into the knowledge frontier and knowledge gaps for planning research in data pricing. We intend to utilize the results for building support for data pricing into the Bonseyes marketplace (www.bonseyes.com).

The paper is structured as follows. Section 2 gives an overview of the research methodology. Section 3 describes the results of reviewing the research literature. Section 4 discusses the obtained results. Section 5 summarizes and concludes.

2 Research Methodology

The study aimed at consolidating the research on the pricing of data products offered on marketplaces. To achieve this aim, we used a systematic approach to reviewing the research literature. We used the following steps to conduct the review. 1) Identify and screen the start set of primary studies with a database search. 2) Identify and screen the final set of primary studies with snowballing. 3) Evaluate the quality of the research based on full texts. 4) Extract and analyze the data for answering the research questions.

We used the snowballing guidelines proposed by Wohlin [8] for paper identification. The snowballing helped us to avoid many false positives that would have been generated by a database search string that is too inclusive. For screening and research quality evaluation, we used the guidelines provided by Kitchenham and Charters [9]. The data extraction and analysis step followed the systematic mapping recommendations of Petersen [10]. We chose to follow Petersen because the results presented by the included papers did not allow any meta-analysis with quantitative statistic methods.

To guide our systematic review, we asked the research questions shown in Table 1. RQ1 is intended to overview how far the state-of-the-art has advanced and where the research gaps are. We followed the ideas of Ivarsson and Gorschek to assess the maturity of the research with the strength of the empirical evaluation [11]. RQ2 is intended to obtain an overview of pricing from the data vendor’s perspective. To understand pricing, we were first interested in what the products were that were priced and which contexts these products targeted. We then described the rules for determining prices, the *pricing models*, and the mechanisms used for applying these rules.

Table 1. Research questions.

Research Question	Description
RQ1: How mature are the researched pricing models?	Maturity is a concern in technology transfer from academia to industry [11]. Maturity is important for practitioners to decide about the adoption of technology, such as pricing models, and for researchers to further mature the technology.
RQ2: How do vendors price data?	The pricing of data is the concern being addressed by the presented research. The answer to this RQ should inform practitioners adopting pricing for the data they offer, trade, or buy and researchers that aim at improving the state-of-the-art.
RQ2.1: Which contexts did the pricing models target?	A context offers the frame for offering and exploiting technology. The contexts for the pricing of data comprise the domains in which the data would be used, the types and storage of data, and scenarios for exploiting that data.
RQ2.2: What kinds of data products were being priced?	A data product is the packaging of data that get a price tag attached. We expect the definition of the data products to consist of the price metrics (i.e. a definition of what is being priced), the quality attributes that are being considered for product definition, and the characteristics of the market for which the product is defined.
RQ2.3: What pricing models were evaluated?	A pricing model is a set of the rules established for defining prices. A pricing model describes how product and context variables are considered to achieve aims of interest, such as profit optimization.
RQ2.4: What mechanisms were proposed to determine a price?	To sell data to a customer the final price for the instance of the data product must be determined by applying a pricing model. With the answer to this RQ, we give an overview of how the pricing model is used to determine a final price.

2.1 Research Process

Start set of primary studies. We built the start set of papers with a keyword search for primary studies in Scopus. Scopus was selected because it offers the largest number of abstracts and citations in science and technology. We searched title, abstract, and keywords fields with the string "*data marketplace*" on January 20, 2017. The string constrained the population while leaving the intervention, comparators, outcomes, and contexts open [9]. These latter parts were used in the analysis for RQ2. We constrained the search to *marketplace*, leaving terms like *databases* and *repositories* out, because of our interest in business with data and not warehousing. The search yielded 181 papers.

We screened the papers based on title, abstract, and meta-information. Following Kitchenham's recommendations [9], we developed the selection criteria based on the research questions and practical issues. We maintained a list of excluded studies together with the reasons for exclusion. Table 2 shows the inclusion and exclusion criteria that resulted from this process. The two authors assessed the exclusion of primary articles by seeking consensus. After screening, the start set of papers contained 11 papers.

Table 2. Study selection criteria (based on the research questions* and practical reasons**).

Inclusion criteria	Exclusion criteria
- Proposal, evaluation, and discussion of a vendor's pricing of data*.	- Short papers of up to 4 pages** - Study report superseded by an ensuing report of the same study**. - Customer or market maker's view of pricing instead of vendor's view*. - Costing, e.g. for cost minimization of data management*. - Units of analysis other than the pricing of data, e.g. market policies*. - Analyses of data value or other variables, rather than data pricing*.

Final set of primary studies. We did backward and forward snowballing by looking at the reference lists of the papers in the start set and by using Scopus to identify papers that cited the papers in the start set. The backward snowballing yielded 66 additional relevant papers. The forward snowballing yielded 6 additional papers that cited the start set. The small number was due to the inclusion of many recent papers in the start set.

We again screened the papers by studying their title and abstract and applying the same selection criteria. After screening, the final set of papers contained 18 papers.

Quality Assessment. We assessed the quality of the so far selected papers with the aim of including only those with research quality sufficient to extract data and answer our research questions reliably. Table 3 shows the quality assessment criteria that we derived from Kitchenham [9] and applied to the full text. Papers with a score of less than 0.6 got removed from further consideration, leaving us with 15 papers for the data extraction and analysis step.

Table 3. Quality assessment criteria.

Quality Criterion	**Assessment Question**	**Evaluation approach**	**Score**
Fulfillment of aims	How well does the research address its original aims?	Identify the aims from the abstract and introduction and compare with the research.	1.0: perfect match 0.5: partial or vague match 0.0: no match
Clarity of background	How clear are the underlying theory and assumptions?	Evaluate the background and related work sections if it fits the performed research.	1.0: well-defined and strong fit 0.5: partial fit 0.0: unclear or not fitting
Quality of the sample	How credible are the data that are used for the research?	Evaluate the data used for validating theories or models.	1.0: representative real-world data 0.5: data well described 0.0: unclear what data was used
Credibility of the research	How clear is the chain of evidence?	Evaluate the match between the method section, data, analysis, and analysis results.	1.0: clear and traceable 0.5: partial chain. 0.0: unclear chain of evidence.
Clarity of synthesis	How clear is the link of analysis results and the related work to the discussed contribution and implications?	Evaluate the traceability of the discussion to the presented results and background literature.	1.0: contribution and both traces clear 0.5: contribution vague or only one trace clear 0.0: no discussion or unclear connection with results and related work

Data Extraction. To answer our research questions, we extracted data with the data extraction form shown in Table 4. The table declares what we extracted, defines how we abstracted the extracts, and offers details about the data extraction.

Table 4. Data extraction form (*: values determined inductively)

Property	**Values**	**Description**
RQ1: Pricing Model Maturity		
Research method	Formal analysis, simulation, laboratory validation, real-world validation	The type of research method influences the readiness of the researched entity. E.g., the European Horizon2020 research program connects research methods[1] to technology readiness levels.

[1] https://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-g-trl_en.pdf

Property	Values	Description
Dataset	No data, synthetic data, synthetic data of justified industrial size, industrial data	The dataset used for analysis or validation influences the readiness of the researched entity. E.g., a synthetic dataset limits the credibility of the research results in comparison to the use of a full-scale industrial dataset.
RQ2.1: Contexts		
Domain	A vertical market like Smart City, Business Administration, or Linguistics.	Different verticals may have different norms, standards, and practices. Trading of data may need to take such contextual factors into consideration.
Type of data*	See column 'Type of Data' in Table 7.	Different types of data may require different types of pricing models to make data sharing attractive.
Data exploitation scenario*	See column 'Data Exploitation Scenario' in Table 7.	Different data exploitation scenarios may require distinct types of pricing models to make data sharing attractive.
Storage mechanism*	See column 'Storage' in Table 7.	Different types of data storage require different types of pricing models to make data sharing attractive.
RQ2.2: Data Products		
Market structure	Perfect competition, oligopoly, monopoly, monopsony	The number of sellers, intermediary market-makers, and buyers influences the market structure and the way the sellers and buyers behave [12].
Price metrics	Free, charging of single requests, volume packages, access to specific data-types, time-based subscription	The price metrics define the unit by which pricing is applied to data product [13]. We use the two taxonomies of metrics described by Muschalle [6] and by Sarkar [14].
Data quality attributes	Accuracy, completeness, time (currency, timeliness, volatility), consistency, other	Data quality is critical in any application using the data and in the processes supported by the data. Data quality may be characterized by a range of attributes [15].
RQ2.3: Pricing Models		
Aims of pricing model*	Internal consistency of pricing model, fairness of prices, profit maximization, social welfare maximization	To understand the rationales behind a pricing model, one must understand its aims.
Pricing model*	Price function with desired properties, game theoretical pricing approach	The categories and description of the pricing models.
Pricing variables*	Price of views, price of tuples, customer profile, data quality, customer bid, data usage, cost of the data	The variables used in the pricing model to determine a price.
RQ2.4: Pricing Mechanisms		
Price determination mechanism*	Algorithm, pricing function	The mechanism used by a party to determine the price for an offer of a data product.
Evaluation results*	Polynomial time (PTIME), Pseudo-PTIME, NP-Complete, N/A	The results of evaluating the pricing mechanism in terms of computational complexity.

Data Analysis. We followed the suggestions from Petersen [10] to systematically map the research literature and aggregate the results. Table 4, column "*Values*" describes categorization schemes that we used for classifying the papers. Our analysis focused on giving an overview of the categories and how common publications were for each category. This analysis made it possible to see which categories have been emphasized and which categories represent gaps in the research. Instead of bubble plots, we used tables and networks to give a visual representation of research focus and intensity.

Some values were not defined with a predefined categorization scheme. Here, we developed the categories inductively by following a conventional content analysis approach [16]. We let insights about categories emerge by studying the papers. We then gave an overview of these categories and defined their meaning with a synthesis of the relevant data extracted from the papers. The results represent a proposal of a categorization scheme that is grounded in the research that we have reviewed.

2.2 Threats to Validity

Kitchenham and Charters suggest the following four criteria for assessing the quality of a systematic literature review [9]: completeness of the literature search, clarity of paper inclusion, transparency of the study quality assessment, and adequacy of the description of the basic studies. These quality criteria were also used by tertiary studies to judge the quality a secondary study like this literature review, e.g. [17].

Our research process used a hybrid approach for literature search: keyword database search followed by snowballing. The combination of the two techniques allowed us to obtain a reasonable sample of the literature. The search efficiency of 6% is a figure that can be found in other literature reviews [8]. For increasing the confidence, one could further increase the start set of primary studies with a wider search string or validate the obtained set of papers with experts in the data marketplace and pricing domains. A consultation of experts could also give us insights about publication bias [9], about which we cannot make any statement with our research process.

We made explicit the inclusion and exclusion criteria that we applied. The criteria were discovered and documented during a pilot search as rationales for our inclusion and exclusion decisions. Inclusion and exclusion were decided by seeking consensus between the two authors. A limitation is that we applied the inclusion and exclusion criteria on titles and abstract only. Thus, we assumed that the authors succeeded to accurately reflect the contents of their papers in title and abstract.

For the study quality assessment, we used explicit rubrics with clear scoring instructions. The scoring results were developed, reviewed, and discussed by both authors and reflect the consensus of the two parties.

Due to the imposed space limitations, we could not offer a comprehensive description of each study. Instead, we decided to list the included papers in the appendix, enrich the analysis with syntheses of the data extracted from the papers, and established traceability of the syntheses to the source papers. This approach allows the reader to appreciate the overall meaning of the papers and obtain details by consulting the cited papers.

3 Results: Pricing of Data Markets

3.1 Quality Assessment

Most papers scored well in the quality assessment, yet no paper in the final set reached a score of 1.0. Of the well-scoring papers, all fulfilled the research aims and offered a clear overview of the research background.

The *quality of sample* and *clarity of synthesis* indicators were difficult to meet. The *quality of sample* indicator was difficult to meet because many papers used formal proofs instead of data for the evaluation or experimented with synthetic data. Few papers used real-world empirical data. *Clarity of synthesis* was hardly met because most papers offered only a limited synthesis of the obtained results with the rest of the literature. Table 5 gives an overview of the detailed scores.

Table 5. Quality assessment of the included studies (italics: papers scoring below 0.6).

Paper	**Assessment Score**	**Fulfillment of aims**	**Clarity of background**	**Quality of sample**	**Credibility of Research**	**Clarity of synthesis**
P04 Koutris 2015	**0.9**	1	1	0.5	1	1
P06 Kushal 2012	**0.9**	1	1	1	1	0.5
P09 Niyato 2016	**0.9**	1	1	1	1	0.5
P05 Koutris 2013	**0.8**	1	1	0.5	1	0.5
P08 Li 2014	**0.8**	1	1	0.5	1	0.5
P10 Stahl 2016	**0.8**	1	1	0.5	1	0.5
P11 Tang 2013 Get	**0.8**	1	1	0.5	1	0.5
P12 Tang 2013 Right	**0.8**	1	1	0.5	1	0.5
P13 Tang 2015	**0.8**	1	1	0.5	1	0.5
P01 Balasubramanian 2015	**0.7**	1	1	0	1	0.5
P02 Golrezaei 2014	**0.7**	1	1	0.5	0.5	0.5
P03 Jiang 2015	**0.7**	1	1	0	1	0.5
P07 Li 2012	**0.7**	1	1	0.5	0.5	0.5
P14 Tang 2016	**0.7**	1	1	0	1	0.5
P15 Wu 2010	**0.7**	1	1	0	1	0.5
P16 Balazinska 2013	***0.5***	*1*	*1*	*0*	*0.5*	*0*
P18 Shen 2016	***0.5***	*0.5*	*1*	*0.5*	*0*	*0.5*
P17 Shapiro 1998	***0.3***	*0*	*1*	*0*	*0.5*	*0*

Three papers scored below the threshold of 0.6 points: P16, P17, and P18. In addition to the two quality indicators that were difficult to meet overall, the three papers scored low in the *credibility of the research* and partially did not meet the stated research aims.

3.2 RQ1: Maturity of the Pricing Models

Most research was of conceptual nature and employed formal analysis or simulation of the proposed pricing models for validation. However, none of the pricing models has been validated in the real world or by deploying it in a laboratory environment. P06 was the only study which used real-world industrial data. P05 did a simulation with synthetic data of industrially relevant size. The other simulations used a random synthetic dataset or did not define the used data. Table 6 gives an overview.

Table 6. Maturity of the pricing models (top-left: low maturity, bottom-right: high maturity).

Dataset **Research method**	**No Data**	**Synthetic Data**	**Synthetic and Industrial Size**	**Industrial**
Formal Analysis	P01, P02, P04, P07, P08, P09, P10, P11, P13, P14, P15			
Simulation	P03	P12	P05	P06

3.3 RQ2: Pricing of Data

RQ2.1: Contexts Targeted by the Pricing Models. Table 7 gives an overview of the domains and types of data considered by the papers. While many domains were covered, some evident ones were missing. When using the Horizon2020 program as a reference[2], the domains of health and wellbeing, food and agriculture, and energy appear to be of relevance but were not considered.

Also, the data being traded and the scenarios of how these data would be exploited are broad. Four papers, P02, P03, P09, and P13, consider the use of sensor data, which could be generated in mobile sensing and Internet of Things contexts. One paper, P08, considers pricing for personal data, a type of data that is sensitive and subject to strict regulations. One paper, P09, considers the exploitation of data for machine learning, a basis for building systems that enable smart decision-making and control.

Eight papers are unspecific in the application domain or data exploitation scenario. For example, P12 just states that the data was intended for decision-making. The lack of specificity also means that the papers do not report any evaluation of their approaches or, in the case of P06, apply their pricing approach on a diversity of data as broad as demographics, weather imagery, DNA sequences, sales and marketing analytics, and financial records.

Table 7. Contexts.

Domain	**Type of Data**	**Paper**	**Storage**	**Data Exploitation Scenario**
Cities	Sensor data	P02	Cloud	Traffic and waste management
		P03	Edge	Environment management
		P13	Cloud	City management
Business management	Demographic data	P07	(not stated)	Financial assessments
	Personal data	P08	Cloud	Monetization
	(unspecific)	P14	Cloud	Market research and advertisement
		P12	Cloud	Decision-making
Engineering	(unspecific)	P01	Cloud	(no scenario defined)
Consumer	Newsfeed	P15	(not stated)	Social networking
Linguistics	Linguistic data	P05	Cloud	Text analysis and translation
(unspecific)	Sensor data	P09	(not stated)	Machine learning
		P06	Cloud	(no scenario defined)
		P10, P11, P14	(not stated)	(no scenario defined)

In most of the papers, the authors assume that data is uploaded to the market maker's cloud for making that data available for trade. Such upload may be efficient for the market maker but could reduce transparency and control of the transactions for the data vendor. One paper assumed the opposite approach, edge computing, in which the data is controlled by the data vendor. Six papers did not state any assumption about where data would be stored.

[2] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges

RQ2.2: Data Products being Priced. The papers covered a broad variety of product definitions. Table 8 gives an overview. Many papers assumed, explicitly or implicitly, a monopoly market structure where the data provider does not care about competing providers. We judged a paper to consider a monopoly implicitly if it assumed that the offered product is so far differentiated that the pricing model does not need to consider competing offerings. Four papers considered a duopoly situation where two data providers compete. No paper generalized a duopoly to an oligopoly situation. Two papers, P05 and P08, considered a monopsony situation, where a buyer requests data from many data providers. We judged a paper to consider a monopsony if the pricing did not consider interactions between multiple customers. Only one paper studied a market situation with a perfect competition where anybody could trade with anybody.

Most papers studied data products with usage- or request-based price metrics, where charging takes place on a fine-grained level. The variants were pay-per-use or unit, pay-per-query or view, or customer-proposed prices. One volume-based pricing model was investigated, step pricing where the customer pays for a given volume of data. Three papers studied flat fee products that allow all data to be accessed without restriction, either continuously or as part of a time-based subscription. The papers P02 and P03 did not state any price metrics used to define the data product.

Table 8. Data product definitions (*: papers comparing multiple products).

Price Metrics **Market**	**Single Requests**	**Volume Packages**	**Time-based Subscription**	**(not stated)**
Monopoly	P01*, P15*: pay-per-use P06*: per-unit P04, P07, P12, P13: query- or view-based P10, P11, P14: customer-proposed price	P06*: step pricing	P01*: unrestricted use P09: subscription fees P15*: flat fee	P02
Duopoly	P01*: pay-per-use P06*: per-unit	P06*: step pricing	P01*: unrestricted use	P02
Oligopoly				
Monopsony	P05, P08: query-based			
Perfect Competition				P03

Three papers compared the attractiveness of usage and flat fee products, P01 and P06 in both monopoly and duopoly market structures, and P15 in a monopoly market structure alone.

A subset of the papers utilized quality in the product definition and, consequently, as an attribute for pricing. Table 9 gives an overview.

Table 9. Quality attributes used in the product definition and pricing model.

Quality Attribute	**Paper**	**Quality Metrics**
Time	P02	Delay: Delay may influence the perceived value of a data product.
	P05	Aging: Data may need to be updated because it gets incorrect over time.
	P10	Freshness: a price should be defined depending on how new the data is.
Accuracy	P08	Perturbations: noise for deteriorating aggregated data quality for privacy.
	P11	Accuracy: distance and likelihood of deviation from the true value.
Completeness	P10	Completeness: parts of the data may be missing.
	P14	Completeness: incompleteness may be traded for discounted prices.
Consistency	-	-

In these papers, quality played a role in price setting, delivering value, and managing privacy. Quality differences may influence a customer's perceived value of a data product. Thus, reduced quality was a counterpart for price reductions: "you pay what you get." Also, quality was considered to deteriorate over time. Thus, data needed to be updated to be of high value or prices be reduced. Quality, finally, was a trade-off with privacy. Perturbations were introduced into the data to avoid unwanted disclosure of information. Alternatively, price increases were used to compensate for disclosure.

RQ2.3: Pricing Models. We identified three approaches to researching pricing models. Some papers designed a price function with desired properties. Most of these papers addressed a single-vendor situation (monopoly). Other papers casted pricing into game theory to identify an optimal pricing approach in a competitive situation. Most of these papers addressed a multi-vendor situation (duopoly and monopsony). A final set of papers compared constellations of price metrics and market to select pricing approaches. Most of these papers addressed both, single-vendor and multi-vendor situations.

Fig. 1 gives an overview of the pricing models for the *single-vendor situation*. We used the function symbol to depict papers designing a price function. The dice symbol was used to denote a game-theoretic analysis.

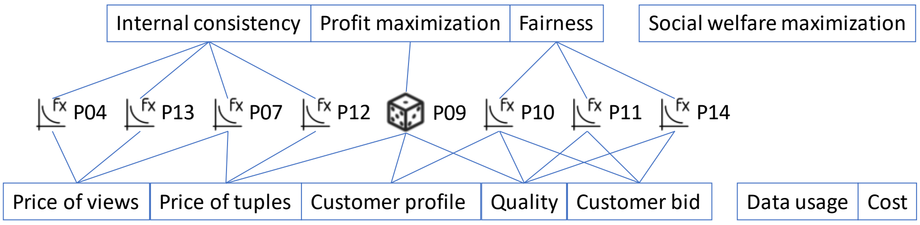

Fig. 1. Papers researching pricing models for single-vendor situations.

The papers proposed and evaluated pricing models for achieving internal consistency of the pricing function, profit maximization, and fairness between customers and vendors. *Internal consistency* meant monotonicity of the pricing function (i.e. higher prices mean more data), usage or volume-based prices are not higher than the price of the whole database, non-disclosiveness (i.e. impossible to infer unpaid query answers), and freedom from arbitrage (i.e. all ways to obtain an insight have the same price), freedom from discounts (i.e. the prices are maximal), and freedom from regret (i.e. all sequences to obtain an insight have the same price). *Profit maximization* meant pricing models that maximized the data vendor's profitability. *Fairness* meant a fair trade-off between quality and price.

Within the single-vendor context, three groups of pricing models could be discerned: customer bid-based pricing, view-based pricing, and tuple-based pricing. *Customer bids* were answered by compensating low bids with the delivery of low-quality data. The compensation was motivated by the customers' understanding that with just a little money only low quality can be bought. For the vendors, the compensation was an aspect of fairness. *View-based pricing*, a variant of usage-based pricing, was based on the idea that the customers' queries could be answered with predefined data views that are stored in the vendor's database. P07 called this approach *deductive pricing*. The price for a query is the price of the cheapest set of views needed to answer the query. *Tuple-based pricing* is another variant of usage-based pricing. Its idea is to charge access to rows in a database. P07 called this approach also *inductive pricing*.

Fig. 2 gives an overview of the pricing models for *multi-vendor situations*. Again, the same symbols were used for price function designs and game-theoretic analyses.

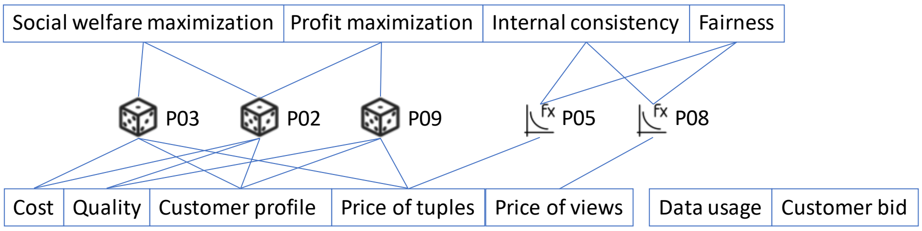

Fig. 2. Papers researching pricing models for multi-vendor situations.

The papers proposed and evaluated pricing models for the additional goal of maximizing social welfare as well as the already mentioned goals of profit maximization, internal consistency, and fairness goals. *Social welfare maximization* meant to maximize the sum of all customers' payoffs. *Profit maximizations* and *internal consistency* had the same meaning as before. *Fairness* meant now a fair split of revenue among sellers.

Two groups of pricing models could be discerned: pricing models that aimed at internal consistency and fairness, and game-theoretic approaches for maximizing social welfare or profit. The *design of the pricing functions* resembled the view- and tuple-based pricing models studied in the monopolistic context, but now extended with a mechanism to fairly compensate a multitude of sources for the data they provided. The *game-theoretic approaches* allowed parties to decide about the role they wanted to adopt in the marketplace, how pricing tactics would affect the equilibria in the market, and how to compute the optimal price.

Fig. 3 gives an overview of pricing model comparisons. We used the tick-box symbol to denote these papers that aimed at offering decision support for selecting an appropriate pricing model. The shaded dices and function indicate secondary contributions of the papers. For example, P01 used game theory to study the duopoly situation.

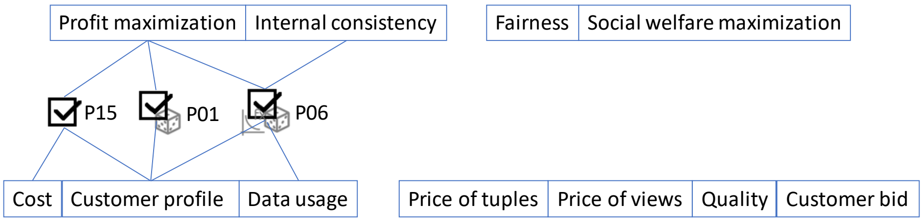

Fig. 3. Papers comparing pricing models.

All three papers compared pricing approaches with the aim of profit maximization. P01 and P06 made this comparison for both the monopoly and duopoly situations. P15 did it for the monopoly situation only. P06 studied arbitrage for the pricing function, thus pursued the secondary goal of achieving internal consistency of the pricing.

According to the three papers, a monopoly requires a different approach for a good product definition than a duopoly. For a monopoly, the papers conclude that usage-based pricing of data is attractive. The pricing may be fine-grained or package-based; the granularity of the price steps does not matter. Thus, no clear preference can be established between usage or volume-based pricing. In a duopoly, the two competing data vendors should offer complementary product definitions, or the profits will rapidly erode to zero.

RQ2.4: Price Determination. Most papers proposed equations or algorithms to calculate prices. Only in P01, P02, and P03, we could not identify any specific price determination mechanisms that could be used by a data vendor. Table 10 gives an overview.

Table 10. Price determination mechanisms for data vendors.

Price Determination	Paper	Specific Result	Complexity
Algorithm	P04	Pricing for chain queries	PTIME
		Pricing for cyclic queries	PTIME
	P06	Multi-step pricing	PTIME
	P07	Cell-based or regret-free inductive pricing	PTIME
		Deductive pricing for continuous price functions	PTIME
	P10	Knapsack pricing	Pseudo-PTIME
	P12	Approximate pricing	PTIME
	P13	Rewriting-based pricing	NP-Complete
	P11	Fair quality distortion	N/A
	P14	Uniform or binary tree sampling	PTIME
Pricing functions	P08	Basic and synthesized pricing	NP-Complete
	P09	Globally optimal pricing	N/A
	P15	Vendor's generic optimization problem	N/A
	P05	ILP-formulation for some conjunctive queries	PTIME

P04, P05, P07, P12, P13, and P14 suggested that pricing is NP-complete in general. Also, P8 suggests that consistency checking of arbitrary price point setting is NP-complete. However, as Table 10 shows, algorithms for specific cases may be designed that are less complex and offer tractable pricing. The algorithms that may be executed in polynomial time (PTIME or Pseudo-PTIME) were considered tractable. Some of the pricing functions may be formulated so that they can be solved as differential equations or by a solver, e.g. an Integer Linear Programming (ILP)-Solver.

4 Discussion

This paper has contributed the first systematic review of research on pricing for data products, thus helping to enable business with the massive amounts of data generated by Mobile Computing and the Internet of Things. Fifteen papers were analyzed that proposed or evaluated pricing models for data product vendors. While earlier work has introduced pricing metrics [14] as well as the structure of a marketplace and its participants [6], no systematic overview had been given of the models and mechanisms used for pricing. The here presented research enables marketplace owners and data vendors to plan how to generate revenue and profit from data. Such thinking is important to make the potentially vast amount of data created by billions of humans and devices available for the development of smart systems and services.

Section 3 gave an overview of the objectives for pricing data products and the attributes that could be considered as inputs for a pricing model. The results suggest that data vendors seek profit maximization and consistency of the pricing model. Further concerns are social welfare and fairness. Some pricing attributes could be used for value-oriented pricing [13] and cover the customer profile, the data usage, and customer bids. Other attributes consider the cost side of data and include the cost of data provision and the price of tuples or views. A special role plays quality of the data that, according to the reviewed research, is a means acceptable for customers to relate to prices.

The here presented research also identified concrete advice on how to act when discovering a competitive situation (i.e. achieving complementarity of product definitions) and what the pricing models are that should be preferred when offering unique data (i.e. usage-based pricing rather than a flat subscription fee).

While the pricing models are appealing from a conceptual point-of-view, the calculation of prices remains challenging. Good pricing model should exhibit a variety of characteristics, such as monotonicity, boundedness, non-disclosiveness, and freedom from arbitrage, discount, and regret. Price determination is NP-complete in general. Only for special cases, approaches of polynomial complexity were proposed.

We have constrained our review to papers that discuss pricing of data products for use in data marketplaces from a vendor's perspective. This strict scope excluded studies that focused purely on value and cost of data without having used these attributes for pricing the data. Also excluded were papers that studied the data consumer's or market maker's perspective, e.g. of procuring data at a minimal cost. Future reviews should expand towards value and cost aspects of data including the customers' view.

Research on pricing for data products is still in its infancy. Most research we identified features microeconomic modeling and formal analysis of the pricing models. When using the 9-level European Horizoon2020 technology readiness (TRL) model as a benchmark, such research is positioned at TRL2 only. Four papers went as far as TRL3 by offering a simulation-based evaluation of the pricing models. With our search, we could not identify any paper at a higher TRL that would have reported applications of pricing in relevant environments. This disconnect of research from practice is surprising as several data marketplaces have been launched (c.f. Section 1) and are confronted with pricing questions. Real-world research is urgently needed to understand the applicability and impact of the pricing models. The work of Schomm could represent a starting point and offer guidance for such practical applications [3].

Also, the surveyed pricing models were developed for simple market situations only. Considered were the monopoly where competition could be ignored and the duopoly where competition is a gameplay between two adversaries. Such simplification is attractive because it makes formal analysis feasible. From a practical perspective, it would be important to understand how to design and differentiate data products to make the offering so unique that it could be considered a monopoly or at least complementary to existing products. Software product management offers such product strategy advice for software products [7]. It would be interesting to understand whether and how such advice can be transferred and applied to data products.

5 Summary and Conclusions

This paper has offered a systematic review of the literature on pricing models for data marketplaces. The papers were identified first with a keyword-based search in Scopus and then complemented with forward and backward snowballing. From initially 181 papers 11 papers were selected for snowballing. The snowballing step yielded 18 papers that were assessed for research quality. 15 papers made it in the final set of papers.

11 papers offered formal analysis of pricing models, while 4 additional papers went as far as simulating the formal models. Cities, business management, engineering, consumer, and linguistics were the contexts addressed by the pricing models. Usage-based, volume-based, and flat fee pricing models were proposed or evaluated for single-vendor and multi-vendor situations. The pricing models aimed at profit maximization, internal consistency, fairness, and social welfare maximization. Pricing attributes included customer bids and profile, data usage, quality, the price of views or tuples, and cost. Price calculation is NP-hard with PTIME approaches existing for special cases.

Our results offer an overview of what in the domain of data pricing has been researched and where the gaps are. It serves as a compact advice for anybody who seeks incentives and rewards for data sharing. However, the presented results should be used with caution. Research is needed to validate the models in the laboratory and real-world settings.

Acknowledgments. The presented work was funded by the European Union’s Horizon 2020 research and innovation program under grant agreement No 732204 (Bonseyes) and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 16.0159. The opinions expressed and arguments employed herein do not necessarily reflect the official views of these funding bodies.

Bibliography

1. Atzori, L., Iera, A., and Morabito, G. (2010) The Internet of Things: A Survey. Computer Networks 54(15):2787-2805.
2. LeCun, Y., Bengio, Y., and Hinto, G. (2015) Deep Learning. Nature 521(7553):436-444.
3. Schomm, F., Stahl, F., and Vossen, G. (2013) Marketplaces for Data: An Initial Survey. ACM SIGMOD Record 42(1):15-26.
4. Schwab, K., et al. (2011) Personal Data: The Emergence of a New Asset Class. World Economic Forum.
5. Koutsopoulos, I., Gionis, A., and Halkidi, M. (2015) Auctioning Data for Learning, in IEEE 15th International Conference on Data Mining Workshops, Sydney, Australia.

6. Muschalle, A., Stahl, F., Löser, A., Vossen, G. (2012) Pricing Approaches for Data Markets, in International Workshop on Business Intelligence for the Real-Time Enterprise.
7. Kittlaus, H.-B. and Clough, P. (2009) Software Product Management and Pricing. Springer.
8. Wohlin, C. (2013) Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering, in 18th International Conference on Evaluation and Assessment in Software Engineering.
9. Kitchenham, B. and Charters, S. (2007) Guidelines for Performing Systematic Literature Reviews in Software Engineering, in EBSE Technical Report.
10. Petersen, K., Feldt, R., Mujtaba, S., Mattson, M. (2008) Systematic mapping studies in software engineering, in 12th International Conference on Evaluation and Assessment in Software Engineering.
11. Ivarsson, M. and Gorschek, T. (2009) Technology Transfer Decision Support in Requirements Engineering Research: A Systematic Review of REj. Requirements Engineering 14(3):155-175.
12. Frank, R. and Cartwright, E. (2013) Microeconomics and Behaviour. McGraw-Hill Education.
13. Nagle, T.T. and Hogan, J.E. (2006) The Strategy and Tactics of Pricing: A Guide to Growing More Profitably. Pearson Prentice Hall.
14. Sarkar, P. (2015) Data as a Service - Framework for Providing Re-Usable Enterprise Data Services. John Wiley & Sons.
15. Battini, C. and Scannapieco, M. (2010) Data Quality: Concepts, Methodologies and Techniques. Springer.
16. Hsieh, H.F. and Shannon, S.E. (2005) Three approaches to qualitative content analysis. Qualitative health research 15(9):1277–1288.
17. Nurdiani, I., Börstler, J., and Fricker, S. (2016) The Impact of Agile and Lean Practices on Project Constraints: A Tertiary Study. Journal of Systems and Software 119:162-183.

Appendix: Bibliography of Included Papers

Full version: https://dl.dropboxusercontent.com/u/21095508/FrickerMaksimov_IncludedPapers.pdf

Shortened overview: **P01** Balasubramanian, S., et al (2015) Pricing information goods: A strategic analysis of the selling and pay-per-use mechanisms. **P02** Golrezaei, N., & Nazerzadeh, H. (2014) Pricing Schemes for Metropolitan Traffic Data Markets. **P03** Jiang, C., et al (2015) Economics of peer-to-peer mobile crowdsensing. **P04** Koutris, P. et al (2015) Query-based data pricing. **P05** Koutris, P. et al (2013) Toward practical query pricing with QueryMarket. **P06** Kushal, A. et al (2012) Pricing for data markets. **P07** Li, C., & Miklau, G. (2012) Pricing Aggregate Queries in a Data Marketplace. **P08** Li, C. et al (2014) A theory of pricing private data. **P09** Niyato, D. et al (2016) Market model and optimal pricing scheme of big data and Internet of Things (IoT). **P10** Stahl, F., & Vossen, G. (2016) Fair Knapsack Pricing for Data Marketplaces. **P11** Tang, R. et al (2013) What you pay for is what you get. **P12** Tang, R. et al (2013) The price is right. **P13** Tang, R. et al (2015) Valuating Queries for Data Trading in Modern Cities. **P14** Tang, R. et al (2016) A Framework for Sampling-Based XML Data Pricing. **P15** Wu, S. Y. et al (2010) Best pricing strategy for information services. **P16** Balazinska, M., et al (2013) A discussion on pricing relational data. **P17** Shapiro, C., & Varian, H. R. (1998) Versioning: the smart way to sell information. **P18** Shen, Y. (2016) A pricing model for Big Personal Data.