Tsinghua Science and Technology

Volume 21 | Issue 5 Article 2

2016

A Pricing Model for Big Personal Data

Yuncheng Shen
College of Computer Science, Sichuan University, Chengdu 610065, China.

Bing Guo
College of Computer Science, Sichuan University, Chengdu 610065, China.

Yan Shen
School of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China.

Xuliang Duan
College of Computer Science, Sichuan University, Chengdu 610065, China.

Xiangqian Dong
College of Computer Science, Sichuan University, Chengdu 610065, China.

See next page for additional authors

A Pricing Model for Big Personal Data

Authors

Yuncheng Shen, Bing Guo, Yan Shen, Xuliang Duan, Xiangqian Dong, and Hong Zhang

TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 02/11 pp482-490
Volume 21, Number 5, October 2016

A Pricing Model for Big Personal Data

Yuncheng Shen, Bing Guo*, Yan Shen*, Xuliang Duan, Xiangqian Dong, and Hong Zhang

Abstract: Big Personal Data is growing explosively. Consequently, an increasing number of internet users are drowning in a sea of data. Big Personal Data has enormous commercial value; it is a new kind of data asset. An urgent problem has thus arisen in the data market: How to price Big Personal Data fairly and reasonably. This paper proposes a pricing model for Big Personal Data based on tuple granularity, with the help of comparative analysis of existing data pricing models and strategies. This model is put forward to implement positive rating and reverse pricing for Big Personal Data by investigating data attributes that affect data value, and analyzing how the value of data tuples varies with information entropy, weight value, data reference index, cost, and other factors. The model can be adjusted dynamically according to these parameters. With increases in data scale, reductions in its cost, and improvements in its quality, Big Personal Data users can thereby obtain greater benefits.

Key words: data tuple; Big Personal Data; positive grading; reverse pricing; pricing model

1 Introduction

As the great value of big data has been recognized and computer memory costs have declined, collection of personal information is reaching unprecedented levels. The economic value of data reflects the fact that many Internet companies benefit from search engines, social media sites, and sale of information gathered through them. Due to privacy concerns, large amounts of potentially useful private data cannot be accessed by stakeholders[1]. Monetizing private data is an improvement to the narrow view of data confidentiality, because it can enable individuals to control their own data by financial means.

• Yuncheng Shen, Bing Guo, Xuliang Duan, Xiangqian Dong, and Hong Zhang are with College of Computer Science, Sichuan University, Chengdu 610065, China. E-mail: guobing@scu.edu.cn.
• Yuncheng Shen is also with College of Information Science and Technology, Zhaotong University, Zhaotong 657000, China. E-mail: 403953413@qq.com.
• Yan Shen is with School of Control Engineering, Chengdu University of Information Technology, Chengdu 610225, China. E-mail: sheny@cuit.edu.cn.
∗ To whom correspondence should be addressed.
Manuscript received: 2016-07-23; revised: 2016-08-03; accepted: 2016-08-22

Personal data is assumed to be the "energy" or "new money" of digital world. The authors in Ref. [2] presented a "user-centric" model, which aims at unlocking such potential, by enabling individuals to control the collection, management, use, and sharing of their own data. It analyzes a new personal data ecosystem centered around the role of "Bank of Individuals' Data" (BID), a provider of "personal data management services" enabling people to exploit their personal data, and defines how personal data are revealed through the use of third-party trust organizations. Personal data has often-inconsistent value for the data owner and organizations that attempt to analyze it. In Ref. [3], the authors designed an effective technical approach to negotiate these competing benefits. The authors in Ref. [4] dealed with the use of digital technology by making the individual become a provider and co-creator of a service and product in an economic system. However, current data products, pricing, and trading mechanisms almost completely bypass the ultimate user, and do not allow stakeholders to provide services by taking advantage of data. In an ideal situation, the Big Personal Data market represents a virtual or physical trading space, where users provide personal data to goods and/or service providers; the providers offer money and/or non-monetary individualized products to users. Data

quality is completely observable and known to all participants in the data market.

This paper proposes a pricing model of positive grading and reverse pricing for Big Personal Data based on tuple granularity. An individual submits personal data to a safe and reliable data trading platform, which processes and reproduces data through cloud computing and sells it to customers. The data demand price less the data cost is the supply price for personal data. To further embody the value of personal data, positive grading and reverse pricing of the data supply price are applied. The individual, the data exchange platform, and the data customer can all benefit from trading data, and achieve a mutually beneficial data transaction ecosystem.

2 Related Work

Researchers in Ref. [5] observed main pricing strategies, including free pricing, usage-based prices, package pricing, flat-fee tariffs, two-part tariffs, and freemium. The authors in Ref. [6] said there are four weaknesses with existing market pricing models: per-query costs are irrelevant; the pricing model can inadvertently lead to arbitrage situations; all tuples have equal value and data providers have no principled way to set the pricing tiers; and the systems provide no guidance. In Ref. [7], the authors reported that information product prices differ greatly from those of tangible goods, and put forward a new theory of pricing information products based on the concept of version. Researchers in Ref. [8] proposed a pricing model that charges for a single query, allowing the seller to set a clear price for few views. The authors in Ref. [9] proposed the concept of the origin of the minimum, which regards the query origin as a whole. Due to the nature of the information product, payment to a tuple occurs at most once, no matter how many times it contributes to the query results. In Ref. [10], the authors proposed the concept of view in the data market, which amounts to a version of the information product. In Ref. [11], researchers thought buyers should have access to unbiased samples of private data in a reasonable data market, and properly compensate individuals according to the privacy attitude of the seller (individual). Researchers in Ref. [12] used an auction approach to sell private data, where the data are either completely hidden or fully disclosed; in either case, the data price is determined by the buyers, and does not take into account the value of personal privacy to the data owner. The authors in Ref. [13] proposed a pricing theory framework according to noise query responses, which divides the price among the data owners, offers due compensation in terms of the loss of their privacy, and points out that privacy valuations may be strongly related to the data itself. In Ref. [14], the authors indicated that the data owner privacy valuation is very complex and difficult to express, and is different from individual to individual. In fact, if there is no specific context or reference, people will find it hard to understand their private data.

In the current personal data trading market, little transparency exists between buyers and sellers regarding how data has been collected and manipulated prior to sale, and how it will be used post-sale. This is in part a competitive strategy for companies, but it can hinder the market. This lack of transparency leads parties involved in the transaction to be misinformed and results in asymmetric information. Thus, we chose to study personal data valuation in order to propose a rigorous and transparent pricing model to enhance the personal data trading market and lessen the likelihood of the "lemon" market asymmetry.

If a standard model for data pricing existed, one that considered many aspects of value, such as the age of the data, the reliability of the sample, and other factors, sellers would be able to price optimally in the market and buyers could make appropriate comparisons across data service providers to get a fair price. If the personal data trading market adopted some of these valuation strategies and standardized a pricing model, the transaction experience for all parties would improve drastically and facilitate more efficient and effective data science.

It is important to recognize previous research in this area. Moody and Walsh[15] addressed the subject of asset valuation of information. They viewed data as a raw material, information systems as the manufacturer, and information as the end product requiring valuation. This paper addresses valuation of data itself rather than focusing on the even more abstract concept of information. This should prove more useful, as the distinction between "information" and "data" often lies in their use, rather than in inherent properties.

3 Description of Big Personal Data

Data is usually the combination of private and sensitive individual data and public business data. It is

necessary to distinguish personal data from statistical and research data that does not involve any private data. Big Personal Data refers to characteristic individual behavior data generally considered private, which is produced in personal life activities or work, and can be owned or controlled by an individual. Big Personal Data has complicated sources and various forms, including basic personal information, personal income, personal property, personal friends, personal health, personal education, personal diaries, personal documents, personal views, and personal perception information. Big Personal Data is important, and its commercial value and data value tend to be underestimated by the owning individual.

3.1 Research justication

The potential impact of constructing a functional pricing model can be realized by examining how this problem is similar to a pricing issue that evolved in traditional financial markets. Black-Scholes defines a stochastic partial differential equation that calculates the theoretical price of an option over time. The model incorporates various factors, including the current value, returns, and volatility of the underlying asset; the strike price and time to expiration of the contract; and the prevailing risk-free interest rate. As either volatility or the time to expiration decreases, the value of the option declines. While not entirely analogous, personal data valuation has similarities to this pricing model, as it is determined by a complex interaction of multiple factors, which could include both a concept of volatility and time decay. A generalizable scientifically rigorous approach to pricing personal data would likewise help to legitimize and standardize the trading market for personal data.

Another impetus for this research is the growing need to value personal data as a data asset. Assessing intangible value is not a new challenge for business. Existing valuation approaches for intangibles like patents and data include cost-based methods, which attempt to determine the expense of generating or replacing an asset, and market-based methods, which rely on previous market transactions of comparable assets. Both methods are unsatisfying because they do not directly assess the value of the asset itself and are subject to externalities such as market fluctuations. It is necessary to develop a model that more directly assesses the intrinsic value of personal data and addresses the use and sale of personal data between parties.

3.2 Attribute selection

There are a number of data characteristics that affect the value of data. With the ultimate goal of identifying a model that can be used to price data in an open market, we examined how other digital assets are traded. This included the pricing strategy for digital media (audio, images, videos), licensing fees for intellectual property assets and patents, pricing variables used for software-as-a-service products, and techniques from software engineering for estimation and pricing. Based on this examination, we identified a set of candidate parameters, falling into three main categories, which could help determine the value of data as follows.

(1) Value-based parameters

- The value of the data in terms of saving time, effort, or money;
- The Return On Investment (ROI) for the customer (or a profit share arrangement with the customer based on the profit derived from the acquired data);
- Risk exposure — Data cleansed of personally identifiable information and privacy violations could be priced higher;
- Data exclusivity — Whether the data is provided on an exclusive basis, nonexclusive basis, or some combination of these two can influence price;
- Level of ownership — Is the customer buying (implying transfer of ownership), leasing (allowing use for a fixed time) or licensing (allowing limited use for a specific purpose)?

(2) Qualitative parameters

- Age of the data;
- Credibility of the data;
- Accuracy of the data elements;
- Quality of the data — Missing fields for certain rows, incorrect types, data precision, etc.;
- Format and level of structure of the data — Plain text, streaming data, tabular datasets, etc.

(3) Fixed and marginal cost parameters

- Cost of collecting the data;
- Cost of data storage, bandwidth, and other operational costs;
- Cost of data-as-a-service offerings — Add-on services to process the data, computing resources for the data, analytic reports, or aggregation on the data;
- Delivery cadence — One-time, batch, or continuous basis.

Currently, the market value of data is mostly

determined through value-based parameters, which are difficult to quantify and model. While it is possible to use the value-based parameters to command a premium price for the data, it will become necessary to move to a set of parameters that can be measured and modeled.

3.3 Data tuples parameter value

This paper takes the data tuple as the basic unit of a data metric, using it to assess the value of personal data in order to calibrate its price in the data market. Combining previous research works and our survey and research, we can determine the parameters that affect the value of a data tuple. These are data cost, value weight, information entropy, credit rating, and data reference index.

(1) Data cost

The product cost refers to all kinds of cost to an enterprise to produce products; these are composed of fixed costs and variable costs[16]. A trading platform collects, organizes, and analyzes data, then forms the final data products to trade with the customer. The cost of a data product consists of a fixed cost and a marginal cost. As the fixed cost of a data product is low, it can be ignored when data expand. Thus, the cost of product data mainly refers to the cost of producing data in a trading platform, which can be easily determined by the marginal cost of producing, storing, and sharing data.

(2) Value weight

As for a data tuple, its value weight has a positive correlation with its value quantity, as well as its price. In order to accurately reflect the value of each data tuple, it is important to set an attribute known as value weight. The greater the weight, the higher its value. So it can embody the value of different tuples.

(3) Information entropy

According to Shannon, information entropy is a probability distribution function, depicting the uncertainty. The entropy of a certainty event is zero. The more uncertain an event, the greater the entropy, and the higher the value. Value and information content are positively correlated. The value goes up as the information content is enriched.

(4) Credit rating

The higher a personal credit rating, the higher the credibility provided by data; the higher the quality of data, the higher its value.

(5) Data reference index

The more data tuples provided by an individual, the more data tuples are referenced (sold), the higher the data reference index, and the greater its data value.

4 Research Design

Personal data refers to basic individual data in this paper, that is, an individual's raw unprocessed data. For convenience of description, the data seller (provider) is referred to as the user. Because this paper discusses personal data pricing, "user" refers to a personal user. It is assumed that the basic sales unit is a data packet, which consists of n data tuples.

4.1 Information entropy

According to Shannon, what can reduce the uncertainty in a given case is called information, and information content is the amount by which the degree of uncertainty can be reduced. Here information content is a relative amount, related to the possibility of things happening; in other words, information content equals the probability of the logarithm of the probability to choose.

If X is a discrete random variable, the value space is $\mathbb{R}$, the probability distribution is $p(x) = P(X = x), x \in \mathbb{R}$. Then the entropy of X, $H(X)$, is defined as

$$H(X) = -\sum_{x\in\mathbb{R}} p(x)\log_2 p(x) \tag{1}$$

where $0\log 0 = 0$ is appointed. $H(X)$ can be written as $H(P)$. For the base of the logarithm is 2 in the definition, the unit of entropy defined by Eq. (1) is binary. Usually $\log_2(P(X))$ is abbreviated as $\log(P(X))$.

Entropy is also known as self-information, which describes the uncertainty measure of a random variable. It expresses the average information provided by information source X, which sends a symbol (no matter what symbol)[17]. The greater the entropy of a random variable, the greater its uncertainty, and the smaller the possibility of correctly estimating its value. The greater the uncertainty of a random variable, the larger the information content that is used to determine its value.

Suppose a data packet has n data tuples, and each data tuple has k attributes, x_{ij} expresses the j-th data item (data attribute values) of i-th data tuple. $p(x_{ij})$ expresses the probability that x_{ij} appears in the packet. The probability of the j-th data item of the i-th data tuple in a data packet can be calculated by the following formula:

$$p(x_{ij}) = \frac{\text{Number of occurrences of } x_{ij}}{\text{Total number of packets tuples } (n)} \tag{2}$$

The entropy of the i-th data tuple in a data packet can

be calculated by the following formula:

$$H(x_i) = -\sum_{j=1}^{k} p(x_{ij}) \log_2 p(x_{ij}) \tag{3}$$

The total entropy of all data tuples in a data packet is

$$H(X) = \sum_{i=1}^{n} H(x_i) = -\sum_{i=1}^{n}\sum_{j=1}^{k} p(x_{ij}) \log_2 p(x_{ij}) \tag{4}$$

According to Shannon's definition, the greater the entropy of a random variable, the greater the uncertainty, and the greater the information content. So the size of the entropy represents the size of the information content. Calculate the entropy of the i-th data tuple $H(x_i)$ according to Eq. (3). Let q_i denote the size of information content of the i-th data tuple, and let $H(x_i)$ be assigned to q_i. Calculate the entropy of the entire data packet $H(X)$ according to Eq. (4). Let q express the size of the information content of the data packet, and $H(X)$ be assigned to q. The weight of information content of the i-th data tuple is determined by $\frac{q_i}{q}$, and satisfies the following constraints:

$$\sum_{i=1}^{n} \frac{q_i}{q} = 1 \tag{5}$$

4.2 Data reference index

The data reference index is derived from "H-index". The H-index was suggested in 2005 by Jorge E. Hirsch, a physicist at UCSD, as a tool for determining theoretical physicists' relative quality and is sometimes called the Hirsch index or Hirsch number. The definition of H-index is that a scholar with an index of h has published h papers, each of which has been cited in other papers at least h times. Thus, the H-index reflects both the number of publications and the number of citations per publication. Using H-index to measure the authority of publication, the greater the H-index, the more the paper is cited, the more citation times are. It measures the authority of publication from paper amount and citation times at the same time. According to the H-index ranking for a publication, the larger the H-index, the higher the ranking.

With reference to the definition of H-index, users are equivalent to publication, so a data tuple of user equals to a paper of publication.

Definition 1 Data reference index refers that at least r data tuples is purchased r times respectively, the maximum is called user data reference index, shortened as "R-index".

With R-index measuring the authority of user, the greater the R-index, the more data tuple is bought, the more times it is purchased. It measures the authority of user from the purchase amount and purchase times of data tuple at the same time. According to the R-index ranking for a user, the larger the R-index, the higher the ranking.

Assume data packet contain m users, the R-index value of the j-th user is r_j, the sum of R-index value of all users in data packet is r. Then the weight of R-index of j-th user is determined by $\frac{r_j}{r}$, and satisfies the following constraints:

$$\sum_{j=1}^{m} \frac{r_j}{r} = 1 \tag{6}$$

4.3 Value weight

For personal data, data classification (data table) is regarded as the unit. First, weight value is set according to experience, and weight value falls into n levels, that is from 1 to n. The greater the value weight is, the more important data is. Assume the value weight of the i-th data tuple is w_i, the sum of value weight of all data tuples in data packet is w. Then the weight of value weight of data tuple is determined by $\frac{w_i}{w}$, and satisfies the following constraints:

$$\sum_{i=1}^{n} \frac{w_i}{w} = 1 \tag{7}$$

4.4 Pricing model

Creating a universal model for all data types would be a monumental task, and data sources may require different pricing models, based both on the type of data and its potential uses. The development of the model would require further exploration of objective independent variables that could have a relation to data value, some of which have been outlined above. Additionally, an appropriate number of sample datasets with their prices and attributes would have to be collected as inputs to the model, ideally ranging from large to small, spanning multiple uses.

We propose a positive grading and reverse pricing model of Big Personal Data based on tuple granularity. Here we define the term positive grading and reverse pricing respectively.

Definition 2 Positive grading refers that data attribute is divided into different factors according to

the importance affecting data quality.

Definition 3 Reverse pricing refers that the exact price of each data duple is calculated according to the factor of data attribute and the supply price of a data packet.

The higher the quality of personal data, the higher supply price users will ask for supplying the data and the higher demand price buyers will be willing to pay for the data. The data exchange platform can analyze and convert data for the buyer, and can reduce data uncertainty thereby and improve data quality.

Suppose the demand price of a data packet is P_{D}, and the cost to collect, analyze, and share trading platform data is C. Then the supply price of a data packet, P_{S}, can be obtained by the following formula:

$$P_{\mathrm{S}} = P_{\mathrm{D}} - C \tag{8}$$

The demand price minus the data cost is the supply price. Then we apply the reverse pricing method to price a data tuple at a fine-grained level. In order to encourage individuals to move from being passive data "sellers" into "data operators", and to encourage individuals to consistently maintain and update the data to improve its quality, it is essential to create an incentive mechanism. In view of this, positive grading and reverse pricing personal data tuples are not unnecessary to embody its value.

As discussed above, there are many factors that can affect the value of a data tuple. It is almost impossible to design a universal model to cover all the impact factors. There are three most important factors: value weight; information entropy; and data reference index (R-index). Let α be the value weight, β be information entropy, and γ be the data reference index (R-index). Let them satisfy the following constraint:

$$\alpha + \beta + \gamma = 1 \tag{9}$$

Let the price of the i-th data tuple in data packet be p_i, based on the assumptions above. We can derive the price calculation equation for the i-th data tuple:

$$p_i = P_{\mathrm{S}} \times \left(\frac{w_i}{w} \times \alpha + \frac{q_i}{q} \times \beta + \frac{r_j}{r} \times \gamma \right) \tag{10}$$

In this equation, $i = 1, \ldots, n$, $j = 1, \ldots, m$, P_{S} denotes the supply price of a data packet, n denotes the number of tuples in data packets, m expresses the number of users in data packets, w_i is the value weight of the i-th data tuple, w is the sum of the value weights of all data tuples in a data packet, q_i is the information entropy of i-th data tuple, q is the sum of information entropy in a data packet, r_j is the data reference index (R-index) of the j-th user, and r is the sum of R-indexes of all users.

Equation (10) should satisfy the following constraints:

$$\sum_{i=1}^{n} p_i = P_{\mathrm{S}} \tag{11}$$

where p_i denotes the price of i-th data tuple, n denotes the number of data tuples in a data packet, and P_{S} denotes the supply price of a data packet.

Equation (11) indicates that the sum of all the data tuples should be equal to the sum of supply prices of data packets.

5 Experimental Analysis

Here we use a specific example to validate the reasonability and effectiveness of the pricing model. Assume there is a data packet in a data exchange platform that contains 10 data tuples, and each tuple has 5 items. This data packet is shown in Table 1.

Assume the cost of gathering, analyzing, and sharing a data packet is 20 Yuan, and the demand price of this data packet is 60 Yuan; then the supply price of the data package is 40 Yuan (obtained by Eq. (8)). Let the value weight factor $\alpha = 0.3$, information entropy factor $\beta = 0.4$, and the R-index factor $\gamma = 0.3$.

After using Eq. (2) to compute the probability of each data item, the result is shown in Table 1.

The entropy of each tuple can be calculated according to Eq. (3), the entropy of this data packet can be calculated according to Eq. (4), the weight of information content of each tuple can be calculated according to Eq. (5), the data reference index of each tuple can be calculated according to Eq. (6), the value weight of each tuple can be calculated according to Eq. (7), and the price of each tuple can be calculated according to Eq. (10). The results are shown in Table 2.

It can be seen from Fig. 1 that the higher the value weight, the information content, and the data reference index, the higher the price of each tuple. It also can be seen from Fig. 2 that if the weighted sum of value weight, the information entropy, and the data reference index of one data tuple is greater than that of another data tuple, then this data tuple should be more valuable than the other data tuple, which is consistent with the conclusion. This proves that the propose model is correct, reasonable, and effective.

At present most data pricing algorithms adopt average price to decide the price of each data tuple.

Table 1 Detailed data item and respective probability in data packet.

UserID	Expenditure (Yuan)	Class	Facilitator	Value weight
001 (30%)	100 (20%)	Shopping (20%)	Trust-Mart (10%)	5 (10%)
001 (30%)	30 (20%)	Entertainment (30%)	University city (30%)	4 (10%)
002 (40%)	50 (10%)	Shopping (20%)	Trust-Mart (10%)	3 (20%)
002 (40%)	200 (10%)	Traffic (10%)	Railway (10%)	6 (10%)
002 (40%)	80 (10%)	Medical (20%)	Hospital (20%)	7 (20%)
002 (40%)	40 (20%)	Entertainment (30%)	University city (30%)	3 (20%)
001 (30%)	40 (20%)	Medical (20%)	Hospital (20%)	2 (20%)
003 (20%)	30 (20%)	Entertainment (30%)	University city (30%)	2 (20%)
003 (20%)	60 (10%)	Dining (10%)	Restaurant (10%)	1 (10%)
004 (10%)	100 (20%)	Phone (10%)	Mobile company (10%)	7 (20%)

Note: The value in the parentheses is the probability of data item.

Table 2 Detailed price statement of each tuple.

Serial number	Value weight	Ratio of value weight	Information content	Ratio of information content	R-index	Ratio of R-index	Weighted sum of each impact factor	Price of each tuple (Yuan)
1	5	0.125	2.114	0.097	3	0.100	0.107	4.267
2	4	0.100	2.362	0.109	3	0.100	0.103	4.104
3	3	0.075	2.120	0.098	4	0.133	0.105	4.204
4	6	0.150	1.856	0.085	4	0.133	0.124	4.958
5	7	0.175	2.252	0.104	4	0.133	0.137	5.477
6	3	0.075	2.500	0.115	4	0.133	0.110	4.414
7	2	0.050	2.378	0.109	3	0.100	0.088	3.513
8	2	0.050	2.436	0.112	2	0.067	0.075	3.012
9	1	0.025	1.792	0.082	2	0.067	0.059	2.356
10	7	0.175	1.924	0.089	1	0.033	0.092	3.696
Total	40	1.000	21.734	1.000	30	1.000	1.000	40.000

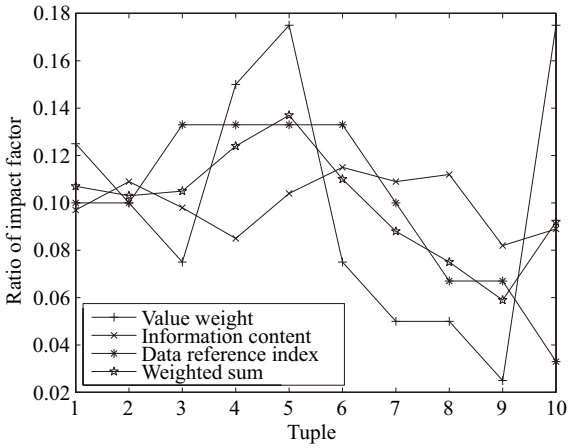


Fig. 1 Impact factor comparison chart.

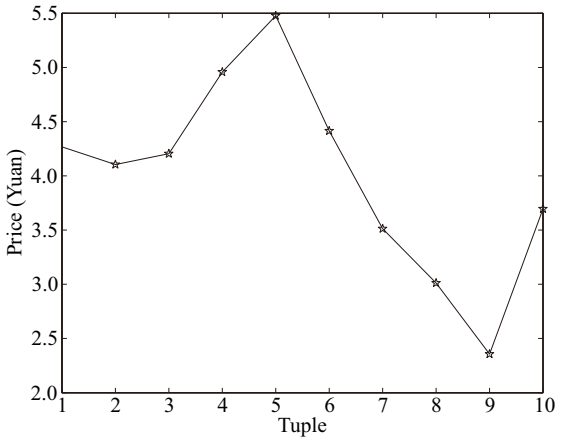


Fig. 2 Data tuple price chart.

That is, the price of a data packet is equally assigned to each data tuple, without taking into account the value contained by the data tuple. Such a pricing method does not reflect the fairness and reasonability of data pricing. We propose a pricing method that can accurately control the price of each data tuple and reflect its due value. It can be seen from Fig. 3 that usual average pricing is a straight line. However, the pricing method we propose is a curve fluctuating around the straight line, which can accurately reflect the intrinsic value of each data tuple.

This pricing model can be adjusted dynamically. Specifically, it allows for the adjustment of four

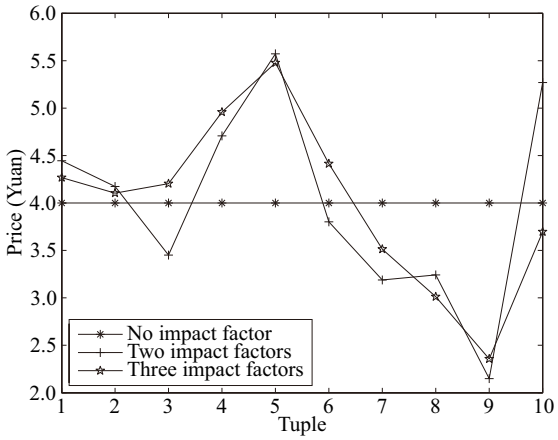


Fig. 3 Data pricing method comparison.

parameters of a data packet: cost of trading platforms to collect, store, and share data; a value weight factor; an information entropy factor; and a data reference index factor. It can work out fair and reasonable market prices for data tuples, reflecting the real value of data and forming a virtuous data market. With data scale increasing, costs gradually decline; and with the continuous improvement of data quality, the demand price of data goes up. The result is that the more the demand price of data increases, the individual user gets more interested, which encourages more individuals to get involved in the data exchange market. In this way, the data trading environment can scale up smoothly, stakeholders can get more material benefits, and a reasonable and efficient data ecosystem can be created.

6 Conclusion

A dynamic, standardized pricing model would revolutionize the existing data market, facilitating transparent transactions and making data science more efficient. Because the value of personal data is dependent on numerous variables, such as information entropy and the value weight of the data, the model will take time to develop, test, and train. However, if successful, this model will be extremely beneficial to any group that produces or uses such data. Based on this, this paper proposes a positive grading and reverse pricing model of Big Personal Data based on tuple granularity, and discusses several factors affecting data value, including cost, value weight, information entropy, and data reference index. This pricing model can be adjusted dynamically; as data scale increases, the cost gradually declines, and with the continuous improvement of data quality, the demand price of data increases, and the supply price of data increases even more. This will form a data trading environment with a scale effect, resulting in stakeholders getting more material benefits, and forming a benign data trading ecosystem.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 61332001, 61272104, and 61472050) and the Science and Technology Planning Project of Sichuan Province (Nos. 2014JY0257, 2015GZ0103, and 2014-HM01-00326-SF).

References

[1] H. Haddadi, R. Mortier, and S. Hand, Privacy analytics, *ACM SIGCOMM Computer Communication Review,* vol. 42, no. 2, pp. 94–98, 2012.

[2] C. Moiso and R. Minerva, Towards a user-centric personal data ecosystem, in *Proc. 16th Conf. on Intelligence in Next Generation Networks*, Berlin, Germany, 2012, pp. 202–209.

[3] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, Privacy-preserving data publishing: A survey of recent developments, *ACM Comput. Surv.*, vol. 42, no. 4, pp. 2623–2627, 2010.

[4] I. C. L. Ng and S. Y. Ho, *Creating New Markets in the Digital Economy: Value and Worth*. Cambridge University Press, 2014.

[5] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, Pricing approaches for data markets, in *Enabling Real-Time Business Intelligence*, M. Castellanos, U. Dayal, and E. A. Rundensteiner, eds. Springer Berlin Heidelberg, 2013, pp. 129–144.

[6] M. Balazinska, B. Howe, and D. Suciu, Data markets in the cloud: An opportunity for the database community, *Proceedings of the VLDB Endowment*, vol. 4, no. 12, pp. 1482–1485, 2011.

[7] C. Shapiro and H. R. Varian, Versioning: The smart way to sell information, *Harvard Business Review*, vol. 76, pp. 107–115, 1998.

[8] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, Query-based data pricing, *ACM Transactions on Database Systems*, vol. 27, no. 4, pp. 438–493, 2012.

[9] R. Tang, H. Wu, Z. Bao, S. E. Bressan, and P. Valduriez, The price is right: Models and algorithms for pricing data, *Lecture Notes in Computer Science*, vol. 8056, no. 2, pp. 380–394, 2013.

[10] M. Balazinska, B. Howe, P. Koutris, S. Dan, and P. Upadhyaya, A discussion on pricing relational data, *Lecture Notes in Computer Science*, vol. 8000, pp. 167–173, 2013.

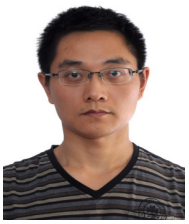
[11] V. Gkatzelis, C. Aperjis, and B. A. Huberman, Pricing private data, *Social Science Electronic Publishing*, vol. 60, no. 5, pp. 1249–1257, 2012.

[12] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, For sale: your data: by: you, in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, 2011, pp. 1–6.
[13] C. Li, D. Y. Li, G. Miklau, and D. Suciu, A theory of pricing private data, in *Proceedings of the 16th International Conference on Database Theory*, 2013, pp. 33–44.
[14] A. Acquisti, L. John, and L. George, What is privacy worth? *Journal of Legal Studies,* vol. 42, no. 2, pp. 249–274, 2009.
[15] D. L. Moody and P. Walsh, Measuring the value of information—An asset valuation approach, presented at European Conference on Information Systems, Ecis 1999, Copenhagen, Denmark, 1999.
[16] H. Sun, Q. W. Tu, and X. W. Wang, Pricing of SaaS in cloud computing based on mathematical models, (in Chinese), *Journal of Shanghai University of Science and Technology*, vol. 36, no. 2, pp. 199–204, 2014.
[17] D. Jang, *Information Theory and Coding*, (in Chinese). Electronic Industry Press, 2013.

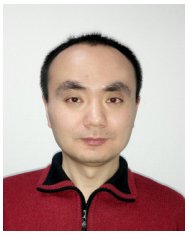
Yuncheng Shen is a PhD candidate in the College of Computer Science, Sichuan University. He received the MS degree in 2011 from Kunming University. His current research interests include big personal data and big data pricing. He is a student member of CCF and IEEE.

Bing Guo is currently a professor of Sichuan University. He received the MS and PhD degrees in computer science, both from University of Electronic Science and Technology of China, in 1999 and 2002, respectively. His current research interests include green computing and big personal data.

Yan Shen is currently a professor of Chengdu University of Information Technology. She received the PhD degree from University of Electronic Science and Technology of China in 2004. Her research interests include smart terminal and instruments.



Xuliang Duan is a PhD candidate in the College of Computer Science, Sichuan University. He received the MS degree in 2008 from Beijing Forestry University. His current research interests include big personal data and big data governance. He is a student member of CCF and IEEE.



Xiangqian Dong is a PhD candidate in the College of Computer Science, Sichuan University. He received the MS degree in 2007 from University of Electronic Science and Technology of China. His current research interests include big personal data, data privacy, and open data.

Hong Zhang is a PhD candidate in the College of Computer Science, Sichuan University. He received the MS degree in 2007 from Southwest Jiaotong University. His current research interests include big personal data and computer network. He is a student member of CCF and IEEE.