THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Computing 1, 13 Computing Drive, Singapore 117417

TRC4/13

What you Pay for is What you Get

**Ruiming Tang, Dongxu Shao, Stephane Bressan
and Patrick Valduriez**

April 2013

Technical Report

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

OOI Beng Chin
Dean of School

What you Pay for is What you Get

Ruiming Tang [#], Dongxu Shao [#], Stéphane Bressan [#], and Patrick Valduriez [⋆]

[#]National University of Singapore
[⋆]INRIA & LIRMM, Montpellier, France

Abstract. In most data markets, prices are prescribed and accuracy is determined by the data. Instead, we consider a model in which accuracy can be traded for discounted prices: "what you pay for is what you get". The data market model consists of data consumers, data providers and data market owners. The data market owners are brokers between the data providers and data consumers. A data consumer proposes a price for the data that she requests. If the price is less than the price set by the data provider, then she gets an approximate value. The data market owners negotiate the pricing schemes with the data providers. They implement these schemes for the computation of the discounted approximate values.
We propose a theoretical and practical pricing framework with its algorithms for the above mechanism. In this framework, the value published is randomly determined from a probability distribution. The distribution is computed such that its distance to the actual value is commensurate to the discount. The published value comes with a guarantee on the probability to be the exact value. The probability is also commensurate to the discount. We present and formalize the principles that a healthy data market should meet for such a transaction. We define two ancillary functions and describe the algorithms that compute the approximate value from the proposed price using these functions. We prove that the functions and the algorithm meet the required principles.

1 Introduction

The three participants to a data market are data providers, data market owners and data consumers ([6]). Data providers bring data to the market and set its base price. Data consumers buy data from the market. A data market owner is a broker. She negotiates the pricing schemes with data providers and manages the market infrastructure that facilitates the transactions between data providers and data consumers.

In the pricing schemes considered in the data market literature (e.g. [9, 4, 2, 8, 7, 3]), prices are prescribed unilaterally and data quality is the best it can be and is non-negotiable. Instead, we consider a model in which accuracy can be traded for discounted prices: "what you pay for is what you get". If a data consumer proposes to buy data at a price less than the price set by the data provider (a price less than the base price) then she can obtain an approximate value, whose quality is commensurate to the discount.

In this paper, we propose a theoretical and practical pricing framework with its algorithms for publishing an approximate value when the price is discounted. In this framework, the value published is randomly determined from a probability distribution. The distribution is computed in such a way that its distance to the actual value is commensurate to the discount sought. The published value comes with a guarantee on the probability to be the exact value. This probability also is commensurate to the discount sought.

The computation of the approximate value for the discounted price requires both a distance function and a probability function. We define an invertible distance function δ, that takes as input the price sought by the data consumer and returns the maximum distance between the degenerate distribution (i.e. the distribution that takes one value only with probability 1) for the exact value and the sampling distribution used to sample the returned value. We define an invertible probability function π, that takes as input the price offered by the data consumer and returns the minimum probability of the returned value.

The data market owner negotiates the distance and probability functions with the data providers. We formalize the principles that a healthy data market should meet for such a transaction. The distance function δ and the probability function π are inverse functions of two pricing functions, δ^{-1} and π^{-1}, respectively. δ^{-1} computes the price given the maximum distance. π^{-1} computes the price given the minimum probability. We formalize the principles of co- and contra-variance, of invariance, of a threshold requirement for these functions with respect to the price. We prove that the two functions meet the principles.

The data market owner uses the distance and probability functions to generate a probability distribution from which the published value is sampled. We argue that, in addition and in order to ensure fairness among consumers without discounting exact values, the data market owner must maximize the distance of the sampling distribution with the degenerate distribution. We devise and present algorithms to generate such an optimal distribution using the distance function δ and the probability function π.

The rest of this paper is organized as follows. Section 2 presents a rapid overview of related work on pricing. In Section 3, we present the data model and basic concepts used in our framework. Two important functions δ and π are introduced in Section 4. We study optimal satisfying distributions in Section 5, and devise algorithms to generate such an optimal distribution in Section 6. Finally, we conclude our framework in Section 7.

2 Related Work

In this section, we categorize the related works based on two dimensions. One dimension categorizes works to be data based pricing or query based pricing. Data based pricing defines prices of information goods or resources, while query based pricing defines prices of queries. The concern of other dimension is that whether a framework allows data consumers' willingness-to-pay (WTP) being less than the price of an item (p), i.e. $WTP < p$. Allowing $WTP < p$ means that a data consumer is free to propose any price for the data she requests.

[9, 4, 2, 8, 7] are in the category of data based pricing and not allowing $WTP < p$. The pricing schemes considered in [9, 4, 2] are static, which means that prices are not updated. The authors of [8, 7] introduce a dynamic pricing scheme in the context of cloud computing. The prices of resources are updated when the utilization varies. However, regardless of a pricing scheme being static or dynamic, consumers have to pay the full price at the moment.

[3, 5] are query based pricing. The authors of [3] propose a pricing model charging for individual queries. Their query-based pricing model allows the seller to set prices for a few views. The price of a query is the price of the cheapest set of views which can determine the query. However, this work also only accepts full payment if a data consumer wants a query. For privacy concerns, the authors of [5] allow a partial payment. They propose a theoretic framework to assign prices to noisy query answers, as a function of a query and standard deviation. For the same query, the more accurate the answer is, the more expensive the query is. If a data consumer cannot afford the price of a query, she can choose to tolerate a higher standard deviation to lower the price.

Our framework is in the category of data based pricing and allowing $WTP < p$. The framework of [5] is similar to our work because both works return a sampled value from a generated distribution to data consumers. Our work differentiates itself from [5] in the following aspects. The framework of [5] focuses on prices of linear queries with only numerical domains, which is the same as researchers do in differential privacy; while in our framework there is a pricing function defining prices of generated distributions on any discrete domains. Moreover, in their framework, the price of a query depends on the standard derivation provided by consumers. Standard derivation is not meaningful without the exact answer of the query. In our framework, we have a probability function which states clearly that what is the probability of getting the exact value if a data consumer pays certain amount of money. It would be much easier for a data consumer to decide her payment. Even worse, in their framework a data consumer has to find an acceptable price by themselves by tuning standard deviation, which may be troublesome. In contrast, in our framework, a data consumer is able to specify the payment directly, and she gets what she pays for.

3 Data Model and Basic Concepts

In this section, we describe the data model and basic concepts of our framework. The data market model consists of data consumers, data providers and data market owners. For the ease of presentation, we consider only one data provider, one data consumer and one data market owner.

The data provider brings data to the market and sets its base price. We consider the data in relational model. A data provider owns a set of tuples. The schema is $(id, \mathbf{v})$, where id is the primary key and $\mathbf{v}$ is the value. We consider the domain of $\mathbf{v}$ to be a discrete ordered or unordered domain. All possible values of $\mathbf{v}$ are $v_1, v_2, ..., v_m$.

Although we only consider two attributes in the schema in our model, cases of more than two attributes can be handled by treating $\mathbf{v}$ as a set of attributes.

Without loss of generality, we consider the case that the data provider only owns one tuple t, $t = (\xi, v_k)$ where $k \in [1, m]$, with a base price pr_{base}. Our model can be easily extended to handle the case that the data provider owns multiple tuples, by considering each tuple independently.

The data consumer wants to buy data from the market. She requests the tuple t. However, she is not forced to pay the base price pr_{base} for it, instead, she is free to propose a price.

The data market owner is a broker between the data provider and data consumer. She negotiates the pricing scheme with the data provider. Receiving a payment from the data consumer, the data market owner returns a value, which is randomly determined from a probability distribution, to the data consumer. The distribution, namely an X-tuple [1], is generated by the data market owner according to the price a consumer is agreeable to pay.

Definition 1. *(X-tuple) An* ***X-tuple*** *is a set of m mutually exclusive tuples with the same id value but different* $\mathbf{v}$ *values, where m is the number of different possible values in the domain of* $\mathbf{v}$*. Each tuple is associated with a probability representing the confidence that this tuple is actual.*

In our framework, we consider that there is only one tuple t owned by the data provider. Under this assumption, all the X-tuples contain the same set of tuples, but with different distributions over possible values of $\mathbf{v}$, therefore, it is possible to represent an X-tuple by its distribution. As a special case, the tuple t can be represented by a degenerate distribution.

Example 1. A tuple $t = (20053046, A)$ telling that the student with *id* 20053046 gets grade A. There are four possible grades A, B, C, D. Based on t, an X-tuple $X = \{< (20053046, A), 0.3 >, < (20053046, B), 0.2 >, < (20053046, C), 0.4 >, < (20053046, D), 0.1 >\}$. X can be represented by its distribution $(0.3, 0.2, 0.4, 0.1)$. The tuple $(20053046, A)$ is sampled from X with a probability 0.3. The tuple t can be represented by a degenerate distribution $(1, 0, 0, 0)$. □

In the rest of the paper, **we represent an X-tuple by its distribution** when there is no ambiguity. **We reserve D_{base} as the degenerate distribution of the tuple t, reserve k as the index where the probability is 1 in D_{base}, reserve m as the number of tuples in an X-tuple**.

Receiving the payment pr_0 from the data consumer, the data market owner generates a distribution (i.e. an X-tuple) X satisfying two constraints:

(1) the distance between X and D_{base} is not larger than $\delta(pr_0)$;

(2) the probability of a tuple sampled from X to be t is at least $\pi(pr_0)$.

There may exist infinitely many distributions satisfying the above two constraints. In order to ensure fairness among consumers, the data market owner generates an optimal satisfying distribution (the reason is in Section 5). A distribution created is optimal in the sense that it has the maximum distance to the degenerate distribution D_{base} under the two constraints. Finally, the data market owner samples a value from an optimal satisfying distribution and returns it to the data consumer.

4 Distance and Probability Functions

In our framework the accuracy of the value is determined from the price given by the data consumer. Therefore the distance function δ and the probability function π are inverse pricing functions. In this section we define the corresponding pricing functions, δ^{-1} and π^{-1}. δ^{-1} computes the price given the maximum distance requirement. π^{-1} computes the price given the minimum probability requirement. We formalize the principles for the exchange in terms of δ^{-1} and π^{-1} in Section 4.1 and Section 4.2, respectively.

We first recall the definition of Earth Mover's Distance, as we use it to define the pricing function δ^{-1}. To define Earth Mover's Distance, a ground distance d_{ij} is needed to measure the difference from cell i to j. We consider a family of ground distance $d_{ij} = |i-j|^q$, where $q \geq 0$. This family of ground distance is a metric, i.e. satisfying non-negative, identity of indiscernible, symmetry and triangle inequality conditions. The ground distance of $q = 0$ suits unordered domains. The ground distance of $q > 0$ suits ordered domains.

Earth mover's distance can be defined on the basis of the ground distance. Given two distributions D_1 and D_2, Earth Mover's Distance from D_1 to D_2, $EMD(D_1, D_2)$, is the optimal value of the following linear programming:
$Minimize: \sum_{i,j} f_{ij} \times d_{ij} \quad s.t.$
$\forall i: \sum_j f_{ij} = D_1(i) \;\; and \;\; \forall j: \sum_i f_{ij} = D_2(j) \;\; and \;\; \forall i,j: f_{ij} \geq 0$
where $D_1(i)$ is the i^{th} cell of distribution D_1 and d_{ij} is the ground distance. Note that in our model $EMD(D_1, D_2) = EMD(D_2, D_1)$ since d_{ij} considered is symmetry. Hence, sometime in this paper, we will say that $EMD(D_1, D_2)$ is the distance **between** D_1 and D_2. In our model, EMD is easier to compute compared to general cases, because D_{base} is special. $EMD(D_{base}, X) = \sum_{i\in[1,m]}(p_i \times d_{ki})$, where p_i is the probability of the i^{th} cell of X.

4.1 The Distance Function δ and Its Inverse Function δ^{-1}

As stated in Section 3, given the payment pr_0, one of the two constraints that a generated distribution X has to satisfy is that the distance between X and D_{base} is not larger than $\delta(pr_0)$. In this section, we introduce the distance function δ, by defining and describing its inverse function δ^{-1}. δ^{-1} is a pricing function, which defines the price of a distance.

Definition 2. *(Distance function and its pricing function) A* distance function *is an invertible function $\delta: Dom_2 \to Dom_1$, where $Dom_1, Dom_2 \subseteq [0, \infty)$. It takes a price as input and returns a distance between D_{base} and the sampling distribution X.*

The inverse function of the distance function is a pricing function δ^{-1} : *$Dom_1 \to Dom_2$, where $Dom_1, Dom_2 \subseteq [0, \infty)$. Its input is a distance (earth mover's distance) between D_{base} and X, and it returns the price.*

We have several principles for δ^{-1} that a healthy data market should meet. **Contra-variance** principle: the larger the distance is, the less expensive it should be, i.e. $\Delta_1 \geq \Delta_2 \;\Rightarrow\; \delta^{-1}(\Delta_1) \leq \delta^{-1}(\Delta_2)$.

For instance, in a gold trading market, a material mixing of gold and silver has a lower price if its veracity of gold is smaller, which means that its "distance" to the pure gold is larger.

Invariance principle: each distance has only one price, i.e. $\mathit{\Delta}_1 = \mathit{\Delta} \Rightarrow \delta^{-1}(\mathit{\Delta}) = \delta^{-1}(\mathit{\Delta}_1)$.

For instance, in a gold trading market, two same materials should have the same price.

Threshold[1] principle: $\mathit{\Delta} \leq \mathit{\Delta}_U$ and $\delta^{-1}(\mathit{\Delta}_U) = pr_{min}$, where U is the uniform distribution, $\mathit{\Delta}_U = EMD(D_{base}, U)$, pr_{min} is minimum amount of money one has to pay.

We assume that X can be charged only if $EMD(D_{base}, X) \leq EMD(D_{base}, U)$. The assumption states that only when the distance between D_{base} and X is not larger than the distance between D_{base} and U, X is valuable. We make this assumption because if X is too far from D_{base}, it would be misleading to the data consumer if she buys it. In a gold trading market, there is a threshold of the veracity of gold to claim that a material is a piece of gold.

Lemma 1. *contra-variance and invariance principle imply that* $\delta^{-1}(\mathit{\Delta}) \leq pr_{base}$.

Based on these three principles, Dom_1 and Dom_2 in Definition 2 have to be updated accordingly: $Dom_1 \subseteq [0, \mathit{\Delta}_U]$ and $Dom_2 \subseteq [pr_{min}, pr_{base}]$.

We aim to design the pricing function δ^{-1} to be not only satisfying the above principles but also as smooth as possible and easy to parameterize by the data market owner. We construct the pricing function based on "sigmoid function"[2]: $S(x) = \frac{1}{1+e^{-x}}$, which is "S" shape and easily parameterized. However, applying the sigmoid function directly does not satisfy the required principles. Moreover, the domain and range of the sigmoid function do not meet our needs as well. Therefore we parameterize the sigmoid function as:

$$S_{\lambda,a}(x) = \left(\frac{1}{1+e^{-\lambda(x-a)}} - \frac{1}{1+e^{\lambda a}}\right) \times \frac{1+e^{\lambda a}}{e^{\lambda a}} \times (pr_{base} - pr_{min}) + pr_{min} \quad (1)$$

where $a > 0$ controls the inflection point and $\lambda > 0$ controls the changing rate. It is easy to verify that the range of $S_{\lambda,a}(x)$ is $[pr_{min}, pr_{base}]$ when $x \in [0, \infty)$. Since $S_{\lambda,a}(x)$ increases as x increases, it violates the contra-variance principle. Therefore we choose a kernel function: $K_1(X) = \frac{1}{EMD(D_{base},X)} - \frac{1}{EMD(D_{base},U)}$.

The pricing function δ^{-1} is a composition of K_1 and $S_{\lambda,a}$:

$$\delta^{-1}(\mathit{\Delta}) = \left(\frac{1}{1+e^{-\lambda(\frac{1}{\mathit{\Delta}}-b-a)}} - \frac{1}{1+e^{\lambda a}}\right) \times \frac{1+e^{\lambda a}}{e^{\lambda a}} \times (pr_{base} - pr_{min}) + pr_{min} \quad (2)$$

where $\mathit{\Delta} \in [0, \frac{1}{b}], b = \frac{1}{EMD(D_{base},U)}$, $a > 0, \lambda > 0$.

Theorem 1. *The pricing function* δ^{-1} *satisfies contra-variance, invariance and threshold principles.*

---

[1] The data owner is able to define another special distribution to be the threshold.

[2] Other functions are also possible. In this paper, we focus on adapting the sigmoid function.

Proof. Due to space limit, we only show the sketch of the proof:

(1) $(\delta^{-1})'(\varDelta) = \lambda \times (1 + \frac{1}{e^{\lambda a}})\frac{e^{-\lambda(\frac{1}{\varDelta}-b-a)}}{(1+e^{-\lambda(\frac{1}{\varDelta}-b-a)})^2} \times (-\frac{1}{\varDelta^2}) < 0$.

(2) It is satisfied since δ^{-1} is a function.

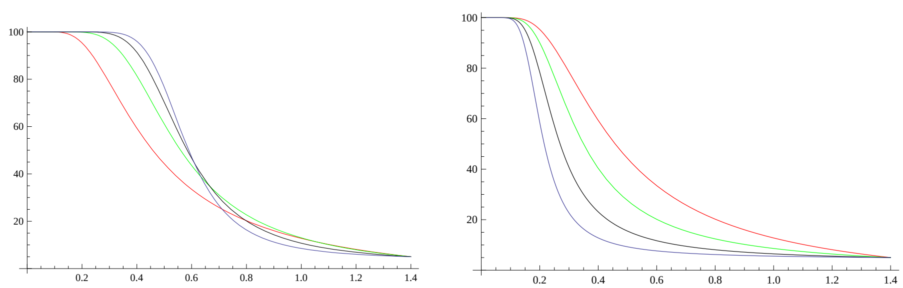
(3) $\delta^{-1}(EMD(D_{base}, U)) = 0 + pr_{min} = pr_{min}$. □

After verifying δ^{-1} satisfying all the principles, we study how to tune the parameters in δ^{-1} according to different needs. There are two parameters in δ^{-1}: λ and a. λ controls the decreasing rate and a adjusts the inflection point.

In Figure 1, we present the curves of δ^{-1} varying λ and a, respectively. We set $pr_{base} = 100, pr_{min} = 5, D_{base} = (0, 1, 0, 0, 0)$ and $q = 1$ for the ground distance. Figure 1(a) shows four curves of δ^{-1} whose a value is fixed to be 1 and λ varies from 1 to 4. When λ is small ($\lambda = 1$), δ^{-1} starts to decrease earlier and decreases more slowly, compare to other curves when λ is larger ($\lambda = 2, 3, 4$).

After the data market owner chooses a proper λ value for herself, the parameter a in δ^{-1} is used to adjust the inflection point. Figure 1(b) shows four curves of δ^{-1} whose λ value is fixed to be 1 and a value varies from 1 to 4. As can be observed, the smaller a value is, the larger the x value of the inflection point is.

To conclude, if the data market owner prefers δ^{-1} to decrease earlier and more slowly, she chooses a smaller λ value; otherwise she chooses a larger λ value. After choosing λ, she can set a smaller a value if she prefers a larger x value of the inflection point; otherwise she chooses a larger a value.



(a) $a = 1$, varying $\lambda(\lambda = 1$ (red), $\lambda = 2$ (green), $\lambda = 3$ (black), $\lambda = 4$ (blue))

(b) $\lambda = 1$, varying $a(a = 1$ (red), $a = 2$ (green), $a = 3$ (black), $a = 4$ (blue))

Fig. 1.

4.2 The Probability Function π and Its Inverse Function π^{-1}

As stated in Section 3, given the payment pr_0, the other one of the two constraints that a generated distribution X has to satisfy is that the probability of a tuple sampled from X to be t is not less than $\pi(pr_0)$. In this section, we introduce the probability function π by presenting its inverse function π^{-1}. π^{-1} is a pricing function, which defines the price of a probability requirement.

Definition 3. *(Probability function and its pricing function) A* probability function *is an invertible function* $\pi : Dom_2 \to Dom_1$*, where* $Dom_1 \subseteq [0, 1]$ *and*

$Dom_2 \subseteq [0, \infty)$. It takes a price as input and returns a probability of a tuple sampled from the sampling distribution X to be t.

The inverse function of the probability function is a pricing function π^{-1} *: $Dom_1 \rightarrow Dom_2$, where $Dom_1 \subseteq [0, 1]$ and $Dom_2 \subseteq [0, \infty)$. Its input is a probability requirement, and it returns the price.*

We have three principles for the pricing function π^{-1}, which are similar to the ones for δ^{-1}.

Co-variance principle: the larger the probability is, the more expensive it should be, i.e. $p_1 \geq p_2 \ \Rightarrow \ \pi^{-1}(p_1) \geq \pi^{-1}(p_2)$.

Invariance principle: for any p, its has only one price, i.e. $p_1 = p \ \Rightarrow \ \pi^{-1}(p) = \pi^{-1}(p_1)$.

Threshold[3] principle: $p \geq \frac{1}{m}$ and $\pi^{-1}(\frac{1}{m}) = pr_{min}$.

Lemma 2. *co-variance and invariance principles imply that $\pi^{-1}(p) \leq pr_{base}$.*

Based on these three principles, Dom_1 and Dom_2 in Definition 3 have to be updated accordingly: $Dom_1 \subseteq [\frac{1}{m}, 1]$ and $Dom_2 \subseteq [pr_{min}, pr_{base}]$.

Using the same strategy as developing δ^{-1}, we choose a kernel function: $K_2(p) = \frac{1}{1-p} - \frac{m}{m-1}$.

The pricing function π^{-1} is a composition of $S_{\lambda,a}$ (Equation 1) and K_2:

$$\pi^{-1}(p) = \left(\frac{1}{1 + e^{-\lambda(\frac{1}{1-p} - \frac{m}{m-1} - a)}} - \frac{1}{1 + e^{\lambda a}}\right) \times \frac{1 + e^{\lambda a}}{e^{\lambda a}} \times (pr_{base} - pr_{min}) + pr_{min} \tag{3}$$

where λ and a have **the same value** as the corresponding value in δ^{-1}. Note that π^{-1} can be derived from δ^{-1} by applying the ground distance of $q = 1$.

Theorem 2. *The pricing function π^{-1} satisfies co-variance, invariance and threshold principles.*

Proof. We only present the sketch of the proof due to space limit.

(1) $(\pi^{-1})'(p) = \lambda \times (1 + \frac{1}{e^{\lambda a}}) \frac{e^{-\lambda(\frac{1}{1-p} - b - a)}}{(1+e^{-\lambda(\frac{1}{1-p} - b - a)})^2} \times (-\frac{1}{(1-p)^2}) \times (-1) > 0$.

(2) It is satisfied since π^{-1} is a function.

(3) $\pi^{-1}(\frac{1}{m}) = 0 + pr_{min} = pr_{min}$. □

5 Optimal $(pr_0, D_{base})-$acceptable distributions

In this section, we study optimal distributions that satisfy the two constraints given by the distance function δ and the probability function π. Recall that given the payment pr_0 by the data consumer, the data market owner generates a distribution X (on the basis of t) satisfying two constraints:

(1) $EMD(D_{base}, X) \leq \delta(pr_0)$;

(2) $p_k \geq \pi(pr_0)$, where k is the dimension that $v_k = t.v$.

We define such satisfying distributions to be $(pr_0, D_{base})-$**acceptable**. D_{base} is a $(pr_0, D_{base})-$acceptable distribution. However, the data market owner should try to avoid generating D_{base}, otherwise no consumer will pay pr_{base} for it.

[3] The data owner is able to define a special probability (other than $\frac{1}{m}$) to be the threshold.

Definition 4. *(Optimal distribution) A $(pr_0, D_{base})-$acceptable distribution X_{opt} is* optimal *if $EMD(D_{base}, X_{opt}) \geq EMD(D_{base}, Y)$, where Y is an arbitrary $(pr_0, D_{base})-$acceptable distribution.*

There may exist infinitely many $(pr_0, D_{base})-$acceptable distributions. In order to ensure fairness among consumers, the data market owner generates an optimal $(pr_0, D_{base})-$acceptable distribution. If she arbitrarily generates a $(pr_0, D_{base})-$acceptable distribution, some unfair cases may happen. It is possible that when a consumer c_1 pays more than a second consumer c_2, the distribution X_1 for c_1 is less expensive than the distribution X_2 for c_2. This is unfair for c_1. We will show that generating an optimal $(pr_0, D_{base})-$acceptable distribution prevents such kind of unfairness.

In the rest of this section, we aim to find $EMD(D_{base}, X_{opt})$ where X_{opt} is an optimal $(pr_0, D_{base})-$acceptable distribution. For the ease of presentation, given pr_0, m and D_{base}, let $s = S^{-1}_{\lambda,a}(pr_0)$, $\alpha = \frac{m}{m-1}$ and $\beta = \frac{1}{EMD(D_{base},U)}$. **These symbols will be used in the theoretic analysis and in the algorithms.** From Equations 1, 2 and 3, we have

$$\frac{1}{s+\beta} = (\delta^{-1})^{-1}(pr_0) = \delta(pr_0) \quad and \quad 1 - \frac{1}{s+\alpha} = (\pi^{-1})^{-1}(pr_0) = \pi(pr_0)$$

Lemma 3. *When $m = 2$ and $k = 1$, the optimal $(pr_0, D_{base})-$acceptable distribution is $(1 - \frac{1}{s+\beta}, \frac{1}{s+\beta})$.*

Proof. When $m = 2$ and $k = 1$, $D_{base} = (1, 0), U = (\frac{1}{2}, \frac{1}{2}), X = (p^X_1, p^X_2)$[4]. Therefore $EMD(D_{base}, X) = p^X_2$ and $\alpha = \beta = 2$.

A distribution X is $(pr_0, D_{base})-$acceptable if and only if

$$\begin{cases} EMD(D_{base}, X) \leq \delta(pr_0) \Leftrightarrow p^X_2 \leq \delta(pr_0) = \frac{1}{s+\beta} \\ p^X_1 \geq \pi(pr_0) \Leftrightarrow p^X_2 \leq 1 - \pi(pr_0) = \frac{1}{s+\alpha} \end{cases}$$

$\alpha = \beta$, therefore the maximum $EMD(D_{base}, X) = \frac{1}{s+\beta}$ is obtained by setting $p^X_2 = \frac{1}{s+\beta}$ and $p^X_1 = 1 - \frac{1}{s+\beta}$. □

Corollary 1. *When $m = 2$ and $k = 2$, the optimal $(pr_0, D_{base})-$acceptable distribution is $(\frac{1}{s+\beta}, 1 - \frac{1}{s+\beta})$.*

Corollary 2. *When $m = 3$ and $k = 2$, an optimal $(pr_0, D_{base})-$acceptable distribution is $(\epsilon, 1 - \frac{1}{s+\beta}, \frac{1}{s+\beta} - \epsilon)$, where $\epsilon \in [0, \frac{1}{s+\beta}]$.*

Lemma 4. *When $m = 3$ and $k = 1$, the distance between D_{base} and an optimal $(pr_0, D_{base})-$acceptable distribution is* $\min\{\frac{1}{s+\beta}, \frac{2^q}{s+\alpha}\}$.

Proof. When $m = 3$ and $k = 1$, $EMD(D_{base}, X) = p^X_2 + 2^q p^X_3$.

A distribution X is $(pr_0, D_{base})-$acceptable if and only if

$$\begin{cases} EMD(D_{base}, X) \leq \delta(pr_0) \Leftrightarrow p^X_2 + 2^q p^X_3 \leq \delta(pr_0) = \frac{1}{s+\beta} \\ p^X_1 \geq \pi(pr_0) \Leftrightarrow p^X_2 + p^X_3 \leq 1 - \pi(pr_0) = \frac{1}{s+\alpha} \end{cases}$$

$EMD(D_{base}, X)$ reaches its maximum $\frac{1}{s+\beta}$ by setting $p^X_2 = \frac{1}{2^q-1}(\frac{2^q}{s+\alpha} - \frac{1}{s+\beta})$ and $p^X_3 = \frac{1}{2^q-1}(\frac{1}{s+\beta} - \frac{1}{s+\alpha})$, in which case the above two equalities hold.

[4] In this paper, we use p^X_l to represent the probability in the l^{th} cell of distribution X.

There is another requirement which is $p_2^X \geq 0, p_3^X \geq 0$. p_3^X is always non-negative since $\alpha \geq \beta$. However, p_2^X will be negative if $\frac{1}{s+\beta} > \frac{2^q}{s+\alpha}$.

– If $\frac{1}{s+\beta} \leq \frac{2^q}{s+\alpha}$, $p_2^X = \frac{1}{2^q-1}(\frac{2^q}{s+\alpha} - \frac{1}{s+\beta}) \geq 0$ and $p_3^X = \frac{1}{2^q-1}(\frac{1}{s+\beta} - \frac{1}{s+\alpha}) \geq 0$. In this case, the maximum $EMD(D_{base}, X)$ is $\frac{1}{s+\beta}$.
– If $\frac{1}{s+\beta} > \frac{2^q}{s+\alpha}$, the maximum $EMD(D_{base}, X)$ is obtained by setting $p_2^X = 0$ and $p_3^X = \frac{1}{s+\alpha}$. $EMD(D_{base}, X) = \frac{2^q}{s+\alpha}$.

The maximum $EMD(D_{base}, X)$ is $\min\{\frac{1}{s+\beta}, \frac{2^q}{s+\alpha}\}$. Therefore the distance between D_{base} and an optimal $(pr_0, D_{base})-$acceptable distribution is $\min\{\frac{1}{s+\beta}, \frac{2^q}{s+\alpha}\}$.

The proof of Lemma 4 can be generalized to cases of $m > 3$. Before going to the generalization, we first prove a lemma which gives a uniform representation for the distributions with the same distance from D_{base}.

Definition 5. *(D_{base} equivalence) X, Y are two distributions. We say $X \equiv_{D_{base}} Y$ if $EMD(D_{base}, X) = EMD(D_{base}, Y)$ and $p_k^X = p_k^Y$.*

Definition 6. *($D_{base}-$form) Let $m > 3$ and X be a distribution. X is in $D_{base}-$**form** if $p_l^X = 0$ for $l \neq k, i, j$, where i is an index nearest to k and j is an index furthest from k. If there exist more than one such i, j, we randomly choose one i and j. For $l = k, i, j, p_l^X = 0$ or $p_l^X \neq 0$.*

Example 2. Assume $D_{base} = (0, 1, 0, 0, 0)$. k is the index where the probability is 1 in D_{base}, i.e. $k = 2$; j is an index furthest from k, i.e. $j = 5$; i is an index nearest to k, i.e. $i = 1$ or $i = 3$. We randomly choose $i = 1$. Therefore, a possible distribution in $D_{base}-$form is $X_1 = (0.2, 0.7, 0, 0, 0.1)$. □

Lemma 5. *Let $m > 3$ and X be an arbitrary distribution. There is always a distribution Y in $D_{base}-$form such that $X \equiv_{D_{base}} Y$.*

Proof. Fix i, j and k as in Definition 6. For each $l \neq i, j, k$ such that $p_l^X \neq 0$, we allocate p_l^X to p_i^Y and p_j^Y. Let $A_{l,i}$ be the flow from p_l^X to p_i^Y. Let $A_{l,j}$ be the flow from p_l^X to p_j^Y. To preserve $X \equiv_{D_{base}} Y$, the following equations hold:

$$\begin{cases} A_{l,i} + |j-k|^q A_{l,j} = |l-k|^q p_l^X \\ A_{l,i} + A_{l,j} = p_l^X \end{cases}$$

Hence we have $A_{l,i} = \frac{|j-k|^q - |l-k|^q}{|j-k|^q - 1} p_l^X, A_{l,j} = \frac{|l-k|^q - 1}{|j-k|^q - 1} p_l^X$.

j being furthest from k infers that $|j-k|^q - |l-k|^q \geq 0$. Since $m > 3$, $|j-k|^q - 1$ is always positive. Therefore, $A_{l,i}$ and $A_{l,j}$ are always non-negative.

To construct Y, let $p_k^Y = p_k^X, p_i^Y = p_i^X + \sum_{l \neq i,j,k} A_{l,i}, p_j^Y = p_j^X + \sum_{l \neq i,j,k} A_{l,j}$ and $p_l^Y = 0$ for $l \neq i, j, k$. □

Extending Lemma 4, the theorem below presents the distance between D_{base} and an optimal (pr_0, D_{base})-acceptable distribution, in the general case of $m > 3$.

Theorem 3. *Let $m > 3$ and X_{opt} be an optimal (pr_0, D_{base})-acceptable distribution.* $EMD(D_{base}, X_{opt}) = \min\{\frac{|j-k|^q}{s+\alpha}, \frac{1}{s+\beta}\}$.

Proof. Due to Lemma 5, we only have to consider X in $D_{base}-$form. Recall that the probability of the k^{th} cell of D_{base} is 1. Fix an index i such that $|i-k| = 1$ and fix an index j maximizing $|j-k|$. $p_l^X = 0$ for $l \neq k, i, j$.

X is (pr_0, D_{base})-acceptable if and only if:

$$\begin{cases} EMD(D_{base}, X) \leq \delta(pr_0) \Leftrightarrow p_i^X + |j-k|^q p_j^X \leq \frac{1}{s+\beta} \\ p_k^X \geq \pi(pr_0) \Leftrightarrow p_i^X + p_j^X \leq \frac{1}{s+\alpha} \end{cases}$$

– If $\frac{|j-k|^q}{s+\alpha} \geq \frac{1}{s+\beta}$, by solving the equations below:

$$\begin{cases} p_i^X + |j-k|^q p_j^X = \frac{1}{s+\beta} \\ p_i^X + p_j^X = \frac{1}{s+\alpha} \end{cases}$$

we have $p_i^X = \frac{1}{|j-k|^q-1}(\frac{|j-k|^q}{s+\alpha} - \frac{1}{s+\beta}), p_j^X = \frac{1}{|j-k|^q-1}(\frac{1}{s+\beta} - \frac{1}{s+\alpha})$.
p_j^X is always non-negative since $\frac{1}{s+\beta} \geq \frac{1}{s+\alpha}$ (the same reason as in Lemma 4).
p_i^X is also non-negative since $\frac{|j-k|^q}{s+\alpha} \geq \frac{1}{s+\beta}$. In this setting $EMD(D_{base}, X) = \frac{1}{s+\beta}$, which is the maximum value that $EMD(D_{base}, X)$ can reach.

– If $\frac{|j-k|^q}{s+\alpha} < \frac{1}{s+\beta}$, let $p_i^X = 0$ and $p_j^X = \frac{1}{s+\alpha}$. Then X is $(pr_0, D_{base})-$acceptable and $EMD(D_{base}, X) = \frac{|j-k|^q}{s+\alpha}$. Below we prove that this value is the maximum value that $EMD(D_{base}, X)$ can reach in this case.
Let Y be an arbitrary $(pr_0, D_{base})-$acceptable distribution in $D_{base}-$form. Then $p_i^Y + |j-k|^q p_j^Y \leq \frac{1}{s+\beta}$ and $p_i^Y + p_j^Y \leq \frac{1}{s+\alpha}$. Hence

$$EMD(D_{base}, Y) = p_i^Y + |j-k|^q p_j^Y = p_i^Y + p_j^Y + (|j-k|^q - 1)p_j^Y$$
$$\leq \frac{1}{s+\alpha} + (|j-k|^q - 1)\frac{1}{s+\alpha} = \frac{|j-k|^q}{s+\alpha} = EMD(D_{base}, X)$$

When $p_i^Y = 0$ and $p_j^Y = \frac{1}{s+\alpha}$, the equality holds.

To conclude, the maximum value of $EMD(D_{base}, X)$ is $\min\{\frac{|j-k|^q}{s+\alpha}, \frac{1}{s+\beta}\}$, for $m > 3$. Therefore $EMD(D_{base}, X_{opt}) = \min\{\frac{|j-k|^q}{s+\alpha}, \frac{1}{s+\beta}\}$. □

Remark 1. Combining the results of Corollary 1, 2, Lemma 3, 4 and Theorem 3, we can conclude that when $m > 1$, $EMD(D_{base}, X_{opt}) = \min\{\frac{|j-k|^q}{s+\alpha}, \frac{1}{s+\beta}\}$. □

Generating an optimal $(pr_0, D_{base})-$acceptable distribution ensures fairness among consumers, as we have stated in the beginning of this section. We prove it by the lemma below.

Lemma 6. *Assume consumer c_1, c_2 pay pr_1, pr_2 respectively and the data market owner generates optimal $(pr_1, D_{base})-$acceptable and $(pr_2, D_{base})-$acceptable X-tuples X_1, X_2 respectively. It is true that $pr_1 \geq pr_2 \Rightarrow \delta^{-1}(EMD(D_{base}, X_1)) \geq \delta^{-1}(EMD(D_{base}, X_2))$.*

Proof. Assume $pr_1 \geq pr_2$. Due to the contra-variance principle of δ^{-1}, $\frac{1}{s_1+\beta} = \delta(pr_1) \leq \frac{1}{s_2+\beta} = \delta(pr_2)$. Due to the co-variance principle of π^{-1}, $1 - \frac{1}{s_1+\alpha} = \pi(pr_1) \geq 1 - \frac{1}{s_2+\alpha} = \pi(pr_2)$, which means $\frac{1}{s_1+\alpha} \leq \frac{1}{s_2+\alpha}$. X_1, X_2 are optimal distributions for pr_1, pr_2 respectively. Therefore, from Theorem 3, we have $EMD(D_{base}, X_1) = \min\{\frac{|j-k|^q}{s_1+\alpha}, \frac{1}{s_1+\beta}\} \leq EMD(D_{base}, X_2) = \min\{\frac{|j-k|^q}{s_2+\alpha}, \frac{1}{s_2+\beta}\}$.

Therefore $\delta^{-1}(EMD(D_{base}, X_1)) \geq \delta^{-1}(EMD(D_{base}, X_2))$. □

6 Algorithms

We present algorithms to generate an optimal $(pr_0, D_{base})-$acceptable distribution in this section. Firstly we present an algorithm which generates an optimal $(pr_0, D_{base})-$acceptable distribution in D_{base}-form. However, D_{base}-form may not always meet the needs. Hence we present another algorithm transforming an optimal $(pr_0, D_{base})-$acceptable distribution X_{opt} in $D_{base}-$form to another optimal distribution Y_{opt} which is not in D_{base}-form, while preserving $X_{opt} \equiv_{D_{base}} Y_{opt}$.

In this section, we consider the case of the domain size $m > 3$. The complete discussion on the cases of $m = 2, 3$ are omitted due to space limit. When $m = 2$, an optimal $(pr_0, D_{base})-$acceptable distribution can be generated according to Lemma 3 and Corollary 1; while an optimal $(pr_0, D_{base})-$acceptable distribution of $m = 3$ can be produced according to Corollary 2 and Lemma 4.

We use X_{opt} to represent an optimal $(pr_0, D_{base})-$acceptable distribution. Algorithm 1 generates X_{opt} in $D_{base}-$form. It is devised based on the proof of Theorem 3.

Algorithm 1: Generating X_{opt} in $D_{base}-$form

Data: $D_{base}, pr_0, \pi, \delta$
Result: X_{opt} in $D_{base}-$form

1 **if** $\frac{|j-k|^q}{s+\alpha} \geq \frac{1}{s+\beta}$ **then**
2 $\quad p_i^{X_{opt}} = \frac{1}{|j-k|^q-1}(\frac{|j-k|^q}{s+\alpha} - \frac{1}{s+\beta})$; $p_j^{X_{opt}} = \frac{1}{|j-k|^q-1}(\frac{1}{s+\beta} - \frac{1}{s+\alpha})$;
$\quad p_k^{X_{opt}} = 1 - \frac{1}{s+\alpha}$;
3 **else**
4 $\quad p_i^{X_{opt}} = 0$; $p_j^{X_{opt}} = \frac{1}{s+\alpha}$; $p_k^{X_{opt}} = 1 - \frac{1}{s+\alpha}$;

However, the data market owner may not want D_{base}-form in some cases. Therefore, we devise an algorithm to transform X_{opt} in $D_{base}-$form to another optimal distribution Y_{opt} in non-D_{base}-form while preserving $X_{opt} \equiv_{D_{base}} Y_{opt}$. Below we discuss the transformation in different cases individually and then summarize them in an algorithm.

Firstly, we consider the case of $\frac{|j-k|^q}{s+\alpha} \geq \frac{1}{s+\beta}$. In this case, we observe from Algorithm 1 that $p_i^{X_{opt}}, p_j^{X_{opt}}, p_k^{X_{opt}} \neq 0$. To preserve $X_{opt} \equiv_{D_{base}} Y_{opt}$, we have to make sure that $p_k^{X_{opt}} = p_k^{Y_{opt}}$ and $EMD(D_{base}, X_{opt}) = EMD(D_{base}, Y_{opt})$. We consider the following sub-cases.

Subcase 1: $p_i^{X_{opt}} < p_i^{Y_{opt}}, p_j^{X_{opt}} < p_j^{Y_{opt}}$. It is impossible to transform to Y_{opt}. Since $p_k^{X_{opt}} = p_k^{Y_{opt}}$ and all the probabilities other than cells i, j, k are 0, no flow can go to $p_i^{Y_{opt}}$ and $p_j^{Y_{opt}}$.

Subcase 2: $p_i^{X_{opt}} < p_i^{Y_{opt}}, p_j^{X_{opt}} > p_j^{Y_{opt}}$. Assume $p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1$ ($\epsilon_1 \in (0, p_j^{X_{opt}})$), and $p_i^{Y_{opt}} = p_i^{X_{opt}} + \epsilon_2$ ($\epsilon_2 \in (0, \epsilon_1)$). Without loss of generality, we assume that there is another cell $p_a^{Y_{opt}} = \epsilon_1 - \epsilon_2$. The probabilities of Y_{opt} are: $p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1, p_i^{Y_{opt}} = p_i^{X_{opt}} + \epsilon_2, p_a^{Y_{opt}} = \epsilon_1 - \epsilon_2, p_k^{Y_{opt}} = p_k^{X_{opt}}$.

$$
\begin{aligned}
&EMD(D_{base}, Y_{opt}) \\
&= (p_j^{X_{opt}} - \epsilon_1)|j-k|^q + (p_i^{X_{opt}} + \epsilon_2)|i-k|^q + (\epsilon_1 - \epsilon_2)|a-k|^q \\
&= p_j^{X_{opt}}|j-k|^q + p_i^{X_{opt}}|i-k|^q + \epsilon_2(|i-k|^q - |j-k|^q) + (\epsilon_1 - \epsilon_2)(|a-k|^q - |j-k|^q) \\
&< p_j^{X_{opt}}|j-k|^q + p_i^{X_{opt}}|i-k|^q = EMD(D_{base}, X_{opt})
\end{aligned}
$$

The reason for "$<$" is that $|j-k| > |i-k|$ and $|j-k| > |a-k|$. Therefore, this case is impossible for the transformation.

Subcase 3: $p_i^{X_{opt}} > p_i^{Y_{opt}}, p_j^{X_{opt}} < p_j^{Y_{opt}}$. This case is also impossible for the transformation, since $EMD(D_{base}, Y_{opt}) > EMD(D_{base}, X_{opt})$. The reason is similar to that of the previous case, therefore we omit it due to space limit.

Subcase 4: when $p_i^{X_{opt}} = p_i^{Y_{opt}}$, and there exists an index a such that $|a-k| = |j-k|$, we can have $EMD(D_{base}, X_{opt}) = EMD(D_{base}, Y_{opt})$ if $p_a^{Y_{opt}} = \epsilon$ and $p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon$ ($\epsilon \in (0, p_j^{X_{opt}})$). Similarly, when $p_j^{X_{opt}} = p_j^{Y_{opt}}$, and there exists an index a such that $|a-k| = |i-k|$, we can have $EMD(D_{base}, X_{opt}) = EMD(D_{base}, Y_{opt})$ if $p_a^{Y_{opt}} = \epsilon$ and $p_i^{Y_{opt}} = p_i^{X_{opt}} - \epsilon$ ($\epsilon \in (0, p_i^{X_{opt}})$).

Subcase 5: $p_i^{X_{opt}} > p_i^{Y_{opt}}, p_j^{X_{opt}} > p_j^{Y_{opt}}$. Assume $p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1$ ($\epsilon_1 \in (0, p_j^{X_{opt}}]$), and $p_i^{Y_{opt}} = p_i^{X_{opt}} - \epsilon_2$ ($\epsilon_2 \in (0, p_i^{X_{opt}}]$). We consider that the flow $\epsilon_1 + \epsilon_2$ is transferred to **only one cell**, namely the cell a. This cell is selected **uniformly at random** from the cells other than i, j and k. Moreover, we have to avoid $|a-k| = |i-k|$ and $|a-k| = |j-k|$. The probabilities of Y_{opt} are:
$p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1, p_i^{Y_{opt}} = p_i^{X_{opt}} - \epsilon_2, p_a^{Y_{opt}} = \epsilon_1 + \epsilon_2, p_k^{Y_{opt}} = p_k^{X_{opt}}$.

By solving the equation $EMD(D_{base}, X_{opt}) = EMD(D_{base}, Y_{opt})$, we have

$$\frac{\epsilon_1}{\epsilon_2} = \frac{|a-k|^q - |i-k|^q}{|j-k|^q - |a-k|^q}$$

$\epsilon_2 = \frac{|j-k|^q - |a-k|^q}{|a-k|^q - |i-k|^q}\epsilon_1 \le p_i^{X_{opt}}$, therefore $\epsilon_1 \le \frac{|a-k|^q - |i-k|^q}{|j-k|^q - |a-k|^q} p_i^{X_{opt}}$. Let $L = \min\{p_j^{X_{opt}}, \dfrac{|a-k|^q - |i-k|^q}{|j-k|^q - |a-k|^q} p_i^{X_{opt}}\}$. We sample ϵ_1 from $(0, L]$ uniformly at random. Then ϵ_2 can be computed using the above equation. Once ϵ_1 and ϵ_2 is computed, Y_{opt} is produced.

Secondly, we consider the case of $\frac{|j-k|^q}{s+\alpha} < \frac{1}{s+\beta}$. In this case, we observe from Algorithm 1 that only $p_j^{X_{opt}}, p_k^{X_{opt}} \neq 0$. In order to keep $X_{opt} \equiv_{D_{base}} Y_{opt}$, we have to make sure that $p_k^{X_{opt}} = p_k^{Y_{opt}}$ and $EMD(D_{base}, X_{opt}) = EMD(D_{base}, Y_{opt})$. We consider the following sub-cases.

Subcase 1: $p_j^{X_{opt}} < p_j^{Y_{opt}}$. It is impossible to transform to Y_{opt}. Since $p_k^{X_{opt}} = p_k^{Y_{opt}}$ and all the probabilities other than cells j, k are 0, no flow can go to $p_j^{Y_{opt}}$.

Subcase 2: $p_j^{X_{opt}} > p_j^{Y_{opt}}$. Assume $p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1$ ($\epsilon_1 \in (0, p_j^{X_{opt}})$), and without loss of generality, we assume ϵ_1 is transferred to $p_a^{Y_{opt}}$. Therefore, the probabilities of Y_{opt} are: $p_i^{Y_{opt}} = 0, p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1, p_a^{Y_{opt}} = \epsilon_1, p_k^{Y_{opt}} = p_k^{X_{opt}}$.

$$
\begin{aligned}
&EMD(D_{base}, Y_{opt}) = (p_j^{X_{opt}} - \epsilon_1)|j-k|^q + \epsilon_1|a-k|^q \\
&= p_j^{X_{opt}}|j-k|^q + \epsilon_1(-|j-k|^q + |a-k|^q) \\
&\leq p_j^{X_{opt}}|j-k|^q = EMD(D_{base}, X_{opt})
\end{aligned}
$$

Only when $|j-k| = |a-k|$, $EMD(D_{base}, X_{opt}) = EMD(D_{base}, Y_{opt})$.

Based on the above discussion, we devise Algorithm 2 transforming an optimal $(pr_0, D_{base})-$acceptable distribution X_{opt} in $D_{base}-$form to a non-D_{base}-form distribution Y_{opt} while preserving $Y_{opt} \equiv_{D_{base}} X_{opt}$.

Finally, the data market owner samples a value from a distribution generated by the algorithms, and returns it to the data consumer.

7 Conclusion

In this paper, we proposed a theoretical and practical pricing framework for a data market in which data consumers can trade data quality for discounted prices. In our framework, the value provided to a data consumer is exact if she offers the full price for it. The value is approximate if she offers to pay only a discounted price. In the case of a discounted price, the value is randomly determined from a probability distribution. The distance of the distribution to the actual value (to the degenerate distribution) is commensurate to the discount. The published value comes with a guarantee on its probability. The probability is also commensurate to the discount. We defined ancillary pricing functions and the necessary algorithms under several principles for a healthy market. We proved the correctness of the functions and algorithms.

References

1. O. Benjelloun, A. D. Sarma, A. Y. Halevy, and J. Widom. Uldbs: Databases with uncertainty and lineage. In *VLDB*, pages 953–964, 2006.
2. B. Birnbaum and A. Jaffe. Relational data markets. 2007.
3. P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. Query-based data pricing. In *PODS*, 2012.

4. A. Kushal, S. Moorthy, and V. Kumar. Pricing for data markets. 2012.
5. C. Li, D. Y. Li, G. Miklau, and D. Suciu. A theory of pricing private data. *CoRR*, abs/1208.5258, 2012.
6. A. Muschalle, F. Stahl, A. Loser, and G. Vossen. Pricing approaches for data markets. In *BIRTE*, 2012.
7. T. Püschel, A. Anandasivam, S. Buschek, and D. Neumann. Making money with clouds: Revenue optimization through automated policy decisions. In *ECIS*, pages 2303–2314, 2009.
8. T. Püschel and D. Neumann. Management of cloud infastructures: Policy-based revenue optimization. In *ICIS*, page 178, 2009.
9. S. yi Wu and R. D. Banker. Best pricing strategy for information services. *J. AIS*, 11(6), 2010.

Algorithm 2: Transforming X_{opt} to non-D_{base}-form Y_{opt}

Data: X_{opt} in $D_{base}-$form

Result: a non-D_{base}-form Y_{opt} such that $X_{opt} \equiv_{D_{base}} Y_{opt}$

if $\frac{|j-k|^q}{s+\alpha} < \frac{1}{s+\beta}$ **then**
    if *there exists an index a that* $|j-k| = |a-k|$ **then**
        sample ϵ_1 from $(0, p_j^{X_{opt}})$;
        the probabilities of Y_{opt} are $p_i^{Y_{opt}} = 0, p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1, p_a^{Y_{opt}} = \epsilon_1, p_k^{Y_{opt}} = p_k^{X_{opt}}$;
    else
        impossible to generate such Y_{opt};
else if $\frac{|j-k|^q}{s+\alpha} \geq \frac{1}{s+\beta}$ **then**
    sample an index a from cells other than i, j, k;
    if $|a-k| = |j-k|$ **then**
        sample ϵ_1 from $(0, p_j^{X_{opt}})$;
        the probabilities of Y_{opt} are $p_i^{Y_{opt}} = p_i^{X_{opt}}, p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1, p_a^{Y_{opt}} = \epsilon_1, p_k^{Y_{opt}} = p_k^{X_{opt}}$;
    else if $|a-k| = |i-k|$ **then**
        sample ϵ_1 from $(0, p_i^{X_{opt}})$;
        the probabilities of Y_{opt} are $p_i^{Y_{opt}} = p_i^{X_{opt}} - \epsilon_1, p_j^{Y_{opt}} = p_j^{X_{opt}}, p_a^{Y_{opt}} = \epsilon_1, p_k^{Y_{opt}} = p_k^{X_{opt}}$;
    else
        sample ϵ_1 from $(0, \min\{p_j^{X_{opt}}, \frac{|a-k|^q - |i-k|^q}{|j-k|^q - |a-k|^q} p_i^{X}\}]$;
        $\epsilon_2 = \frac{|j-k|^q - |a-k|^q}{|a-k|^q - |i-k|^q} \epsilon_1$;
        the probabilities of Y_{opt} are $p_i^{Y_{opt}} = p_i^{X_{opt}} - \epsilon_2, p_j^{Y_{opt}} = p_j^{X_{opt}} - \epsilon_1, p_a^{Y_{opt}} = \epsilon_1 + \epsilon_2, p_k^{Y_{opt}} = p_k^{X_{opt}}$;