An Incentive Mechanism for Trading Personal Data in Data Markets

Sayan Biswas(✉), Kangsoo Jung, and Catuscia Palamidessi

Inria and École Polytechnique, Palaiseau, France
{sayan.biswas,gangsoo.zeong}@inria.fr, catuscia@lix.polytechnique.fr

Abstract. With the proliferation of the digital data economy, digital data is considered as the crude oil in the twenty-first century, and its value is increasing. Keeping pace with this trend, the model of data market trading between data providers and data consumers, is starting to emerge as a process to obtain high-quality personal information in exchange for some compensation. However, the risk of privacy violations caused by personal data analysis hinders data providers' participation in the data market. Differential privacy, a de-facto standard for privacy protection, can solve this problem, but, on the other hand, it deteriorates the data utility. In this paper, we introduce a pricing mechanism that takes into account the trade-off between privacy and accuracy. We propose a method to induce the data provider to accurately report her privacy price and, we optimize it in order to maximize the data consumer's profit within budget constraints. We show formally that the proposed mechanism achieves these properties, and also, validate them experimentally.

Keywords: Data market · Differential privacy · Incentive mechanism · Game theory

1 Introduction

Nowadays, digital data is becoming an essential resource for the information society, and the value of personal data is increasing. In the past, data broker companies such as Acxiom collected personal data and sold them to companies that needed them. However, as the value of personal data is becoming clear to the data providers, and concern about their privacy is increasing among them, people are less and less willing to let their data to be collected for free. In this scenario, the model of *data market* is starting to emerge, as a process to obtain high-quality personal information in exchange of a compensation. Liveen [1] and Datacoup [2] are examples of prototypes of data market services, where the data providers can obtain additional revenue from selling their data, and the consumers can collect the desired personal data.

The problem of privacy violation by personal data analysis is one of the major issues in such data markets. As the population becomes more and more aware of the negative consequences of privacy breaches, such as the Cambridge Analytica

A. Cerone and P. C. Ölveczky (Eds.): ICTAC 2021, LNCS 12819, pp. 197–213, 2021.
https://doi.org/10.1007/978-3-030-85315-0_12

scandal, people are reluctant to release their data, unless they are properly sanitised. In order to solve this problem, techniques like noise insertion [3], synthetic data [4], secure multi-party computation (SMC) [5], and homomorphic encryption [6] are being actively studied. Differential privacy [3], a de-facto standard for privacy protection, is one of the techniques to prevent privacy violations in the data market.

Differential privacy provides a privacy protection framework based on solid mathematical foundations, and enables quantified privacy protection according to the amount of noise insertion. However, like all privacy-protection methods, it deteriorates the data utility. If the data provider inserts too much noise because of privacy concern, the data consumer cannot proceed with the data analysis with the required performance. This trade-off between privacy and utility is a long-standing problem in differential privacy. The privacy protection and data utility depend on the amount of noise insertion while applying differential privacy, and the amount of noise insertion is determined by the noise parameter ϵ. Thus, determining the appropriate value of the parameter ϵ is a fundamental problem in differential privacy. It is difficult to establish the appropriate ϵ value because it depends on many factors that are difficult to quantify, like the attitude towards privacy of the data provider, which may be different from person to person.

We propose an incentive mechanism to encourage the data providers to join in the data market and motivate them to share more accurate data. The amount of noise insertion depends on the data providers' privacy preference and the incentives provided to them by data consumers, and the data consumers decide on incentives to pay to the data provider by considering the profit to be made from the collected data. By sharing some of the consumers' profit with the data provider as incentive, the data provider can get fair prices for providing her data. The proposed mechanism consists of the truthful price report mechanism and an optimization method within budget constraints. The truthful price report mechanism guarantees that the data provider takes the optimal profit when she reports her privacy price to the data consumer honestly. Based on a data provider's reported privacy price, a data consumer can maximize her profit within a potential budget constraint.

1.1 Contribution

The contributions of this paper are as follows:

(i) Truthful price report mechanism: We propose an incentive mechanism that guarantees that the data provider maximizes her benefit when she reports her privacy price honestly.

(ii) Optimized incentive mechanism within the budget constraints: We propose an optimization method to maximize the data consumer's profit and information gain in the setting where the data consumer has a fixed financial budget for data collection.

(iii) Optimized privacy budget splitting mechanism: We propose a method of splitting the privacy budget for the data providers, that allows them to maximize her utility-gain within a fixed privacy budget, in a multiple data consumer environment.

The properties of our methods are both proved formally and validated through experiments.

1.2 Structure of the Paper

The structure of this paper is as follows: we explain the related works and preliminaries in Sects. 2 and 3, respectively. We describe the proposed incentive mechanism in Sect. 4 and validate the proposed incentive mechanism through experiments in Sect. 5. Our conclusion and some potential directions of future work are discussed in Sect. 6.

2 Related Work

2.1 Methods for Choosing ϵ

In differential privacy concept, parameter ϵ is the knob to control the privacy-utility trade off. The smaller the ϵ, the higher is the privacy protection level and the more it deteriorates the data utility. Conversely, a larger ϵ decreases the privacy protection level and enhances the data utility. However, there is no gold standard to determine the appropriate value of ϵ. Apple has been promoting the use of differential privacy to protect user data since iOS 10 was released, but the analysis of [7] showed the ϵ value was set at approximately 10 without any particular reason. The work of [8] showed that the privacy protection level set by an arbitrary ϵ can be infringed by inference using previously disclosed information and proposed an ϵ setting method considering posterior probability. This matter is the main factor that undermines the claim that personal information is protected by differential privacy. Much research have been conducted to study and solve this problem [9–12]. Although a lot of research is being done in this area, the problem of determining a reasonable way of choosing an optimal value for ϵ still remains open, as there are many factors to consider in deciding the value of ϵ, and more studies are still needed. In this paper, we propose a technique to determine an appropriate value of ϵ by setting a price of the privacy of the data provider.

2.2 Pricing Mechanism

One of the solutions to find an appropriate value of ϵ is to price it according to the data accuracy [13–18]. In [13], strength of the privacy guarantee and the accuracy of the published results are considered to set the ϵ value, and a simple ϵ setting model that can satisfy data providers and consumers was suggested. In [14], the author proposed a compensation mechanism via auction in which data providers are rewarded based on data accuracy and data consumer's budget when they provide data with differential privacy. It is the most similar work to our study. The main differences between our paper and Ghosh and Roth's work are as follows:

(i) We define a truthful price report mechanism that a data provider get a best profit when she reports her privacy price honestly, and prove it.

(ii) We propose an optimized incentive mechanism to maximize the data consumer's profit with a fixed expense budget, and a privacy budget splitting method to maximize the data provider's utility-gain in a multi-data consumer environment.

In [17] the authors design a mechanism that can estimate statistics accurately without compromising the user's privacy. They propose a Bayesian incentive and privacy-preserving mechanism that guarantees privacy and data accuracy. The study of [18] proposes a Stackelberg game to maximize mobile users who provides their trajectory data.

Several techniques for pricing data assuming a data market environment have been studied in [19–26].

In [19] the authors suggested a data pricing mechanism to make the balance between privacy and price in data market environment. In [20], the authors propose the data market model in the IoT environment and show the proposed pricing model has a global optimal point. In [21] the authors proposed a theoretical framework for determining prices to noisy query answer in the differentially private data market. However, this research cannot flexibly reflect the requirements of the data market. In the study of [23], the author proposed an ϵ-choosing method based on Rubinstein bargaining and assumes a market manager that mediates a data provider and consumer in the data trading.

It is realistic to consider personal data as a digital asset, and reasonable to attempt to find a bridge between privacy protection level and price according to the value of ϵ in differential privacy, as has been done in this paper. Existing studies are attempting to find an equilibrium between data providers and consumers under the assumption that both are reasonable individuals. In this paper, we follow a research direction similar to existing studies, and focus on the incentive mechanism that motivates a data provider report her privacy price honestly. In particular, we consider that the value of differentially private data increases non-linearly with respect to the increase of the value of ϵ.

3 Preliminaries

In this section, we explain the basic concepts of differential privacy. Differential privacy is a mathematical model that guarantees the privacy protection at a specified level ϵ. For all datasets D_1 and D_2 differing exactly at a single element, it is defined to satisfy ϵ-differential privacy, if the probability distribution difference of the result of a specific query K on two databases is less than or equal to the threshold e^ϵ. The definition of the differential privacy is as follows:

Definition 1 (Differential privacy [3]). *A randomized function $\mathcal{K}$ provides* ϵ-differential privacy *if all datasets, D_1 and D_2, differing by one only element, and all subsets, $S \subseteq$ Range($\mathcal{K}$),*

$$\mathbb{P}[\mathcal{K}(D_1) \in S] \leq e^\epsilon \mathbb{P}[\mathcal{K}(D_2) \in S]$$

The Laplace mechanism [3] is one of the most common methods for achieving the ϵ-differential privacy.

One of the important properties of differential privacy is the compositionality that allows query composing to facilitate modular design [3].

Sequential compositionality. For any database D, let we query on the randomization mechanism K_1 and K_2 which is independent for each query. The results of $K_1(D)$ and $K_2(D)$ whose guarantees are the ϵ_1 and ϵ_2-differential privacy, is $(\epsilon_1 + \epsilon_2)$-differentially private.

Parallel compositionality. Let A and B be the partition of any database $D(A \cup B = D, A \cap B = \phi)$. Then, the result of the query on the randomization mechanism $K_1(A)$ and $K_2(B)$, is the $\max(\epsilon_1, \epsilon_2)$-differentially private.

Recently, a variant of differential privacy called *local differential privacy* has been proposed [27–30]. In this model, data providers obfuscate their own data by themselves. Local differential privacy has an advantage that it does not need a trusted third party to satisfy the differential privacy. The properties of parallel and sequential compositionality hold for the local model as well.

In the rest of this paper, we consider the local model of differential privacy.

4 Incentive Mechanism for Data Markets

4.1 Overview of the Proposed Technique

The data market aims at collecting personal data legally with the consent of the provider. A data provider can sell her own data and get paid for it, and a data consumer can collect the personal data for analysis by paying a price, resulting in a win-win situation.

Naturally, the data consumer wants to collect personal data as accurately as possible at the lowest possible price, and the data provider wants to sell her data at a price as high as possible while protecting sensitive information. In general, every effective protection technique affects the utility of the data negatively. In the particular case of differential privacy, the levels of utility and privacy are determined by the parameter ϵ; thus, the data price is affected directly by the value of ϵ.

Determining the appropriate value of ϵ and the actual price of the data are critical to the success of the data market. However this is not an easy task, also because each data provider has different privacy needs [30].

We propose an incentive mechanism to find the price of the data and the value of ϵ that can satisfy both the data provider and the data consumer. The proposed method consists of two parts: an incentive mechanism encouraging the data provider to report her privacy price honestly to the data consumer, and an optimization scheme to maximize both the data consumer and provider's profit within a budget constraint.

We consider a scenario with n data providers, $u_1, \ldots, u_n$, and m data consumers, $D_1, \ldots, D_m$, and where each provider and consumer proceeds with the deal independently (we use the term "data provider" and "data producer" interchangeably, in the same sense). The term "ϵ unit price" (e.g., 1$ per ϵ value 0.1) will be used to express the price, where ϵ is the parameter of differential privacy, which is a measure of the accuracy of information. We recall that, as ϵ increases, the data becomes less private and more information can be obtained from it, and vice versa. Thus, the price per unit ϵ represents the "value" of the provider's information[1]. The price of ϵ is expected to differ from one data provider to another, because each individual has a different privacy need. We denote the ϵ unit price reported by u_i as p_i and her true ϵ unit price as π_i.

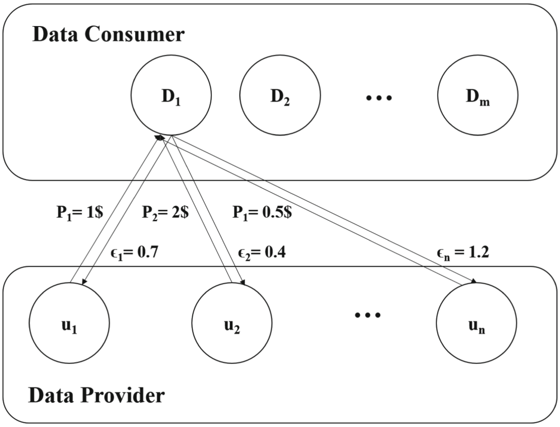

Fig. 1. An example of data trading process. In this figure, u_i means the i^{th} data provider and D_j means the j^{th} data consumer.

Figure 1 illustrates how the process works. At first, every data consumer broadcasts a function f to the data providers, which represents the amount of data (expresses in ϵ units) the consumer is willing to buy for a given ϵ unit price. Each consumer has her own such function, and it can differ from one consumer to another. We will call it *ϵ-allocating function*. We assume f to be monotonically decreasing, as the consumers naturally prefer to buy more data from those data producers who are willing to offer them for less. Note that the product $p_i f(p_i)$ represents the total amount that will be payed by the data producer to the consumer if they agree on the trade. The function f however has also a second purpose: as we shown in Sect. 4.2, it is designed to encourage providers to demand the price that they really consider the true price of their privacy, rather than asking for more.

Then, thanks to the truthful price report mechanism (cf. Sect. 4.2), the data providers report the prices of their data honestly to the data consumers in accordance with the published f. In the example in Fig. 1, u_1 reports her ϵ price per

[1] The ϵ unit price can be of any form, including a monetary one. The method we propose is independent from the nature of the price, so we do not need to specify it.

0.1 as 1\$ and u_2 reports her ϵ price per 0.1 as 2\$. Finally, the data consumer checks the price reported by the data provider and determines the total price and value of ϵ to be obtained from each provider using f. In this example, the data consumer D_1 determines ϵ_1 to be 0.7 and ϵ_2 to be 0.4.

Then, the data providers select the consumers to whom to sell their data in order to maximize their profits, and confirm with them the values of their ϵ and the total price they would receive. In the example in Fig. 1, D_1 pays 7\$ to u_1 and 8\$ to u_2. Finally, the data providers add noise to their data based on the determined ϵ and share the sanitized data with the respective consumers, and the consumers pay the corresponding prices to the providers. We assume that data providers and consumers keep the promise of the value of ϵ and compensation decided in the deal, once confirmed.

This process can be repeated until the data consumers exhaust all their budget or achieve the targeted amount of information. The task of allocating a suitable budget in each round and the how to determine the amount of needed information are also important topics, but they are out of the scope of this paper and are left for future work.

4.2 Truthful Price Report Mechanism

For the correct functioning of the data trading, the data provider should be honest and demand her true privacy price. However, she may be motivated to report a higher price, in the hope to persuade the data consumer that the information is "more valuable", and be willing to pay more. Note also that the true privacy price of each data provider is a personal information that only the provider herself knows and is not obliged to disclose.

To solve this problem, we propose a truthful price report mechanism to ensure that the data providers report their ϵ unit prices honestly. The purpose of the mechanism is to provide incentive so that the providers are guaranteed to get the greatest profit when they report their true price.

When the data provider reports her price p_i, the data consumer determines the amount of ϵ to purchase using $f(p_i)$, where f is the ϵ-allocating function introduced in Sect. 4.1. We recall that f is a monotonically decreasing function, chosen by the consumer. We assume that the domain of f, the ϵ price unit, is normalized to take values in the interval $[0, 1]$. The total price for the data estimated by the consumer is the product of the ϵ price unit and the amount to be purchased, namely, $p_i f(p_i)$. To this value, the consumer adds an *incentive* $\int_{p_i}^{\infty} f(z)\, dz$, the purpose of which is to make convenient for the data producer to report the true price (we assume that the data producer knows f and the strategy of the consumer in advance). The consumer should of course choose f so to be happy with the incentive. In particular, the incentive should be finite, so the contribution of $f(z)$ should vanish as z goes to ∞. An example of such a function is illustrated in Fig. 2.

Thus the data consumer sets the *offer* $\mu(p_i)$ to the provider u_i as follows:

Definition 2 (Payment offer). *The* offer $\mu(p_i)$ *is defined as:*

$$\mu(p_i) = p_i f(p_i) + \int_{p_i}^{\infty} f(z)\, dz$$

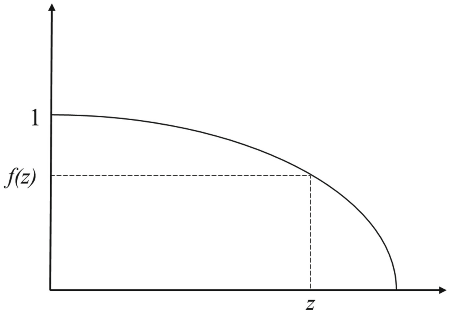

Fig. 2. An example of a monotonically decreasing function $f(z)$. Let c be a parameter representing the "reported value-to-admitted ϵ value" ratio. For $z \geq 0$, we set $f(z)$ as $f(z) = ln(e - cz)$ if $(e - cz) \leq 1$, and $f(x) = 0$ otherwise.

We now illustrate how this strategy achieves its purpose of convincing the consumer to report her true price. We start by defining the *utility* that the data provider obtains by selling her data as the difference between the offer and the true price of her data, represented by the product of the true ϵ unit price and the amount to be sold, namely $\pi_i f(p_i)$:

Definition 3 (Utility of the data provider). *The* utility $\rho(p_i)$*, of the provider* u_i*, for the reported price* p_i*, is defined as:*

$$\rho(p_i) = \mu(p_i) - \pi_i f(p_i)$$

We are now going to show that he proposed mechanism guarantees truthfulness. The basic reason is that each provider u_i achieves the best utility when reporting the true price. Namely, $\rho(\pi_i) \geq \rho(p_i)$ for any $p_i \in \mathbb{R}^+$, where we recall that π_i is the true price of the provider u_i. The only technical condition is that the function f is monotonically decreasing. Under this assumption, we have the following results (see also Fig. 3 to get the intuition of the proof):

Lemma 1. *If* u_i *reports a price greater than her true price, i.e.,* $p_i \geq \pi_i$*, then her utility will be less than the utility for the true price, i.e.,* $\rho(p_i) \leq \rho(\pi_i)$*.*

Proof. The proof can be found in the full version of this paper, available at [31].

Lemma 2. *If* u_i *reports a price smaller than her true price, i.e.,* $p_i \leq \pi_i$*, then her utility will be less than the utility for the true price, i.e.,* $\rho(p_i) \leq \rho(\pi_i)$*.*

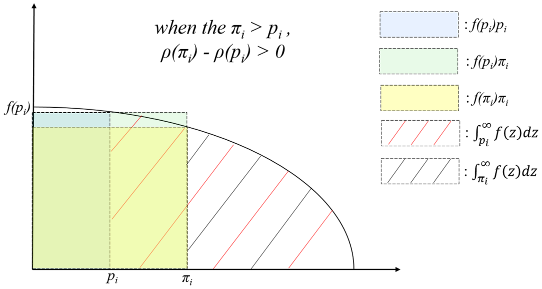
when the $\pi_i > p_i$,
$\rho(\pi_i)$ - $\rho(p_i) > 0$
$f(p_i)$
p_i
π_i
: $f(p_i)p_i$
: $f(p_i)\pi_i$
: $f(\pi_i)\pi_i$
: $\int_{p_i}^{\infty} f(z)dz$
: $\int_{\pi_i}^{\infty} f(z)dz$

Fig. 3. Graphical illustration of Theorem 1. We prove that $\rho(\pi_i)$ (blue hatching area) is always larger than $\rho(p_i)$ (blue rectangle area+red hatching area−green rectangle area).

Proof. The proof can be found in the full version of this paper, available at [31].

Combining Lemma 1 and Lemma 2 gives the announced result. We assume of course that each data producer is a rational individual, i.e., capable of identifying the best strategy to maximize her utility.

Theorem 1. *If every data producer acts rationally, then the proposed incentive mechanism guarantees the truthfulness of the system.*

Proof. Immediate from Lemma 1 and Lemma 2. □

4.3 Optimizing the Incentive Mechanism

In this section, we propose an optimization mechanism to identify an optimal function f for the data consumer with respect to the following two desiderata:

(i) *Maximum Information:* maximize the total information gain of the data consumer with a fixed budget.
(ii) *Maximum Profit:* maximize the total profit of the data consumer with a fixed budget.

By "budget" here we mean the budget of the data consumer to pay the data providers.

We start by introducing the notions of total information and profit for the consumer. Note that, *by the sequential compositionality of differential privacy*, the total information is the sum of the information obtained from each data provider.

Definition 4 (Total information). *The* total information $\mathcal{I}(\boldsymbol{u})$ *obtained by the data consumer by concluding trades with each of the data providers of the tuple* $\boldsymbol{u} = (u_1, \ldots, u_n)$ *is defined as*

$$\mathcal{I}(\boldsymbol{u}) = \sum_{i=1}^{n} f(p_i)$$

As for the profit, we can reasonably assume to be monotonically increasing with the amount of information obtained, and that the total profit is the sum of the profits obtained with each individual trading. The latter is naturally defined as the difference between the benefit (aka *payoff*) obtained by re-selling or processing the data, and the price payed to the data provider.

Definition 5 (Payoff and profit).

- *The payoff function for the data consumer, denoted by $\tau(\cdot)$, is the benefit that the data consumer receives by processing or selling the information gathered from the different data providers. The argument of $\tau(\cdot)$ is ϵ, the amount of the information received. We assume $\tau(\epsilon)$ to be monotonically increasing with ϵ.*
- *The total profit for the data consumer is given by $\sum_{i=1}^{n}(\tau(\epsilon_i) - \mu(\epsilon_i))$, where $\epsilon_i = f(p_i)$, i.e., the ϵ-value allocated to u_i.*

We will consider a family of functions $\mathcal{F}$ indexed by a parameter c, to which the ϵ-allocating function f belongs. The parameter c reflects the data consumer's will to collect the information and, for technical reasons, we assume f to be continuous, differentiable and concave with respect to it. For each data provider, different values of c will give different f, that, in turn, will give rise to a different incentive-curve as per equation (2), which the data consumer should adhere to for compensating for the information obtained from that data provider.

As described in previous sections, the ϵ-allocating functions should be monotonically decreasing with the ϵ unit price, as the consumer is motivated to buy more information from the consumers that offer it at a lower price. This property also ensures, by Theorem 1, that the prices reported by the data producers will be their true prices. Hence we impose the following constraint on $\mathcal{F}$:

$$\begin{aligned}\mathcal{F} \;\subseteq\; \{f(\cdot,\cdot) : c, p \in \mathbb{R}^+, f(c,p) \text{ is continuous, differentiable} \\ \text{and concave on } c, \text{and decreasing with } p\}.\end{aligned} \tag{1}$$

Note that we have added the parameter c as an additional argument in f, so f has now two arguments.

Example 1. An example of such class $\mathcal{F}$ is that of Fig. 2:

$$\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}.$$

Example 2. Another example is:

$$\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}.$$

After the prices $p_1, \ldots, p_n$ have been reported by the data producers $u_1, \ldots, u_n$, the data consumer will try to choose an optimal c maximizing her profit. Figure 4 illustrates an example with two data provider's incentive graph and payoff for data consumer.

We will analyse the possibility to choose an optimal c, that, in turn, leads to an optimal $f(c,\cdot)$ addressing scenarios (i) and (ii).

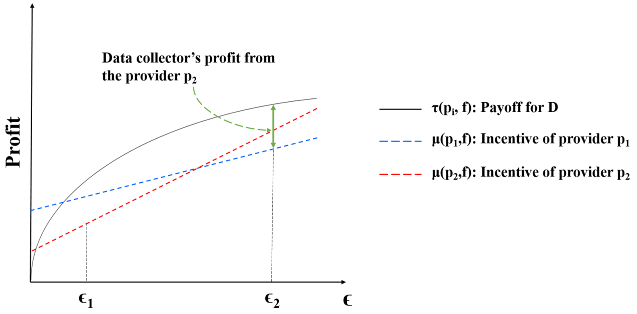

Fig. 4. Illustrating the payoff for c and the incentive-plots for the data consumer involving two data providers reporting p_1, p_2. The Y-intercept of μ_1 is $\int_{p_1}^{\infty} f(z)dz$ and that for μ_2 is $\int_{p_2}^{\infty} f(z)dz$.

In the context of differential privacy, we may assume that τ (the data consumer's payoff function) is additive, i.e.,

$$\textbf{Additivity} \quad \tau(a+b) = \tau(a) + \tau(b) \quad \text{for every } a, b \in \mathbb{R}^+ . \tag{2}$$

This is a reasonable assumption that goes well along with the sequential compositionality property of differential privacy, at least for small values of ϵ[2].

We start by showing that the two desiderata (i) and (ii) are equivalent:

Theorem 2. *If $\tau(\cdot)$ is additive, then maximizing information and maximizing profit (desiderata (i) and (ii)) are equivalent, in the sense that a ϵ-allocating function $f(\cdot,\cdot)$ that maximizes the one, maximizes also the other.*

Proof. The proof can be found in the full version of this paper, available at [31].

Corollary 1. *If $\tau(\cdot)$ is additive, then the optimal choice of $f(\cdot,\cdot)$ w.r.t. the selected family of functions will maximize both the information gain and the profit for the data consumer.*

Proof. Immediate from Theorem 2. □

We now consider the complexity problem for finding the optimal $f(\cdot,\cdot)$. Due to the assumptions made in Eq. 1, and to the additivity of τ, we can apply the method of the Lagrangians to find such $f(\cdot,\cdot)$ (cf. Appendix A of the full version of this paper, available at [31]).

Theorem 3. *If τ is additive, then there exists a c that gives an optimal* **profit-maximizing** *function $f(c,\cdot) \in \mathcal{F}$, for a fixed budget, and we can derive such c via the method of the Lagrangians.*

[2] From a technical point of view, the additive property holds also for large values of ϵ. However, from a practical point of view, for large values of ϵ, for instance 200 and 400, then the original information is almost entirely revealed in both cases, and would not make sense to pay twice the price of 200 ϵ units to achieve 400 ϵ units.

Proof. The proof can be found in the full version of this paper, available at [31]. □

Theorem 4. *There exists a c that gives an optimal* **information-maximizing** *function $f(c, \cdot) \in \mathcal{F}$, for a fixed budget, and we can derive such c via the method of the Lagrangians.*

Proof. The proof can be found in the full version of this paper, available at [31].

To demonstrate how the method works, we show how to compute the specific values of c on the two classes $\mathcal{F}$ of Examples 1 and 2. Such c gives the optimal ϵ-allocating function $f(c, \cdot)$, maximizing $\mathcal{I}(\boldsymbol{u})$ for a given budget. The derivations are described in detail in the full version of this paper, available at [31]. In each example, p_i is the reported ϵ unit price of u_i.

Example 3. Let $\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}$. The optimal parameter c is the solution of the equation $\ln(\prod_{i=1}^{n} e^{p_i}(e - cp_i)^{\frac{e}{c}}) = B + \frac{n(e-1)}{c}$.

Example 4. Let $\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}$. The optimal parameter c is the solution of the equation $c^2 \sum_{i=1}^{n} p_i^2 + 2Bc - n = 0$.

4.4 Discussion

In our model, for the scenario we have considered so far, the parameter c is determined by the number of providers and the budget. We observe that, in both Examples 3 and 4, if n increases than c increases, and vice versa. This seems natural, because in the families of both these example c the incentive that the consumer is going to propose decreases monotonically with c. This means that the larger is the offer, the smaller is the incentive that the consumer needs to be paying. In other words, the examples confirm the well known market law according to which the price decreases when the offer increases, and vice versa.

We note that we have been assuming that there is enough offer to satisfy the consumer's demand. If this hypothesis is not satisfied, i.e., if the offer is smaller than the demand, then the situation is quite different: now the data producer can choose to whom to sell hid data. In particular, the data consumer who sets a lower c will have a better chance to buy data because, naturally, the provider prefers to sell her data to the data consumers who give a higher incentive. In the next section we explore in more detail the process, from the perspective of the data provider, in the case in which the demand is higher than the offer.

4.5 Optimized Privacy Budget Splitting Mechanism for Data Providers

After optimizing an incentive mechanism for a given data consumer dealing with multiple data providers, we focus on the flip side of the setup. We assume a scenario in which a given data provider has to provide her data to multiple data consumers, and that there is enough demand so that she can sell all her data.

Let there be m data consumers, $D_1, \ldots, D_m$ seeking to obtain data from the user u. By truthful price report mechanism, as discussed in Sect. 4.2, u reports her true price to each D_i. As discussed in Sect. 4.3, D_i computes her optimal ϵ-allocating function f_i and requests data from u, differentially privatized with $\epsilon = f_i(\pi)$. After receiving $f_1, \ldots, f_m$, u would like to provide her data in such a way that maximizes her utility received after sharing her data.

Definition 6. *We say that the data provider has made a* deal *with the data consumer D_i if, upon reporting the true per-unit price of her information, π, she agrees to share her data privatized with privacy parameter $\epsilon = f_i(\pi)$.*

It is important to note here that u is not obliged to deal with any data consumer D_i, even after receiving f_i. Realistically, u has a privacy budget of ϵ_{total}, which she would not exceed at any price. Let $S = \{i_1, \ldots, i_k\}$ be an arbitrary subset $\{1, \ldots, m\}$. By the sequential composition property of differential privacy, the final privacy parameter achieved by u by sharing her data to an arbitrary set of data consumers $D_{i_1}, \ldots, D_{i_k}$ is $\epsilon_S = \sum_{j \in S} f_j(\pi)$. u's main intention is to share her data in such a way that ensures $\epsilon_S \leq \epsilon_{\text{total}}$ for all subset S of $\{1, \ldots, n\}$, while maximizing $\sum_{j \in S} \rho_i(\pi, f_j)$, i.e., the total utility received. Reducing it down to the 0/1 knapsack problem, we propose that u should be dealing with $\{D_{i_1}, \ldots, D_{i_k}\}$ where $S^* = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$, chosen as

$$S^* = \arg\max_S \{ \sum_{j \in S} \rho(\pi, f_j) | S \subseteq \{1, \ldots, m\}, \textstyle\sum_{j \in S} f_j(\pi) \leq \epsilon_{\text{total}} \} \quad (3)$$

We show the pseudocode for the ϵ allocation algorithm and the entire process in Algorithms 1 and 2.

Algorithm 1: Optimized privacy budget splitting algorithm

```
Input: {$\epsilon_1, \ldots, \epsilon_n$} stored in array w, {$p_1, \ldots, p_n$} stored in array v, $\epsilon_{\text{total}}$ ;
Output: List of data consumer {$D_1 \ldots D_k$} that is selected to sell data;
initiate Two-dimension array m;
while $i \leq n$ do
    while $j \leq \epsilon_{total}$ do
        if $w[i] > \epsilon_{total}$ then
            m[i, j] := m[i-1, j]
        else
            m[i, j] := max(m[i-1, j], m[i-1, j-w[i]] + v[i])
        end
    end
end
backtrack using the final solution m and find the index of the data consumer ;
return List of selected data consumer ;
```

Algorithm 2: The proposed data trading process

```
Input: the data provider $\{u_1, \ldots, u_n\}$,the data consumer $\{D_1, \ldots, D_m\}$;
Output: List of the data provider and consumer pair that trade is completed;
while $i \leq m$ do
    $D_i$ calculate the parameter c to optimize the $f_i(\cdot)$;
    $D_i$ inform the $f_i(\cdot)$ to the data provider
while $j \leq n$ do
    $u_j$ report price $p_j$ to the data consumer
while $i \leq m$ do
    while $j \leq n$ do
        $D_i$ calculate the $\epsilon_j$ based on $p_j$;
        $D_i$ inform the $\epsilon_j$ to the $u_j$
while $j \leq n$ do
    $u_j$ perform the Optimized $\epsilon$ allocation algorithm to maximize the utility
```

5 Experimental Results

In this section we perform some experiments to verify that the proposed optimization method can find the best profit for the data consumer. For the experiments, we consider the families $\mathcal{F}$ of Examples 3 and 4, namely $\mathcal{F} = \{\ln(e - cp) : c \in \mathbb{R}^+\}$ and $\mathcal{F} = \{1 - cp : c \in \mathbb{R}^+\}$. For these two families the optimal parameter c is also derived formally, as shown in Appendix B of the full version of this paper, available at [31].

The experimental variables are set as follows: we assume that there are 10 data consumers, and the total number of data providers n is set from 1000 to 2000 at an interval of 500. The data provider's ϵ unit price is distributed normally with mean 1 and standard deviation 1, i.e., $\mathcal{N}(1, 1)$, and convert ϵ unit price less than 0 or more than 2 to 0 and 2 respectively. We set the unit value ϵ to 0.1, and the maximum ϵ value of data provider to 3. We set the budgets as 60, 90, and 120 and the number of the data provider as $1,000$, $1,500$, $2,000$. We assumed that the data consumer earned a profit of 10 per 0.1 epsilon and set the parameter c to 1 and 10 for comparison.

The results are shown in Fig. 5. For instance, in the case of the log family $\ln(e - cp)$, the optimal parameter c is 5.36, and in the case of the linear family $1 - cp$, the optimal parameter c is 4.9. It is easy to verify that the optimal values of c correspond to those determined by solving Equations (8) and (13) in Appendix B of the full version of this paper, available at [31], of Examples 3 and 4, respectively.

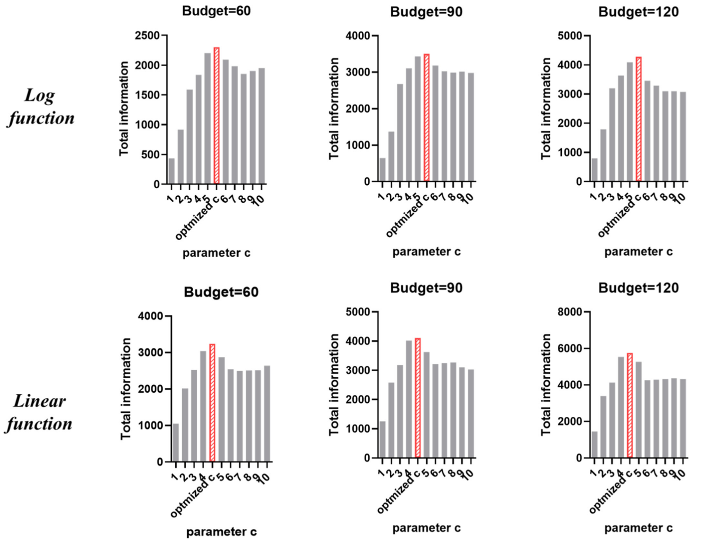

Fig. 5. Experimental result of profit under a fixed budget. Log function is the family $ln(e - cp)$ and Linear function is the family $1 - cp$. We let the parameter c range from 0 to 1. The red bin represents the optimal value of c, namely the c that gives maximum information.

6 Conclusion and Future Work

As machine learning and data mining are getting more and more deployed, individuals are becoming increasingly aware of the privacy issues and of the value of their data. This evolution of people's attitude towards privacy induces companies to develop new approaches to obtain personal data. We envision a scenario where data consumers can trade private data directly from the data provider by paying the price for the respective data, which has the potential to obtain personal information that could not be obtained in the traditional manner. In order to ensure a steady offer in the data market, it is imperative to provide the privacy protection that the data providers deem necessary. Differential privacy can be applied to meet this requirement. However, the lack of standards for setting an appropriate value for the differential privacy parameter ϵ, that determines the levels of data utility and privacy protection, makes it difficult to apply this framework in the data market.

In order to address this problem, we have developed a method, based on incentives and optimization, to find an appropriate value for ϵ in the process of data trading. The proposed incentive mechanism motivates every data provider to report her privacy price honestly in order to maximize her benefit, and the proposed optimization method maximizes the profit for the data consumer under a fixed financial budget. Additionally, in an environment involving multiple data consumers, our mechanism suggests an optimal way for the data providers to split the privacy budgets, maximizing their utility. Through experiments, we have verified that the proposed method provides the best profits to the provider and consumer.

Along the lines of what we have studied in this paper, there are many interesting research issues still open in this area. In future work, we plan to study the following issues:

1. Mechanism for a fair incentive share in an environment where the data providers make a federation for privacy protection
2. Maximization of the data consumers' profits by estimating privacy price distribution of the data providers in an environment where demand of the data providers may change dynamically.

References

1. Liveen. https://www.liveen.com/
2. Datacoup. https://datacoup.com/
3. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3–4), 211–407 (2014)
4. Bowen, C., Snoke, J.: Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge, pp. 1–32. arXiv preprint arXiv:1911.12704 (2019)
5. Volgushev, N., et al.: Conclave: secure multi-party computation on big data. In: Proceedings of the 14th EuroSys Conference, pp. 1–18 (2019)
6. Acar, A., et al.: A survey on homomorphic encryption schemes: theory and implementation. ACM Comput. Surv. (CSUR) **51**(4), 1–35 (2018)
7. Tang, J., et al.: Privacy loss in Apple's implementation of differential privacy on macOS 10.12, pp. 1–12. arXiv preprint arXiv:1709.02753 (2017)
8. Lee, J., Clifton, C.: How much is enough? Choosing ε for differential privacy. In: Lai, X., Zhou, J., Li, H. (eds.) ISC 2011. LNCS, vol. 7001, pp. 325–340. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24861-0_22
9. Chen, Y., et al.: Truthful mechanisms for agents that value privacy. ACM Trans. Econ. Comput. **4**(3), 1–30 (2016)
10. Ligett, K., Roth, A.: Take it or leave it: running a survey when privacy comes at a cost. In: Goldberg, P.W. (ed.) WINE 2012. LNCS, vol. 7695, pp. 378–391. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35311-6_28
11. Xiao, D.: Is privacy compatible with truthfulness? In: Proceedings of the 4th Conference on Innovations in Theoretical Computer Science, pp. 67–86 (2013)
12. Nissim, K., Orlandi, C., Smorodinsky, R.: Privacy-aware mechanism design. In: Proceedings of the 13th ACM Conference on Electronic Commerce, pp. 774–789 (2012)
13. Hsu, J., et al.: Differential privacy: an economic method for choosing epsilon. In: Proceedings of the 27th IEEE Computer Security Foundations Symposium, pp. 1–29 (2014)
14. Ghosh, A., Roth, A.: Selling privacy at auction. Games Econ. Behav. **91**(1), 334–346 (2015)
15. Dandekar, P., Fawaz, N., Ioannidis, S.: Privacy auctions for recommender systems, pp. 1–23 (2012). https://arxiv.org/abs/1111.2885
16. Roth, A.: Buying private data at auction: the sensitive surveyor's problem. ACM SIGecom Exch. **11**(1), 1–8 (2012)

17. Fleischer, L.K., Lyu, Y.H.: Approximately optimal auctions for selling privacy when costs are correlated with data. In: Proceedings of the 13th ACM Conference on Electronic Commerce, pp. 568–585 (2012)
18. Li, W., Zhang, C., Liu, Z., Tanaka, Y.: Incentive mechanism design for crowdsourcing-based indoor localization. IEEE Access **6**, 54042–54051 (2018)
19. Nget, R., Cao, Y., Yoshikawa, M.: How to balance privacy and money through pricing mechanism in personal data market, pp. 1–10. arXiv preprint arXiv:1705.02982 (2018)
20. Oh, H., et al.: Personal data trading scheme for data brokers in IoT data marketplaces. IEEE Access **7**(2019), 40120–40132 (2019)
21. Li, C., Li, D.Y., Miklau, G., Suciu, D.: A theory of pricing private data. ACM Trans. Database Syst. **39**(4), 34–60 (2013)
22. Aperjis, C., Huberman, B.A.: A market for unbiased private data: paying individuals according to their privacy attitudes, pp. 1–17 (2012). SSRN: https://ssrn.com/abstract=2046861
23. Jung, K., Park, S.: Privacy bargaining with fairness: privacy-price negotiation system for applying differential privacy in data market environments. In: Proceedings of the International Conference on Big Data, pp. 1389–1394 (2019)
24. Krehbiel, S.: Choosing epsilon for privacy as a service. Proc. Priv. Enhanc. Technol. **2019**, 192–205 (2019)
25. Zhang, T., Zhu, Q.: On the differential private data market: endogenous evolution, dynamic pricing, and incentive compatibility, pp. 1–30. arXiv preprint arXiv:2101.04357 (2021)
26. Jorgensen, Z., Yu, T., Cormode, G.: Conservative or liberal? Personalized differential privacy. In: Proceedings of the 31St International Conference on Data Engineering, pp. 1023–1034. IEEE (2015)
27. Erlingsson, U., Pihur, V., Korolova, A.: Rappor randomized aggregatable privacy-preserving ordinal response. In: Proceedings of International Conference on Computer and Communications Security, pp. 1054–1067 (2014)
28. Cormode, G., et al.: Privacy at scale: local differential privacy in practice. In: Proceedings of the International Conference on Management of Data, pp. 1655–1658 (2018)
29. Thông, T.N., Xiaokui, X., Yin, Y., et al.: Collecting and analyzing data from smart device users with local differential privacy, pp. 1–11. https://arxiv.org/abs/1606.05053 (2016)
30. Kasiviswanathan, S.P., et al.: What can we learn privately. SIAM J. Comput. **40**(3), 7903–8826 (2011)
31. Biswas, S., Jung, K., Palamidessi, C.: An incentive mechanism for trading personal data in data markets, pp. 1–22. https://arxiv.org/abs/2106.14187 (2021)