Published in final edited form as:
Decis Support Syst. 2014 March 1; 59: 63–73. doi:10.1016/j.dss.2013.10.006.

Pricing and disseminating customer data with privacy awareness

Xiao-Bai Li[a] and **Srinivasan Raghunathan**[b]

Xiao-Bai Li: xiaobai_li@uml.edu; Srinivasan Raghunathan: sraghu@utdallas.edu

[a]Department of Operations & Information Systems, University of Massachusetts Lowell, United States

[b]School of Management, University of Texas at Dallas, United States

Abstract

Organizations today regularly share their customer data with their partners to gain competitive advantages. They are also often requested or even required by a third party to provide customer data that are deemed sensitive. In these circumstances, organizations are obligated to protect the privacy of the individuals involved while still benefiting from sharing data or meeting the requirement for releasing data. In this study, we analyze the tradeoff between privacy and data utility from the perspective of the data owner. We develop an incentive-compatible mechanism for the data owner to price and disseminate private data. With this mechanism, a data user is motivated to reveal his true purpose of data usage and acquire the data that suits to that purpose. Existing economic studies of information privacy primarily consider the interplay between the data owner and the individuals, focusing on problems that occur in the *collection* of private data. This study, however, examines the privacy issue facing a data owner organization in the *distribution* of private data to a third party data user when the real purpose of data usage is unclear and the released data could be misused.

Keywords

Privacy; Pricing; Incentive compatibility; Data mining; Data analytics

1. Introduction

In recent years, there has been a rapid increase in collecting and sharing personal data, owing to widespread use of the Internet and database technologies. Along with this unprecedented growth of activities to collect and share data, data mining and analytics technologies have gained popularity in a wide variety of domains, including database marketing, credit and loan evaluation, web usage/clickstream analysis, medical research and crime analysis. As a result, the buying and selling of customer data have become a multibillion-dollar business [29]. While organizations have benefited from information

Correspondence to: Xiao-Bai Li, xiaobai_li@uml.edu.

sharing and successful application of data mining and analytics, there are increasing concerns about invasions to privacy caused by these practices.

Data marketers such as Acxiom, LexisNexis and ChoicePoint (acquired by Reed Elsevier in 2008) are among the major players in the business of buying and selling consumer data. These companies collect and combine personal data from multiple public and private sources and sell them to retailers, banks, insurance firms, and government agencies. Protecting individual privacy is essential for the survival and success of these businesses. Credit bureaus, such as Equifax, Experian and TransUnion, are another category of firms that are heavily involved in the business of buying and selling consumer data, albeit in a more regulated environment. In addition to disseminating data to external parties, it is a common practice for organizations to share customer data among affiliations and supply-chain partners.

Non-profit organizations have also taken advantage of the value of personal information. The College Board, which organizes standardized tests for college admission, provided about 1700 colleges and universities with lists of students who matched requested SAT and PSAT test score ranges and other demographic information, at a cost of 28 cents per student [29]. The Center for Medicare and Medicaid Services, a federal agency, sells individual Medicare and Medicaid claims data to third parties, which include an individual's medical, demographic, geographic, and financial information (http://www.resdac.org/). The center's operations follow the guidelines of the Health Insurance Portability and Accountability Act (HIPAA). However, studies have shown that the HIPAA rules may be insufficient in protecting patient privacy [2,27].

Privacy-related data sharing and distribution also take place without a direct monetary context, particularly when governments are involved. In 2005, the Department of Justice (DOJ) asked Google Inc. to turn over millions of users' search queries in order to pursue a study into Internet pornography. Google rejected the DOJ's demand, citing that this would undermine the users' trust in Google's ability to keep their information private. This incident aroused intense public debates. A poll [12] revealed that 51% of the respondents believed Google should not release search data to the DOJ, while 43% thought otherwise. The case eventually ended up with a federal judge ruling that Google should provide 50,000 URLs to the DOJ, with individual URLs being randomly selected [7]. The amount of the data is significantly smaller than that of requested by the DOJ initially, and no identity attributes were included.

Clearly, it is not easy for an organization to make a right decision that provides adequate protection for the privacy of the individuals involved while still benefiting from sharing or selling the data, or meeting the requirement for data release. A main source of difficulty is that the real purpose of the data user is usually unknown to the data owner organization and thus the released data could be misused. For example, the Google search query data initially requested by the DOJ can be used either to study general browsing patterns or to help investigate individual cases. Similarly, the same consumer data acquired from a data provider like Acxiom could be used either for macro-level marketing research or for personalized target marketing.

This study takes an economics-based approach to the privacy problem in data sharing and dissemination. We consider a data sharing or disseminating activity as an economic transaction where personal data is viewed as an economic good. The tradeoff between privacy and data utility is analyzed and modeled from the standpoint of a data owner organization. We develop a pricing mechanism for the data owner to collect and distribute sensitive data. This mechanism takes into consideration the differences in the utility of different types of data users for data of different sensitivity levels and provides incentive for a data user to reveal his true purpose of data usage and acquire the data that suits to that purpose. To the best of our knowledge, this work is the first economics-based study that addresses the privacy issue facing a data owner regarding how to disseminate sensitive data to a third party data user. This is a main contribution of the paper.

The rest of the paper is organized as follows. Section 2 presents a review of related research work. In Section 3, we formulate our privacy problem, present our decision models, and discuss their analytical properties. We then provide an illustrative example in Section 4, followed by practical insights in Section 5. Section 6 concludes the paper by discussing limitations of this work and offering directions for future research.

2. Literature review

Information privacy has been studied extensively from different perspectives at individual, organizational and societal levels [3,22,26]. There are typically three parties involved in the privacy problem in data dissemination: (i) the *data owner* (the organization that owns the data) who wants to benefit from disseminating data while fulfilling the obligation of protecting privacy; (ii) *individuals* who provide their personal data to the data owner and want their privacy protected; and (iii) the third party *data user* who acquires data from the data owner; this third party can be either a legitimate user or a privacy invader.

From a privacy viewpoint, the attributes (or variables) in data involving people can be classified into three categories: (i) *identifying attributes*, which explicitly describe the identity of an individual, such as social security number, name, phone number and credit card number; (ii) *confidential attributes*, which contain private information that an individual typically does not want revealed, such as salary, medical test results, and sexual orientation; and (iii) *quasi-identifier attributes* [27], which are normally not considered as confidential by individuals, such as age, gender, race, education, and occupation. However, the values of these attributes can often be used to match the values of identifying attributes from different data sources, resulting in disclosure of individual identities. For instance, Sweeney [27] found out that 87% of the population in the United States can be uniquely identified with three attributes – gender, date of birth, and 5-digit zip code –which are accessible from voter registration records available to the public. The identifying attributes alone typically do not cause privacy problems. For example, the name, phone number and address of an individual can usually be found from a white-page telephone book. Privacy concerns arise when the identifying or quasi-identifier attributes are released together with confidential attributes. In this paper, we use the term *sensitivity* to refer to the risk of disclosing *both* identity and confidential attributes.

In the area of data privacy research, computing *technology-based* approaches attempt to resolve the conflict between privacy protection and data sharing at operational levels. The majority of these approaches use a data masking technique, such as perturbation, swapping, generalization, and suppression, to alter the original data such that, while the individuals in the data are well protected, the utility of the data is also reasonably preserved in the masked version for distribution [1,11,21,27,30]. There are two related limitations in technology-based approaches. First, these approaches conservatively assume that a data user should be considered as a potential privacy invader. When this is not true, that is, when the released data are used for legitimate purposes, data utility is more or less weakened due to masking of data to protect privacy. Second, these approaches typically assume that the data released will be used to find aggregate information or collective patterns in the data. Therefore, identifying attributes are almost always removed or encrypted in the released data processed by a technology-based method. However, in some data sharing and mining applications, such as in database marketing, disease outbreak detection, and crime analysis, individual identities and sensitive data have to be released in order for the data to be useful. Technology-based approaches are generally not applicable to such situations. It is difficult to overcome these limitations with a technology-based approach alone because the problems involve not only operational-level but also policy-level issues.

Economics-based studies focus more on the policy-level privacy issues. Laudon [17] introduces the idea of a regulated National Information Market (NIM) that could allow personal information to be bought and sold like a commodity. In NIM, individuals would decide whether or how much their personal information can be released for secondary use, and data collectors and users would pay for the collection and use of this information. However, no mechanism to implement NIM is provided by Laudon [17] and the idea has not been put into practice so far.

Based on a comprehensive review of the literature [14], the main body of economics-based privacy research generally deals with the relationship between a data owner and individuals, focusing on problems occurring in the *collection* of private data. In particular, an interesting study [13] concerns monetary incentive related to privacy. The study finds that individuals are generally willing to trade off their privacy for economic benefits. The study also provides the experimental results in terms of dollar value for secondary use of personal information. The findings of this study validate the assumption that personal data can be viewed as economic goods and the transaction of such data can be analyzed with economics tools.

Garfinkel et al. [10] propose a mechanism that integrates a data masking technique into an economic model. The mechanism allows individuals to dynamically specify and revise their privacy protection levels, which are tied to their compensation amount to be paid by the data user. The mechanism is, however, closely related to a specific data masking technique [11] and thus is somewhat limited in its application domains. The mechanism we propose is more general and is not tied to any specific technology-based approach.

In summary, economics-based research focuses on analysis and modeling of privacy problems at the individual level. There is a lack of privacy research at the organizational level [3]. To fill in this gap, this research develops an economics mechanism to address the privacy issue that arises at policy and organizational level when the data owner organization *disseminates* individuals' data to a third party data user, after individual data have been collected. This is new to the literature. This work provides economics analysis and models that facilitate the implementation of Laudon's idea of information market. The proposed mechanism alleviates the limitations of the technology-based approach mentioned above. Our approach should be viewed as a mechanism to complement, rather than to substitute, the existing technology-based approaches. In fact, our approach requires that technology exists to mask data and offer them at different protection levels. Practically, a data owner can use this mechanism together with a technology-based approach to simultaneously protect individual privacy and enable effective data usage, at both policy and operational levels.

Our approach is based on the vertical differentiation framework in economics and marketing [23,24,9]. In that framework, a seller offers products with multiple quality levels at different prices to consumers who are heterogeneous in their valuation of quality. Since the seller does not know about an individual customer's type, she has to design an incentive-compatible quality-price combination so that a customer self-selects the product targeted for that customer. In our model, the sensitivity level of the information offered is analogous to the product quality. In addition to sensitivity, however, we also consider the amount of data as a choice variable for a data consumer. Particularly, due to the nature of our data privacy problem, our model departs from the "single crossing" condition that is typically assumed in the conventional vertical differentiation model. As a result, the models we develop are richer than those in the basic vertical differentiation literature.

3. The proposed pricing scheme

Like the basic vertical differentiation framework, we assume that the data owner acts as a monopoly. This is indeed true in most data sharing cases, because customer data owned by an organization is typically tied to the organization's business and thus is unique in its characteristics. We consider situations where the data acquired by the data consumer *cannot be resold* to another party. This is a reasonable scenario because the data owners can include this condition in the contract with the data consumer [5,6]. We should also point out that the term "consumer" in existing privacy studies typically refers to the individuals who want their privacy protected, while in this study the "consumer" refers to the third party data user. As such, we will use the term "(third party) data user" and "data consumer" interchangeably in this paper.

Many data analysis and data mining tasks focus on finding aggregate information, or discovering collective patterns in the data and relationships between attributes. For these tasks, identifying attributes are generally not useful because it is unlikely that a collective pattern will be related to identities. For example, it might be interesting to study the relationship between a disease and age and occupation, but not that between the disease and patient name. In other cases, such as in database marketing, disease outbreak detection, and crime investigation, the data user is interested in tracing and linking individuals with certain characteristics. In these cases, the data user has a higher incentive to acquire identifying information. Note that both types of data analysis involve using confidential attributes. Otherwise, privacy concerns are negligible.

Based on the account above, in our modeling and mechanism design, we consider two types of data consumers: a type *A* data user who is more interested in *A*ggregate information and patterns in data rather than personal information, and a type *I* data user who is more interested in *I*ndividual identity and personal information. The two types of data users are defined in a relative sense — the setup merely assumes that the type *I* user is more interested in individual information than the type *A* user. In practice, of course, there can be more than two types of data users in terms of the levels of interest in personal information. We consider this simple two-type setting in order to make the analysis manageable. The results of the analysis based on this simplified scenario can provide practically valuable insights, as we elaborate later. We note that restricting to two types of consumers is common in the related literature [9,16]. We discuss issues and possible approaches concerning situations with more than two types of consumers in the final section.

3.1. The cost models

The cost of personal data to the data owner can be classified into three categories. The first is the cost of collecting data, such as discounts offered to customers with membership cards, coupons and sweepstake prizes to draw survey participants, and financial incentives for customers to provide more personal information when they purchase products online [4,13]. From a data distribution viewpoint, this type of cost can be considered sunk cost. The second category is the cost of processing (including masking) the data for distribution. This cost is negligible since it is small and is essentially fixed. The third type of cost is tied to the loss in privacy incurred by disseminating the data. When personal data is disseminated to a third party, the individual subjects involved should be compensated for their privacy loss. This type of cost has been considered in many privacy studies [10,16,17]. We adopt in this study the compensation scheme proposed by Laudon [17], which allows individuals to provide their personal data to a data collector with financial compensation. The compensation amount, of course, varies with different individuals and the type of information the individuals provide.

With the cost of the data determined by the above compensation scheme, we consider two methods of selecting and releasing data for the data owner. With the first method, individual records are *randomly* selected and disseminated by the data owner to the data user, when only a subset of the data is released. Sometimes, the data user may be interested in only a segment of the individuals with a certain characteristic. In this case, the pool of the data should be reduced to include only those individuals with such a characteristic, and random sampling would still apply to the reduced pool. For instance, suppose the data owner has a data set that includes all customers in a region who have bought a new product. A data user may be interested only in the customers who bought the product using a promotion coupon. In this case, the pool should include only those customers who used the coupon, and random sampling would be applied within this pool. With random sampling, each unit has an equal chance of being selected.

Let the amount of data be represented by the number of individual records, denoted by *x*. Let the sensitivity level of data be represented by the number and type of attributes. Based on the compensation scheme above, the sensitivity level *s* can be measured for a specified set of attributes (requested by the data user). Let *c* be the mean unit cost for an attribute value of a record. Then, with random sampling the total cost of *x* records at sensitivity level *s* can be written as

$$c(s,x)=\int_0^x\int_0^s c\,du\,dv=csx. \quad (1)$$

That is, the above cost function in the case of *random selection* is linear in the following sense.

Linear cost function: For a given sensitivity level, the cost of data with random sampling is an increasing linear function of the amount of the data. Similarly, for a given amount of data, the cost of data is increasing linearly in the sensitivity of the data.

Next, we consider the second method, called *ordered selection*, which selects individual records based on the cost of the individuals, in ascending order from the least expensive one to the most. The above discussion regarding the pool of the data also applies to this case. With ordered selection, each additional unit of amount or sensitivity will cost more for the data owner. As such, the cost function will be convex in the following sense.

Convex cost function: For a given sensitivity level, the cost of data with ordered selection is an increasing and convex function of the amount of the data. Similarly, for a given amount of data, the cost of data is increasing and convex in the sensitivity of the data.

The convexity of the cost function can be expressed as:

$$\partial C(s,x)/\partial x>0, \partial C^2(s,x)/\partial x^2 \geq 0, \partial C(s,x)/\partial s>0, \partial C^2(s,x)/\partial s^2 \geq 0. \quad (2)$$

Note that the linear function is mathematically a special case of the convex function.

3.2. The utility models

In contrast to the linear or convex behavior of the cost function, the utility of data typically exhibits a concave behavior. This means, in our problem context, as the amount and sensitivity of data increase, the data consumer's utilities increase, but at a decreasing rate. This observation is well-grounded on the results of numerous prior analytical and empirical studies in the same or a similar context [23,24,9,19]. A typical example is the use of poll to estimate public opinion. It is well-known that estimation accuracy increases (i.e., the margin of error decreases) with sample size; however, the improvement in estimation accuracy diminishes as the sample size continues to increase [25]. In terms of sensitivity, the more detailed break-down (by age, gender, race, income, etc.) the poll offers – i.e., the more sensitive the data is – the more valuable the poll results are. But the added value generally diminishes as the break-down becomes more detailed (e.g., a break-down of income into ten groups will not likely to be five times as valuable as into two groups). Similar observations have been made in studies involving data mining tasks such as classification, feature selection and association rules mining [15,18,8]. Based on these observations, we state the utility function behavior below (we will discuss later how the analysis will be affected if utility functions are linear or even convex).

Increasing concave utility function: A data consumer's utility is (i) an increasing and concave function of the amount of the data, and (ii) an increasing and concave function of the sensitivity of the data.

Let $U_t(s, x)$ be the utility of user t, where $t \in \{A, I\}$ is the type of the user. Then,

$$\begin{gathered}\partial U_t(s,x)/\partial x>0, \partial U_t^2(s,x)/\partial x^2<0, \partial U_t(s,x)/\partial s>0,\\ \partial U_t^2(s,x)/\partial s^2<0, t\in\{A,I\}.\end{gathered} \quad (3)$$

We assume that the utility of the data user is zero if no data is provided by the data owner or if the sensitivity level of the data is zero (one could interpret zero sensitivity level data as perfectly protected data such as encrypted data). That is,

$$U_t(s,0)=0 \text{ and } U_t(0,x)=0. \quad (4)$$

Because the type I consumer is more interested in more sensitive data than the type A consumer, it is reasonable to argue that *given an increase* (*decrease*) *in sensitivity level of data*, *the increase* (*decrease*) *in utility for a type I data consumer is greater than or equal to that for a type A consumer*. This "single crossing" condition is quite common in similar prior studies [9,24]. The condition can be expressed as

$$\partial U_I(s,x)/\partial s>\partial U_A(s,x)/\partial s. \quad (5)$$

since $U_I(0, x) = U_A(0, x) = 0$, inequality (5) implies that

$$U_I(s,x) \geq U_A(s,x). \quad (6)$$

Different sensitivity levels mentioned above can be achieved by using a technology-based approach to mask the data with different values of a model parameter, such as the k parameter in k-anonymity [27], the upper and lower bounds in data swapping [20], and the variance of noise in data perturbation [21]. Since such a parameter is typically continuous, the sensitivity variable s is considered continuous. Conceptually, we divide s into two types, s_H and s_L, with s_H having higher sensitivity levels than that of s_L. In our modeling, we distinguish s_H and s_L in a very practical way. The released data is said to be of s_H type if it contains explicit identifying attributes; otherwise, it is of s_L type. Clearly, the data with explicit identifiers and confidential attributes are more sensitive than those without. Both s_H and s_L data must also contain confidential attributes in order to be sensitive. For the s_H type, different degree of sensitivity represents different set of explicit identifying attributes and/or different scale of masking to the confidential attributes. Similarly for the s_L type, different degree of sensitivity represents different scale of masking to the quasi-identifier and confidential attributes. Therefore, both s_H and s_L are continuous variables and variable s in the cost and utility functions above can be legitimately replaced with s_H or s_L.

It turns out that the utility function for a type *I* user behaves rather differently given s_H or s_L data. The type *I* user is interested in individual identity and personal information. This is available in the s_H data; so the utility function with s_H is concave and monotonic increasing, as described earlier. With the s_L data, which does not have explicit identifiers such as name and phone number, the user has to use the quasi-identifier attributes, such as age, gender and zip code, to match the records in the data with those in an external source (e.g., voter registration records) to re-identify the individuals [27,30]. So, there is a cost involved for the type *I* user (but not for the type *A* user) in order to use the s_L data. This cost is increasing and convex with respect to the number of records because, given a fixed s_L level, it is increasingly more difficult to re-identify additional individuals (and some individuals may not be re-identified at all). As a result, the net utility (after deducting this cost) will first increase with the number of records, and then decrease after a certain point. The utility function will exhibit an inverted U-shaped behavior.

Non-monotonic concave utility function: The utility of a type *I* data consumer for the s_L data, $U_I(s_L, x)$, is a non-monotonic concave function of x; it first increases with x, and then declines after x is greater than its maximizing point.

Consequently, with the s_L data there will be “double crossing” between the utility function of a type *I* user and that of a type *A* user. This behavior departs from the “single crossing” condition that is assumed in the conventional vertical differentiation models. As a result, the properties associated with the “single crossing” condition (e.g., Eqs. (3), (5) and (6)) will hold only up to a certain point. Beyond that point, new analysis will be required and the corresponding results will be different from those of the conventional models. (It can also be argued that the utility will never decrease even in this case, because the data user can stop using more data once the utility reaches its maximum. We point out that, as shown later in the paper, the analysis and the subsequent models remain the same even if we assume the utility function stops at the maximum.)

3.3. The pricing models

The proposed mechanism works as follows. In response to a request of data from a third party user, the data owner first selects different sensitivity types to offer. The data owner then offers a menu of different price schedules with different sensitivity types, based on incentive compatibility and individual rationality conditions to be discussed later. This will effectively force the data user to reveal his type. The user then selects the corresponding price, and attempts to maximize his net payoff. The sequence of these events is shown in Fig. 1, which include four stages. The models for optimal pricing are developed in a reverse order. We first derive the optimal amount of data for the two types of users respectively, with the chosen price function. The pricing models are then formulated to maximize the data owner’s benefit, given the users’ optimal values.

3.3.1. Price function and optimal data quantities—

We use the popular "two-part tariff" pricing scheme in our modeling, as stated below.

Two-part tariff pricing: For a given sensitivity level, the price of the data includes a fixed charge and a variable charge that is an increasing linear function of the amount of data.

Let *R*(*s*, *x*) be the total price charged for *x* amount of data with sensitivity level *s*. The price function can be written as

$$R(s,x)=\alpha_s+\beta_s x, \quad (7)$$

where α_s and β_s are the fixed and variable price respectively and both are function of *s*. The "two-part tariff" pricing is widely used in practice. It serves particularly well for our purposes because the fixed charge represents the "access fee" or effort that the data user must pay in order to obtain the data at all. Due to the characteristics of sensitive data, this fixed charge may be non-monetary, such as the effort to obtain a security clearance or a judge order [28]. Although this kind of fixed charge is not paid to the data owner directly, it reduces the risk from potential privacy violation penalties by governments. It is thus considered as a benefit to the data owner.

Table 1 summarizes the notation used in this paper.

The net payoff for a user of type *t* that chooses to acquire x_s amount of data with sensitivity level *s* is $U_t(s, x_s) - R(s, x_s)$. Given the concave utility in (3) and two-part tariff pricing in (7), the net payoff has a unique non-corner maximum for the type *A* and type *I* users, respectively (if the cost function is linear and the utility function is linear or convex, then the maximum net payoff will generally occur at the upper bound of *x*, which is less interesting analytically). In stage 4, the user's objective is to maximize his net payoff:

$$\max_{x_s^t} U_t\left(s, x_s^t\right) - \alpha_s - \beta_s x_s^t, \ t \in \{A, I\}. \quad (8)$$

The optimal solution x_s^{t*} can be obtained from the first-order condition below:

$$\frac{\partial U_t\left(s, x_s^t\right)}{\partial x_s^t}\Big|_{x_s^t = x_s^{t*}} = \beta_s, t \in \{A, I\}. \quad (9)$$

3.3.2. Constraints—

Constraints are specified in terms of individual rationality (IR) and incentive compatibility (IC) for different data users.

3.3.2.1. Individual rationality constraints: The net payoff for the type *A* user to use low sensitivity data must be non-negative:

$$U_A\left(s_L, x_L^A\right) - R\left(s_L, x_L^A\right) \geq 0 \quad (10)$$

Similarly, the net payoff for the type *I* user to use high sensitivity data must be non-negative:

$$U_I\left(s_H, x_H^I\right) - R\left(s_H, x_H^I\right) \geq 0. \quad (11)$$

3.3.2.2. Incentive compatibility constraints: The price should be set such that a type *A* user's net payoff using low sensitivity data is greater than or equal to that using high sensitivity data:

$$U_A\left(s_L, x_L^A\right) - R\left(s_L, x_L^A\right) \geq U_A\left(s_H, x_H^A\right) - R\left(s_H, x_H^A\right). \quad (12)$$

For a type *I* user, the incentive for high sensitivity data should be greater than or equal to that for low sensitivity data:

$$U_I\left(s_H, x_H^I\right) - R\left(s_H, x_H^I\right) \geq U_I\left(s_L, x_L^I\right) - R\left(s_L, x_L^I\right). \quad (13)$$

The data owner wants $R(s_L, x_L^A)$ and $R(s_H, x_H^I)$ in the above constraints as large as possible. In the traditional incentive compatibility mechanism setting, it has been shown that only (10) and (13) will be binding. In our problem, the type *I* user's utility function is non-monotonic concave with the s_L data. Therefore, which constraints will be binding depend on the first order conditions for $U_A(s_L, x_L^A)$ and $U_I(s_L, x_L^I)$ determined by (9). Proposition 1 below provides the binding IR and IC constraints with different scenarios.

Proposition 1: Maximizing $R(s_L, x_L^A)$ and $R(s_H, x_H^I)$ results in the following binding IR and IC constraints:

i. If $U_I(s_L, x_L^{I*})/\partial x_L^{I*} > U_A(s_L, x_L^{A*})/\partial x_L^{A*}$, then (10) and (13) will be binding; i.e.,

$$R\left(s_L, x_L^A\right) = U_A\left(s_L, x_L^A\right), \quad (14)$$

$$R\left(s_H, x_H^I\right) = U_I\left(s_H, x_H^I\right) - U_I\left(s_L, x_L^I\right) + U_A\left(s_L, x_L^A\right) + \beta_L\left(x_L^I - x_L^A\right). \quad (15)$$

ii. If $U_I(s_L, x_L^{I*})/\partial x_L^{I*} = U_A(s_L, x_L^{A*})/\partial x_L^{A*}$, then (10), (11) and (13) will be binding; i.e., (14) and (15) will hold, and binding (11) can be written as

$$R\left(s_H, x_H^I\right) = U_I\left(s_H, x_H^I\right). \quad (16)$$

iii. If $U_I(s_L, x_L^{I*})/\partial x_L^{I*} < U_A(s_L, x_L^{A*})/\partial x_L^{A*}$, then (10) and (11) will be binding; i.e., (14) and (16) will hold.

The proofs of all propositions are provided in the Appendix A.

3.3.3. The model for optimal prices—In stage 2, the data owner's objective is to maximize her total net benefit:

$$\max_{\alpha_s,\beta_s} \alpha_s+\beta_s x_s^{t*}-C\left(s, x_s^{t*}\right), t\in\{A, I\}, \quad (17)$$

subject to the binding constraints specified in Proposition 1 for type *A* and type *I* users respectively. These constraints ensure that in stage 3 type *A* and type *I* users select s_L and s_H respectively. We consider in order the three scenarios described in Proposition 1.

Scenario (i) For type *A* user, substituting (14) into (17), the data owner's optimal solution for β_L can be obtained by

$$\frac{\partial}{\partial\beta_L}\left[U_A\left(s_L, x_L^{A*}\right)-C\left(s_L, x_L^{A*}\right)\right]=0.$$

That is,

$$\frac{\partial U\left(s_L, x_L^{A*}\right)}{\partial x_L^{A*}}\frac{\partial x_L^{A*}}{\partial\beta_L}-\frac{\partial C\left(s_L, x_L^{A*}\right)}{\partial x_L^{A*}}\frac{\partial x_L^{A*}}{\partial\beta_L}=0.$$

Substituting (9) into the above equation, we have

$$\beta_L^*=\partial C\left(s_L, x_L^{A*}\right)/\partial x_L^{A*}, \quad (18)$$

$$\alpha_L^*=U_A\left(s_L, x_L^{A*}\right)-\left[\partial C\left(s_L, x_L^{A*}\right)/\partial x_L^{A*}\right]x_L^{A*}, \quad (19)$$

where x_L^{A*} is determined by (9).

For type *I* user, substituting (15) into (17), the data owner's optimal solution for β_H can be obtained by

$$\frac{\partial}{\partial\beta_H}\left[U_I\left(s_H, x_H^{I*}\right)-U_I\left(s_L, x_L^{I*}\right)+U_A\left(s_L, x_L^{A*}\right)+\beta_L^*\left(x_L^{I*}-x_L^{A*}\right)-C\left(s_H, x_H^{I*}\right)\right]=0.$$

Given a fixed β_L value, $U_I(s_L, x_L^{I*}), U_A(s_L, x_L^{A*})$, and $\beta_L^*(x_L^{I*}-x_L^{A*})$ will not change with respect to a small change in β_H. So the above expression simplifies to

$$\frac{\partial}{\partial\beta_H}\left[U_I\left(s_H, x_H^{I*}\right)-C\left(s_H, x_H^{I*}\right)\right]=0.$$

Thus, similar to (18), we have

$$\beta^*_H = \partial C\left(s_H, x^{I*}_H\right)/\partial x^{I*}_H. \quad (20)$$

It follows from (15), (17), (18) and (20) that

$$\alpha^*_H = U_I\left(s_H, x^{I*}_H\right) - U_I\left(s_L, x^{I*}_L\right) + U_A\left(s_L, x^{A*}_L\right) + \frac{\partial C\left(s_L, x^{A*}_L\right)}{\partial x^{A*}_L}\left(x^{I*}_L - x^{A*}_L\right) - \frac{\partial C\left(s_H, x^{I*}_H\right)}{\partial x^{I*}_H} x^{I*}_H, \quad (21)$$

where x^{A*}_L, x^{I*}_L and x^{I*}_H are determined by (9).

Scenarios (ii) and (iii) For type *A* user, optimal solutions α^*_L and β^*_L are the same as in (18) and (19). For type *I* user, β^*_H is the same as in (20). It follows from (16), (17) and (20) that

$$\alpha^*_H = U_I\left(s_H, x^{I*}_H\right) - \left[\partial C\left(s_H, x^{I*}_H\right)/\partial x^{I*}_H\right] x^{I*}_H. \quad (22)$$

For all three scenarios above, Proposition 2 below states a relationship between β^*_L and β^*_H.

Proposition 2: The optimal variable price for low sensitivity data β^*_L is always smaller than the optimal variable price for high sensitivity data β^*_H.

Note that there is no analogous relationship between α^*_H and α^*_L; i.e., α^*_H can be larger or smaller than α^*_L

3.3.4. The model for optimal sensitivity levels—The data owner's problem in stage 1 is to maximize the expected net benefit with respect to different sensitivity types s_L and s_H. This objective can be written as

$$\max_{s_L, s_H} p\left(\alpha^*_L + \beta^*_L x^{A*}_L - \frac{\partial C\left(s_L, x^{A*}_L\right)}{\partial x^{A*}_L} x^{A*}_L\right) + (1-p) \times \left(\alpha^*_H + \beta^*_H x^{I*}_H - \frac{\partial C\left(s_H, x^{I*}_H\right)}{\partial x^{I*}_H} x^{I*}_H\right), \quad (23)$$

where *p* is the probability that a data user is type *A*. We assume that the data owner can estimate this input parameter fairly accurately, based on the historical data (e.g., the proportion of the data releases with/without personal identifiers). Substituting (18) and (20) into (23), this objective simplifies to

$$\max_{s_L, s_H} p\alpha^*_L + (1-p)\alpha^*_H, \quad (24)$$

where α^*_L and α^*_H are functions of s_L and s_H, as shown in (19), (21) and (22). The optimal solution (s^*_L, s^*_H) can be found from the first-order conditions of (24) with respect to s_L and

s_H. We note that the data owner can choose to make only one type of data available. In this case, Eq. (24) can be easily adapted for s_L and s_H separately. We discuss scenarios when offering only one type of data in the next section with an example.

4. An illustrative example

As a benchmark, we first consider a case where both s_L and s_H types are provided. It is also possible for the data owner to make only one type of data available. We discuss this situation and compare it with the benchmark in the second part of this section, using the same example.

4.1. Offering both low- and high-sensitivity data

Consider a type *A* data user whose utility function is

$$U_A\left(s, x_s^A\right) = s^{1/2}\left[x_s^A - \frac{1}{2}\left(x_s^A\right)^2\right], 0 \le s \le 1, 0 \le x_s^A \le 1. \quad (25)$$

It is easy to verify that this function satisfies (3). To ease the illustration, we consider the linear cost function as in (1), which is, as mentioned earlier, a special case of the convex cost function in (2). The idea for the convex case is the same but mathematical manipulations become more cumbersome. For the linear case, Eqs. (18) and (20) simplify to

$$\beta_L^* = cs_L \text{ and } \beta_H^* = cs_H. \quad (26)$$

We first compute x_s^{A*}, the optimal solution to the user's net payoff *P*(·):

$$P_A\left(x_s^A\right) = s^{1/2}\left[x_s^A - \frac{1}{2}\left(x_s^A\right)^2\right] - \alpha_s - \beta_s x_s^A.$$

Setting $\partial P_A / \partial x_s^A = 0$ and using (26), we have

$$x_s^{A*} = 1 - \frac{\beta_s}{s^{1/2}} = 1 - cs^{1/2}. \quad (27)$$

$$U_A\left(s, x_s^{A*}\right) = s^{1/2}\left[\left(1 - cs^{1/2}\right) - \frac{1}{2}\left(1 - cs^{1/2}\right)^2\right] = \frac{1}{2}s^{1/2} - \frac{c^2}{2}s^{3/2}. \quad (28)$$

Substituting (26), (27) and (28) into (19), we have

$$a_L^* = \frac{c^2}{2}s_L^{3/2} - cs_L + \frac{1}{2}s_L^{1/2}. \quad (29)$$

Now, consider a type *I* data user whose original utility function is

$$U_I\left(s,x_s^I\right)=ks^{1/2}\left[x_s^I-\frac{1}{2}\left(x_s^I\right)^2\right],0\le s\le 1,0\le x_s^I\le 1,k>1.\quad(30)$$

Again, this utility satisfies (3). Note that the condition $k>1$ ensures that the relationship between $U_A(s,x_s^A)$ and $U_I(s,x_s^I)$ satisfy (5) and (6). For low sensitivity data, there is a re-identification cost. We first consider a scenario (i) case, which has a re-identification cost of $ks_L^{1/2}(x_L^I)^2/12$ for the type *I* user. Then, the utility with s_L is

$$U_I\left(s_L,x_L^I\right)=ks_L^{1/2}\left[x_L^I-\frac{7}{12}\left(x_L^I\right)^2\right],0\le s_L\le 1,0\le x_L^I\le 1,k>1,\quad(31)$$

(which peaks at x = 6/7, before reaching boundary x = 1). Eq. (30) will be used for the s_H data only. Following the same procedure above, we can get results below for this user:

$$x_H^{I*}=1-\frac{c}{k}s_H^{1/2},x_L^{I*}=\frac{6}{7}\left(1-\frac{c}{k}s_L^{1/2}\right).\quad(32)$$

$$U_I\left(s_H,x_H^{I*}\right)=\frac{k}{2}s_H^{1/2}-\frac{c^2}{2k}s_H^{3/2},U_I\left(s_L,x_L^{I*}\right)=\frac{3k}{7}s_L^{1/2}-\frac{3c^2}{7k}s_L^{3/2}.\quad(33)$$

$$\alpha_H^*=\frac{c^2}{2k}s_H^{3/2}-cs_H+\frac{k}{2}s_H^{1/2}-\frac{3c^2}{7k}s_L^{3/2}+\frac{c^2}{2}s_L^{3/2}-\frac{c}{7}s_L-\frac{3k}{7}s_L^{1/2}+\frac{1}{2}s_L^{1/2}.\quad(34)$$

Substituting (29) and (34) into (24), we have

$$\max_{s_L,s_H}p\left(\frac{c^2}{2}s_L^{3/2}-cs_L+\frac{s_L^{1/2}}{2}\right)+(1-p)\times\left(\frac{c^2}{2k}s_H^{3/2}-cs_H+\frac{k}{2}s_H^{1/2}-\frac{3c^2}{7k}s_L^{3/2}+\frac{c^2}{2}s_L^{3/2}-\frac{c}{7}s_L-\frac{3k}{7}s_L^{1/2}+\frac{s_L^{1/2}}{2}\right).$$

The first-order conditions for this problem lead to the following results:

$$(s_H^*)^{1/2}=k/(3c),\quad(35)$$

$$(s_L^*)^{1/2}=\frac{(2k+12pk)-\left[(2k+12pk)^2-(21k-18+18p)\left(7k-6k^2+6pk^2\right)\right]^{1/2}}{c(21k-18+18p)}.\quad(36)$$

Substituting (35) and (36) into (26) through (34), the optimal solutions $\beta_{s^*}^*,\alpha_{s^*}^*,x_{s^*}^{t}{}^*$ and $U(s^*,x_{s^*}^{t}{}^*)$ $(s^*\in\{s_L^*,s_H^*\},t\in\{A,I\})$ can be expressed in terms of known parameters *p*, *c* and *k*.

Let *p* = 0.5, *c* = 1, *k* = 1.25. Then $(s_L^*)^{1/2}=0.263$, $(s_H^*)^{1/2}=0.417$, and

$$\beta^*_L|_{s^*_L}=0.069, \beta^*_H|_{s^*_H}=0.174, \alpha^*_L|_{s^*_L}=0.071, \alpha^*_H|_{s^*_H}=0.099,$$
$$x^{A*}_L|_{s^*_L}=0.737, x^{I*}_L|_{s^*_L}=0.677, x^{I*}_H|_{s^*_H}=0.667.$$

Figs. 2 and 3 show the utility, cost and price functions for data with sensitivity levels s^*_L and s^*_H, respectively. These are the two optimal scenarios out of numerous possible scenarios for the data owner (s_L and s_H are continuous). The two figures cannot be plotted together because they are based on two data sets with different sensitivity levels. As shown in Fig. 2, $U_A(x)$ and $R(x)$ are tangent at $x = 0.737$. This is the only point where the type *A* user has a non-negative net payoff. To provide some positive incentives for the user, the data owner may set a fixed charge slightly smaller than $\alpha^*_L|_{s^*_L}=0.071$ or a variable charge slightly smaller than $\beta^*_L|_{s^*_L}=0.069$. Fig. 3 shows that the type *A* user cannot afford to buy the high sensitivity data as his utility function is lower than the price function for such data over the entire range of *x*. The type *I* user can have positive incentive with either types of data since his utilities are higher than the type *A* user. The maximum net payoff occurs at $x = 0.677$ for the low sensitivity data and at $x = 0.667$ for the high sensitivity data, both resulting in the same amount of net payoff (as shown by the equal maximum gap between $U_I(x)$ and $R(x)$ in the two figures). In order to facilitate the type *I* self-revelation, the data owner can set a slightly lower price for the high sensitivity data, as long as it is higher than the type *A* user's utility. Note that the absolute values of the variables are not important. For example, $x = 0.737$ can be interpreted as 737 records or 737,000 records. Similarly, sensitivity values should be interpreted in a relative sense.

We now consider a scenario (iii) case. Let the re-identification cost for the type *I* user be $ks^{1/2}_L(x^I_L)^2/3$. Then, the utility with s_L is

$$U_I\left(s_L, x^I_L\right)=ks^{1/2}_L\left[x^I_L-\frac{5}{6}\left(x^I_L\right)^2\right], 0\leq s_L\leq 1, 0\leq x^I_L\leq 1, k>1. \quad (37)$$

It can be verified that $U_I(s_L, x^{I*}_L)/\partial x^{I*}_L<U_A(s_L, x^{A*}_L)/\partial x^{A*}_L$. Using the analytical results derived earlier for scenario (iii), we obtain the following optimal solutions:

$$x^{I*}_L=\frac{3}{5}\left(1-\frac{c}{k}s^{1/2}_L\right), U_I\left(s_L, x^{I*}_L\right)=\frac{3k}{10}s^{1/2}_L-\frac{3c^2}{10k}s^{3/2}_L, \quad (38)$$

$$\alpha^*_H=\frac{c^2}{2k}s^{3/2}_H-cs_H+\frac{k}{2}s^{1/2}_H, \quad (39)$$

$$(s^*_L)^{1/2}=1/(3c). \quad (40)$$

The solutions for the other decision variables are the same. The utility, cost and price functions with the same *c*, *p* and *k* values are shown in Figs. 4 and 5 for data with sensitivity levels s_L^* and s_H^*, respectively. It is observed from Fig. 4 that *R*(*x*) for the low sensitivity data dominates the type *I* user's utility $U_I(x)$. So, the data owner is able to raise the price for the high sensitivity data to match the maximum utility of the type *I* user (Fig. 5) and still induce proper self-selection. If the $U_I(x)$ curve in Fig. 4 is higher such that it tangents with *R*(*x*), we have a scenario (ii) case (this actually occurs when $U_I(s_L, x_L^I) = ks_L^{1/2}[x_L^I - (121/160)\,(x_L^I)^2]$.

For all scenarios, the type *A* user will have no incentive to use the high sensitivity data while the type *I* user will have no incentive to use the low sensitivity data. It is observed that both fixed and variable charges for the high sensitivity data are considerably larger than those for the low sensitivity data. As mentioned earlier, the fixed charge could include non-monetary element such as the effort to obtain a security clearance or a judge order to access such data at all. The high barrier presents little or no problem for a legitimate investigator but can serve as a "protection shield" to prevent a privacy invader to access high sensitivity data.

4.2. Offering either low- or high-sensitivity data

We consider the scenario (i) case first. When offering only one type of data, only the IR constraints, as described in Section 3.3.2, should be considered; there is no IC constraint. When offering only low sensitivity data, the data owner can set the price such that it is affordable to both type *A* and type *I* users (denoted as *L*2) or to type *I* user only (denoted as *L*1). Note that any prices affordable to type *A* will be affordable to type *I* because of the higher utility of type *I*; so it is not possible to have a price affordable to type *A* only. In the *L*2 case, α_{L2}^* and β_{L2}^* have the same expressions as in Eqs. (26) and (29), respectively, because they are derived based on the IR constraints only. So,

$$\alpha_{L2}^* = \frac{c^2}{2} s_{L2}^{3/2} - cs_{L2} + \frac{1}{2} s_{L2}^{1/2}. \quad (41)$$

The data owner's expected net payoff is

$$\begin{aligned} E(P_{L2}) &= p\left(\alpha_{L2}^* + \beta_{L2}^* x_{L2}^{A*} - \frac{\partial C\left(s_{L2}, x_{L2}^{A*}\right)}{\partial x_{L2}^{A*}} x_{L2}^{A*}\right) + (1-p)\left(\alpha_{L2}^* + \beta_{L2}^* x_{L2}^{I*} - \frac{\partial C\left(s_{L2}, x_{L2}^{I*}\right)}{\partial x_{L2}^{I*}} x_{L2}^{I*}\right). \\ &= p\alpha_{L2}^* + (1-p)\alpha_{L2}^* = \alpha_{L2}^*. \end{aligned} \quad (42)$$

Substituting (41) into (42) and taking the first-order condition with respect to s_{L2}, we have

$$(s_{L2}^*)^{1/2} = 1/(3c). \quad (43)$$

Next, consider the *L*1 case; i.e., offering the low sensitivity data to type *I* user only. The data owner's net payoff is maximized when the related IR constraint is bounding:

$$\alpha_{L1}^* + \beta_{L1}^* x_{L1}^{I*} = U_I\left(s_{L1}, x_{L1}^{I*}\right). \quad (44)$$

Substituting the relevant expressions in (26), (32) and (33), which are derived independent of the IC constraints, into (44), we have

$$\alpha^*_{L1}=\frac{3c^2}{7k}s^{3/2}_{L1}-\frac{6c}{7}s_{L1}+\frac{3k}{7}s^{1/2}_{L1}. \quad (45)$$

The data owner's expected net payoff is

$$\begin{aligned}E(P_{L1})=&(1-p)\left(\alpha^*_{L1}+\beta^*_{L1}x^{I_*}_{L1}-\left[\partial C\left(s_{L1},x^{I_*}_{L1}\right)/\partial x^{I_*}_{L1}\right]x^{I_*}_{L1}\right)\\ &=(1-p)\alpha^*_{L1}.\end{aligned} \quad (46)$$

Substituting (45) into (46) and taking the first-order condition with respect to s_{L1}, we have

$$(s^*_{L1})^{1/2}=k/(3c). \quad (47)$$

When offering only high sensitivity data, the data owner can set the price such that it is affordable to both type *A* and type *I* users (denoted as *H*2) or to type *I* user only (denoted as *H*1). In the *H*2 case, the data owner's net payoff is maximized when

$$\alpha^*_{H2}+\beta^*_{H2}x^{A_*}_{H2}=U_A\left(s_{H2},x^{A_*}_{H2}\right). \quad (48)$$

Substituting (26), (27) and (28), which are derived independent of the IC constraints, into (48), we have

$$\alpha^*_{H2}=\frac{c^2}{2}s^{3/2}_{H2}-cs_{H2}+\frac{1}{2}s^{1/2}_{H2}. \quad (49)$$

Similar to (42), the data owner's expected net payoff is

$$E(P_{H2})=p\alpha^*_{H2}+(1-p)\alpha^*_{H2}=\alpha^*_{H2}. \quad (50)$$

Substituting (49) into (50) and taking the first-order condition with respect to s_{H2}, we have

$$(s^*_{H2})^{1/2}=1/(3c). \quad (51)$$

It turns out that s^*_{H2} is the same as s^*_{L2} in (43). This occurs because no boundary between s_L and s_H is specified (if a boundary value between 1/(3*c*) and *k*/(3*c*) is specified, then s^*_{H2} will be equal to the boundary value).

Now consider the *H*1 case; i.e., offering the high sensitivity data to type *I* user only. The data owner's net payoff is maximized when

$$\alpha^*_{H1}+\beta^*_{H1}x^{I_*}_{H1}=U_I\left(s_{H1},x^{I_*}_{H1}\right). \quad (52)$$

Substituting the relevant expressions in (26), (32) and (33) into (52), we have

$$\alpha^*_{H1}=\frac{c^2}{2k}s^{3/2}_{H1}-cs_{H1}+\frac{k}{2}s^{1/2}_{H1}. \quad (53)$$

The data owner’s expected net payoff is

$$E(P_{H1})=(1-p)\alpha^*_{H1}. \quad (54)$$

Substituting (53) into (54) and taking the first-order condition with respect to s_{H1}, we have

$$(s^*_{H1})^{1/2}=k/(3c). \quad (55)$$

Again, it turns out that s^*_{H1} is the same as s^*_{L1} in (47), because no boundary between s_L and s_H is specified.

Table 2 provides the results of different data offering strategies using the same parameters (p = 0.5, c = 1, k = 1.25). It is observed that when offering one type of data to both type *A* and type *I* users (Strategies 2 and 4), the optimal sensitivity value is the same (s^*= 0.333) even though they are labeled as s^*_L and s^*_H respectively. If a boundary value is specified to divide between s_L and s_H, the s^*_L and s^*_H values will be different in general. Similarly, when offering one type of data to type *I* user only (Strategies 3 and 5), the optimal sensitivity value is the same. In this situation, however, the expected net payoff values are different. This is caused by the “re-identification cost” associated with s_L (but not with s_H). In practice, the data owner should be able to divide between s_L and s_H data so that the s^*_L and s^*_H values will be different.

It can be observed that the data owner has the largest expected net payoff with Strategy 1 — offering s_L to type *A* user and s_H to type *I* user for the example parameter values. Next, we analyze how the result changes with parameters *p*, *k* and *c*.

It follows from (41), (42), (43), (49), (50) and (51) that $E(P_{L2})$ and $E(P_{H2})$ do not depend on *p* and *k*. The expected net pay off with Strategy 1 is

$$E(P_{L\&H})=p\alpha^*_L+(1-p)\alpha^*_H. \quad (56)$$

Since $\alpha^*_L<\alpha^*_H$ in this example, $E(P_{L\ \&H})$ decreases as p increases. When p = 1, Eq. (36) simplifies to $(s^*_L)^{1/2}=1/(3c)$, resulting in $E(P_{L\ \&H}) = E(P_{L2}) = E(P_{H2})$ = 0.0741. Therefore, Strategy 1 is at least as good as Strategy 2 and Strategy 4 for any $p \in [0,1]$ and k >1.

It is also observed that in general Strategy 1 is better than Strategy 3 and Strategy 5 (offering data to type *I* only). However, the situation may be different when *p* is very small (i.e., the probability of the user being type *I* is very large). As *p* decreases, $E(P_{L\ \&H})$, $E(P_{L1})$ and $E(P_{H1})$ all increase. When *p* reaches a certain critical value, s_L^* will be zero, which implies that the data owner will not offer s_L data. To find out this critical value of *p*, set s_L^* in (36) to zero and solve for *p*, we have

$$p=1-\frac{7}{6k}. \quad (57)$$

For this example, *k* = 1.25. So, when *p* ≤1/15, the data owner no longer offers s_L data, and Strategy 1 is the same as Strategy 5. That is, for *k* = 1.25, if *p* ≤1/15, then it is optimal for the data owner to offer one type of data and target only the type *I* user for this data. It is also clear from (57) that an increase in *k* increases the critical value of *p* below which it is optimal to offer only one data type.

Finally, consider parameter *c* (mean unit cost). It is clear from (35), (36), (43), (47), (51) and (55) that the optimal sensitivity level for any strategy is inversely proportional to *c*. Substituting s^* in these equations into (29), (34), (41), (45), (49) and (53), respectively, we find that the optimal fixed price for any strategy is also inversely proportional to *c*; i.e.,

$$\alpha_s^* \propto 1/c, \ \ s=\{s_L^*, s_H^*, s_{L1}^*, s_{H1}^*, s_{L2}^*, s_{H2}^*\}. \quad (58)$$

It follows from (42), (46), (50), (54) and (56) that the data owner's net payoff for any strategy depends on α_s^* but not on β_s^*. Consequently, as *c* increases (or decreases), the data owner's net payoffs for all strategies decrease (or increase) proportionally. That is, a change in *c* will not affect the results in relative comparison between different strategies.

We have analyzed different strategies for scenario (i). The same process of analysis can be applied to scenarios (ii) and (iii). We will not pursue that exercise due to the length constraints. We should point out that the result that the best strategy for the data owner is to offer both types of data is derived based on the utility functions given in this example. The result could be different with different utility functions.

5. Practical insights

The buying and selling of customer data are a common practice in database marketing. There are in general two types of data usages in database marketing. In the early analysis stage, the focus is on the use of statistical and data mining techniques to develop models of customer behavior. This corresponds to type *A* usage. The results of the data analysis are then used in the later stage to select target customers for communications, which can be considered as type *I* usage. The identifying attributes, such as name, phone number and email address, are typically not useful at the analysis stage but are necessary at the communication stage. The models developed in this study offer valuable insights for data owners who provide the data for database marketing (e.g., Acxiom and credit bureaus). For example, the data owners can offer a lower price (α_L^* and β_L^*) for data without identifying attributes, to be used for analysis/modeling purposes, and a higher price (α_H^* and β_H^*) for data with identifying attributes, used for communications with individuals. In this way, the data buyer's purpose is self-revealed, and the buyer will have no incentive to buy the data that is not designed for that purpose. Note that α_L^*, β_L^*, α_H^* and β_H^* are optimal prices to the data buyers given their intended purposes. Without this optimality, even if the high sensitivity data are priced higher than the low sensitivity data, the data user does not necessarily end up buying what the user initially intended to buy.

On the buyer's side, the marketers can also benefit from this pricing scheme. They can buy a large amount of de-identified data at lower price for data analysis and mining, and then acquire a smaller amount of data with identifying attributes based on the results of the data analysis. This enables the marketers to focus on the customers who are more suitable targets, and thus reduce the cost of marketing and the risk of privacy infringement.

The models developed in this study also provide helpful insights for data owners in determining the cost of collecting private data. Since the cost is tied to the price in the model, sensitivity analyses can be performed to more accurately evaluate the cost of data with respect to privacy variables. The College Board, for example, provides students' standardized test score data to colleges and universities at a minor charge [29]. The data is offered at a highly aggregated level due to privacy concern. Consequently, colleges and universities buy excessive amount of data and mail more brochures than necessary to prospective students. However, it is quite likely that many students are not very sensitive about letting universities know their detailed test score information. If the College Board provides a financial incentive for students to permit more detailed disclosure of their test data, and likewise charges a higher price for acquiring and using such data, it will be economically more efficient for both the universities and the students to find their matches.

The proposed incentive-compatible mechanism also helps understand the rationale behind the decision on the DOJ vs. Google case. In this event, the DOJ stated that its purpose of getting data is to find patterns in Internet pornography and it had no intention to investigate individual cases. As such, the DOJ played a type *A* user's role here. The judge ruled that Google provides to the DOJ a small sample set of URLs, which contain website contents but not Google users' identities and thus are of low sensitivity. The judge, however, denied the DOJ's motion to acquire users' search queries from Google's databases, which contain both user identities and website contents and thus are of high sensitivity. This decision is consistent with the proposed incentive-compatible mechanism.

6. Future research

Our model considers two types of data consumers, resulting in two corresponding sets of optimal prices and choices. If there are more than two types of data consumers, the other types will be forced to choose from one of the two prices offered, causing non-optimal choices (on the positive side, restricting to two types enables incentive compatible self-selection even if data users' utility functions are not estimated very accurately). Our modeling framework can be extended to more than two types by adding corresponding IR and IC constraints for the additional types, and adjusting estimated distribution for the type proportion. However, the problem of finding an optimal set of sensitivity levels might not have a feasible solution when the number of types is greater than two. Moorthy [23] shows that for the market segmentation problem some conditions regarding the relationships in utility functions and type proportions must be satisfied in order to guarantee the existence of the optimal solution. Perhaps due to this difficulty, related economics studies typically assume two types of consumers [16,9]. Our problem with more than two types will be more difficult than that in Moorthy [23] since it involves two decision variables (*s* and *x*) as opposed to only one in Moorthy [23].

Acknowledgments

Xiao-Bai Li's research was supported in part by the National Library of Medicine of the National Institutes of Health under Grant Number R01LM010942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Biographies

Xiao-Bai Li is a Professor of MIS in the Department of Operations and Information Systems, Manning School of Business at the University of Massachusetts Lowell. He received his Ph.D. in management science from the University of South Carolina. Dr. Li's research focuses on data mining, information privacy, and information economics. He has received funding for his research from National Institutes of Health (NIH) and National Science Foundation (NSF). His work has appeared or is forthcoming in *Decision Support Systems*, *Information Systems Research*, *Management Science*, *MIS Quarterly*, *Operations Research*, *IEEE Transactions* (*TKDE*, *TSMC*, *TAC*), *Communications of the ACM*, *INFORMS Journal on Computing*, and *European Journal of Operational Research*, among others.

Srinivasan Raghunathan is a Professor of Information Systems in the School of Management, The University of Texas at Dallas. He obtained B.Tech degree in Electrical Engineering from IIT, Madras, Post Graduate Diploma in Management from IIM, Calcutta, and Ph.D. in Business Administration from the University of Pittsburgh. His current research interests are in the economics of information security and the value of collaboration in supply chains. His papers have been published in journals such as *Management Science*, *Operations Research*, *Information Systems Research*, *Decision Analysis*, *Journal of MIS*, *Decision Support Systems*, *IEEE transactions*, *IIE transactions*, *European Journal of Operational Research, and Production and Operations Management*, among others.

References

1. Adam NR, Wortmann JC. Security-control methods for statistical databases: a comparative study. ACM Computing Surveys. 1989; 21(4):515–556.
2. Bai X, Gopal R, Nunez M, Zhdanov D. A decision methodology for managing operational efficiency and information disclosure risk in healthcare processes. Decision Support Systems. 2014; 57:406–416.
3. Bélanger F, Crossler RE. Privacy in the digital age: a review of information privacy research in information systems. MIS Quarterly. 2011; 4(35):1017–1041.

4. Cheng HK, Dogan K. Customer-centric marketing with Internet coupons. Decision Support Systems. 2008; 44(3):606–620.

5. Centers for Medicare and Medicaid Services. Agreement for use of Centers for Medicare and Medicaid Services (CMS) data containing individual identifiers. Retrieved May 5, 2013 http://www.resdac.org/sites/resdac.org/files/RIF_DataUseAgreement.pdf

6. College Board. Guidelines for the release of data. Retrieved May 5, 2013 http://www.collegeboard.com/prod_downloads/research/RDGuideforReleaseData.pdf

7. DOJ vs. Google. Ruling on Department of Justice vs. Google Inc. case. 2006. Retrieved April 23, 2007 http://www.google.com/press/images/ruling_20060317.pdf

8. Farquad MAH, Bose I. Preprocessing unbalanced data using support vector machine. Decision Support Systems. 2012; 53(1):226–233.

9. Fudenberg, D.; Tirole, J. Game Theory. The MIT Press; Cambridge, MA: 1991.

10. Garfinkel R, Gopal RD, Nunez M, Rice DO. Secure electronic markets for private information, IEEE Transactions on Systems, Man, and Cybernetics. Part A. 2006; 36(3):461–471.

11. Gopal R, Garfinkel R, Goes P. Confidentiality via camouflage: the CVC approach to disclosure limitation when answering queries to databases. Operations Research. 2002; 50(3):501–516.

12. KDnuggets. Google subpoena: child protection vs. privacy. 2006. Retrieved April 23, 2007 http://www.kdnuggets.com/polls/2006/google_subpoena.htm

13. Hann, I-H.; Hui, KL.; Lee, SYT.; Png, IPL. Online Information Privacy: Measuring the Cost-Benefit Trade-Off. Proceedings of the 23rd International Conference on Information Systems; Association for Information Systems; 2002. p. 13-42.

14. Hui, KL.; Png, IPL. The Economics of Privacy. In: Hendershott, T., editor. Handbooks in Information Systems. Vol. 1. Elsevier; 2006. p. 471-498.

15. Ishibuchi, H.; Nakashima, T.; Nii, M. Genetic-Algorithm-Based Instance and Feature Selection. In: Liu, H.; Motoda, H., editors. Instance Selection and Construction for Data Mining. Kluwer Academic; Norwell, MA: 2001. p. 95-112.

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

17. Laudon KC. Markets and privacy. Communications of the ACM. 1996; 39(9):92–104.

18. Li XB. A scalable decision tree system and its application in pattern recognition and intrusion detection. Decision Support Systems. 2005; 41(1):112–130.

19. Li XB, Jacob VS. Adaptive data reduction for large-scale transaction data. European Journal of Operational Research. 2008; 188(3):910–924.

20. Li XB, Sarkar S. Protecting privacy against record linkage disclosure: a bounded swapping approach for numeric data. Information Systems Research. 2011; 22(4):774–789.

21. Li XB, Sarkar S. Class-restricted clustering and microperturbation for data privacy. Management Science. 2013; 59(4):796–812.

22. Li Y. Empirical studies on online information privacy concerns: literature review and an integrative framework. Communications of the Association for Information Systems. 2011; 28(1) (Article 28).

23. Moorthy K. Market segmentation, self-selection, and product line design. Marketing Science. 1984; 3(4):288–307.

24. Mussa M, Rosen S. Monopoly and product quality. Journal of Economic Theory. 1978; 18(2):301–317.

25. Stokes, L.; Belin, T. What is a Margin of Error?. In: Scheuren, F., editor. What Is a Survey?. American Statistical Association; Alexandria, VA: 2004. p. 63-67.

26. Smith HJ, Dinev T, Xu H. Information privacy research: an interdisciplinary review. MIS Quarterly. 2011; 35(4):989–1015.

27. Sweeney L. *k*-anonymity: a model for protecting privacy, International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems. 2002; 10(5):557–570.

28. Sweeney L. Privacy-preserving surveillance using database from daily life. IEEE Intelligent Systems. 2005; 20(5):83–84.

29. Whiting, R. Everybody, Information Week. Jul 10. 2006 Who's Buying and Selling Your Data?; p. 30-32.

30. Zhu D, Li XB, Wu S. Identity disclosure protection: a data reconstruction approach for privacy-preserving data mining. Decision Support Systems. 2009; 48(1):133–140.

Appendix A. Proofs of propositions

Proof of proposition 1

Maximizing $R(s_L, x_L^A)$ will cause at least one of the constraints (10) and (12), which involves $R(s_L, x_L^A)$, to be binding. Similarly, maximizing $R(s_H, x_H^I)$ will cause one of constraints (11) and (13) to be binding.

i. If $U_I(s_L, x_L^{I*})/\partial x_L^{I*} > U_A(s_L, x_L^{A*})/\partial x_L^{A*}$, then this is the traditional scenario. So,

$$U_I\left(s_L, x_L^I\right) > U_A\left(s_L, x_L^A\right) \text{ and } U_I\left(s_H, x_H^I\right) > U_A\left(s_H, x_H^A\right), \quad \text{(A.1)}$$

where $x_s^t (t \in \{A, I\}, s \in \{s_L, s_H\})$ is within a small neighborhood of x_s^{t*}. Now, suppose (11) is binding; i.e., $R(s_H, x_H^I) = U_I(s_H, x_H^I)$. Then, it must be that $R(s_L, x_L^I) \geq U_I(s_L, x_L^I)$, because otherwise the type I user will select the s_L data, which is designed for the type A user. However, if $R(s_H, x_H)$ and $R(s_L, x_L)$ are set this way, then it follows from (A.1) that neither of them will be attainable by $U_A(\cdot)$. Therefore, (11) is not binding and (13) is binding. That is,

$$R\left(s_H, x_H^I\right) = U_I\left(s_H, x_H^I\right) - U_I\left(s_L, x_L^I\right) + R\left(s_L, x_L^I\right). \quad \text{(A.2)}$$

Next, because (11) is not binding, (10) must be binding. Otherwise, the data owner can increase $R(s_L, x_L)$ and $R(s_H, x_H)$ by the same small amount, which would keep the IC constraints (12) and (13) always satisfied. Continue increasing $R(s_L, x_L)$ and $R(s_H, x_H)$ in this way, eventually one of the IR constraints (10) and (11) will be binding. Since (11) will not be binding, (10) must be binding. Therefore, (14) holds.

It follows from the price function (7) that

$$R\left(s_L, x_L^I\right) = R\left(s_L, x_L^A\right) + \beta_L\left(x_L^I - x_L^A\right). \quad \text{(A.3)}$$

Substituting (14) and (A.3) into (A.2), we have (15).

ii. Given that $U_I(s_L, x_L^I)$ is non-monotonic concave and $U_A(s_L, x_L^A)$ is increasing concave, condition $U_I(s_L, x_L^{I*})/\partial x_L^{I*} = U_A(s_L, x_L^{A*})/\partial x_L^{A*}$ will occur only once. It follows from (9) that $\beta_L = U_I(s_L, x_L^{I*})/\partial x_L^{I*} = U_A(s_L, x_L^{A*})/\partial x_L^{A*}$; that is,

$$R\left(s_L, x_L^{A*}\right) = U_A\left(s_L, x_L^{A*}\right), \quad \text{(A.4)}$$

$$R\left(s_L, x_L^{I*}\right) = U_I\left(s_L, x_L^{I*}\right). \quad \text{(A.5)}$$

With (A.4), Eq. (14) holds. Substituting (A.4) and (A.5) into (12) and (13) respectively, we have

$$U_A\left(s_H, x_H^{A*}\right) - R\left(s_H, x_H^{A*}\right) \leq 0, \quad \text{(A.6)}$$

$$U_I\left(s_H, x_H^{I*}\right) - R\left(s_H, x_H^{I*}\right) \geq 0. \quad \text{(A.7)}$$

As $R(s_H, x)$ increases, (A.7) will eventually be binding while (A.6) will be further away from binding. Therefore, (13) will be binding and (12) will not. Also, with (A.7) binding, (16) holds.

iii. If $U_I(s_L, x_L^{I*})/\partial x_L^{I*} < U_A(s_L, x_L^{A*})/\partial x_L^{A*}$, then β_L from (9) based on $U_A(s_L, x_L^A)$ will be greater than that based on $U_I(s_L, x_L^I)$. So, the price function based on the former will dominate that based on the latter; that is, $R(s_L, x_L^A) > R(s_L, x_L^I)$. As a result, to maximize prices, both (10) and (11) must be binding, which is the best possible choice for the data owner out of all feasible binding alternatives. Note that when both (10) and (11) binding, $R(s_L, x_L)$ is not attainable by $U_I(s_L, x_L^I)$ and $R(s_H, x_H)$ is not attainable by $U_A(s_H, x_H^A)$. So the right hand sides of the IC constraints (12) and (13) are negative and both constraints are satisfied. In other words, the data consumers will have no incentive to select the price designed for the other type.

Proof of proposition 2

It follows from (9) that β_L^* is the slope of $U_A(s_L, x_L^A)$ at x_L^{A*}. On the other hand, it follows from(18) that β_L^* is also the slope of $C(s_L, x_L^A)$ at x_L^{A*}. In other words, if we “raise” the convex curve $C(s_L, x_L^A)$ such that it eventually tangents with the concave curve $U_A(s_L, x_L^A)$, then x_L^{A*} is the tangent point and β_L^* is the slope of the price line that passes x_L^{A*}. Thus,

$$\partial U_A\left(s_L, x_L^A\right)/\partial x_L^A\big|_{x_L^A = x_L^{A*}} = \beta_L^* = \partial C\left(s_L, x_L^A\right)/\partial x_L^A\big|_{x_L^A = x_L^{A*}}. \quad \text{(A.8)}$$

Similarly, it follows from (9) and (20) that

$$\partial U_I\left(s_H, x_H^I\right)/\partial x_H^I|_{x_H^I=x_H^{I*}}=\beta_H^*=\partial C\left(s_H, x_H^I\right)/\partial x_H^I|_{x_H^I=x_H^{I*}}. \quad \text{(A.9)}$$

Consider $x_L^{A*} \leq x_H^{I*}$. Let $x_L^{A*} \leq x \leq x_H^{I*}$. It follows from (2) that $\partial C(s_L, x)/\partial x < \partial C(s_H, x)/\partial x$. Thus, by comparing the right sides of (A.8) and (A.9), we have

$$\partial C\left(s_L, x_L^A\right)/\partial x_L^A|_{x_L^A=x_L^{A*}}<\partial C\left(s_H, x_H^I\right)/\partial x_H^I|_{x_H^I=x_H^{I*}} \Rightarrow \beta_L^*<\beta_H^*.$$

Now, if $x_L^{A*}>x_H^{I*}$, let $x_L^{A*} \geq x \geq x_H^{I*}$. It follows from (3) that $\partial U_A(s_L, x)/\partial x < \partial U_I(s_H, x)/\partial x$. Thus, by comparing the left sides of (A.8) and (A.9), we have

$$\partial U_A\left(s_L, x_L^A\right)/\partial x_L^A|_{x_L^A=x_L^{A*}}<\partial U_I\left(s_H, x_H^I\right)/\partial x_H^I|_{x_H^I=x_H^{I*}} \Rightarrow \beta_L^*<\beta_H^*.$$

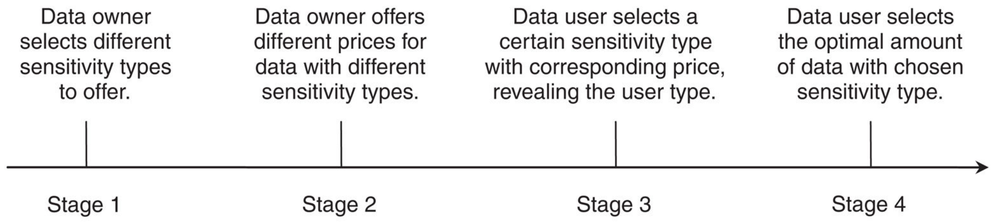
Data owner selects different sensitivity types to offer.
Data owner offers different prices for data with different sensitivity types.
Data user selects a certain sensitivity type with corresponding price, revealing the user type.
Data user selects the optimal amount of data with chosen sensitivity type.
Stage 1
Stage 2
Stage 3
Stage 4

Fig. 1.
Sequence of events.

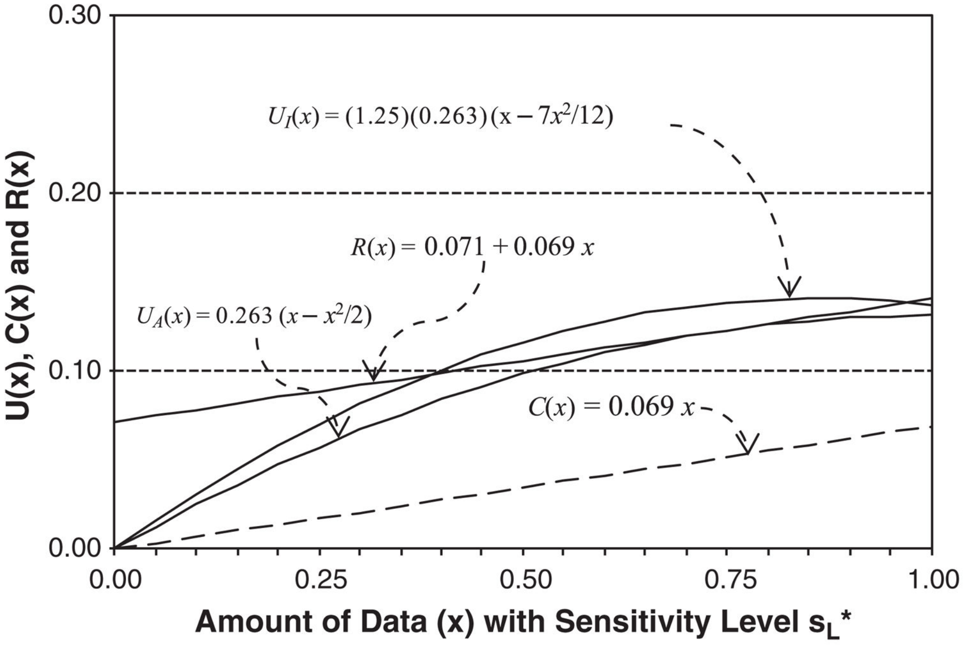

Fig. 2.

Utility, cost and price for low sensitivity data – scenario (i).

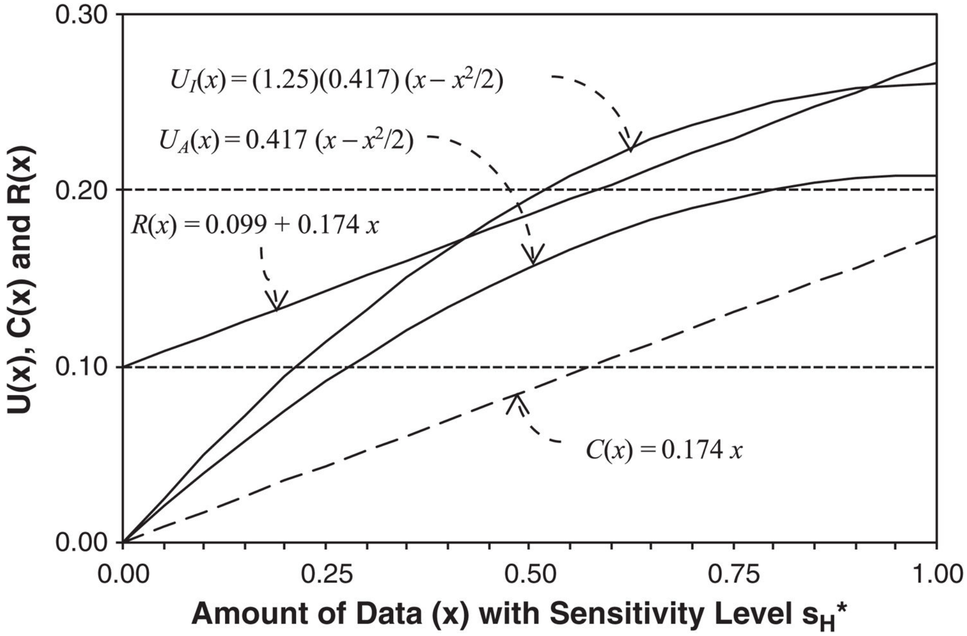

Fig. 3.

Utility, cost and price for high sensitivity data – scenario (i).

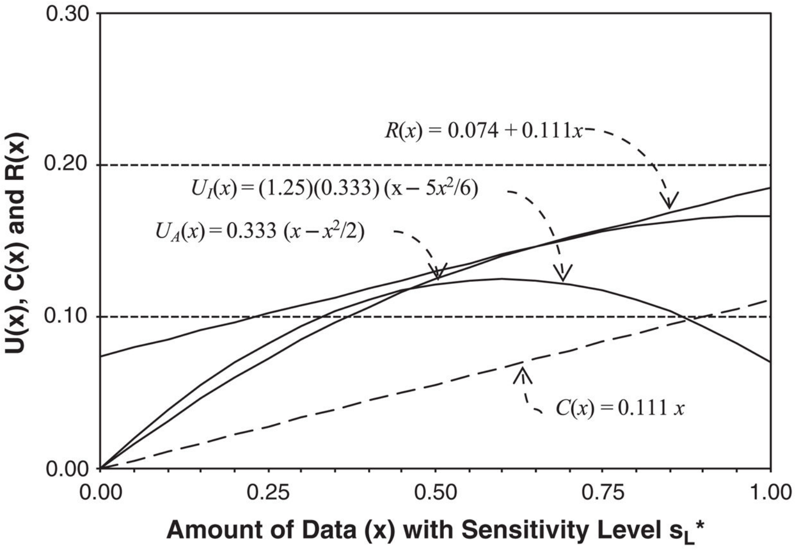

Fig. 4.

Utility, cost and price for low sensitivity data – scenario (iii).

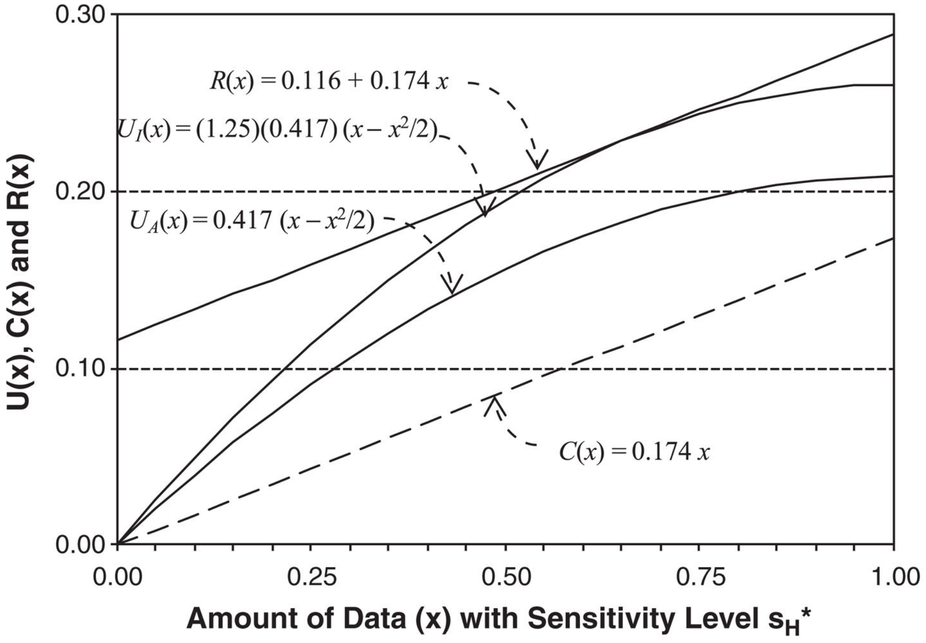

Fig. 5.

Utility, cost and price for high sensitivity data – scenario (iii).

Table 1

Table of notation.

x, x_s^t	Amount of data; and amount of data acquired by user type $t \in \{A, I\}$ for a given sensitivity level s
s_L, s_H	Sensitivity type of data
$U_A(s, x)$, $U_I(s, x)$	Utility of type A and type I data consumers respectively
$C(s, x)$	Total cost of data for data owner
$R(s, x)$	Total price charged to data consumer
α_s	Fixed charge for a given sensitivity level s
β_s	Rate of variable charge for a given sensitivity level s

Table 2

Results of different data offering strategies (p = 0.5, c = 1, k = 1.25).

Strategy	$(s_L^*)^{1/2}$	$(s_H^*)^{1/2}$	α_L^*	α_H^*	***E*(*P*)**
1. Offer s_L to type *A* and s_H to type *I*	0.2626	0.4167	0.0714	0.0994	0.0854
2. Offer s_{L2} to both type *A* and type *I*	0.3333		0.0741		0.0741
3. Offer s_{L1} to type *I* only	0.4167		0.0992		0.0496
4. Offer s_{H2} to both type *A* and type *I*		0.3333		0.0741	0.0741
5. Offer s_{H1} to type *I* only		0.4167		0.1157	0.0579