information

MDPI

Article

Personal Data Market Optimization Pricing Model Based on Privacy Level

Jian Yang [1,*] and Chunxiao Xing [2]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] Research Institute of Information, Beijing National Research Center for Information Science and Technology, Department of Computer Science and Technology, Institute of Internet Industry, Tsinghua University, Beijing 100084, China; xingcx@tsinghua.edu.cn
* Correspondence: yjian180@gmail.com

Received: 3 March 2019; Accepted: 30 March 2019; Published: 3 April 2019

Abstract: In the era of the digital economy, data has become a new key production factor, and personal data represents the monetary value of a data-driven economy. Both the public and private sectors want to use these data for research and business. However, due to privacy issues, access to such data is limited. Given the business opportunities that have gaps between demand and supply, we consider establishing a private data market to resolve supply and demand conflicts. While there are many challenges to building such a data market, we only focus on three technical challenges: (1) How to provide a fair trading mechanism between data providers and data platforms? (2) What is the consumer's attitude toward privacy data? (3) How to price personal data to maximize the profit of the data platform? In this paper, we first propose a compensation mechanism based on the privacy attitude of the data provider. Second, we analyze consumer self-selection behavior and establish a non-linear model to represent consumers' willingness to pay (WTP). Finally, we establish a bi-level programming model and use genetic simulated annealing algorithm to solve the optimal pricing problem of personal data. The experimental results show that multi-level privacy division can meet the needs of consumers and maximize the profit of data platform.

Keywords: privacy data; compensation; willingness to pay (WTP); pricing; multi-level

1. Introduction

In recent years, due to the widespread use of the Internet of Things and big data technologies, the amount of personal information collected by some large internet companies and social networking services has reached an unprecedented level [1]. These personal data are very valuable for the public and private sectors to improve their products or services. However, raw data may contain sensitive information about individuals, access to personal data is strictly limited. In particular, some companies and organizations have collected a large amount of personal information that is very useful to third parties. For example, banks have personal assets and credit information that P2P lenders want to know to decide whether to issue a loan, or search engines, such as Google, have search queries for millions of users that research institutions want to know to study internet pornography. Due to the protection of personal privacy by relevant laws [2], these entities are usually not willing to allow others to access such data. Meanwhile, the demand for individual data by third parties is increasing for research and commercial purposes. However, there is actually no safe and effective supply of personal data. To solve the contradiction between supply and demand of personal data in practical applications, the most effective solution is to establish a personal data market [3] to achieve a balance between supply and demand. The data publisher can sell the personal data to the data platform according to

the privacy preference [4] to obtain the corresponding privacy compensation or remuneration, and the data consumer can pay a certain subscription fee to the data platform to obtain the data resource of interest. Therefore, data market services require privacy awareness and pricing mechanisms to achieve maximum profit by jointly optimizing privacy levels and subscription fees.

In fact, some startups are currently developing applications to support this trend. For example, Datacoup [5] have created the world's first personal data market, which contains thousands of personal information about its users (i.e., demographic, education, location, profession, spending, health, interests, etc.). However, they did not provide an effective pricing mechanism. The subscription fee for Datacoup is fixed. In addition, the data provider's privacy attitude has not been effectively considered. Because of individual differences, each person's privacy attitude is different, so the utility [6,7] of publishing data is different, which involves compensation for data providers during the transaction and is closely related to the operating costs of the data market.

In this paper, we consider three important issues in the pricing model of personal data market. Firstly, from the perspective of data market cost, how to establish a fair privacy compensation mechanism for different privacy attitudes of data providers. Secondly, from the perspective of consumer self-selection, how to effectively balance the relationship between privacy level and willingness to pay [8]? Finally, from the perspective of data market revenue, what is the optimal subscription fee and privacy level required to maximize data market profits? In this paper, we propose a personal data market pricing model based on privacy level, which allows the data market to compensate data publishers with different privacy attitudes and provide subscription services for data consumers to maximize their profits. The main contributions of this paper are summarized as follows:

- We propose a fair privacy compensation mechanism. From the perspective of data publishers, they can be compensated according to their privacy attitudes. From the perspective of data market, their operating costs are effectively controlled.
- We present a nonlinear mathematical model that describes the relationship between consumer self-selection behavior and privacy-aware data utility.
- We developed a bi-level programming [9,10] model to maximize the profit of the data market. The data market collects personal data from privacy-aware people and provides subscription services to data consumers. Maximize profits through optimization of subscription fees and privacy levels.

The remainder of this study is organized as follows. Section 2 first describes the basic structure of the personal data pricing framework. Then, a fair compensation mechanism is described. Finally, the consumer willingness to pay based on the privacy data utility function is analyzed, and a nonlinear model is established to describe the relationship between them. In Section 3, considering the two layers of decision making of the data platform and the consumer, we built a bi-level programming model to maximize the profitability of the data platform. To make the problem solvable, in Section 4, we use genetic simulated annealing algorithms to solve complex data pricing problems. Section 5 presents and analyzes the numerical experiment results. Section 6 concludes the paper.

2. Framework Description

2.1. Stakeholder Description

Before we proceed any further, it will be helpful to define a few notations. Table 1 summarizes the key parameters and descriptions used in the model.

Table 1. Frequently used notations.

Notation	Description
i	The number of privacy data sensitivity levels i = 1, 2,..., N
j	The number of data consumers, j = 1, 2,..., M
p	Unit price of privacy data
τ	The maximum tolerance privacy loss by data providers
C_1	The privacy compensation of risk taker data providers
C_2	The privacy compensation of risk averse data providers
s	The sensitivity level of privacy data
h	The type of heterogeneous customer
s_h	The consumer h's demand for data sensitive level s
$W(h,s)$	The willingness to pay of consumer type h for data sensitivity level s
U_{ij}	The utility of consumer j for data sensitivity level i
x_{ij}	The purchasing decision of consumer j for data sensitivity level i , x_{ij} = 0 or 1

Next, we describe the basic framework of the private data pricing, illustrated in Figure 1. The framework consists of three stakeholders, i.e., data providers, data market, data consumers.

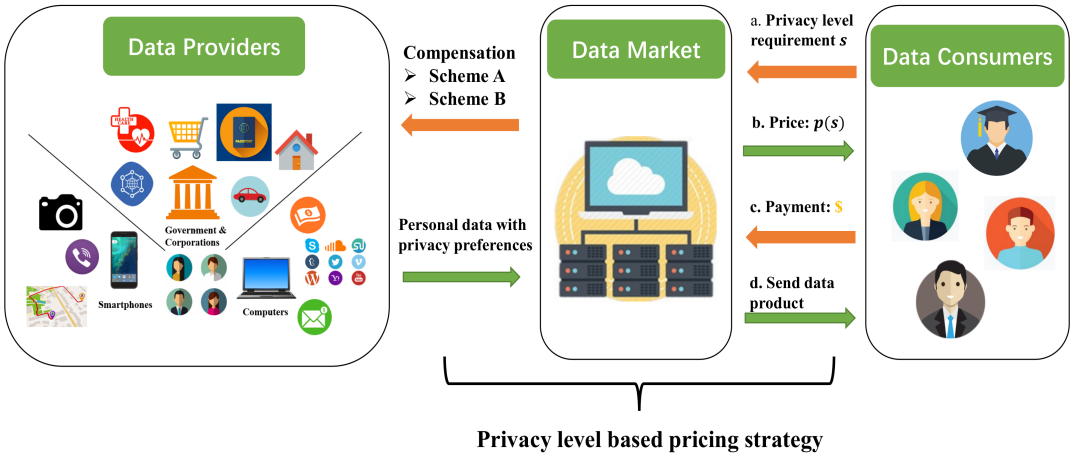

Figure 1. A typical personal data market framework.

- **Data provider** Let K denote the number of data providers who are willing to sell their private data. Each data provider $i \in K$ is modeled with a non-decreasing function (i.e., compensation scheme) C_i: $\tau \rightarrow R^+$, representing a promise between a data platform and a data provider on how much data provider should be compensated for their privacy loss τ (Privacy loss is the extent to which individual sensitive information is disclosed. The quantification of privacy loss to an individual is defined by differential privacy [11,12]), where τ is the quantification of privacy loss of data provider as τ and money are correlated [13,14]. In addition, the compensation scheme is determined by the data market according to the type of data provider (i.e., privacy preference). Therefore, a good compensation scheme can not only help data owners understand and determine their τ, but also make the cost of the data platform more transparent and controllable. In Section 2.2, we discuss how to design such a compensation scheme.
- **Data market** (a.k.a., data platform) is an intermediary between data consumers and data providers. It is trusted by both parties and collects privacy-aware personal data from data providers to compensate them appropriately. Instead of asking for their valuations, data providers are given a fixed number of options. Based on the compensation model selected by the data provider, the data platform determines privacy level and compensation amount for the data they provide. Meanwhile, the data platform also determines the data price for different privacy levels to cover

its operating costs. Therefore, the data platform needs to jointly optimize purchase fees and privacy compensation fees to maximize profits.

- **Data consumer** is the terminal of the entire framework. For research and commercial purposes, the terminal consumer can decide whether to purchase the data or services provided by the data platform based on their willingness to pay (WTP). It is also worth mentioning that the purchase task can only be completed when the consumer's willingness to pay is greater than the utility of the data provided. Otherwise, the purchase process is considered to be a failure.

In summary, our framework works as follows. Data owners provide personal data to the data market based on their privacy preferences and receive appropriate compensation. The data provided by the data owner is stored by the trusted data market. To make a profit, the data market needs to charge the data consumer a certain fee, and the data consumer can purchase data products based on their willingness to pay.

2.2. Fair Privacy Compensation Mechanism

Much like traditional commodity trading, the most important focus in the data market is fairness and truth. Of course, this is a basic requirement for all trading processes. For data market and data providers, unrealistic privacy assessment is an important reason for the significant increase in data platform collection cost. Without a well-designed privacy compensation mechanism, some shrewd data owners will always try to choose any solution that will bring them more benefits, who may deliberately report unreasonably high privacy assessments [15]. For example, literature [12] applies a linear compensation scheme ($C = \tau \cdot p$) and allows each data provider to define the unit price of private data p. Under the same τ, most data providers will always set very high p for maximum benefit. If the data platform sets the compensation scheme to a fixed value in order to show its monopoly position, then the situation will be worse. For example, literature [16] set a fixed compensation scheme, which is unfair to data owners with different privacy attitudes. Because it is possible that sellers offering different levels of sensitivity receive the same amount of compensation, which is unfair to data owners who provide highly sensitive information. Furthermore, it will accelerate the loss of the providers who offer the stable and high quality data sources to the data platform, which is not conducive to the long-term stable development of the platform. To establish a fair privacy compensation mechanism, encourage real privacy assessments, and without compromising the interests of highly sensitive information providers, the data platform should provide appropriate compensation schemes that correspond to the data provider's privacy attitude.

In this paper, we used the compensation scheme proposed by [12], which allows individuals to provide personal data to data platform and receive financial compensation. Of course, the amount of compensation varies depending on the privacy attitude of the individual and the sensitivity of the information provided by the individual. Specifically, the data platform can classify data providers into two types, i.e., risk averse [17] and risk taker, as shown in Figure 2. Meanwhile, we established corresponding compensation mechanisms for these two types of data providers.

- **A. Risk Taker** Such data providers are subjectively willing to take risks and have strong risk tolerance. A privacy compensation mechanism based on sublinear functions is designed to support such data providers.

$$C_1 = 1 - \frac{1}{(1 + 0.25s)} \tag{1}$$

- **B. Risk Averse** Such data providers are relatively conservative and unwilling to take risks. A privacy compensation mechanism based on a logarithm function is designed to support such data providers.

$$C_2 = \frac{log(10^3 s + 1)}{35} \tag{2}$$

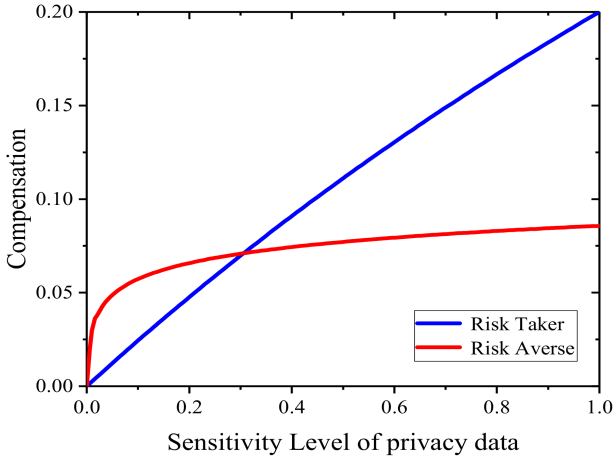

Figure 2. Compensation mechanism for two different privacy attitudes.

Here s is the sensitivity level of the data or the privacy level, which is proportional to the loss of privacy τ, i.e., $s = k \cdot \tau$ ($k > 0$). In other words, the greater the loss of privacy, the more information is disclosed, and thus the sensitivity level of the data is higher. For risk taker, moderate privacy losses, there is a small amount of compensation, but for significant privacy losses, a huge amount of compensation is required. For the risk averse, even the smallest loss of privacy has a non-zero small compensation. However, even the largest compensation is far less than the risk taker's compensation.

2.3. Customer's Self-Selection Behavior Analysis

In the privacy data pricing optimization model, consumer self-selection behavior analysis [18,19] is an important component and foundation. In an actual market environment, customer purchases of data products are complex and can be affected by a variety of factors, such as consumer preferences, interests, and data product prices. In this paper, assuming that the consumer is rational, the consumer's purchase selection behavior for the data product is determined by the utility function of the privacy data.

In previous work, most studies assumed that the willingness to pay was a linear function [18,20]. However, the linear value function does not accurately capture the estimates of actual data and consumer's self-selection behavior. A typical example is the use of opinion polls to estimate public opinion in the world. It is commonly known that as the sensitivity items (i.e., age, gender, race, sexual orientation, income) increase, the results of the survey will be more valuable, but as the entries become more detailed, the increase in estimated accuracy will decrease. For instance, the probability of breaking up revenue into ten groups would not be five times that of the two groups. Similar observations have been made in the task of designing machine learning and data mining. In this paper, we propose a willingness to pay function based on the sensitivity level of privacy data.

For a given consumer, there is a willingness to pay function: $W(h, s)$, where h represents the type of heterogeneous consumer and s is the sensitive level of privacy data that the consumer wants to purchase. We have $h \in (0, h_{max})$, $s \in (0, s_{max})$, where h_{max}, s_{max} corresponds to the maximum value. Consumers get utility at price p^*, so $U(h, s^*, p^*) = W(h, s^*) - p^*$, where privacy level s^* is priced as p^*.

$$W(h,s) = \begin{cases} 0, & s < \lambda s_h \\ 2(\frac{s-\lambda s_h}{s_{h_{max}} - \lambda s_h})^2, & \lambda s_h < s < s_h \\ 1 - 2(\frac{s_{h_{max}} - s}{s_{h_{max}} - \lambda s_h})^2, & s_h < s < s_{h_{max}} \end{cases} \tag{3}$$

Assuming that a consumer has a specific data sensitivity level requirement, $s_h = r(h)$, and $r(h) > 0$, $r'(h) > 0$ indicate that a high-type consumer has a higher demand for data sensitivity. In this paper, we use $s_h = (h/h_{max}) s_{h_{max}}$, which means that the level of data sensitivity required by consumers corresponds to its type. Assuming that s_h is the data sensitivity level required by consumer h, the evaluation function is defined as Equation (3).

There is an infimum point (λs_h). When $s < \lambda s_h$, it indicates that the sensitivity level of the data is too low, and can not bring useful value to consumers, so consumers judge the value of such data as 0. When $\lambda s_h < s < s_h$, it indicates that as the level of data sensitivity increases, the resulting utility will also increase, which will increase the willingness of consumers to pay. When $s_h < s < s_{h_{max}}$, due to the marginal utility be a decreasing function [21], as the sensitivity level increases further, the increase in data utility will decrease. Figure 3 plots the image of the evaluation function.

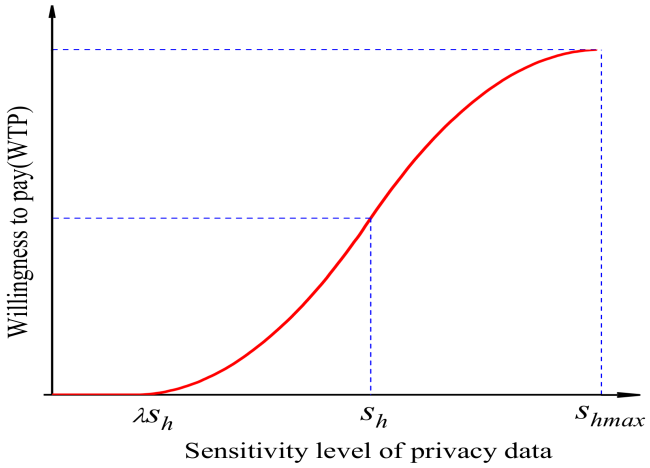

Figure 3. Consumers' willingness to pay for privacy levels.

3. Bi-Level Programming Model of Pricing Privacy Data

Consider a real-life scenario that monopolizes the personal data market. In such a market, the data platform first determines the level of sensitivity and corresponding price of the privacy data, with the aim of maximizing profits by providing services that sell data. We made some assumptions about the leader level, including:

- Provide N sensitivity levels for privacy data. When $N = 1$, the platform monopolist provides a non-hierarchical service, and when $N \geq 2$, the platform monopolist provides a level difference service.
- The different levels of sensitivity of the data correspond to different utility U_i, indicating that the i-th level provides data utility for the consumer, and the corresponding price for each level is labeled p_i.
- The price p_i increases as the level of data sensitivity increases.

At the following level, potential data consumers determine whether to purchase the appropriate level of privacy data through a self-selection process to maximize their utility. The consumer's self-selection is assumed as follows:

- The utility of consumer j is for products with data sensitivity level i. If there are multiple privacy data sensitivity levels that meet expectations, consumer j will choose the data product that can produce the most utility, i.e., $U_{ij} = \arg\max_i \{U(h_j, s_i, p_i)\}$. If $U_{ij}(u_i, p_i) < 0$, the customer wouldn't subscribe to data product with sensitivity level i.
- Each consumer can only subscribe to no more than one type of data product. Suppose we measure the consumer's self-selection process through x_{ij}, Where i represents the sensitivity level of the privacy data. If $x_{ij} = 1$, the consumer accepts subscription for product with a sensitivity level i. If $x_{ij} = 0$, or not.
- The marginal willingness to pay for all potential data consumers h is between 0 and 1 ($h \in [0, h_{max}]$) subject to a Gaussian distribution, and once a choice is made, there is no change.

In summary, the above model can be expressed as follows:

- Leader level: data markplace's decision

$$Maximize\ \mathcal{G}(s,p,x) = \sum_{j=1}^{M}\sum_{i=1}^{N}(p_i - c_i)x_{ij} \tag{4}$$

$$s.t. \quad p_i > c_i, \qquad i = 1,2,\ldots,N \tag{5}$$

$$s_{i+1} > s_i, \quad i = 1,2,\ldots,N-1 \tag{6}$$

$$p_{i+1} > p_i, \quad i = 1,2,\ldots,N-1 \tag{7}$$

- Following level: data customer's decision

$$Maximize\ \mathcal{U}_j(x) = \sum_{j=1}^{M} U_{ij}(u_i,p_i)x_{ij} \tag{8}$$

$$s.t. \quad x_{i_1 j}x_{i_2 j} = 0\ if\ i_1 \neq i_2,\ i_1,i_2 = 1,2,\ldots,M \tag{9}$$

$$x_{ij} = 0\ if\ U_{ij}(u_i,p_i) < 0 \tag{10}$$

$$x_{ij} = 0\ or\ 1 \tag{11}$$

We consider the consumer's self-selection process and the monopolist's decision-making behavior through a bi-level programming model. Generally speaking, the online trading process of the data platform includes two parts. First, the division of the privacy data sensitivity level and its corresponding price. Second, the appropriate level of sensitivity data is selected based on the consumer's willingness to pay. It is also worth noting that since the data platform has the data pricing power, its pricing strategy will directly affect the consumer's selection and have a great impact on the profit of the data platform. If the data product is not priced reasonably, the data platform needs to change the pricing strategy. In our bi-level programming model, Equation (4) represents the total profit of the data platform, and the constraints of (5)–(7) ensure that the pricing and sensitivity levels are relatively reasonable. Equation (8) indicates the utility of the consumer. The constraints of (9)–(11) indicate that each consumer can only select one level of data products and require their utility to be non-negative.

4. BLGASA for Obtaining Optimal Solutions

The uncertainty of the Bi-Level Programming model(BLP) is very high and the difficulty of solving it is very large. At present, the solution for bi-level programming mainly has the following aspects: (1) extreme point search method [22], (2) KKT(Karush-Kuhn-Tucker method) [23], and (3) decent method [24]. When applying these methods, the functional formula must satisfy the characteristics of continuity, convexity and differentiability. In fact, as [25] have shown, even if all the functions involved are linear, it is NP-hard. Therefore, dealing with nonlinear versions of these models is quite difficult.

In recent years, some non-numeric optimization methods, such as genetic algorithms (GA) [26,27], have been widely applied to BLP, i.e., supply chain, traffic planning, vehicle scheduling, and commodity pricing. Co-evolution of groups is realized through evolutionary operations such as selection, crossover, mutation, etc. GA has strong global search ability, but its local search ability is poor, which is prone to "premature" convergence problem [28]. In this paper, we propose a personal data pricing problem based on privacy levels. For different levels of privacy data, the price is uncertain. Hence, the range of values of each gene in the chromosome is large. Furthermore, the crossover operation will increase the price of the same level of data and result in a lower quality of the solution. To overcome the shortcomings of genetic algorithm in solving the problem of personal data pricing, we combine genetic algorithm with simulated annealing algorithm to learn the length and complement each other, and a hybrid genetic simulated annealing algorithm(SA) is proposed to solve the bi-level programming

problem, named as **BLGASA**. SA [29] is a simulation of the annealing process in thermodynamics. It is an optimization method that randomly finds the global optimal solution of the objective function in the solution space by slowly descending the temperature parameter at a given initial temperature and according to the corresponding probability acceptance criterion. Although SA has strong local search ability, its global grasp ability is weak. GASA is an optimization algorithm composed of GA and SA, which overcomes the shortcomings of slow convergence of GA and SA easy to fall into local optimum, thus improving the quality of the problem.

The privacy data pricing process based on BLGASA is shown in Figure 4.

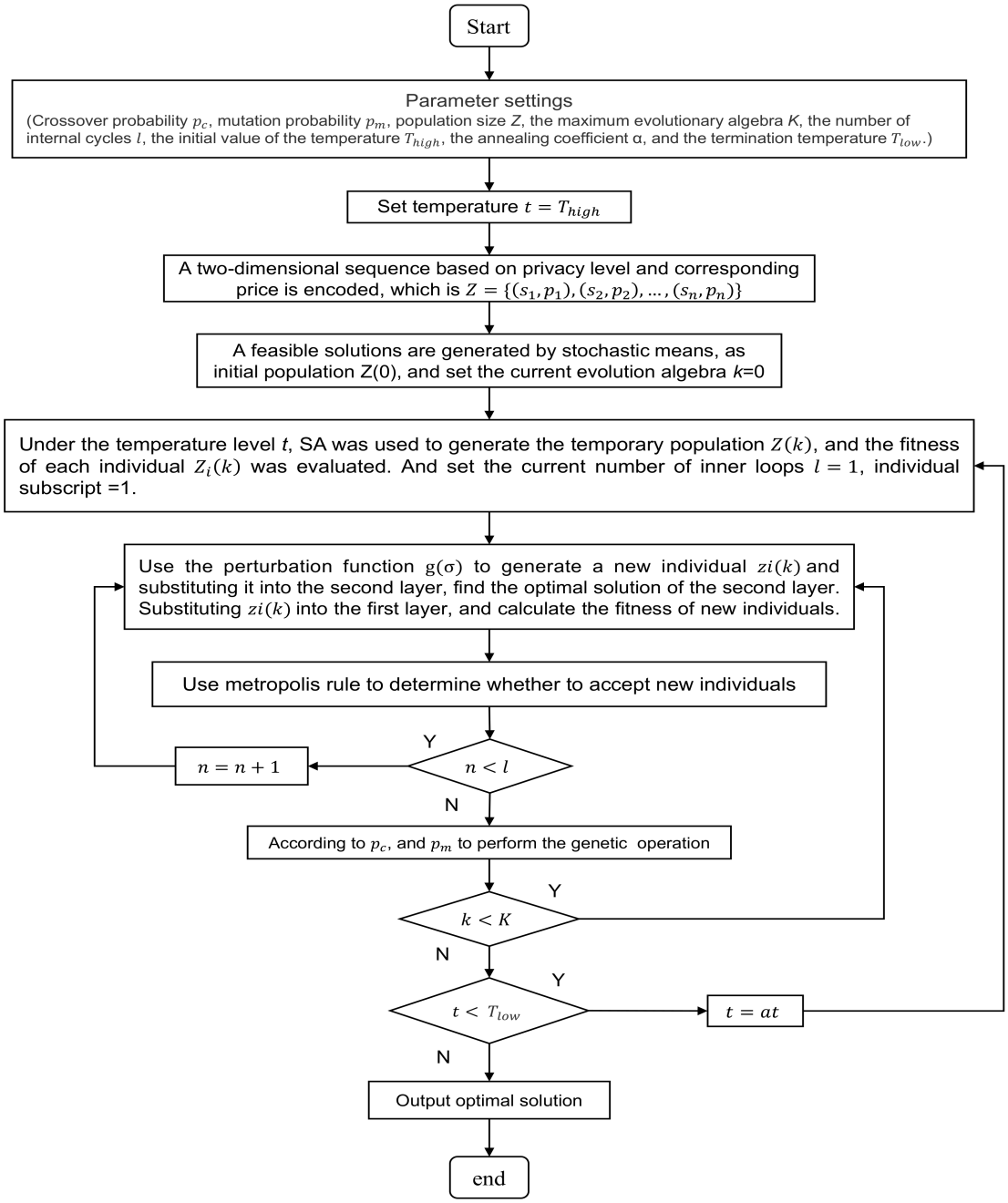

Figure 4. Process diagram of private data pricing based on BLGASA.

- **Step.1** Select the parameters of $BLGASA$. At this stage, the evaluation parameters of GA and SA need to be determined. In the genetic algorithm, take the crossover and mutation probability as p_c and p_m, the population size Z, the maximal generation K, and in the simulated annealing

algorithm, take the number of internal cycles as l, the initial value of the temperature T_{high}, the annealing coefficient as α, and the termination temperature as T_{low}.

- **Step.2** Set temperature $t = T_{high}$.
- **Step.3** A two-dimensional sequence based on privacy level and corresponding price is encoded, i.e., $Z = \{(s_1, p_1), (s_2, p_2),...,(s_n, p_n)\}$. According to the objective function of the leader level, the fitness of an individual is the total profit.
- **Step.4** A feasible solutions are generated by stochastic method, taking it as the initial population $Z(0)$, and set the current evolution algebra $k = 0$.
- **Step.5** Under the temperature level t, SA was used to generate the temporary population $Z(k)$, and the fitness of each individual $Z_i(k)$ was evaluated. Then, set the current number of inner loops $l = 1$, individual subscript $i = 1$.
- **Step.6** Generate a new individual $\tilde{z}_i(k)$ by using the perturbation function $g(\sigma)$ and substituting it into the second layer, find the optimal solution of the second layer, let $\tilde{z}_i(k) = argmaxU(k)$. Substituting $\tilde{z}_i(k)$ into the first layer, and calculate the fitness of new individuals. Among them, the perturbation function $g(\sigma)$ is a random number in $[-\sigma z, \sigma z]$, σ is the perturbation coefficient, and the default is 0.05.
- **Step.7** Use metropolis rule to determine whether to accept new individuals $\tilde{z}_i(k)$, $\Delta f = g(\tilde{z}_i(k))$ - $g(z_i(k))$:

  1. If $\Delta f > 0$, replace $\tilde{z}_i(k)$ with $z_i(k)$.
  2. Otherwise, generate a random number $\eta \in (0, 1)$, if $exp(-\Delta f/(\alpha t)) > \eta$, replace $\tilde{z}_i(k)$ with $z_i(k)$

- **Step.8** Set $n = n + 1$, if $n < l$, return Step.6. Otherwise, execute Step.9.
- **Step.9** The crossover and mutation operations are performed according to p_c and p_m, $z(k + 1)$ is generated, and the fitness of each individual is calculated.
- **Step.10** Set $k = k + 1$, if $k < K$, return Step.6. Otherwise, execute Step.11.
- **Step.11** Set $t = \alpha t$, if $t > T_{low}$, return Step.5. Otherwise, execute Step.12.
- **Step.12** Output optimal solution

5. Numerical Experiment

In this Section, we use the BLGASA to solve the bi-level programming model of personal data pricing. The purpose is to make the data market obtain the optimal number of privacy levels, and then maximize the total marginal profit. The cost of the data market is determined by the payment overhead of the personal information provided by the data owner of Risk Taker (C_1) and Risk Aversion (C_2) in Section 2. This section contains two subsections. The first subsection describes the various parameters in the experiment. Tables 2 and 3 list the characteristics of the consumer and data products. Table 4 shows the parameter settings of BLGASA. The second subsection shows our experimental results and related conclusions.

Table 2. Parameter settings for customers.

Related Variable	Parameterization on the Customer Side
Customer type	$h \in [0, h_{max}]$, $h_{max} = 1$
Customer distributions	$h \sim N(\mu, \sigma^2)$ $\mu = 0.5$, $\sigma = 0.25$
Customer specific privacy level requirement	$s_h = r(h) = h$
Customer willingness to pay function	$\lambda = 0.2$, $s_h = 0.6$
The total number of potential data consumers	M = 10,000

Table 3. Parameter settings for privacy data.

Related Variable	Parameterization on the Data Product Side	
Maximum level number	$N \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$	
Highest data sensitivity level	$s_{h_{max}} = 1$	
Cost function	Sublinear	$C_1 = 1 - \frac{1}{(1+0.25s_n)}, n = 1, 2, ..., N$
	Logarithm	$C_2 = \frac{log(10^3 s_n+1)}{35}, n = 1, 2, ..., N$

Table 4. The BLGASA parameters and their values.

Parameter	Notation	Value
Population size	Z	50
Crossover rate	p_c	0.8
Crossover operator	-	Uniform
Mutation rate	p_m	0.01
Mutation operator	-	Bit-flip
Maximum generation	K	500
Initial temperature	T_{hith}	250
Annealing coefficient	α	0.9
Number of internal cycles	l	8
Termination temperature	T_{low}	0.01

5.1. Experimental Design

Table 2 lists the relevant parameter settings for heterogeneous customers. Assuming that customer type in the market are obey Gaussian distribution and set the total number of potential data consumers is 10,000. Table 3 shows the relevant parameters for the personal data product, which is assumed to be divided into 10 levels based on the maximum number of the privacy level and the corresponding price. In addition, we consider two cost functions in Section 2, i.e., the sub-linear cost function and the logarithm cost function. Table 4 lists the relevant parameters and optimal values of the BLGASA.

All experiments were coded in python and run independently 30 times to obtain more accurate and stable optimal numerical solutions.

5.2. Experimental Results

As can be seen from Figure 5, the total profit of data platform increases in logarithm with the increase of the maximum number of privacy levels, indicating that the introduction of a new lower-privacy level contributes less to the total profit of muti-level privacy division strategy. For the sublinear and logarithmic cost functions, the total profit from the single level to the largest 10 level increases from the initial 103 and 133 to 196 and 327, respectively, with an increase of 90% and 146%, indicating that the multi-level division strategy's contribution to total profit is significantly greater than the less-level division strategy. In addition, the sub-linear cost function C_1 has a 62.5% increase in profit relative to the logarithmic cost function C_2. The reason is that risk taker provides more sensitive personal information. For data platform, although the cost of purchasing data increases, the increase of sensitive information further strengthens consumers'willingness to pay, thereby offsetting the amount of profit lost due to increased costs. Due to the high privacy level of the data, the data platform can set a higher price to increase profits.

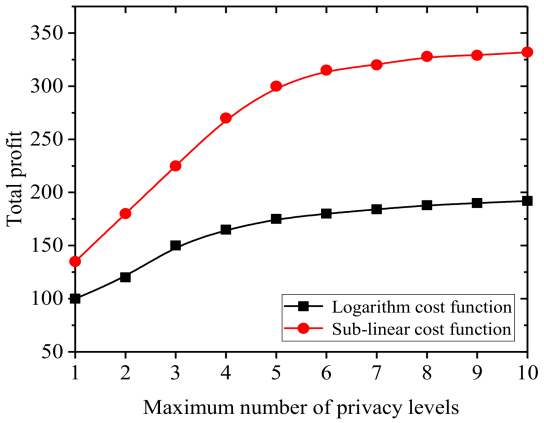

Figure 5. Total profit with the maximal number of privacy levels.

In microeconomics, market coverage is one of the most important indicators for assessing the sales of an enterprise's products. In this article, market coverage is defined as the percentage of customers who purchase private data products to all potential customers, whose formulation is given in (12).

$$Market\ coverage = \frac{The\ number\ of customers\ who\ purchased\ data\ products}{The\ total\ number\ of\ potential\ data\ consumers} \times 100\% \quad (12)$$

When the total potential market is considered, the data platform will realize larger market coverage by introducing more privacy levels. As shown in Figure 6, for both cost functions, market coverage increases in logarithm scale as the number of maximum privacy level increases. The reason is that the multi-level division strategy has expanded the range of consumer choices. The market coverage of the cost function C_1 is greater than the cost function C_2. To a large extent, this is because the more sensitive the information disclosed, the greater its utility, and the wider the consumer.

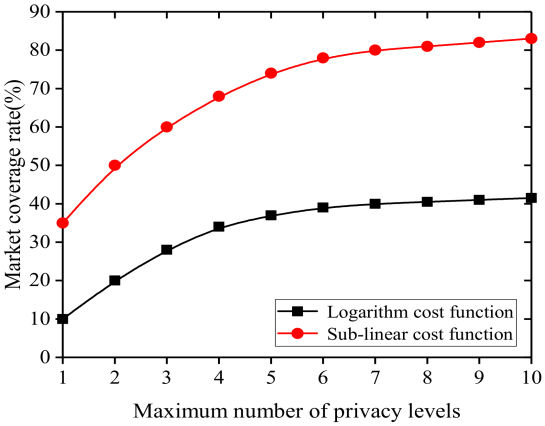

Figure 6. Market coverage with the maximum number of privacy levels.

Setting multiple privacy levels is profitable because data products with more privacy levels make the data market better segmented and the price of the product increases with the number of privacy levels. Figure 7a,b report the ratio of the best price to the privacy level when a different maximum number of privacy levels are assumed (*for simplicity, we only plot the case with a maximum of 6 privacy levels, and so on*). As we have seen, a lower privacy level always has a smaller optimal price, and the ratio monotonically decreases as the privacy level decreases. When the data products offered to the market have a lower level of privacy, the price of the higher privacy level is higher than the previous solution. This means that data owners can get extra profit from each level.

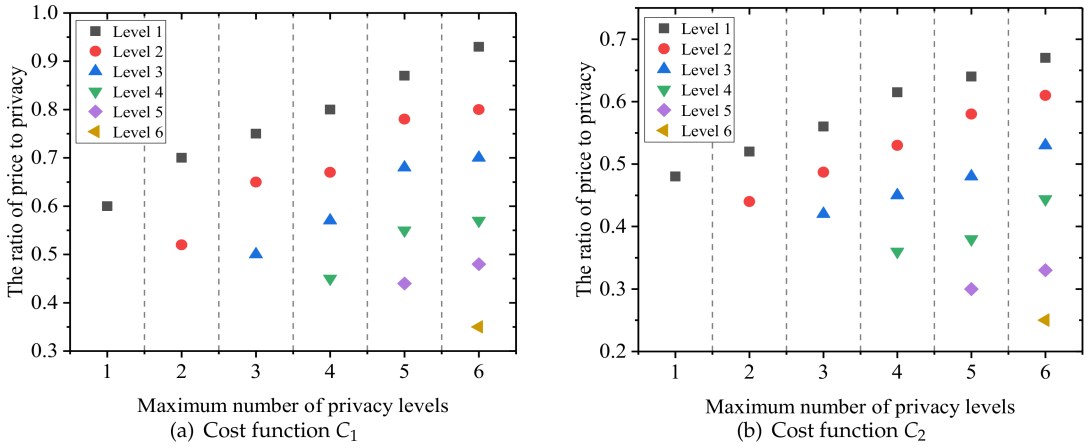

Figure 7. Ratio of price to privacy with different maximal number of levels under two cost function.

Figure 8 reports the percentage of each level's contribution to total profit in the optimal 5-levels division strategy. For both cost functions, the highest level of privacy is the largest percentage of total profit due to its higher price. In the experiment, although we assume that the customer type obeys the Gauss distribution, it also means that the proportion of high-valued customers is less than the average, even so, the highest level of privacy still accounts for the largest percentage of total profits. Our observations are consistent with the literature [19]. Except for the highest level of privacy, the other four levels contribute more than 55% of the total profit. Therefore we can conclude that multi-level privacy division strategy can segment customers well, and make data platforms more profitable in the optimal pricing.

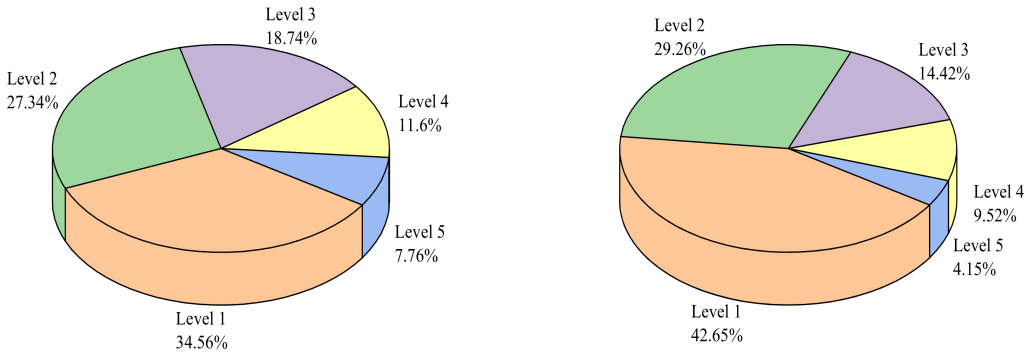

Figure 8. Percentage of all levels in the total profit of optimal 5-levels division strategy.

Figures 9 and 10 show the division of the optimized five privacy levels and their corresponding prices, respectively. In Figure 9, the two cost functions have the highest privacy level of 1, and the lowest privacy level is greater than 0.2. According to the consumer's self-selection behavior, a low privacy level does not bring benefits to the consumers, so the privacy level is set to be greater than a certain threshold. In Figure 10, C_1 's privacy level is more evenly divided than C_2, and the price of each privacy level of C_1 is much higher than C_2. Risk-taker provide more sensitivity information, therefore, increases the cost of the data platform, and the price of data increases.

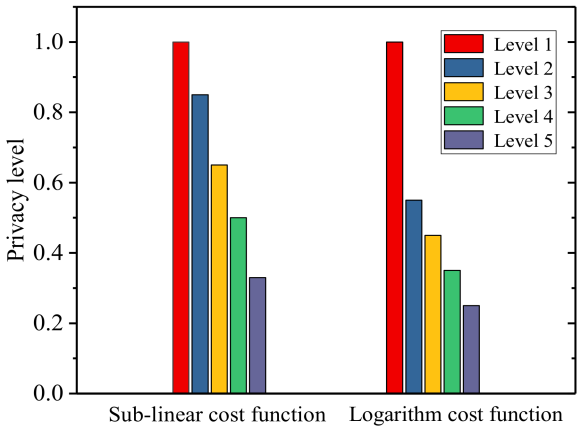

Figure 9. Optimal five-level privacy partitioning scheme with two cost functions.

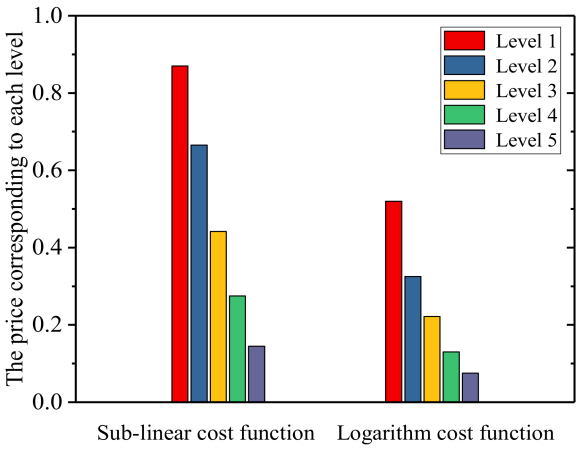

Figure 10. The price corresponding to the best five-level privacy partitioning with two cost functions.

6. Conclusions and Future Work

From the above numerical experiments, we can draw the following conclusions.

The utility function based on consumer self-selection more accurately represents the consumer's willingness to pay for personal data products. Compared with the traditional linear utility function, the nonlinear payment utility function proposed by us can better express consumers' inherent consumer willingness, and this new utility function ensures an optimal solution in privacy data products based on multi-level partitioning strategy. In addition, a multi-level privacy segmentation strategy can maximize the profit of data platform.

Data platform and data consumers as interest decision makers of two levels, using traditional analysis methods is extremely difficult, and the bi-level programming model provides a universal and feasible solution to solve this problem. Which provide a computing platform for multi-level private data division and corresponding pricing. In addition, the genetic simulated annealing algorithm provides a useful idea for solving the bi-level programming model. It not only takes into account the strong global search ability of the genetic algorithm, but also overcomes the shortcomings of slow convergence of GA and SA, thus improving the quality of problem solving.

In practice, the customer's self-selection behavior is very complex and is influenced by many factors, such as customer preferences, interests, and prices. In future, we will consider more variables that affect the customer's purchase decision, and build a more complex customer willingness to pay function. In addition, customer distribution can be diverse, such as Gaussian distribution, uniform distribution, exponential distribution, etc., and these distributions are critical to the privacy level division and pricing of data products. We will consider more customer distribution and discuss how different distributions affect privacy level division strategies and pricing.

Author Contributions: Conceived and designed the experiments, J.Y.; methodology, J.Y. and C.X.; Performed the experiments/Wrote the paper, J.Y.; Supervision and funding acquisition, C.X.

Funding: This research was supported in part by National Nature Science Foundation of China (Grant no. 91646202), National Key R&D Program of China (SQ2018YFB140235).

Acknowledgments: The authors are deeply thankful to the editor and reviewers for their valuable suggestions to improve the quality of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mai, J.E. Big data privacy: The datafication of personal information. *Inf. Soc.* **2016**, *32*, 192–199. [CrossRef]
2. Carey, P. *Data Protection: A Practical Guide to UK and EU Law*; Oxford University Press: Oxford, UK, 2018.
3. Stahl, F.; Schomm, F.; Vossen, G. Data marketplaces: An emerging species. In *Databases and Information Systems VIII*; Haav, H.-M., Kalja, A., Robal, T., Eds.; IOS Press: Amsterdam, The Netherlands, 2014; pp. 145–158.
4. Jorgensen, Z.; Yu, T.; Cormode, G. Conservative or liberal? Personalized differential privacy. In Proceedings of the IEEE 31st International Conference on Data Engineering, Seoul, Korea, 13–17 April 2015; pp. 1023–1034.
5. Datacoup. Available online: http://datacoup.com/ (accessed on 27 February 2019).
6. Gerber, H.U.; Pafum, G. Utility functions: From risk theory to finance. *N. Am. Actuar. J.* **1998**, *2*, 74–91. [CrossRef]
7. Gao, S.; Ma, J.; Sun, C.; Li, X. Balancing trajectory privacy and data utility using a personalized anonymization model. *J. Netw. Comput. Appl.* **2014**, *38*, 125–134. [CrossRef]
8. Homburg, C.; Koschate, N.; Hoyer, W.D. Do satisfied customers really pay more? A study of the relationship between customer satisfaction and willingness to pay. *J. Mark.* **2004**, *69*, 84–96. [CrossRef]
9. Amirtaheri, O.; Zandieh, M.; Dorri, B.; Motameni, A.R. A bi-level programming approach for production-distribution supply chain problem. *Comput. Ind. Eng.* **2017**, *110*, 527–537. [CrossRef]
10. Wang, G.; Ma, L.; Chen, J. A bilevel improved fruit fly optimization algorithm for the nonlinear bilevel programming problem. *Knowl. Based Syst.* **2017**, *138*, 113–123. [CrossRef]
11. Ligett, K.; Neel, S.; Roth, A.; Waggoner, B.; Wu, Z.S. Accuracy first: Selecting a differential privacy level for accuracy-constrained ERM. *arXiv* **2017**, arXiv:1705.10829v1.
12. Li, C.; Li, D.Y.; Miklau, G.; Suciu, D. A theory of pricing private data. *ACM Trans. Database Syst.* **2012**, *39*, 1–28. [CrossRef]
13. Parra-Arnau, J. Optimized, direct sale of privacy in personal data marketplaces. *Inf. Sci.* **2018**, *424*, 354–384. [CrossRef]
14. Malgieri, G.; Custers, B. Pricing privacy—The right to know the value of your personal data. *Comput. Law Secur. Rev.* **2017**, *34*, 289–303. [CrossRef]
15. Nget, R.; Cao, Y.; Yoshikawa, M. How to balance privacy and money through pricing mechanism in personal data market. *arXiv* **2018**, arXiv:1705.02982v2.
16. Aperjis, C.; Huberman, B.A. A market for unbiased private data: Paying individuals according to their privacy attitudes. *arXiv* **2012**, arXiv:1205.0030v1.
17. Donoghue, T.O.; Somerville, J. Modeling risk aversion in economics. *J. Econ. Perspect.* **2018**, *32*, 91–114. [CrossRef]
18. Yu, H.; Zhang, M. Data pricing strategy based on data quality. *Comput. Ind. Eng.* **2017**, *112*, 1–10. [CrossRef]
19. Li, M.; Feng, H.; Chen, F.; Kou, J. Optimal versioning strategy for information products with behavior-based utility function of heterogeneous customers. *Comput. Oper. Res.* **2013**, *40*, 2374–2386. [CrossRef]
20. Sitepu, R.; Puspita, F.M.; Pratiwi, A.N.; Novyasti, I.P. Utility function-based pricing strategies in maximizing the information service provider's revenue with marginal and monitoring costs. *Int. J. Electr. Comput. Eng.* **2017**, *7*, 877–887. [CrossRef]
21. Greene, J.; Baron, J. Intuitions about declining marginal utility. *J. Behav. Decis. Mak.* **2001**, *255*, 243–256. [CrossRef]
22. Bard, J.F. An efficient point algorithm for a linear two-stage optimization problem. *Oper. Res.* **1983**, *31*, 670–684. [CrossRef]
23. Dempe, S.; Zemkoho, A.B. On the Karush-Kuhn-Tucker reformulation of the bilevel optimization problem. *Nonlinear Anal. Theory Methods Appl.* **2012**, *75*, 1202–1218. [CrossRef]

24. Vicente, L.; Savard, G.; Júdice, J. Descent approaches for quadratic bilevel programming. *J. Optim. Theory Appl.* **1994**, *81*, 379–399. [CrossRef]
25. Ben-Ayed, O.; Blair, C.E. Computational difficulties of bilevel linear programming. *Informs* **1990**, *38*, 374–566. [CrossRef]
26. Li, H.; Zhang, L.; Jiao, Y. Solution for integer linear bilevel programming problems using orthogonal genetic algorithm. *J. Syst. Eng. Electron.* **2014**, *25*, 443–451. [CrossRef]
27. Kuo, R.J.; Lee, Y.H.; Zulvia, F.E.; Tien, F.C. Solving bi-level linear programming problem through hybrid of immune genetic algorithm and particle swarm optimization algorithm. *Appl. Math. Comput.* **2015**, *266*, 1013–1026. [CrossRef]
28. Yan, X.; Zhu, Z.; Wu, Q. Hybrid genetic algorithm for engineering design problems. *J. Comput. Theor. Nanosci.* **2016**, *13*, 6312–6319. [CrossRef]
29. Assad, A.; Deep, K. A hybrid harmony search and simulated annealing algorithm for continuous optimization. *Inf. Sci.* **2018**, *450*, 246–266. [CrossRef]