A Reasonable Data Pricing Mechanism for Personal Data Transactions with Privacy Concern

Zheng Zhang, Wei Song(✉), and Yuan Shen

School of Computer Science, Wuhan University, Wuhan, China
{zhangzheng,songwei,shenyuan}@whu.edu.cn

Abstract. In the past few years, more and more data marketplaces for personal data transactions sprung up. However, it is still very challenging to estimate the value of privacy contained in the personal data. Especially when the buyer already has some related datasets, he is able to obtain more privacy by combining and analyzing the bought data and the data he already has. The main research motivation of this work is to reasonably price the data with privacy concern. We propose a reasonable data pricing mechanism which prices the personal privacy data from three aspects and is different from the existing work, we propose a new concept named 'privacy cost' to quantitatively measure the privacy information increment after a data transaction rather than directly measuring the privacy information contained in a single dataset. In addition, we use the information entropy as an important index to measure the information content of data. And we conduct a set of experiments on our personal data pricing method, and the results show that our pricing method performs better than the alternatives.

Keywords: Data pricing · Differential privacy · Data marketplace

1 Introduction

Data commodities and related analysis services are increasingly offered by the online data marketplaces in recent years, which collect personal data with privacy from data owners, process and sell them to data consumers. The privacy contained in data reflects not only the unique value but also the key information of individual like his name, age, gender, even his credit card number, therefore, the access to it should be highly restricted. As for the privacy protection, differential privacy is a standard for data releasing [10]. But we must admit that the introduced noise will perturb the personal data and lead to the inaccuracy.

What is more important, data buyer may have bought some datasets before, which may be related to the dataset he wants to buy this time and are called background datasets. Obviously, the consumer with background datasets could do some operations to obtain more privacy than another data buyer who spends

L. H. U et al. (Eds.): APWeb-WAIM 2021, LNCS 12859, pp. 64–71, 2021.
https://doi.org/10.1007/978-3-030-85899-5_5

the same amount of money but does not have any background dataset. It is unfair and we call this as "privacy increment issue". At present, there is not existing a pricing mechanism that can address this issue.

Based on the problems above, we propose a novel personal data pricing mechanism based on differential privacy, which takes the privacy concern into account. For the first time, we regard the background dataset as an important factor affecting the privacy cost and introduce a new personal data pricing concept named privacy cost to quantitatively measure the privacy increment caused by the union of new and old datasets.

2 Related Work

The general pricing method is subscription, however, this methods can't meet the diverse needs of users. Therefore, Koutris et al. proposed a query-based data pricing framework [4] which allows data buyers to issue different queries for the view. However, the query-based data pricing model does not give guidance on how to price the basic view. Niyato et al. combined the Stackelberg model and the classification algorithm [8]. By using the utility function, the service provider can determine the amount of data to be bought from the data provider, thereby maximizing their own profits. In addition, information entropy, as an important indicator to measure the amount of information contained in the data, has also been introduced into the data pricing model [9]. Li et al. proposed to use information entropy as a new data pricing indicator [6].

As to methods with privacy pricing, Jung et al. [3] introduced a negotiation mechanism, in which data providers and purchasers negotiate on noise scale and unit data price. Nget et al. [7] proposed the concept of data mart based on differential privacy data publishing mechanism. Li et al. [5] proposed a framework for assigning prices to noisy query answers, as a function of their accuracy, and for dividing the price amongst data owners who deserve compensation for their loss of privacy.

However, the above pricing mechanisms are not perfect, especially for the privacy increment issue brought by data union, none of the above mechanisms consider it.

3 Personal Data Pricing Mechanism

3.1 System Model

In this section we describe the basic architecture of proposed pricing mechanism, illustrated in Fig. 1.

The data publisher u_i sends a personal dataset D_i to the trusted data marketplace $\mathbf{M}$. Then $\mathbf{M}$ inserts different scales of noises into raw personal datasets to do differential privacy with different privacy budgets. Finally, the data buyer b_j issues a request $\mathbf{Q}_j(f_j, \epsilon_j)$ which includes an analysis function f_j and a data accuracy ϵ_j he can accept.

Definition 1 (data accuracy). *A privacy mechanism M gives ϵ-differential privacy, where $\epsilon \in (0, 1)$ means privacy budget. Less privacy budget means more noises and implies the personal datasets will be less accurate. Therefore, the privacy budget has the same change tend with data accuracy and is positively correlated to it. So, in some extensis, data accuracy could be represented by privacy budget.*

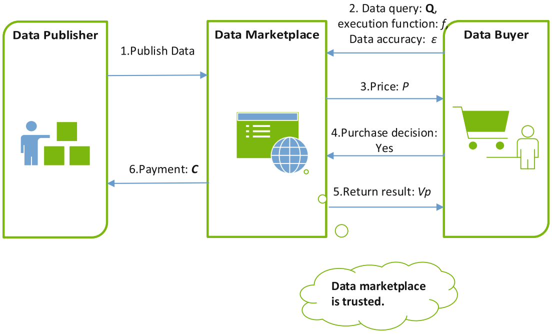


Fig. 1. Trading framework for personal data

One thing needs to be noted is that based on differential privacy [2], the risk of privacy leakage is related to the analysis function. Therefore, our pricing scheme considers not only data accuracy ϵ but also the analysis function f.

After receiving a data request $\mathbf{Q}_j(f_j, \epsilon_j)$, $\mathbf{M}$ will first find the personal dataset V_P with right privacy budget version buyer is interested in. Then the dataset price P is calculated which will be described in details in the next subsection.

3.2 Personal Privacy Data Pricing Function

In this subsection, we will explain our pricing mechanism by detailing every of three prices and the corresponding computing methods for them.

$$P = P_d + P_p + \mathit{profit}. \tag{1}$$

Data Value. P_d is the use value. According to [6], information entropy $H(V)$ is a more reasonable factor to measure information content and data value P_d is positively correlated with $H(V)$.

Also, we must attention one important thing. As we do differential privacy with different data accuracies ϵ, the data marketplace will insert different scales of noises V_s to dataset V, so the data has become not accuracy as it was at first [1]. There must be a accuracy loss δ after inserting noises. We use normalization of root mean square error(RMSE) to describe the accuracy loss δ and give the definition as follows:

Definition 2 (accuracy loss). *For a dataset $D^{m \times n}$, the data in it is x_{ij}, and the D' is obtained by inserting some noises D_s to the D as Eq. (9), the data in D' is x'_{ij}, and the function f is a normalized function, the accuracy loss δ we define as Eq. (7):*

$$\boldsymbol{D'} = \boldsymbol{D} + D_s, \tag{2}$$

$$RMSE = \sqrt{\frac{\sum_{i=0}^{m}\sum_{j=0}^{n}(x'_{ij} - x_{ij})^2}{m \times n}}, \delta = f(RMSE). \tag{3}$$

According to Eq. (7), $\delta \in [0,1]$. In our paper, inserted noise obeys Laplace distribution, namely $D_s \sim Lap(\Delta(f)/\epsilon)$.

We use $H(V)$ to represent the use value of V, and the data value P_d can be obtained by $H(V)$ subtracts the accuracy loss which is brought by inserted noises. And the function $P_d = D(H(V), \delta)$, we design as follows:

$$P_d = 100 \cdot (1 - \delta) \cdot log_2(H(V) + 1). \tag{4}$$

Privacy Cost. P_p indicates the privacy content of personal dataset. We have to pay attention to another thing that different data buyers, who bought the same personal dataset, may obtain different amounts of privacy. Because different data buyers may own different background datasets. When they merge the new dataset they bought and the background dataset, they may get different privacy increments.

Because of the background dataset, different data buyers will obtain different amount of privacy increments, that means the risks of data owners' privacy disclosure are different. Therefore, data buyer who gets more privacy increments $\Delta\theta$, should pay more privacy cost P_p, and we give initial definition of privacy content as follows:

Definition 3. *For any random function f and a dataset $\boldsymbol{D}$ with n tuples $\{t_i | i = 1, ..., n\}$, the privacy contents of t_i and $\boldsymbol{D}$ are defined as:*

$$\theta(t_i) = sup_{S,D}|log\frac{Pr(f(\boldsymbol{D}) \in S}{Pr(f(t_i)) \in S}|, \tag{5}$$

$$\theta(\boldsymbol{D}) = \sum_{i=1}^{n}\theta(t_i), \tag{6}$$

where S is all possible outputs of f.

However it is difficult to compute the privacy content by Definition 3, because the possibility is hard to evaluate. Chao et al. compared the output of a function with and without one data item x_i and imposed a upper bound for privacy loss [5]. The privacy loss they proposed has the same meaning with our privacy content θ, therefore, we transform the formula and introduce it into our paper. We define the function to measure θ as follow:

Definition 4 (privacy content). *For any random function f and a dataset* $\boldsymbol{D}$, *we assume the function f will execute on one attribute X, the privacy content of* $\boldsymbol{D}$ *is defined as:*

$$\theta(\boldsymbol{D}) \leq \frac{\gamma}{\Delta(f)/\epsilon}|D|, \tag{7}$$

where $\gamma = sup_{x \in X}|X|$.

Now let us compute privacy increment $\Delta\theta_j$. Let's suppose that data buyer b_j owns background dataset B_j and wants to buy dataset V_p, and then after this transaction, he will own three datasets: B_j, V_p and U_j which is obtained by doing some operations on B_j and V_p (in our paper, we restrict the operation as union which is a commonly used operation), and also owns three privacy content: $\theta(B_j)$, $\theta(V_p)$ and $\theta(U_j)$. However, b_j have paid for $\theta(B_j)$ when he bought dataset B_j. So the privacy increment $\Delta\theta_j$ he obtains in this transaction is as follows:

$$\Delta\theta_j = \theta(U_j) + \theta(V_p). \tag{8}$$

There is no doubt that P_p is positively related with $\Delta\theta_j$, and the more privacy increment $\Delta\theta_j$ buyer gets, the more he should pay. In our paper, we design the function $P_p = P(\Delta\theta_j)$ as follows:

$$P_p = \frac{\sqrt{50 + 50\Delta\theta_j}}{100}. \tag{9}$$

Profit. The data marketplace should get some remuneration as the middleman between the data publisher and data buyer. In our paper, *profit* represents the income of data marketplace, we just define *profit* as follows:

$$\mathit{profit} = (P_d + P_p) * l, \tag{10}$$

where $l \in (0, 1)$ is a coefficient and is decided by the data marketplace itself. In our paper, we set l as 0.25.

4 Experiments

4.1 Experimental Data and Setup

We use two personal datasets from UCI[1] contain 14 attributes as the data commodities listed on data marketplace. One is the dataset D_1 with 7840 records and the second one is the dataset D_2 with 14720 records, which are both about annual income in the USA.

There are two data buyers. b_1 wants to know the average age of the people in D_1 and b_2 wants to learn the age dispersion in D_2. We assume b_1 has no background datasets and b_2 has a background dataset B with 10000 records. For simplicity, B has the same attributes with D_2 and that means b_2 can easily merge D_2 with B. And we name the transaction on D_1 as experiment 1 and the other one is experiment 2. We compare our pricing mechanism with the baseline method and other alternatives.

[1] https://archive.ics.uci.edu/ml/datasets.php.

Baseline Pricing Mechanism. Just as the analysis in Definition 1, data accuracy ϵ is positively related with personal dataset price P. For simplicity, we consider the relationship between ϵ and P in the baseline pricing mechanism as direct ratio, and the function of it is defined as follows:

$$P = m * \epsilon, \tag{11}$$

where m is a coefficient and in our paper we set m as 1000.

Comparison Pricing Mechanism. We use two pricing mechanisms in our comparison experiments, one is information entropy-based data pricing mechanism [6] and the other is balanced pricing mechanisms [5].

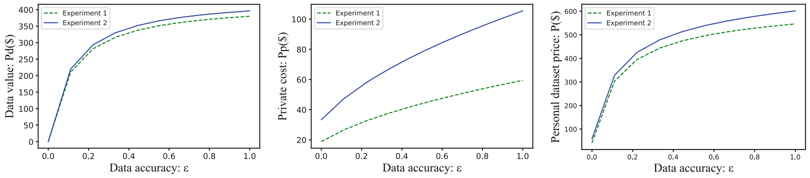
4.2 Experimental Results

Simulation Experiment. We first simulate personal dataset transactions (Fig. 2) when data buyers choose different data accuracy ϵ. Figure 2a shows that data value P_d increases as ϵ increases. And the data value P_d increases dramatically when ϵ is 0∼0.4 but then increases slightly when ϵ is 0.4∼1.0. This pattern is reasonable in practice. We consider that with inserting noises into original personal dataset, the scale of noise may reach a certain threshold, then the availability of dataset will be greatly reduced, and even the dataset is no longer available.

Figure 2b shows the correlation between privacy cost P_p and ϵ. There is no doubt privacy cost P_p increases as the ϵ increases, for that higher ϵ means less privacy protection and data buyer will obtain more privacy. Remarkably, we can see that two curves in Fig. 2b are not exactly the same. When ϵ approaches 0, P_p of two transactions are particularly close, with a difference of less than \$10. When ϵ is close to 1.0, there is a large gap between P_p of the two transactions. We consider that when ϵ is low, even if the data buyer has background datasets, it is still difficult to obtain a large privacy by the background dataset. But when personal dataset is accurate, the data buyer with background datasets can easily to obtain more privacy, so they should pay more.

The last Fig. 2c shows that P increases as ϵ increases. According to Eq. (1), P is the sum of data value P_d, privacy cost P_p and transaction profit. Because $profit$ is constant, so P change trend is the function synthesis of P_d and P_p.

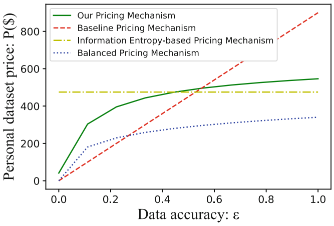
Comparison Experiment. We next compare the result of our personal data pricing mechanism with these of baseline pricing mechanism and other pricing mechanisms described before (Fig. 3 and Fig. 4). We can see no matter how ϵ changes, the P of information entropy-based pricing mechanism remains unchanged. Obviously, from the perspective of data accuracy, it is not reasonablepaper5 for that if two data buyers bought the same personal data with different data accuracies, and they spent the same amount of money. Also, it is not reasonable that P is just linearly related to ϵ just as what baseline pricing



(a) Data value P_d and ϵ (b) Privacy cost P_p and ϵ (c) Query price P and ϵ

Fig. 2. Our pricing mechanism simulation

mechanism shows. When ϵ gets closer and closer to zero, the use value of personal dataset has plummeted, like what our personal data pricing mechanism and balanced pricing mechanism show. That means personal dataset has no meaning for data buyers, when data accuracy is too small, so in our pricing mechanism, it is not recommended data buyers choose too smaller ϵ.



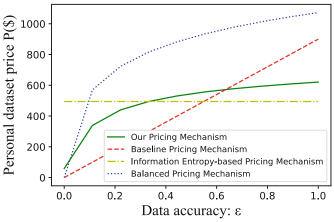


Fig. 3. Query price vs. ϵ on Experiment 1 **Fig. 4.** Query price vs. ϵ on Experiment 2

At last, we do simulations about the above mechanisms described before based on Experiment 2 to show how P changes when data buyers have the same $\mathbf{Q}(f, \epsilon)$ but different scales of background datasets.(Fig. 5). We can see that no matter how the scale of background dataset changes, the P of baseline pricing mechanism and other pricing mechanisms remain unchanged. However, from the perspective of privacy increment, this is not reasonable.

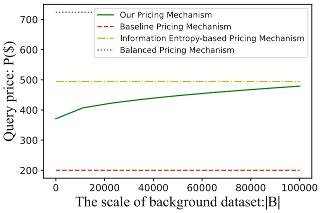


Fig. 5. Trading framework for personal data

5 Conclusion

In this paper, a reasonable data pricing mechanism for the personal data transactions from many aspects is proposed. In the pricing mechanism, we allow data buyers to choose data accuracy, which will meet their different demands. Moreover, to solve the problem of privacy increment brought by background datasets, for the first time, we propose a new concept, privacy cost, and provide the measurement method for it, which is based on differential privacy. Additionally, we consider the influence of inserted noises on the data value, which pricing data value from the perspective of information entropy and accuracy loss. Our data pricing mechanism satisfies the three requirements proposed in Section I and the rationality of it was validated by the simulation and comparison experiments.

Acknowledgement. This work is partially supported by National Key Research and Development Project of China No. 2020YFC1522602, National Natural Science Foundation of China Nos. 62072349, U1811263, and Technological Innovation Major Program of Hubei Province No. 2019AAA072.

References

1. Aperjis, C., Huberman, B.A.: A market for unbiased private data: paying individuals according to their privacy attitudes. First Monday **17**(5) (2012)
2. Dwork, C.: Differential privacy: a survey of results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.) TAMC 2008. LNCS, vol. 4978, pp. 1–19. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79228-4_1
3. Jung, K., Lee, J., Park, K., Park, S.: PRIVATA: differentially private data market framework using negotiation-based pricing mechanism. In: Proceedings of CIKM, pp. 2897–2900 (2019)
4. Koutris, P., Upadhyaya, P., Balazinska, M., Howe, B., Suciu, D.: Query-based data pricing. J. ACM (JACM) **62**(5), 1–44 (2015)
5. Li, C., Li, D.Y., Miklau, G., Suciu, D.: A theory of pricing private data. Commun. ACM **60**(12), 79–86 (2017)
6. Li, X., Yao, J., Liu, X., Guan, H.: A first look at information entropy-based data pricing. In: Proceedings of ICDCS, pp. 2053–2060 (2017)
7. Nget, R., Cao, Y., Yoshikawa, M.: How to balance privacy and money through pricing mechanism in personal data market. In: Proceedings of the SIGIR Workshop on eCommerce (eCOM@SIGIR) (2017)
8. Niyato, D., Alsheikh, M.A., Wang, P., Kim, D.I., Han, Z.: Market model and optimal pricing scheme of big data and Internet of Things (IoT). In: IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2016)
9. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mob. Comput. Commun. Rev. **5**(1), 3–55 (2001)
10. Wang, Q., Zhang, Y., Lu, X., Wang, Z., Qin, Z., Ren, K.: Real-time and spatio-temporal crowd-sourced social network data publishing with differential privacy. IEEE Trans. Dependable Secur. Comput. **15**(4), 591–606 (2016)