Privacy or Utility in Data Collection? A Contract Theoretic Approach

Lei Xu, Chunxiao Jiang, *Member, IEEE*, Yan Chen, *Senior Member, IEEE*, Yong Ren, and K. J. Ray Liu, *Fellow, IEEE*

***Abstract*—With the growing popularity of data mining, privacy has become an issue of growing importance. Privacy can be seen as a special type of goods, in a sense that it can be traded by the owner for incentives. In this paper, we consider a private data collecting scenario where a data collector buys data from multiple data owners and employs anonymization techniques to protect data owners' privacy. Anonymization causes a decline of data utility; therefore, the data owner can only sell his data at a lower price if his privacy is better protected. Can one pursue higher data utility while maintaining acceptable privacy? How to balance the trade-off between privacy protection and data utility is an important question for big data. Considering that different data owners treat privacy differently, and their privacy preferences are unknown to the collector, we propose a contract theoretic approach for data collector to deal with the trade-off. By designing an optimal contract, the collector can make rational decisions on how to pay the data owners, and more importantly, how he should protect the owners' privacy. We show that when the collector requires a large amount of data, he should ask data owners who care privacy less to provide as much as possible data. We also find that whenever the collector requires higher utility of data or the data becomes less profitable, the collector should provide a stronger protection of the owners' privacy. Performance of the proposed contract is evaluated by both numerical simulations and real data experiments.**

***Index Terms*—Privacy preserving, data collecting, data anonymization, contract theory, optimal control.**

I. Introduction

A. Data Mining and Privacy Concerns

In the “big data” era, data mining has attracted much attention from both academia and industry. The key to developing a successful data mining-based application is to prepare a sufficient amount of data, which may contain private information about individuals. If such data is disclosed or used for purposes other than those initially intended, individual's privacy will be compromised. Thus, there is now an increasing concern about the privacy threats posed by data mining.

To deal with the privacy issues, substantial work has been done in the field of privacy-preserving data publishing (PPDP) [1] and privacy-preserving data mining (PPDM) [2]. Viewing privacy issues from a data collector's perspective, PPDP mainly studies how to *anonymize* data in such a way that after the data is published, individual's identity and sensitive information cannot be re-identified [3]–[5]. And PPDM studies how to prevent sensitive data from being directly used in data mining as well as how to exclude sensitive mining results [6], [7].

B. Privacy Auction

Aside from using PPDP and PPDM techniques, the conflict between individual's demand for privacy safety and commercial application's need for accessing personal data can be solved in an economic manner [8]. By seeing privacy as a type of goods, a data collector, who has a need for personal data, can trade with individuals by paying them compensations. However, since different individuals have different privacy preferences, e.g., someone cares about privacy very much while someone cares less, it is difficult for the data collector to decide how to make proper compensations to different individuals.

A feasible approach to deal with the diversity of individual's privacy preference is to set up an auction for privacy [9]. At a privacy auction, each individual reports his valuation on privacy to the data collector. The collector applies some mechanism to decide how many data he should buy from each individual and how much he should pay. Ghosh and Roth [10] initiated the study of privacy auction. Based on their work, a few improved mechanisms have been proposed [11]–[13]. Current privacy auction mechanisms are mainly proposed for the *sensitive surveyor's problem* [9], where a data collector collects individuals' data to obtain an estimate of a simple population statistic. The private data that an individual owns is represented by a single bit $b_i \in \{0,1\}$ indicating whether the individual meets a specified condition, and the individual's privacy cost is quantified by differential privacy [14]. The objective of the data collector is to make an accurate estimation of the sum of bits at a low cost of payments. However, in practice, individual's data is usually represented by a relational record which consists of multiple attributes. Such representation of data is the most basic assumption of anonymization algorithms [1]. Therefore, simply using one bit to represent private data will make the derived auction mechanism less practical. It is necessary to model the problem with more proper formalizations.

Manuscript received September 30, 2014; revised February 14, 2015; accepted April 17, 2015. Date of publication April 23, 2015; date of current version September 14, 2015. This work was supported by the National Natural Science Foundation of China under Grants 61371079, 61271267, 61471025, and 91338203, and a Postdoctoral Science Foundation funded project. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Radha Poovendran.

L. Xu, C. Jiang, and Y. Ren are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: xul04ster@gmail.com; jchx@tsinghua.edu.cn; reny@tsinghua.edu.cn).

Y. Chen and K. J. Ray Liu are with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: yan@umd.edu; kjrliu@umd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2425798

C. Contract Theoretic Approach

In this paper, we study the private data collecting problem in a setting where a data collector needs to collect a set of data records from multiple data owners. Each data owner provides a certain number of data records to the collector and gets paid accordingly. To protect data owners' privacy, the data collector applies anoymization algorithms to the collected data. The anonymized data will then be used in some data mining task. A high level of anonymization means the data owners' privacy can be well protected, thus the owners are willing to provide more data or require less compensation. In that sense, anonymization is beneficial to the collector. However, a high level of anonymization also causes a large decrease in data utility, which means the collector will get less benefit from the data. Therefore, the data collector needs to make a trade-off between data utility and privacy protection level. Besides, since different data owners have different privacy preferences, they will react differently to the collector's decision on privacy protection. Considering that the owners' privacy preferences are unknown to the collector, or in other words, there is *information asymmetry* between the owners and the collector, it is quite difficult for the collector to make a good trade-off.

Information asymmetry is a common phenomenon in economic life. For example, when hiring a new employee in the job market, the employer is unable to know exactly the true ability of the employee. As a result, the employer may hire someone who pretends to be capable of the job. A useful tool to deal with the problems caused by information asymmetry is *contract theory* [15]. In the aforementioned example, the employer can sign a contract with the employee to clearly define what kind of work results he expects from the employee and how he will pay the salary. In this paper, we propose a contract-based approach to handle the trade-off between privacy and utility. Specifically, in the context of private data collection, a contract is signed by both the data owner and the data collector to define how many data that the data owner should provide, how much compensation the owner can receive, and to what extent the owner's privacy should be protected. By designing an optimal contract, the data collector can induce the data owners to act in a way that benefits him most. The idea of applying contract theory has adopted in Yang *et al.*'s work [16], [17], where a contract-based mechanism is proposed for an aggregator to incentivize self-interested electric vehicles to participate in ancillary services to power grid. Despite the fact that their problem has little connection with privacy protection, the contract formulation does provide us some useful inspirations.

To solve the optimization problem embedded in the design of optimal contract, we propose a two-step approach which first determines the optimal transfer function for a given level of privacy protection and then optimizes the collector's payoff with respect to the protection level. Due to the complexity of the resulting payoff function, we are unable to explicitly solve the second optimization problem. Instead, based on numerical simulation results, we qualitatively analyze how those external factors, e.g., the data's value to the collector, influence the design of optimal contract. We show such analysis can provide meaningful insight into the data collector's trade-off problem. In addition, by conducting experiments on real data, we have demonstrated that the proposed contract is more beneficial to both the data collector and data owners, when compared to a simple-formed contract which requires the data utility contributed by a data owner to be proportional to how the owner values his privacy,.

The rest of the paper is organized as follows. In Section II, we introduce the system model and the contract-theoretic formulation. An elaborative description of the design of optimal contract is presented in Section III. In Section IV, we conduct qualitative analysis of the optimal contract, and evaluate the performance of the two types of contracts through simulations. Finally, we draw conclusions in Section V.

II. System Model and Problem Formulation

A. Private Data Collecting

Consider the data collecting scenario shown in Fig. 1. A data collector, on the request of some data miner, collects data from N individuals. Each individual, referred to as the data owner, owns a number of data records. The data owner is free to decide how many and what kind of data he would like to provide to the collector. Once handing over his data, the data owner may suffer a loss in privacy. Different data owners may provide same data to the collector. However, when privacy disclosure happens, owners who treat privacy seriously will perceive more loss than those who have little concern about privacy. We use a parameter $\theta \in [\underline{\theta}, \bar{\theta}](\underline{\theta} \geq 0)$ to describe a data owner's privacy preference. A large θ means the owner cares much about privacy. One thing we do not clarify here is that how the value of the privacy parameter is defined. Quantifying privacy is non-trivial, since complicated sociological and psychological factors may be involved. Here in this paper, following the conventions of contract theory [15], we think each data owner's θ is decided by the *nature*. The privacy parameter can also be interpreted as the unit cost that the data owner pays for producing data. Let q denote the quantity and quality, together referred to as *utility*, of the data provided by the owner. Then the owner with parameter θ will suffer a monetized loss θq if privacy disclosure happens. Correspondingly, the owner receives a transfer, denoted by t, from the collector as a compensation.

Once the collector has collected enough data, he applies some anonymization technique to the data. After being anonymized, the data becomes more secure, in a sense that the possibility that a data owner is re-identified by an attacker decreases. While in the meantime, the utility of data declines. We use $d(q, \delta)$ to denote the utility of anonymized data, where $\delta \in [0, 1]$ denotes the level of privacy protection that is realized by anonymization. Intuitively, a large δ causes a large decrease in data utility. To embody this intuition, we define $d(q; \delta)$ as

$$d(q, \delta) = [\alpha_1(1-\delta)^{\alpha_2} + \alpha_3]q, \tag{1}$$

where α_1,α_2 and α_3 are positive constants. This formulation is actually obtained from anonymization experiments on real data (see Section IV-B1 for more details). According to the experiment results, there is $0 < \alpha_1 < 1$, $0 < \alpha_2 < 0.5$ and $0 < \alpha_3 < 1$. Here we define $\alpha_3 = 1 - \alpha_1$ to capture the intuition that if no privacy protection measure is taken, i.e., $\delta = 0$, there should be no utility loss.

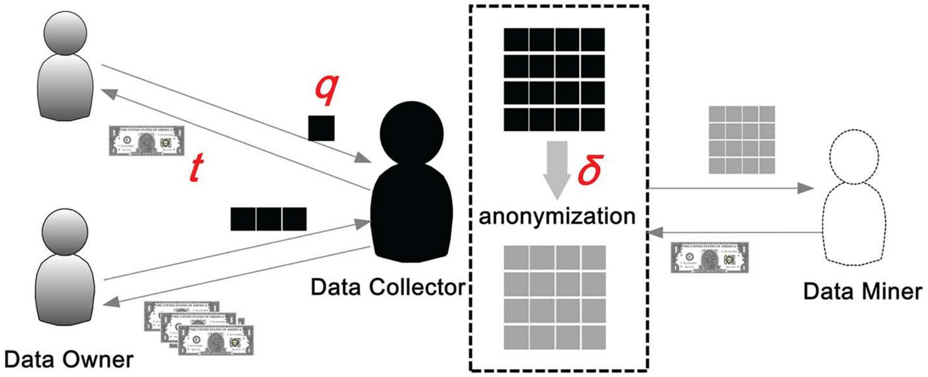


Fig. 1. The data collecting scenario.

After finishing the anonymization process, the collector releases the data to a data miner and gets paid, or conducts some analysis by himself. Either way, the collector obtains an income from the data. Let $S(q)$ denote the income, and we assume $S(0) = 0$, $\frac{dS}{dq} > 0$, and $\frac{d^2S}{dq^2} < 0$, which means the marginal value of data decreases as the collector has obtained more data. Furthermore, to ease the analysis and without loss of generality, we define $S(q)$ as

$$S(q) = \lambda\sqrt{q}, \tag{2}$$

where the positive constant λ indicates how valuable the data is to the collector. The parameter λ is an exogenous parameter, in a sense that its value is determined by some factors that are out of the control of the data collector and data owners. For example, conditions of the data market will have strong influence on λ. Suppose that a data collector makes profit by selling the collected data to a data miner. If the data miner can buy data from other collectors, then the collector may have to sell the data at a lower price, which implies the data become less valuable to the collector.

Based on above discussions, the payoff to a data owner with parameter θ can be defined as

$$u_\theta = t - (1-\delta)\theta q, \tag{3}$$

where $(1-\delta)\theta q$ represents the expected value of privacy loss. The payoff that the data collector obtains from the trade with one data owner is

$$u_C = S(d(q,\delta)) - t. \tag{4}$$

To maximize the payoff, the data collector needs to carefully decide the transfer paid to the owner and the privacy protection level he should guarantee. However, when trading with a data owner, the collector does not know for sure how the owner values his privacy, since the owner's privacy parameter is only known to himself. In other words, from the perspective of the collector, the privacy parameter θ is a random variable. Here for simplicity and without loss of generality, we make the following assumption.

Assumption 1: The data owner's privacy parameter θ is unknown to the data collector. Each owner's θ is drawn independently and identically from $[\underline{\theta}, \overline{\theta}]$, and the corresponding probability density function $f(\theta)$ is known to the collector.

TABLE I
NOTATIONS

N	The number of data owners.
θ	The privacy preference of a data owner.
q	The utility of the data provided by a data owner.
t	The transfer paid to a data owner.
δ	The level of privacy protection realized by the data collector.
β	The probability of privacy disclosure. $\beta = 1 - \delta$.
λ	A parameter indicating the data's value to the data collector.
q_{req}	The data collector's requirement on the total utility of data.
q_{max}	The maximal data utility that a data owner can provide.
ρ	The ratio of the data utility provided by a data owner to the maximal data utility that a data owner can provide. $\rho = {}^{q}/_{q_{\max}}$.
$U(\theta)$	The payoff to a data owner with parameter θ.
U_C	The payoff to the data collector.

Realizing that both the data owners and the collector want to get maximal payoff, and there is an information asymmetry between the two parities, we resort to principle-agent theory [15] to solve the collector's problem. More specifically, we study how to design a contract for the collector, so that the collector can induce data owners to act in a way that can bring him the maximal payoff. Next we will present the formulation of the contract design problem. For convenience, we summarize some important notations used in the formulation in Table I.

B. Contract-Theoretic Formulation

Following the contract theory terminology, above data collecting scenario can be described as follows. A data collector, who plays the role of the *principal*, delegates a data producing task to multiple *agents*, namely the data owners. Each owner's *type* θ is unobservable to the collector. The collector offers a menu of contracts $\{(\delta, t, q)\}$ to each owner. If the owner chooses to accept the contract (δ, t, q), then he will provide the collector with data of utility q, and in return, the collector should pay transfer t to the owner and make sure that the probability of privacy disclosure is no higher than $1 - \delta$. We assume that the data utility that one data owner can contribute is no more than $q_{\max}$. To make the contract more interpretable, hereafter we use $\rho \triangleq {}^{q}/_{q_{\max}}$ as a replacement of the contract item q.

According to the revelation principle [15], it is sufficient for the collector to consider only the direct revelation mechanism $\{(\delta(\theta), t(\theta), \rho(\theta))\}$, where the contract $(\delta(\theta), t(\theta), \rho(\theta))$ is designated for data owner with type θ. Considering that most anonymization algorithms do not support personalized privacy protection [1], that is, they exert the same amount of privacy preservation for all individuals, we define $\delta(\theta) = 1 - \beta$ for all $\theta \in [\underline{\theta}, \bar{\theta}]$ with $\beta \in (0,1]$ denoting the probability of privacy disclosure. Upon choosing the contract $(1-\beta, t(\theta), \rho(\theta))$, the payoff to a data owner with type θ can be written as

$$U(\theta) = t(\theta) - \beta\theta\rho(\theta)q_{\max}, \tag{5}$$

In the study of contract theory, the agent's payoff is usually referred to as *information rent*, which emphasizes that it is because of the information asymmetry that the agent can get extra benefit.

To ensure that the data owner will accept the contract designated for him rather than choosing other contracts or refusing any contract, the menu of contracts must be *incentive feasible*. That is, it should satisfy both the *incentive compatibility* constraints and the *participation* constraints defined below.

Definition 1: A menu of contracts $\{(1-\beta, t(\theta), \rho(\theta))\}$ is incentive compatible if the best response for the data owner with type θ is to choose the contract $(1-\beta, t(\theta), \rho(\theta))$ rather than other contracts, i.e., $\forall(\theta, \tilde{\theta}) \in [\underline{\theta}, \bar{\theta}]^2$,

$$t(\theta) - \beta\theta\rho(\theta)q_{\max} \geq t(\tilde{\theta}) - \beta\theta\rho(\tilde{\theta})q_{\max}. \tag{6}$$

Definition 2: A menu of contracts $\{(1-\beta, t(\theta), \rho(\theta))\}$ satisfies the participation constraints if it yields to each type of data owner a non-negative payoff, i.e., $\forall\theta \in [\underline{\theta}, \bar{\theta}]$,

$$t(\theta) - \beta\theta\rho(\theta)q_{\max} \geq 0. \tag{7}$$

In addition, to make sure that meaningful results can be obtained in subsequent data mining tasks, the data collector usually has a minimum requirement on the total utility of the collected data. Here we assume that a feasible menu of contracts should satisfy the following *isoperimetric* constraint:

$$N\int_{\underline{\theta}}^{\bar{\theta}} q_{\max}\rho(\theta)f(\theta)d\theta = q_{req}, \tag{8}$$

where q_{req} denotes the data collector's requirement. Apparently, the requirement is attainable only if it is no higher than $Nq_{\max}$. Above equation also implies that the total utility of the collected data is assumed to be the summation of the utility of each owner's data. It should be noted that in practice, the relationship between the total utility of data and the utility of each data record is usually application-dependent. Here we define the total utility as a summation, so that we can ease the analysis and meanwhile reflect the general understanding of "total".

Another implicit constraint on the contracts is that the data utility contributed by one data owner is bounded, i.e.,

$$0 \leq \rho(\theta) \leq 1. \tag{9}$$

The data collector offers contracts to data owners before knowing the owners' types, hence the payoff that a menu of contracts brings to the collector is evaluated in expected terms. The collector's objective is to find an optimal menu of contracts which satisfies all the constraints listed above and maximizes the expected payoff. The collector's problem can be formulated as

$$(\mathbf{P}) \qquad \max_{\{(1-\beta, t(\cdot), \rho(\cdot))\}} N\int_{\underline{\theta}}^{\bar{\theta}} U_C(\theta;\beta)f(\theta)d\theta, \quad \text{subject to (6)} \sim \text{(9)}.$$

The function $U_C(\theta;\beta)$ in the integrand is defined as

$$U_C(\theta;\beta) = S(d(q_{\max}\rho(\theta), 1-\beta)) - t(\theta). \tag{10}$$

Next we will discuss how to solve this optimization problem.

III. Contract Designs

A. Method Overview

As defined in the previous section, the contract offered by the collector is formed as a tuple $(1-\beta, t(\theta), \rho(\theta))$, where the first item is independent of the owner's type. To find the optimal menu of contracts $\{(1-\beta^*, t^*(\theta), \rho^*(\theta))\}$, we propose a two-step approach. First, we find the optimal *transfer function* $t^*_\beta(\cdot)$ and *production function* $\rho^*_\beta(\cdot)$ for a given privacy protection level. Specifically, given $\beta \in [0,1]$, we solve the following problem

$$(\mathbf{P1}) \qquad \max_{\{(t(\cdot), \rho(\cdot))\}} \int_{\underline{\theta}}^{\bar{\theta}} U_C(\theta;\beta)f(\theta)d\theta, \quad \text{subject to (6)} \sim \text{(9)}.$$

Both $t^*_\beta(\cdot)$ and $\rho^*_\beta(\cdot)$ can be seen as parametric functions with β being the parameter. By plugging these two functions into the objective function of problem $\mathbf{P}$, we can rewrite the data collector's payoff as a function of β, denoted by $U_C(\beta)$. Thus the second step of optimal contract design is to solve the following optimization problem

$$(\mathbf{P2}) \qquad \max_{\beta \in (0,1]} \int_{\underline{\theta}}^{\bar{\theta}} U^*_C(\theta;\beta)f(\theta)\,d\theta.$$

The function $U^*_C(\theta;\beta)$ in the integrand is defined as

$$U^*_C(\theta;\beta) = S\left(d\left(q_{\max}\rho^*_\beta(\theta), 1-\beta\right)\right) - t^*_\beta(\theta). \tag{11}$$

Let β^* denote the optimal solution to above problem, then the optimal menu of contracts is given by $\{(1-\beta^*, t^*_{\beta^*}(\theta), \rho^*_{\beta^*}(\theta))\}$.

B. Simplifying Constraints

Solving problem $\mathbf{P1}$ is non-trivial, since it involves optimizing a functional with respect to a pair of functions, also the constraints are complicated. Before we explore solutions to the functional optimization problem, we first need to find a concise way to express the incentive constraints and participation constraints.

Though described with one simple inequality, (6) actually implies an infinity of constraints, each of which corresponds to a certain pair of θ and $\tilde{\theta}$. Similarly, (7) should be treated as an infinity of participation constraints, each of which corresponds to a certain $theta$. To identify the set of feasible solutions to problem $\mathbf{P1}$, first we need to simplify theses constrains as much as possible.

Following a similar approach proposed in [15], we reduce the infinity of incentive constraints in (6) to a differential equation

$$\frac{dt(\theta)}{d\theta} - \beta q_{\max}\theta\frac{d\rho(\theta)}{d\theta} = 0 \tag{12}$$

and a monotonicity constraint

$$-\frac{d\rho(\theta)}{d\theta} \geq 0. \tag{13}$$

Details of the simplification process are presented in the Appendix. Further, by using (5) we can express (12) in a simpler way:

$$\dot{U}(\theta) = -\beta q_{\max}\rho(\theta). \tag{14}$$

Due to the simplicity of above expression, hereafter we focus on the design of $U(\,\cdot\,)$ instead of $t(\,\cdot\,)$, after all the optimal $t(\,\cdot\,)$ can be easily determined once the optimal $U(\,\cdot\,)$ and $\rho(\,\cdot\,)$ are found.

Base on (9) and (14), participation constraints in (7) can be simplified to $U(\bar{\theta}) \geq 0$. Further, we can predict that this constraint must be binding at the optimum, i.e.,

$$U_\beta^*(\bar{\theta}) = 0. \tag{15}$$

Suppose that $U_\beta^*(\bar{\theta}) > 0$, then the collector could reduce $U_\beta^*(\bar{\theta})$ by a small amount while keeping $\rho_\beta^*(\,\cdot\,)$ unchanged. As a result, the collector's payoff is increased, which contradicts with the optimality of $U_\beta^*(\bar{\cdot})$.

Based on above simplifications, problem $\mathbf{P1}$ can be rewritten as

$$(\mathbf{P1}') \qquad \max_{\{(U(\,\cdot\,),\rho(\,\cdot\,))\}} \int_{\underline{\theta}}^{\bar{\theta}} U_C(\theta;\beta)f(\theta)d\theta.$$
$$\text{subject to } (13), (14), (15), (8) \text{ and } (9).$$

The function $U_C(\theta;\beta)$ in the integrand is now written as

$$\begin{aligned} U_C(\theta;\beta) =& S(d(q_{\max}\rho(\theta), 1-\beta)) \\ & - U(\theta) - \beta q_{\max}\theta\rho(\theta). \end{aligned} \tag{16}$$

C. Optimal Control-Based Approach

Problem $\mathbf{P1}'$ fits the general formulation of the optimal control problem [18], hence methods developed for optimal control can be applied. Let $y(\theta) \triangleq \rho(\theta)$ be the control variable and $x_1(\theta) \triangleq U(\theta)$ be the state variable. To handle the isoperimetric constraint (8), a new state variable $x_2(\theta)$ is defined, and it satisfies the following differential equation:

$$\dot{x}_2(\theta) = q_{\max}y(\theta)f(\theta), \tag{17}$$

The boundary conditions of $x_2(\theta)$ are $x_2(\bar{\theta}) = {}^{q_{req}}/_{Nq_{\max}}$. The *Hamiltonian* is

$$\begin{aligned} & H(\mathbf{x}(\theta), y(\theta), \mathbf{p}(\theta), \theta) \\ & = [S(d(q_{\max}y(\theta); 1-\beta)) - x_1(\theta) - \beta q_{\max}\theta y(\theta)]f(\theta) \\ & \quad - \beta q_{\max}p_1(\theta)y(\theta) + p_2(\theta)q_{\max}f(\theta)y(\theta), \end{aligned} \tag{18}$$

where $p_1(\theta)$ and $p_1(\theta)$ are co-state variables. To simplify notations, we define $\mathbf{x}(\theta) = (x_1(\theta), x_2(\theta))^T$ and $\mathbf{p}(\theta) = (p_1(\theta), p_2(\theta))^T$.

According to *Pontryagin minimum principle* [18], the optimal solution $(\mathbf{x}^*(\theta), y^*(\theta))$ to problem $\mathbf{P1}'$ should satisfy the following six conditions:

$$\begin{aligned} \dot{x}_1^*(\theta) &= \frac{\partial H(\mathbf{x}^*(\theta), y^*(\theta), \mathbf{p}^*(\theta), \theta)}{\partial p_1(\theta)} \\ &= -\beta q_{\max}y^*(\theta), \end{aligned} \tag{19}$$

$$\begin{aligned} \dot{x}_2^*(\theta) &= \frac{\partial H(\mathbf{x}^*(\theta), y^*(\theta), \mathbf{p}^*(\theta), \theta)}{\partial p_2(\theta)} \\ &= d(q_{\max}y^*(\theta), 1-\beta)f(\theta), \end{aligned} \tag{20}$$

$$\dot{p}_1^*(\theta) = -\frac{\partial H(\mathbf{x}^*(\theta), y^*(\theta), \mathbf{p}^*(\theta), \theta)}{\partial x_1(\theta)} = f(\theta), \tag{21}$$

$$\dot{p}_2^*(\theta) = -\frac{\partial H(\mathbf{x}^*(\theta), y^*(\theta), \mathbf{p}^*(\theta), \theta)}{\partial x_2(\theta)} = 0, \tag{22}$$

$$H(\mathbf{x}^*(\theta), y^*(\theta), \mathbf{p}^*(\theta), \theta) \geq H(\mathbf{x}^*(\theta), y(\theta), \mathbf{p}^*(\theta), \theta), \tag{23}$$

$$p_1^*(\underline{\theta}) = 0. \tag{24}$$

From (21) and (24) we can get

$$p_1^*(\theta) = F(\theta). \tag{25}$$

From (22) we know that for any $\theta \in [\underline{\theta}, \bar{\theta}]$, there is

$$p_2^*(\theta) = \gamma, \tag{26}$$

where the γ will later be determined by using the boundary condition $x_2(\bar{\theta}) = {}^{q_{req}}/_{Nq_{\max}}$.

Having determined $p_1^*(\theta)$ and $p_2^*(\theta)$, now we need to optimize the Hamiltonian with respect to $y(\theta)$. In order to derive the analytic expression of $y^*(\theta)$, we assume that data owner's type is uniformly distributed within $[\underline{\theta}, \bar{\theta}]$. The probability density function is

$$f(\theta) = \frac{1}{\bar{\theta} - \underline{\theta}}, \forall\theta \in [\underline{\theta},\ \bar{\theta}], \tag{27}$$

and the cumulative density function is

$$F(\theta) = \frac{\theta - \underline{\theta}}{\bar{\theta} - \underline{\theta}}, \forall\theta \in [\underline{\theta},\ \bar{\theta}]. \tag{28}$$

Given above assumption, the optimal production function $\rho_\beta^*(\,\cdot\,)$ can be derived via following two steps. First, we ignore the boundary constraint (9) and solve the unbounded $\tilde{y}_\beta(\,\cdot\,)$ that maximizes the Hamiltonian. By using the first order condition $\left.\frac{\partial H(\mathbf{x}^*(\theta), y(\theta), \mathbf{p}^*(\theta), \theta)}{\partial y}\right|_{y=\tilde{y}} = 0$ we get

$$\tilde{y}_\beta(\theta) = \frac{(\alpha_1\beta^{\alpha_2} + \alpha_3)\,\lambda^2}{4[(2\theta - \underline{\theta})\,\beta - \gamma]^2 q_{\max}}, \forall\theta \in [\underline{\theta}, \bar{\theta}]. \tag{29}$$

It can be easily verified that

$$\left.\frac{\partial^2 H\,(\mathbf{x}^*(\theta), y(\theta), \mathbf{p}^*(\theta), \theta)}{\partial y^2}\right|_{y=\tilde{y}} < 0, \tag{30}$$

hence $\tilde{y}_\beta(\,\cdot\,)$ does maximize the Hamiltonian. The β-specific constant γ in (29) can be determined by using the monotonicity

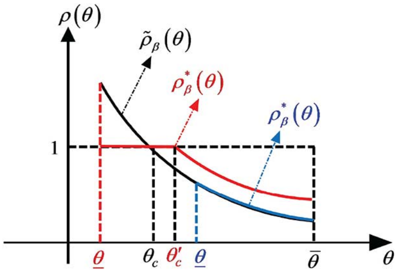


Fig. 2. Optimal production functions under different settings of q_{req}.

constraint (13) and the isoperimetric constraint (8). Take the derivative of $\tilde{y}_\beta(\theta)$ with respect to θ and use (13) we get $\gamma < \beta\underline{\theta}$. Then, plug (29) into the right-hand side of (8) and solve the equation for γ, we get

$$\gamma = \beta\bar{\theta} - \frac{1}{2}\sqrt{4\left(\bar{\theta}-\underline{\theta}\right)^2\beta^2 + \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)N\lambda^2}{q_{req}}}. \quad (31)$$

With $\tilde{y}_\beta(\,\cdot\,)$ determined, the next step to find $\rho^*_\beta(\,\cdot\,)$ is to check whether the boundary constraint (9) can be satisfied.

Let us first consider a special case, that is, the data collector can offer a perfect protection of privacy, namely $\beta = 0$. In such a case, $\tilde{y}_\beta(\,\cdot\,)$ becomes a constant function, i.e.,

$$y^*_0(\theta) = \frac{q_{req}}{Nq_{\max}},\ \forall\theta \in [\underline{\theta},\bar{\theta}]. \quad (32)$$

Then, according to (14) and (15), each data owner will receive zero information rent. The intuition behind this result is that when no privacy disclosure will happen, there is no privacy loss to data owners. Thus the data collector is indifferent to how each owner values his privacy, and the task of data producing is equally assigned to different owners. As for the data owner, since his type θ does not matter to the collector, namely his information advantage over the collector no longer exists, he will receive no information rent. The total payoff to the data collector is

$$U^*_C(0) = \lambda\sqrt{\alpha_3 Nq_{req}}. \quad (33)$$

Given $\beta = 0$, the optimal contract $(1-\beta^*, t^*_\beta(\,\cdot\,), \rho^*_\beta(\,\cdot\,))$ has a very simple form, which is $\left(0, 0, \frac{q_{req}}{Nq_{\max}}\right)$. However, in practice, perfect privacy protection can hardly be realized, thus such a contract is impractical. It is more important to explore the cases when privacy disclosure is inevitable.

Given $\beta \in (0,1]$, $\tilde{y}_\beta(\,\cdot\,)$ can be depicted by the curve segment shown in Fig. 2. As we can see, if the curve segment intersects with the line $y(\theta) = 1$ at some point, then $\tilde{y}_\beta(\,\cdot\,)$ cannot be taken as a feasible production function. Let $(\theta_c, 1)$ denote the intersection point (possibility exists), where θ_c is defined as

$$\theta_c = \frac{1}{2}(\bar{\theta}+\underline{\theta}) - \frac{1}{4\beta}\sqrt{4\left(\bar{\theta}-\underline{\theta}\right)^2\beta^2 + \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)N\lambda^2}{q_{req}}} + \frac{\lambda}{4\beta}\sqrt{\frac{\alpha_1\beta^{\alpha_2}+\alpha_3}{q_{\max}}}. \quad (34)$$

If θ_c lies outside the interval $[\underline{\theta},\bar{\theta}]$, then $\tilde{y}_\beta(\,\cdot\,)$ is the optimal production function we are looking for. Through a simple analysis of (34) we learn that, as long as $q_{req} \le Nq_{\max}$, there is $\theta_c < \frac{1}{2}\left(\bar{\theta}+\underline{\theta}\right) < \bar{\theta}$. However, it is uncertain whether there is $\theta_c < \underline{\theta}$.

As defined in (34), given β and other exogenously specified parameters $\{\lambda, N, q_{\max}, \bar{\theta}, \underline{\theta}, \alpha_1, \alpha_2, \alpha_3\}$, the value of θ_c is fully determined by the collector's requirement q_{req}. With the increase of q_{req}, θ_c increases. At some point when q_{req} is higher than a threshold q_{maxreq}, θ_c will exceed $\underline{\theta}$. The threshold q_{maxreq} can determined by setting $\theta_c = \underline{\theta}$, and clearly it depends on β as follows

$$q_{\text{maxreq}}(\beta) = \frac{\lambda Nq_{\max}\sqrt{\alpha_1\beta^{\alpha_2}+\alpha_3}}{\lambda\sqrt{\alpha_1\beta^{\alpha_2}+\alpha_3} + 4\left(\bar{\theta}-\underline{\theta}\right)\beta q_{\max}}. \quad (35)$$

Considering that $q_{req} \in (0, Nq_{\max}]$ is specified before the contract is formed, the following three situations need to be analyzed respectively.

1) If $0 < q_{req} \le q_{\text{maxreq}}(\beta)$, then for all $\theta \in [\underline{\theta},\bar{\theta}]$, $\tilde{y}_\beta(\theta)$ lies within the boundaries. In such a case, the optimal production function $\rho^*_\beta(\,\cdot\,)$ has exactly the formulation with $\tilde{y}_\beta(\,\cdot\,)$ as defined in (29). Then, by using (14) and (15) we can determine the optimal information rent function, that is

$$U^*_\beta(\theta) = \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)\lambda^2}{8\left[(2\theta-\underline{\theta})\beta-\gamma\right]} - \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)\lambda^2}{8\left[\left(2\bar{\theta}-\underline{\theta}\right)\beta-\gamma\right]}. \quad (36)$$

2) If $q_{\text{maxreq}}(\beta) < q_{req} < Nq_{\max}$, then for $\theta \in [\underline{\theta},\theta_c]$, $\tilde{y}_\beta(\theta)$ lies outside the boundary. In such a case, we define $\rho^*_\beta(\,\cdot\,)$ as a piecewise function, i.e.,

$$\rho^*_\beta(\theta) = \begin{cases} 1, & \underline{\theta} \le \theta \le \theta'_c \\ \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)\lambda^2}{4\left[(2\theta-\underline{\theta})\beta-\gamma'\right]^2 q_{\max}}, & \theta'_c < \theta \le \bar{\theta} \end{cases} \quad (37)$$

where γ' is determined by using (8) and (13). Specifically, γ' is given by

$$\gamma' = \beta\bar{\theta} - \frac{\lambda}{2}\sqrt{\frac{\alpha_1\beta^{\alpha_2}+\alpha_3}{q_{\max}}} + \frac{\left(\bar{\theta}-\underline{\theta}\right)q_{req}\beta}{Nq_{max}} - \frac{\sqrt{\Delta'}}{Nq_{\max}}, \quad (38)$$

where Δ' is defined as

$$\Delta' = \left(\bar{\theta}-\underline{\theta}\right)^2\left(Nq_{\max}-q_{req}\right)^2\beta^2 + N\lambda\beta\sqrt{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)q_{\max}}\left(\bar{\theta}-\underline{\theta}\right)\left(Nq_{\max}-q_{req}\right). \quad (39)$$

Based on the formulation of γ', θ'_c can be determined by

$$\theta'_c = \frac{1}{2}\underline{\theta} + \frac{\gamma'}{2\beta} + \frac{\lambda}{4\beta}\sqrt{\frac{\alpha_1\beta^{\alpha_2}+\alpha_3}{q_{\max}}}. \quad (40)$$

A geometric interpretation of (37) is given below. As shown in Fig. 2, the area under the black curve segment is proportional to the collector's requirement q_{req}. When θ_c lies on the right side of $\underline{\theta}$, the curve segment can be divided into two subsegments, namely the one lies on the left side of the point $(\theta_c, 1)$ and the one lies on the right side. For points on the left-hand segment, we has to "pull" them down until they reach the boundary. By doing so, the area between the original segment and the boundary is discarded. In order to keep the total area unchanged, points on the right-hand segment must be "pushed up", and those who lie close to $(\theta_c, 1)$ may be pushed up to the boundary. For a given β, as q_{req} decreases, the whole curve segment moves towards the left, which means fewer points need to be pulled down. By the time q_{req} decreases to $q_{\mathrm{maxreq}}(\beta)$, the curve intersects with the boundary at $(\underline{\theta}, 1)$. In such a case, no point needs to be pulled down, and this is when the piecewise $\rho^*_\beta(\,\cdot\,)$ degenerates to that defined in (29). On the contrary, as q_{req} increases, the whole curve segment moves towards the right, and more points lie above the boundary. Consequently, the pulling-down operation causes a larger loss in area, which means points on the right-hand segment should be pushed higher. In an extreme case, all the points on the left-hand segment are pushed to the boundary. This is exactly the third case that we will discuss later.

With $\rho^*_\beta(\,\cdot\,)$ defined in (37), we can derive the optimal information rent function $U^*_\beta(\,\cdot\,)$ by using (14) and (15), that is

$$U^*_\beta(\theta) = \begin{cases} -\beta q_{\max}\theta + \Gamma_\beta,\ \theta \in [\underline{\theta}, {\theta'}_c] \\ \frac{(\alpha_1\beta^{\alpha_2}+\alpha_3)\left(\bar{\theta}-\theta\right)\lambda^2\beta}{4[(2\theta-\underline{\theta})\beta-2\gamma'][\left(2\bar{\theta}-\underline{\theta}\right)\beta-2\gamma']},\ \theta \in \left({\theta'}_c, \bar{\theta}\right] \end{cases} \tag{41}$$

where Γ_β is defined as

$$\Gamma_\beta = \beta q_{\max}{\theta'}_c + \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)\left(\bar{\theta}-{\theta'}_c\right)\lambda^2\beta}{4\left[\left(2{\theta'}_c-\underline{\theta}\right)\beta-2\gamma'\right]\left[\left(2\bar{\theta}-\underline{\theta}\right)\beta-2\gamma'\right]}. \tag{42}$$

3) If $q_{req} = Nq_{\max}$, similar to above case, θ_c lies to the right side of $\underline{\theta}$, hence the optimal production function $\rho^*_\beta(\,\cdot\,)$ has the same form as that defined in (32). But in this case, there is ${\theta'}_c = \bar{\theta}$, and (37) becomes a constant function, i.e.,

$$\rho^*_\beta(\theta) = 1,\ \forall\theta \in [\underline{\theta}, \bar{\theta}]. \tag{43}$$

Again, by using (14) and (15) we get the optimal information rent function, which is defined as

$$U^*_\beta(\theta) = (\bar{\theta}-\theta)\beta q_{\max}. \tag{44}$$

Then the optimal transfer function $t^*_\beta(\,\cdot\,)$ can be written as

$$t^*_\beta(\theta) = U^*_\beta(\theta) + \theta\beta q_{\max}\rho^*_\beta(\theta) = \beta q_{\max}\bar{\theta}. \tag{45}$$

This result coincides with the intuition that when different data owners provide the same amount of data, they will be paid equally.

Above we have discussed how to design the optimal production function $\rho^*_\beta(\,\cdot\,)$ and optimal information rent function $U^*_\beta(\,\cdot\,)$ for a given privacy protection level. As we have clarified, different forms of these two functions should be adopted in accordance with different values of q_{req}. It should be noted that as q_{req} approaches $q_{\mathrm{maxreq}}(\beta)$ (or $Nq_{\max}$), the piecewise production function defined in (37) will degenerate to a smooth form.

D. Determining the Optimal Privacy Protection Level

The production function $\rho^*_\beta(\,\cdot\,)$ and the information rent function $U^*_\beta(\,\cdot\,)$ derived in the above subsection are optimal for a given privacy protection level. In other words, both the functions are parameterized by β. With these optimal functions, the data collector can determine the optimal privacy protection level by solving the ordinary optimization problem $\mathbf{P2}$. Similar to previous discussions, in this subsection we study the optimization problem by considering two cases, namely $0 < q_{req} < Nq_{\max}$ and $q_{req} = Nq_{\max}$.

1) $0 < q_{req} < Nq_{\max}$: As discussed in Section III.C, for each $\beta \in (0, 1]$ and $q_{req} \in (0, Nq_{\max}]$, there exists a threshold $q_{\mathrm{maxreq}}(\beta)$ which determines the maximal data requirement that can be realized by the production function defined in (29). According to (35), $q_{\mathrm{maxreq}}(\beta)$ monotonically decreases with β. Thus, given $q_{req} \in (0, Nq_{\max}]$, there exists a threshold ${q_{\mathrm{maxreq}}}^{-1}(q_{req})$, where ${q_{\mathrm{maxreq}}}^{-1}(\,\cdot\,)$ denotes the inverse function of $q_{\mathrm{maxreq}}(\,\cdot\,)$, such that when $\beta \leq {q_{\mathrm{maxreq}}}^{-1}(q_{req})$, $\rho^*_\beta(\theta)$ takes the form defined in (29), and when ${q_{\mathrm{maxreq}}}^{-1}(q_{req}) < \beta \leq 1$, $\rho^*_\beta(\theta)$ takes the form defined in (37). Notice that when $q_{req} = Nq_{\max}$, there is ${q_{\mathrm{maxreq}}}^{-1}(q_{req}) = 0$. We will discuss this special case later.

Given $q_{req} \in (0, Nq_{\max})$, the data collector's expected payoff $U_C(\beta)$ is defined as follows:

i) If $\beta = 0$, as we've discussed in Section III.C, there is $U_C(\beta) = \lambda\sqrt{\alpha_3 Nq_{req}}$.

ii) If $0 < \beta \leq {q_{\mathrm{maxreq}}}^{-1}(q_{req})$, substituting (29) and (36) into the objective function of problem $\mathbf{P2}$ and calculating the integral yields

$$U_C(\beta) = \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)N\lambda^2}{8\left(\bar{\theta}-\underline{\theta}\right)\beta}\ln\frac{\left(\bar{\theta}-\underline{\theta}\right)\beta+\frac{1}{2}\Delta}{\left(\underline{\theta}-\bar{\theta}\right)\beta+\frac{1}{2}\Delta} + \left(\frac{1}{2}\Delta-\beta\bar{\theta}\right)q_{req}, \tag{46}$$

where $\Delta = \sqrt{4\left(\bar{\theta}-\underline{\theta}\right)^2\beta^2+\frac{(\alpha_1\beta^{\alpha_2}+\alpha_3)N\lambda^2}{q_{req}}}$.

iii) If ${q_{\mathrm{maxreq}}}^{-1}(q_{req}) < \beta \leq 1$, substituting (37) and (41) into the objective function of problem $\mathbf{P2}$ and calculating the integral yields

$$U_C(\beta) = \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)N\lambda^2}{8\left(\bar{\theta}-\underline{\theta}\right)\beta}\ln\frac{\left(2\bar{\theta}-\underline{\theta}\right)\beta-\gamma'}{\left(2{\theta'}_c-\underline{\theta}\right)\beta-\gamma'} + N\left(\lambda\sqrt{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)q_{\max}}-\beta q_{\max}{\theta'}_c\right)\frac{{\theta'}_c-\underline{\theta}}{\bar{\theta}-\underline{\theta}} + \frac{\left(\alpha_1\beta^{\alpha_2}+\alpha_3\right)\left({\theta'}_c-\bar{\theta}\right)N\lambda^2\gamma'}{4\left(\bar{\theta}-\underline{\theta}\right)\left[\left(2\bar{\theta}-\underline{\theta}\right)\beta-\gamma'\right]\left[\left(2{\theta'}_c-\underline{\theta}\right)\beta-\gamma'\right]}, \tag{47}$$

where γ' and ${\theta'}_c$ are defined in (38) and (40) respectively.

It can be verified that as β approaches 0, (46) approaches (33). And as β approaches ${q_{\mathrm{maxreq}}}^{-1}$ from right-hand side, (47) approaches (46). Thus, though described in a piecewise form, the collector's payoff changes continuously with β. Let β^* denote the probability of privacy disclosure that maximizes the collector's payoff. Since both β and $U_C(\beta)$ are bounded, the existence of β^* is guaranteed. From a practical perspective, neither $\beta^* = 0$ nor $\beta^* = 1$ is desirable. If β^* does can be found in the interior, the following two conditions must hold:

$$\left.\frac{dU_C(\beta)}{d\beta}\right|_{\beta=\beta^*} = 0, \tag{48}$$

$$\left.\frac{d^2U_C(\beta)}{d\beta^2}\right|_{\beta=\beta^*} < 0. \tag{49}$$

Due to the complicated form of $U_C(\beta)$, it is hardly to derive the analytic form of β^* from (48). Instead, we propose a simple yet useful method to approximately determine the optimal protection level. Suppose that the data collector employs some k-anonymity algorithm [3] to protect data owners' privacy. For a given k, the probability of privacy disclosure can be roughly defined as $\beta = \frac{1}{k}$. Since the total number of collected data records is limited, k can only be chosen from a finite set, e.g., $\{2, \cdots, Nq_{\max}\}$. Given $q_{req} \in (0, Nq_{\max})(q_{req} < Nq_{\max})$, the optimal k can be determined in a following way. For each possible k, the collector first checks whether the condition $q_{req} \leq q_{\text{maxreq}}\left(\frac{1}{k}\right)$ holds. If it does, the collector computes his expected payoff $U_C\left(\frac{1}{k}\right)$ by using (46). Otherwise, the payoff is computed according to (47). After obtaining all the possible payoffs, the collector can decide which k is optimal.

2) $q_{req} = Nq_{\max}$: As discussed in Section III-C, when the collector requires the maximal data utility, i.e., $q_{req} = Nq_{\max}$, different data owners provide the same amount of data and receive the same transfer. In such a case, the collector's payoff is

$$U_C(\beta) = N\left[\lambda\sqrt{(\alpha_1\beta^{\alpha_2} + \alpha_3)\, q_{\max}} - \beta\bar{\theta}q_{\max}\right]. \tag{50}$$

Note that all the parameters, except λ, in the right-hand side of above equation are generally fixed. Thus, whether there exists a $\beta^* \in (0, 1)$ fully depends on λ. Later, by conducting numerical simulations, we will discuss how λ influences the choice of β^*.

E. Non-Optimal Contracts

The contract proposed above is the optimal solution to problem $\mathbf{P}$, i.e., among all the feasible contracts, it should bring the collector the maximal payoff. Despite the fact that it is impossible to explicitly compare this contract to all the other feasible contracts, here we propose a simple-formed contract, which we refer to as *linear-production contract*, with the purpose of obtaining more insight of the optimal contract. The linear-production contract is designed as follows.

Given $\beta \in (0, 1]$, the production function $\hat{\rho}_\beta(\,\cdot\,)$ is defined as

$$\hat{\rho}_\beta(\theta) = (\theta - \underline{\theta})\kappa + 1, \tag{51}$$

where κ is defined as

$$\kappa = \frac{2(q_{req} - Nq_{\max})}{(\bar{\theta} - \underline{\theta})Nq_{\max}}. \tag{52}$$

This linear production function implies that a data owner who does not care about privacy (i.e., $\theta = \underline{\theta}$) should hand over all his data, and for a data owner who cares about privacy, the data utility he contributes should be proportional to his privacy preference. The information rent function is defined as

$$\hat{U}_\beta(\theta) = -\frac{1}{2}\beta q_{\max}\kappa\theta^2 - (1 - \kappa\underline{\theta})\,\beta q_{\max}\theta + \left(\frac{1}{2}\kappa\bar{\theta}^2 - \kappa\underline{\theta}\bar{\theta} + \bar{\theta}\right)\beta q_{\max}. \tag{53}$$

It can be verified that if the collector has a relatively high requirement on data, i.e.,

$$\frac{1}{2}Nq_{\max} \leq q_{req} \leq Nq_{\max}, \tag{54}$$

then $(\hat{U}_\beta(\,\cdot\,), \hat{\rho}_\beta(\,\cdot\,))$ is a feasible solution to problem $\mathbf{P1'}$.

Substitute (51) and (53) into the objective function of problem $\mathbf{P1}$, then we get

$$\hat{U}_C(\beta) = \frac{2N\lambda\sqrt{(\alpha_1\beta^{\alpha_2} + \alpha_3)\, q_{\max}}}{3\left(\bar{\theta} - \underline{\theta}\right)\kappa}\left\{\left[\left(\bar{\theta} - \underline{\theta}\right)\kappa + 1\right]^{\frac{3}{2}} - 1\right\} - \frac{\left(\bar{\theta}^3 - \underline{\theta}^3\right)\kappa Nq_{\max}\beta}{6\left(\bar{\theta} - \underline{\theta}\right)} - \left(\frac{1}{2}\kappa\bar{\theta}^2 - \kappa\bar{\theta}\underline{\theta} + \bar{\theta}\right)N\beta q_{\max} \tag{55}$$

Similar to the case of optimal contract, it is quite difficult to derive the analytic form of β^* that maximizes (55). Considering that this contract is proposed for comparison purpose, we use a experimental method to approximately determine the value of β^* for both the optimal contract and the linear-production contract. More details will be presented in Section IV-A1.

IV. Contract Analysis and Simulation

In the previous section we have presented an elaborate description of the design of optimal contract. Analytic forms of the production function $\rho^*_\beta(\,\cdot\,)$ and information rent function $U^*_\beta(\,\cdot\,)$, which are optimal for a given β, are proposed. The expected payoff to the data collector is explicitly formulated as a function of β. Though we do not provide an explicit formulation of the optimal privacy protection level, we can utilize the derived formulation of $U_C(\beta)$ to provide a general insight into the trade-off between privacy protection and utility preserving.

In this section, by conducting numerical simulations, we qualitatively analyze how the optimal privacy protection level relates to the collector's requirement on data utility and the exogenously determined value of data. Moreover, in order to evaluate whether the optimal contract can bring the collector a good payoff, we conduct real data experiments and make a comparison of the two types of contracts proposed in Section III. In the following part, we first describe how we determine the optimal privacy protection level via simulations. Then based on simulation results, we present a qualitative analysis of the optimal contract. After that, we introduce the settings of real data experiments and present the results.

A. Contract Analysis

1) Determining the Optimal Privacy Protection Level Experimentally: To observe how the two parameters q_{req} and λ influence the choice of privacy protection level in the optimal contract, we conduct the following simulations. First, we set those invariable parameters as follows: $N = 3000$, $q_{\max} = 10$, $\alpha_1 = 0.4804$, $\alpha_2 = 0.2789$, $\alpha_3 = 1 - \alpha_1$, $\bar{\theta} = 1$, and $\underline{\theta} = 0$. Then, for each pair of $q_{req} \in \left\{\frac{1}{20}Nq_{\max}, \frac{2}{20}Nq_{\max}, \cdots, Nq_{\max}\right\}$ and $\lambda \in \{0.1, 0.2, \cdots 0.9, 1, 2, \cdots, 100\}$, we compute a group of $\{U_C(\beta)\}$, each of which corresponds to a $\beta \in \left\{\frac{0}{100}, \frac{1}{100}, \cdots, \frac{100}{100}\right\}$. For each β, we first compare q_{req} with $q_{\text{maxreq}}(\beta)$. Then based on the comparison result, $U_C(\beta)$ is computed according to (33), (46) or (47). After that, the maximal $U^*_C(\beta)$ is picked from $\{U_C(\beta)\}$, and the corresponding privacy protection level $\delta^* \triangleq 1 - \beta^*$ is recorded. As for

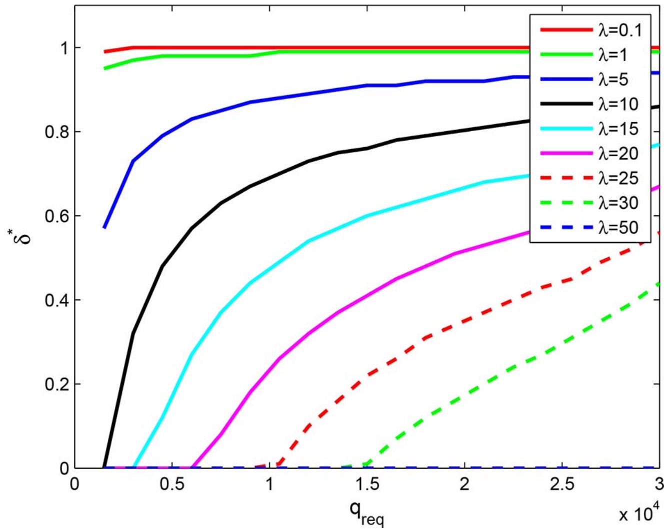


Fig. 3. Relationship between the optimal privacy protection level and the requirement on total data utility.

the linear-production contract proposed in Section III-E, the optimal privacy protection level is determined in a similar way.

2) Data Requirement and Privacy Protection: As discussed in Section III-C, how the optimal contract should be formed largely depends on the collector's requirement on data. From the results shown in Fig. 3 we can see that, for a given λ, δ^* increases with q_{req}, as long as λ is neither too high nor too low. This phenomenon implies that if the data collector wants to get more data from data owners, he should offer better protection for data owners' privacy. Or in other words, knowing that his privacy can be better protected, the data owner will feel less unsafe to hand over his private data, thus he is willing to provide more data.

To better understand above implication, we rewrite the collector's expected payoff $U_C(\beta)$ as a sum of two terms, i.e.,

$$U_C(\beta) = S(\beta) - T(\beta), \tag{56}$$

where $S(\beta)$ denotes the expected income, i.e.,

$$S(\beta) = N \int_{\underline{\theta}}^{\bar{\theta}} \lambda \sqrt{(\alpha_1 \beta^{\alpha_2} + \alpha_3)\, q_{\max} \rho_\beta^*(\theta)}\, f(\theta) d\theta, \tag{57}$$

and $T(\beta)$ denotes the expected transfer, i.e.,

$$T(\beta) = N \int_{\underline{\theta}}^{\bar{\theta}} \left[U_\beta^*(\theta) + \beta\theta q_{\max} \rho_\beta^*(\theta) \right] f(\theta) d\theta. \tag{58}$$

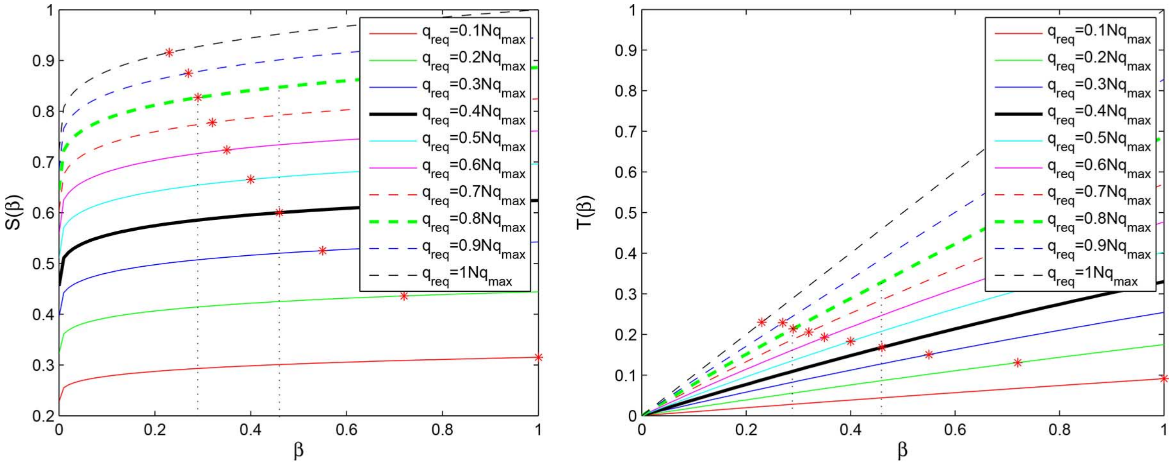
During the simulations, we compute $S(\beta)$ and $T(\beta)$ together with $U_C(\beta)$. As shown in Fig. 4, with λ being fixed at a moderate value (e.g., $\lambda = 15$), for any given q_{req}, both the income and the transfer increase with β. This coincides with the intuition that when privacy protection level decreases, the collector can obtain more benefit from the less anonymized data, meanwhile, data owners face a higher risk of privacy disclosure, hence they require more transfer to compensate the privacy loss. From Fig. 4 we can see that, compared to the income $S(\beta)$, the transfer $T(\beta)$ is more sensitive to β. And as q_{req} becomes higher, $T(\beta)$ grows faster with β, while $S(\beta)$ grows at almost the same rate. According to (2), the marginal value of data decrease with the utility of anonymitized data which, according to (1), grows slower as the utility of collected data increases. This may explain why $S(\beta)$ is insensitive to β. While as for $T(\beta)$, it is roughly proportional to β, which means even a small change of β can be captured by $T(\beta)$.

Suppose that currently the collector's data requirement is $q_{req} = 0.4Nq_{\max}$, and the optimal privacy protection level he adopts is about 0.57. When the collector has a higher requirement, say $q_{req} = 0.8Nq_{\max}$, he has to pay much more transfer if he sticks with original privacy protection level. However, if the collector chooses a higher protection level, despite that he'll lose a small amount of income, the transfer he needs to pay can be largely reduced. Therefore, when the collector desires data of high utility, he should put more effort to protect data owners' privacy.

Fig. 3 also shows that when q_{req} is relatively small and λ is large, the collector does not need to take care of data owners' privacy. This is because that when data is very valuable, the income from the data is far beyond sufficient to compensate data owners' privacy loss, thus there is no need to take privacy protection measures. However, as the collector has collected more data to meet a higher q_{req}, the marginal value of data decreases, and the income may be insufficient to compensate the privacy loss. Therefore, the collector should take privacy protection measures, so that the transfer paid to data owners can be reduced to an affordable level.

3) The Value of Data and Privacy Protection: The parameter λ in data collector's income function 2 indicates whether the data is valuable to the collector. From the simulation results shown in Fig. 5 we can see that, when λ is quite small ($\lambda < 1$), the optimal privacy protection level equals 1, which means the data collector must offer a perfect protection of privacy. The reason why this result appears is that we have defined the privacy parameter θ takes values from 0 to 1. Considering that θ can be interpreted as the unit cost that a data owner spends on producing the data, the transfer that the data collector pays to the owner should be at an equivalent level. Then when $\theta \in [0, 1]$ and $\lambda < 1$, the benefit that the collector gets from the data may be even less than the owner's cost, which means the collector cannot afford any compensation for data owners' privacy loss. As a result, providing perfect privacy protection may be the only feasible choice for the collector.

As shown in Fig. 5, the optimal privacy protection level decreases as λ increases. This implies that as data becomes more valuable, sacrificing data utility for privacy protection becomes less beneficial to the collector. In such cases, though increasing privacy protection level can reduce the transfer paid to data owners, the resulting decrease of data utility will cause a larger loss to the collector. When λ is quite large, data is so valuable to the collector that even a minor decrease in data utility, which is caused by a weak anonymization, will cause a large loss to the collector. As a result, the collector prefers to do nothing to protect privacy. From Fig. 5 we can see that, as q_{req} becomes higher, in a larger range of λ, protecting privacy is more preferred by the collector than providing no protection. This result coincides with the observation we've got in Section IV-A2, that



(a) (b)

Fig. 4. An illustration of how data collector's income and transfer change with the privacy protection level: (a) expected income $S(\beta)$; (b) expected transfer $T(\beta)$. The plots, which denote the optima, are obtained via the simulations described in Section IV-A1 where λ is set to 15. Values of $S(\beta)$ shown in the figure have been normalized by dividing original value by the maximum among all results. Values of $T(\beta)$ have been normalized in a similar way. Red stars denote the income (or transfer) at the optimum, i.e., $(\beta^*, S(\beta^*))$ (or $(\beta^*, T(\beta^*))$).

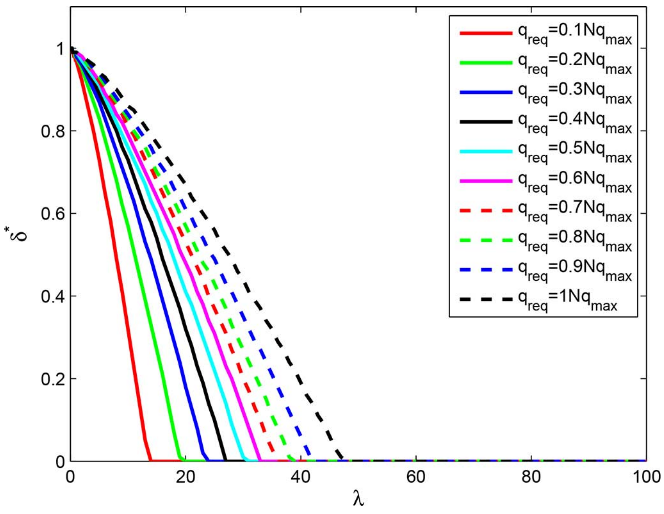


Fig. 5. Relationship between the optimal privacy protection level and the value of data.

is, a better protection of privacy is required if the collector wants to get data of higher utility.

More insight about how the parameter λ influences the design of privacy protection level can be obtained by analyzing the second case discussed in Section III-D2, i.e., $q_{req} = Nq_{\max}$. As mentioned earlier, whether the data collector's payoff can reach its maximum at an interior β can be determined by evaluating the derivation as follows.

When λ meets the following condition

$$\lambda \geq \max_{\beta\in[0,1]} \frac{2\bar{\theta}\beta^{1-\alpha_2}}{\alpha_1\alpha_2}\sqrt{(\alpha_1\beta^{\alpha_2}+\alpha_3)\,q_{\max}} = \frac{2\bar{\theta}\sqrt{q_{\max}}}{\alpha_1\alpha_2}, \tag{59}$$

there is $\frac{dU_C(\beta)}{d\beta} \leq 0$, and $\frac{dU_C(\beta)}{d\beta} = 0$ iff $\beta = 1$. In such a case, the collector's payoff increases as the privacy protection level decreases, thus $\beta^* = 1$.

When $0 < \lambda < \frac{2\bar{\theta}\sqrt{q_{\max}}}{\alpha_1\alpha_2}$, $U_C(\beta)$ reaches its maximum at an interior $\beta^* \in (0,1)$ which satisfies $\frac{dU_C(\beta)}{d\beta}\Big|_{\beta=\beta^*} = 0$. It can be verified that the second order condition $\frac{dU_C^2(\beta)}{d\beta^2}\Big|_{\beta=\beta^*} < 0$ also holds. From Fig. 4 we know that, as the privacy protection level increases (i.e., β decreases), both the income and the transfer decreases. When λ is relatively small, the reduced transfer caused by one-unit increase of protection level is comparable with the corresponding income loss. At some point, a small increase of protection level causes no change to the payoff, that's when the payoff is maximized. Moreover, notice that $U_C(1) = N\sqrt{q_{\max}}(\lambda - \bar{\theta}\sqrt{q_{\max}})$ and $\bar{\theta}\sqrt{q_{\max}} < \frac{2\bar{\theta}\sqrt{q_{\max}}}{\alpha_1\alpha_2}$. Hence, if $\lambda < \bar{\theta}\sqrt{q_{\max}}$, the data collector cannot get a positive payoff unless a certain level of privacy protection can be realized. From Fig. 5 we can see that, as λ becomes smaller, β^* moves towards 0. This phenomena implies that as the data becomes less valuable to the collector, the collector has to put more effort to protect data owners' privacy, so that a low transfer will be accepted by data owners and the collector can still keep his payoff stay at a certain level.

B. Experiments on Real-World Data

1) Dataset and Anonymization Configurations: To evaluate the performance of the contracts in a context where anonymization is performed on real data, we conduct experiments on the Adult data set [19], which is widely used in the study of data anonymization. The original data set consists of 32,561 records from a census database, and each record consists of 15 attributes. After removing records with unknown values, we randomly choose 30,000 records for experiment. Similar to previous study on anonymization [20], [4], only 9 attributes, namely *age*, *workclass*, *education*, *marital-status*, *occupation*, *race*, *sex*, *native-country*, and *salary-class*, are kept for experiment.

To perform anonymization, we develop a Java project based on the open source anonymization framework ARX [21], which supports different types of privacy criteria and provides multiple methods for measuring information loss [1]. Here we choose the most widely applied privacy criterion, i.e., k-anonymity, to conduct experiments. A simple explanation to this privacy criterion

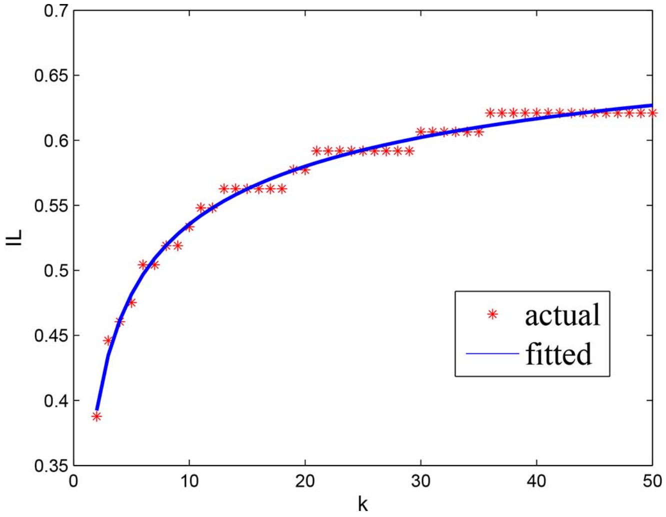


Fig. 6. Relationship between the privacy criterion and information loss. The information loss IL is measure by *precision* [3]. The blue curve is fitted by using the data denoted by red stars. By using MATLAB curve fitting toolbox, we choose a power function to formulate the fitted curve, which is $IL = -0.4804k^{-0.2789} + 0.7883$. From the reported R-square (coefficient of determination) index, which is 0.9896, we know that such formulation is appropriate.

is that after anonymization, the probability that an individual being re-identified by an attacker is no higher than $^1/_k$. Hence, if the k-anonymity criterion is met by the anonymized data, the realized privacy protection level can be defined as $\delta \triangleq 1 - \frac{1}{k}$.

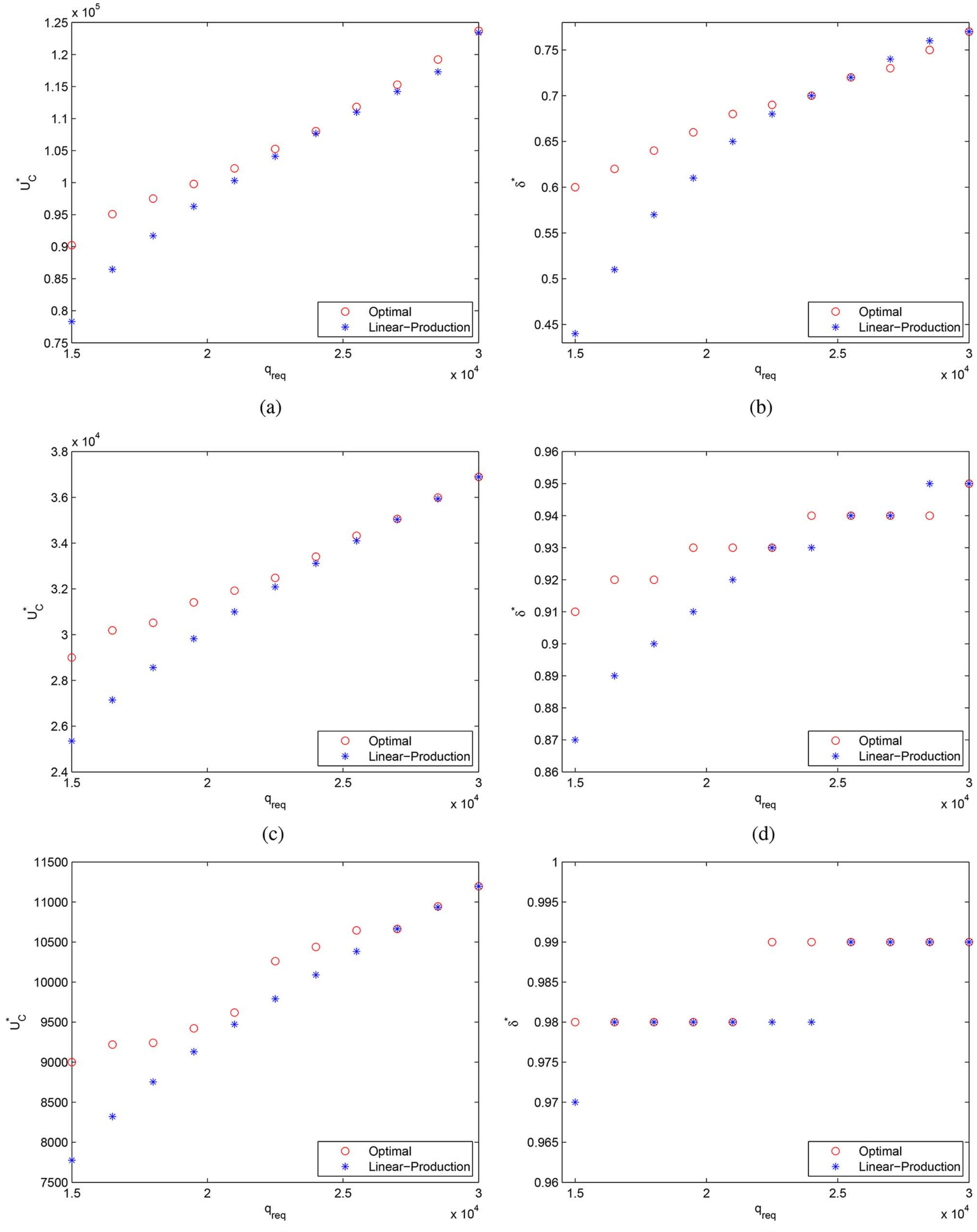
After anonymization, the utility of data decreases. The decrease of utility, also referred to as *information loss*, can be measured in different ways. Here we choose the *precision* metric [3], which ranges from 0 to 1. Intuitively, if a larger k is chosen as the privacy criterion, the information loss will becomes higher. In order to quantitatively analyze how the information loss changes with k, we conduct a group anonymization experiments on aforementioned data set. All the 9 attributes are treated as quasi-identifiers, namely each of them can be generalized according to a domain generalization hierarchy [20]. For each $k \in \{2, \cdots, 50\}$, we run the anonymization program and record the reported information loss IL. Experiment results are shown in Fig. 6. By using the curve fitting toolbox provided in MATLAB, we formulate IL as a power function of k. Then, by defining $IL = \frac{q-d(q,\delta)}{q}$, which means IL is interpreted as the ratio of the decreased utility to that of the original data, we get the formulation defined in (1).

2) Contract Simulation: To demonstrate the superiority of the optimal contract over the linear-production contract, we conduct multiple experiments to simulate data owners' response to different contracts, and check whether the optimal contract can bring the data collector a higher payoff. Experiments are configured in the following way. First, we randomly divide the 30,000 records into N groups, where N is set to 3000, 300, and 30 respectively. Each group of records corresponds to a data owner. That is to say, we set $q_{\max} = 10, 100, 1000$ respectively. The privacy parameter θ of each data owner is set by uniformly sampling in the interval $[0, 1]$. The rest parameters are set as $\lambda = 15$, $\alpha_1 = 0.4804$, $\alpha_2 = 0.2789$, $\alpha_3 = 0.5196$, and $q_{req} = \frac{m}{20} N q_{\max} (m = 10, 11, \cdots, 20)$.

Given the value of N and the value of q_{req}, the maximal payoff that the data collector can get from a certain contract is computed as follows. First, we determine the optimal privacy protection level β^* by using the method described in Section IV.A1. Then, based on the production function $\rho^*_{\beta^*}(\cdot)$ defined in the contract, we determine the number of records that each data owner i will provide. Let n_i denote the number of records and θ_i denote the owner's type. We set $n_i = \left\lceil \rho^*_{\beta^*}(\theta_i) q_{\max} \right\rceil$, where $\lceil a \rceil$ denote the smallest integer that is no less than a. Based on n_i and the information rent function $U^*_{\beta^*}(\cdot)$, the information rent u_i paid to owner i can be determined. After above computation, we construct a new data set by randomly choosing n_i records from the 10 records corresponding to each owner i. To run anonymization experiments on this data set, we set $k = \left\lceil \frac{1}{\beta^*} \right\rceil$. Then, based on the reported information loss and each owner's (n_i, u_i), we can determine the collector's payoff U^*_C. Considering that records in the new data are randomly chosen, for a given q_{req} and a contract, we repeat above procedure 5 times and report the average results.

3) Comparison Results: Simulation results are shown in Fig. 7. As we can see, in all settings of q_{req}, the optimal contract exhibits a better performance than the linear-production contract. As shown in Fig. 7(a), (c) and (e), in all settings of q_{req}, the optimal contract can bring the collector a higher payoff than, if not equal to, that brought by the linear-production contract, especially when q_{req} is close to $0.5Nq_{max}$. As q_{req} increases, the difference between the two contracts, in terms of payoff, becomes insignificant. On the other hand, Fig. 7(b), (d) and (f) show that when $q_{req} < 0.75Nq_{max}$, the optimal contract can offer the data owners a better protection of privacy. However, as q_{req} becomes quite high, the optimal contract can only realize a similar, even lower, privacy protection level as that realized by linear-production contract. It should be noted that a lower privacy protection level does not mean the optimal contract is worse than the linear-production contract, since the optimal contract is designed to maximize the data collector's expected payoff rather than maximizing the privacy protection level. Also, from Fig. 7(f) we can see that when the number of data owners is quite small ($N = 30$), the optimal privacy protection level approaches to 1 in all settings of q_{req}. This result can be explained in a following way. In the setting where $N = 30$, each data owner owns 1000 records. In order to meet the collector's requirement, say $q_{req} = 20000$, on average each data owner has to provide more than 600 records. The data owners will show great concern about privacy when they are asked to provide so many private data. Thus it is necessary for the collector to offer strong protection to the owners' privacy.

To understand why the optimal contract loses its advantage when q_{req} is high, we can recall the geometrical interpretation presented in Section III.C. As illustrated in Fig. 2, a high q_{req} means a large part of the curve segment defined by the production function lies on the boundary. And as q_{req} becomes higher, the rest part of the curve becomes more “flat”. As for the linear-production contract, it uses a liner production function. According to (52), when q_{req} approaches Nq_{max}, the production decreases with θ at a very low rate. To sum up, when q_{req} is close to the maximum Nq_{max}, the two types of con-



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 7. Performance evaluation of the optimal contract and the linear-production contract: (a)(c)(e) data collector's payoff; (b)(d)(f) optimal privacy protection level, $\delta^* = 1 - \beta^*$. (a) $N = 3000$. (b) $N = 3000$. (c) $N = 300$. (d) $N = 300$. (e) $N = 30$. (f) $N = 30$.

tracts will make no obvious difference in their production functions, hence they exhibits similar performance. In addition to above results, it should be noted that the linear-production contract can be applied only when $0.5Nq_{max} \leq q_{req} \leq Nq_{max}$, while the optimal contract can also be applied to cases when q_{req} is low. In that sense, the optimal contract is more practical than the linear-production contract.

V. Conclusion

To deal with the information asymmetry problem emerging in private data collecting, in this paper we proposed a contract theoretic approach to help the data collector make a rational decision on how to pay the data owners. Considering that the data collector also needs to carefully adjust the privacy protection level, we treated the privacy protection level as a contract item, and explicitly solved the optimal production functions and information rent functions for any given protection level. We've shown that as the collector's requirement on data changes, the optimal functions may be formed in a different way. As for the optimal privacy protection level, we've analyzed how it should be adjusted when the collector faces a different requirement on data utility or has a new valuation of data. Such analysis can provide a practical guidance for private data collecting.

The optimal contract proposed in this paper is mainly based on the assumptions we've made on data collector's income function as well as the relationship between data utility and privacy protection level. Whether there are more reasonable formulations of these two functions needs to be further investigated. Besides, in our study we have assumed that the distribution of data owners' privacy preference is known to the collector. In future work, we will study the contract design problem in a context where the distribution knowledge is unavailable to the collector. Moreover, currently we assume that the data owner's privacy parameter is pre-specified by the *nature*, yet it is important to explore practical ways to quantify individual's preference on privacy. Whether we can learn one's valuation of his privacy from one's historical behavior is a problem worth further studying.

Appendix A
Simplification of the Incentive Constraints

According to (6), for any $(\theta, \tilde{\theta}) \in [\underline{\theta},\ \bar{\theta}]^2$, the following two inequalities hold:

$$t(\theta) - \beta\theta\rho(\theta)q_{\max} \geq t(\tilde{\theta}) - \beta\theta\rho(\tilde{\theta})q_{\max}, \qquad (60)$$

$$t(\tilde{\theta}) - \beta\tilde{\theta}\rho(\tilde{\theta})q_{\max} \geq t(\theta) - \beta\tilde{\theta}\rho(\theta)q_{\max}. \qquad (61)$$

Adding (60) and (61) yields

$$(\tilde{\theta} - \theta)(\rho(\theta) - \rho(\tilde{\theta}))\beta q_{\max} \geq 0. \qquad (62)$$

Above inequality should hold for any $\beta \in [0, 1]$, which means $\rho(\cdot)$ has to be a non-increasing function of θ. Furthermore, (62) implies that both $\rho(\cdot)$ and $t(\cdot)$ are differentiable almost everywhere. Hence, we can restrict the analysis to piecewise differentiable functions. Given θ, (60) implies that the function $g(\tilde{\theta}) \triangleq t(\tilde{\theta}) - \theta\beta q_{\max}\rho(\tilde{\theta})$ reaches its maximum at $\tilde{\theta} = \theta$, thus θ must satisfy the following two conditions:

$$\frac{dt(\theta)}{d\theta} - \beta q_{\max}\theta\frac{d\rho(\theta)}{d\theta} = 0, \qquad (63)$$

$$\frac{d^2t(\theta)}{d\theta^2} - \beta q_{\max}\theta\frac{d^2\rho(\theta)}{d\theta^2} \leq 0. \qquad (64)$$

By differentiating (63), (64) can be written as:

$$-\frac{d\rho(\theta)}{d\theta} \geq 0. \qquad (65)$$

The (63) and (65) constitute the local incentive constraints. Then, by using (63) we can write the data owner's information rent as

$$\begin{aligned} t(\theta) - \theta\beta q_{\max}\rho(\theta) =& t(\tilde{\theta}) - \theta\beta q_{\max}\rho(\tilde{\theta}) + \\ & \beta q_{\max}\int_{\tilde{\theta}}^{\theta}[\rho(\tilde{\theta}) - \rho(\tau)]d\tau. \end{aligned} \qquad (66)$$

The non-increasing property (65) ensures that the third item in the right-hand side of above equation is non-negative, which means the local incentive constraints imply also the global incentive constraints. Hence, we can reduce the infinity of incentive constraints in (6) to a differential equation (63) and a monotonicity constraint (65).

References

[1] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, 2010.

[2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *SIGMOD Rec.*, vol. 29, no. 2, pp. 439–450, 2000.

[3] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," *Int. J. Uncertainty, Fuzziness, Knowl.-Based Syst.*, vol. 10, no. 05, pp. 571–588, 2002.

[4] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," in *Proc. 22nd Int. Conf. Data Eng. (ICDE'06)*, Apr. 2006, pp. 24–24.

[5] N. Li, T. Li, and S. Venkatasubramanian, "T-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. ICDE*, 2007, vol. 7, pp. 106–115.

[6] C. C. Aggarwal and S. Y. Philip*, A General Survey of Privacy-Preserving Data Mining Models and Algorithms*. New York, NY, USA: Springer, 2008.

[7] S. Matwin, "Privacy-preserving data mining techniques: Survey and challenges," in *Discrimination and Privacy in the Information Society*. New York, NY, USA: Springer, 2013, pp. 209–221.

[8] A. Acquisti, "The economics of privacy: Theoretical and empirical aspects," 2013 [Online]. Available: http://cusp.nyu.edu/wp-content/uploads/2013/09/C03-acquisti-chapter.pdf

[9] A. Roth, "Buying private data at auction: The sensitive surveyor's problem.," *SIGecom Exchanges*, vol. 11, no. 1, pp. 1–8, 2012.

[10] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proc. 12th ACM Conf. Electron. Commerce*, 2011, pp. 199–208.

[11] L. K. Fleischer and Y.-H. Lyu, "Approximately optimal auctions for selling privacy when costs are correlated with data," in *Proc. 13th ACM Conf. Electron. Commerce ACM*, 2012, pp. 568–585.

[12] K. Ligett and A. Roth, "Take it or leave it: Running a survey when privacy comes at a cost," in *Internet and Network Economics*. New York, NY, USA: Springer, 2012, pp. 378–391.

[13] K. Nissim, S. Vadhan, and D. Xiao, "Redrawing the boundaries on purchasing data from privacy-sensitive individuals," in *Proc. 5th Conf. Innovat. Theoret. Comput. Sci.*, 2014, pp. 411–422.

[14] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. New York, NY, USA: Springer, 2006, pp. 1–12.

[15] J.-J. Laffont and D. Martimort*, The Theory of Incentives: The Principal-Agent Model*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[16] Y. Gao, Y. Chen, C.-Y. Wang, and K. J. R. Liu, "Optimal contract design for ancillary services in vehicle-to-grid networks," in *Proc. IEEE 3rd Int. Conf. Smart Grid Commun. (SmartGridComm)*, Nov. 2012, pp. 79–84.

[17] Y. Gao, Y. Chen, C.-Y. Wang, and K. J. R. Liu, "A contract-based approach for ancillary services in v2g networks: Optimality and learning," in *Proc. IEEE INFOCOM*, Apr. 2013, pp. 1151–1159.

[18] D. Kirk*, Optimal Control Theory: An Introduction*, ser. Dover Books on Electrical Engineering. New York, NY, USA: Dover, 2012.

[19] K. Bache and M. Lichman, "UCI Machine Learning Repository," 2013 [Online]. Available: http://archive.ics.uci.edu/ml

[20] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, New York, NY, USA, 2005, pp. 49–60, ser. SIGMOD'05, ACM..

[21] F. Kohlmayer, F. Prasser, C. Eckert, A. Kemper, and K. Kuhn, “Flash: Efficient, stable and optimal k-anonymity,” in *Proc. Privacy, Security, Risk and Trust (PASSAT), Int. Conf. and 2012 Int. Conf. Social Comput. (SocialCom)*, Sep. 2012, pp. 708–717.

Lei Xu received the Bachelor's degree from Tsinghua University in 2008, and the Ph.D. degree from Tsinghua University in 2015. Her current research interests include privacy issues in data mining, text mining, and game theory.

Chunxiao Jiang (S'09–M'13) received his B.S. degree in information engineering from Beijing University of Aeronautics and Astronautics (Beihang University) in 2008 and the Ph.D. degree from Tsinghua University (THU), Beijing, in 2013, both with the highest honors. During 2011–2013, he visited the Signals and Information Group (SIG) at Department of Electrical & Computer Engineering (ECE) of University of Maryland (UMD) with Prof. K. J. Ray Liu. Dr. Jiang is currently a post-doctor in the EE department of THU with Prof. Yong Ren. His research interests include the applications of game theory and queuing theory in wireless communication and networking and social networks.

Dr. Jiang received Best Paper Award from IEEE GLOBECOM in 2013, the Beijing Distinguished Graduated Student Award, Chinese National Fellowship and Tsinghua Outstanding Distinguished Doctoral Dissertation in 2013.

Yan Chen (SM'14) received the Bachelor's degree from University of Science and Technology of China in 2004, the M.Phil. degree from Hong Kong University of Science and Technology (HKUST) in 2007, and the Ph.D. degree from University of Maryland College Park in 2011. His current research interests are in data science, network science, game theory, social learning and networking, as well as signal processing and wireless communications.

Dr. Chen is the recipient of multiple honors and awards including best paper award from IEEE GLOBECOM in 2013, Future Faculty Fellowship and Distinguished Dissertation Fellowship Honorable Mention from Department of Electrical and Computer Engineering in 2010 and 2011, respectively, Finalist of Deans Doctoral Research Award from A. James Clark School of Engineering at the University of Maryland in 2011, and Chinese Government Award for outstanding students abroad in 2011.

Yong Ren received his B.S., M.S., and Ph.D. degrees in electronic engineering from Harbin Institute of Technology, China, in 1984, 1987, and 1994, respectively. He was a Post Doctor at the Department of Electronics Engineering, Tsinghua University, China, from 1995 to 1997. Now he is a Professor of Department of Electronics Engineering and the Director of the Complexity Engineered Systems Lab (CESL) in Tsinghua University. He holds 12 patents, and has authored or co-authored more than 100 technical papers in the behavior of computer network, P2P network and cognitive networks. He has serves as a reviewer of *IEICE Transactions on Communications*, *Digital Signal Processing*, *Chinese Physics Letters*, *Chinese Journal of Electronics*, *Chinese Journal of Computer Science & Technology*, *Chinese Journal of Aeronautics* and so on. His current research interests include complex systems theory and its applications to the optimization and information sharing of the Internet, Internet of Things and ubiquitous network, cognitive networks and Cyber-Physical Systems.

K. J. Ray Liu (F'03) was named a Distinguished Scholar-Teacher of University of Maryland, College Park, in 2007, where he is Christine Kim Eminent Professor of Information Technology. He leads the Maryland Signals and Information Group conducting research encompassing broad areas of information and communications technology with recent focus on future wireless technologies, network science, and information forensics and security.

Dr. Liu was a recipient of IEEE Signal Processing Society 2014 Society Award, IEEE Signal Processing Society 2009 Technical Achievement Award, and various best paper awards. Recognized by Thomson Reuters as a Highly Cited Researcher, he is a Fellow of IEEE and AAAS.

Dr. Liu is a Director-Elect of IEEE Board of Director. He was President of IEEE Signal Processing Society, where he has served as Vice President C Publications and Board of Governor. He has also served as the Editor-in-Chief of *IEEE Signal Processing Magazine*.

He also received teaching and research recognitions from University of Maryland including university-level Invention of the Year Award; and college-level Poole and Kent Senior Faculty Teaching Award, Outstanding Faculty Research Award, and Outstanding Faculty Service Award, all from A. James Clark School of Engineering.