The Privacy Paradox and Optimal Bias-Variance Trade-offs in Data Acquisition

Guocheng Liao *
CUHK

Yu Su *
Caltech

Juba Ziani
University of Pennsylvania

Adam Wierman
Caltech

Jianwei Huang
CUHK, SZ

ABSTRACT

While users claim to be concerned about privacy, often they do little to protect their privacy in their online actions. One prominent explanation for this “privacy paradox” is that when an individual shares her data, it is not just her privacy that is compromised; the privacy of other individuals with correlated data is also compromised. This information leakage encourages oversharing of data and significantly impacts the incentives of individuals in online platforms. In this extended abstract, we discuss the design of mechanisms for data acquisition in settings with information leakage and verifiable data. We summarize work designing an incentive compatible mechanism that optimizes the worst-case trade-off between bias and variance of the estimation subject to a budget constraint, where the worst-case is over the unknown correlation between costs and data. Additionally, we characterize the structure of the optimal mechanism in closed form and study monotonicity and non-monotonicity properties of the marketplace.

1. INTRODUCTION

There is a fundamental discrepancy between privacy attitudes and the behaviors of users: users claim to be concerned about their privacy but do little to protect privacy in their actions. This phenomenon is known as the *privacy paradox*. Understanding the reasons behind this paradox and its consequences for the design of online platforms is an important goal for both computer scientists and economists.

We try to explain the *privacy paradox* from the perspective of data correlation. In particular, when an individual shares her data, it is not just her privacy that is compromised; the privacy of other individuals whose data is correlated with hers is also compromised. Thus, these other individuals are more likely to share their own data given that some has already been leaked [1]. Information leakage due to correlation has been shown to lead to oversharing since each individual overlooks their own privacy concerns as a result of the negative externalities created by others’ revelation decisions.

In this extended abstract, we study the impact of privacy concerns and information leakage on the design of data markets. Specifically, we study the task of designing mechanisms for obtaining verifiable data from a population for a statisti-

cal estimation task, such as estimating the expected value of some function of the underlying data (e.g., salary and BMI).

A growing line of work has focused on the design of such optimal data acquisition mechanisms, e.g., [3, 2]. Studies in [3, 2] focused on mechanism design for unbiased estimation with minimal variance. However, this literature assumed that all individuals will participate, thus unbiased estimation is possible. Further, this line of work has not considered the issues created by information leakage due to correlation between the participants. Information leakage creates significant incentives for increased data sharing [1] and thus mechanisms that do not consider it directly will suffer from undetected bias and increased variance in the obtained estimates.

Contributions. In this extended abstract, we provide the first characterization of an optimal mechanism for data acquisition in the setting where agents are concerned about privacy due to data correlation. Data correlation causes information leakages to non-data reporters. Additionally, the mechanism allows, for the first time, a trade-off between the bias and variance of the estimator when privacy cost is considered.

Specifically, we propose a novel model for data acquisition. The novelty of our model consists of three important components. First, we introduce the privacy cost to model impacts of data correlation. We divide the agents into different groups and assume that agents within the same group share a same correlation strength. This gives us the power to work with any granularity of choice with regard to data correlation. Second, in reality, not every agent always decides to join the platform. Thus, we introduce the notion of participation rate as the ratio of the number of agents who join the platform to the number of total agents. This further allows us to study equilibria with respect to participation rate, which is crucial since the mechanism impacts the participation rate. Third, given that not every agent joins the platform, it is not always realistic to aim for an unbiased estimator. Instead, we minimize a linear combination of bias and variance of the estimator. Via a choice of constant weights for bias and variance, we are able to balance between these two metrics of the estimator.

Our main theoretical results provide a closed form solution of payment and allocation rules under a choice of equilibrium participation rate in order to achieve a truthful mechanism. More specifically, we aim to minimize a linear combination of bias and variance subject to budget and truthfulness constraints. Moreover, we provide conditions for the optimality of an unbiased estimator in the case when it is possible to

*Equal Contribution.

achieve a full participation rate.

Our results offer some interesting insights about mechanisms for data acquisition. First, an unbiased estimator is possible even if the budget is relatively small. Second, incorporation of privacy cost due to information leakage makes it possible for the analyst to underpay the agents to acquire the same data set and encourages more agents to join the platform. This aligns with the *privacy paradox* frequently observed in platforms today.

2. SYSTEM MODEL

We consider an online platform consisting of an analyst and many agents. The analyst aims to design a mechanism to purchase private data from agents to perform a statistical estimation task, e.g., estimate the mean of the agents' data. There are s agents that hold data of interest to the analyst. Every agent owns a data point. By reporting her data to the platform's survey, the agent incurs an overall cost c , which is known to her but not to the analyst. The overall cost consists of a combination of reporting cost and privacy cost, where the reporting cost results from the act of reporting the data while the privacy cost comes from both data sharing and data correlation.

We consider a family of pricing mechanisms that presents a menu to the agents. The menu consists of pairs of payments A and probabilities P of having agents' data selected for use in the estimation task. Both A and P are functions of agent's reported cost $\tilde{c}$. The mechanism works as follows. First, each agent is presented the menu by the analyst. Second, given the menu, an agent decides if she would like to join the platform. An agent who decides to join the platform is asked to report her cost, which determines the payment and selection probability. The payment P is given to the agent if she joins the platform *and* her data is selected (used) by the analyst. More specifically, her data is selected with probability A , and she receives the payment only if her data is selected.

2.1 Agent Model

We introduce the agent's utility function. If the agent does not join the platform, she suffers a privacy cost $g(c)$ due to information leakage as a result of correlation between her data and the data of those agents who do join. If the agent joins the platform but is not selected for data contribution, she experiences the benefit of participation w and privacy cost. The privacy cost is from the agent herself sharing her data in the platform and the privacy cost due to her friends' sharing of possibly correlated data. We use $h(c)$ to denote this combination. If an agent joins the platform and is selected for data contribution, she obtains the benefit of participation and payment, and suffer true cost c . We use $\mathcal{N}$ to denote the set of agents who join the platform. In summary, an agent k 's utility is:

$$u_i(\tilde{c}|c) = \begin{cases} -g(c), & \text{if } k \notin \mathcal{N}, \\ -h(c) + w(\cdot), & \text{w/ prob. } 1 - A(\tilde{c}), \text{ if } k \in \mathcal{N}, \\ P(\tilde{c}) - c + w(\cdot), & \text{w/ prob. } A(\tilde{c}), \text{ if } k \in \mathcal{N}. \end{cases}$$

Next, we elaborate each term in the utility in details.

Participation Benefit. Agents receives participant benefits by using services provided by the platform. The more participants the platform attracts, the more valuable the platform's service becomes. Let $\bar{\theta}$ be the average participa-

tion rate of the population, i.e., the ratio of number of agents on the platform to the population of agents. We use $w(\bar{\theta})$ to denote the participation benefit, which is a non-decreasing function of the average participation rate $\bar{\theta}$.

Correlation Strength. Data correlation results in information leakage. A stronger correlation naturally leads to more leakage and induces a larger privacy cost. Moreover, some agents might share a stronger correlation with each other within a group of agents than with others outside the group. In order to capture the inter-group versus intra-group difference, we divide the s agents into I groups, and agents within the same group i share a common correlation strength α_i . The correlation strength vector α_i is further defined as $\alpha_i \triangleq (\alpha_i, \alpha_{-i})$, where α_i and α_{-i} are used to respectively denote the correlation strength induced by agents inside group i and those outside group i .

Privacy Cost. For an agent who does not join the platform, her privacy cost $g(\cdot)$ comes entirely from information leakage through her peers' data sharing on the platform due to data correlation. In contrast, for an agent who joins the platform but does not report her data, not only her peers' actions but also her own interactions with the platform result in her privacy cost, which is denoted by $h(\cdot)$. As joining the platform means more privacy leakage, we assume $g(\cdot) > h(\cdot)$. Moreover, recall that we consider I groups of agents parameterized by data correlation. We model both privacy cost functions of group i , i.e., $g_i(\cdot)$ and $h_i(\cdot)$ with index i , that depend on intra-group cost (related to α_i) and inter-group cost (related to α_{-i}).

2.2 Problem Statement

In this section, we introduce the analyst optimization problem. The analyst needs to design the payment function and selection probability for each group of agents, subject to two desirable economic constraints: individuals' truthfulness and expected budget feasibility.

DEFINITION 1 (TRUTHFULNESS). *A mechanism is truthful if for every participant with cost c , she can maximize her expected utility if she truthfully reports her cost, i.e.,*

$$u_i(c|c) \geq u_i(\tilde{c}|c), \quad \forall \tilde{c} \neq c, \forall i. \quad (1)$$

Definition 1 guarantees that rational agents on the platform will truthfully report their costs.

DEFINITION 2 (EXPECTED BUDGET CONSTRAINT). *A mechanism with payment function P and selection probability A satisfies the expected budget constraint B if and only if*

$$\sum_{k:i=i(k)} \mathbb{E}_{c \sim f_i} [P_i(c) A_i(c) \cdot \mathbb{1}(k \in \mathcal{N}_i)] \leq B. \quad (2)$$

Definition 2 limits the expected payments of the analyst to the agents for their data. In (2), the index i (of $P_i(c)$ and $A_i(c)$) denotes the association with agents of group i , $1 \leq i \leq I$. And $i(k)$ denotes the group index of agent k .

The objective of the analyst is to optimize a trade-off between bias $\mathbb{B}(\hat{\mu}; \mathcal{D}, A)$ and variance $\mathbb{V}(\hat{\mu}; \mathcal{D}, A)$ of the estimator. Specifically, the analyst wishes to learn an underlying parameter μ of the whole population, and he obtains an estimator $\hat{\mu}$ based on participants' data. The estimator $\hat{\mu}$ is viewed as a random variable, and the randomness comes from the allocation rule A and the joint distribution $\mathcal{D}$ of agents' data and cost.

Since the analyst does not know the joint distribution $\mathcal{D}$, he cannot directly optimize over bias and variance of the estimator. Instead, we consider the goal of minimizing the worst-case linear combination (with parameter γ) of bias and variance over all instantiations of $\mathcal{D}$ that are consistent with the marginal cost distribution f :

DEFINITION 3 (MECHANISM DESIGN TASK). *The analyst aims to minimize a worst-case linear combination of bias and variance by designing payment function P and allocation rule A subject to truthfulness and budgetary constraints:*

$$\inf_{A, P \text{ consistent with } \mathcal{D}} \sup_{\gamma \cdot \mathbb{V}(\hat{\mu}; \mathcal{D}, A) + (1 - \gamma)\mathbb{B}(\hat{\mu}; \mathcal{D}, A)} \quad \text{s.t. Truthfulness constraint in (1),}$$

$$\text{Expected budget constraint in (2).}$$

2.3 Equilibrium Participation Rate

We define participation rate profile θ as a vector consisting of participation rates for all I groups of agents, i.e., $\theta \triangleq [\theta_i]_{1 \leq i \leq I}$. Notice that agents' utilities depend on participation rate profile, which affect their privacy costs.

Next, we introduce an agent's decision on whether or not to participate in the platform, for a given participation rate profile. For an agent with cost c in group i , we use a binary variable $d_i(c|\theta)$ to denote her decision, where $d_i(c|\theta) = 1$ means participation and $d_i(c|\theta) = 0$ means non-participation. She decides to participate in the platform only if her utility of non-participation is not lower than that of participation:

$$d_i(c|\theta) = \begin{cases} 1, & \text{if } \max_{\tilde{c}} \bar{u}_i(\tilde{c}|c) \geq -g(c, \theta_i; \alpha_i), \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

We now define the equilibrium concept formally. This notion of equilibrium guarantees that agents' decisions are consistent with the participation rate profile.

DEFINITION 4 (EQUILIBRIUM). *A participation rate profile $\theta^* = [\theta_i^*]_{1 \leq i \leq I}$ is an equilibrium, if for each group i , $1 \leq i \leq I$, the fraction of participating agents is exactly θ_i^* , i.e.,*

$$\int_c \mathbb{1}[d_i(c|\theta^*) = 1] \cdot f_i(c) dc = \bar{\theta}_i^*, \quad \forall i. \quad (4)$$

3. RESULTS

Our main results characterize the payment rule and the optimal allocation rule. In the full version of this paper, we provide additional results such as characterizations of properties of the mechanism and the participation ratio.

Payment Rule. To solve the analyst's mechanism design problem in Definition 3, we first analyze the structure of the payment function in the mechanism associated with the equilibrium θ^* . We assume that for each group $i \in [I]$, the privacy cost $h_i(c; \theta^*)$ given participation rate profile θ^* is linear in cost c with parameter $b_i(\theta^*)$. Theorem 1 shows the payment function given monotonic allocation rule A that induces agents' truthfulness reporting of cost. The term τ in (5) is a constant given a fixed θ^* . Due to space limitations, we do not present its detailed definition.

THEOREM 1. *The mechanism is truthful (strictly truthful, respectively) and induces participation profile θ^* at equilibrium if and only if for every group $i \in [I]$, both of the following statements hold:*

- (i) allocation rule $A_i(\tilde{c})$ is a non-increasing (decreasing, respectively) function of the reported cost $\tilde{c}$;
- (ii) payment function is given as in Equation (5) (as a function of A), where θ^* is the desired participation rate profile.

$$P_i(\tilde{c}) = \frac{1}{A_i(\tilde{c})} \left((1 - b_i(\theta^*)) \int_{\tilde{c}}^{c_{\max}} A_i(z) dz + \tau \right) + \tilde{c} - h_i(\tilde{c}; \theta^*). \quad (5)$$

Optimal Allocation Rule. After deriving the payment that induces agents' truthfulness in Section 3, it remains to characterize the allocation rule that optimizes the worst-case bias-variance trade-off, subject to budget constraint. The optimal allocation rule of each group is non-monotonic in the virtual cost.

DEFINITION 5 (VIRTUAL COST). *In group i , the virtual cost of an agent with cost c is*

$$\phi_i(c; \theta^*) = c - h_i(c; \theta^*) + (1 - b_i(\theta^*)) \frac{F_i(c)}{f_i(c)}.$$

F_i and f_i are the CDF and PDF of cost in group i , respectively.

THEOREM 2. *Assume the virtual cost $\phi_i(c; \theta^*)$ in group i , $i \in [I]$ is non-decreasing. Given a desired participation rate profile θ^* , the optimal allocation rule of group i is*

$$A_i(c) = \begin{cases} \chi, & \text{if } \phi_i(c; \theta^*) \leq \hat{\phi}, \\ \frac{\eta}{\sqrt{\phi_i(c; \theta^*)}}, & \text{if } \hat{\phi} < \phi_i(c; \theta^*) \leq \phi_i(\hat{c}_i; \theta^*). \end{cases}$$

The characterizations of the constants η , χ , and $\hat{\phi}$ are omitted due to space limit.

The optimal allocation rule can have two possible structures: *Fixed then Decreasing* (FtD) or *Strictly Decreasing* (SD), depending on system parameters such as budget B . The FtD structure has an allocation rule that is firstly fixed in low-cost region and then strictly decreasing in the high-cost region (inversely proportional to the square root of the virtual cost). The other structure, SD, has an allocation rule that is strictly decreasing (inversely proportional to the square root of the virtual cost) in the whole region.

4. REFERENCES

- [1] D. Acemoglu, A. Makhdoumi, A. Malekian, and A. Ozdaglar. Too much data: Prices and inefficiencies in data markets. Technical report, National Bureau of Economic Research, 2019.
- [2] Y. Chen, N. Immorlica, B. Lucier, V. Syrgkanis, and J. Ziani. Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 27–44. ACM, 2018.
- [3] A. Roth and G. Schoenebeck. Conducting truthful surveys, cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 826–843, 2012.