2015 IEEE 28th Computer Security Foundations Symposium

A Game-Theoretic Study on Non-Monetary Incentives in Data Analytics Projects with Privacy Implications

Michela Chessa
EURECOM
Sophia Antipolis, France
michela.chessa@eurecom.fr

Jens Grossklags
The Pennsylvania State University
University Park, PA, USA
jensg@ist.psu.edu

Patrick Loiseau
EURECOM
Sophia Antipolis, France
patrick.loiseau@eurecom.fr

***Abstract*—The amount of personal information contributed by individuals to digital repositories such as social network sites has grown substantially. The existence of this data offers unprecedented opportunities for data analytics research in various domains of societal importance including medicine and public policy. The results of these analyses can be considered a public good which benefits data contributors as well as individuals who are not making their data available. At the same time, the release of personal information carries perceived and actual privacy risks to the contributors. Our research addresses this problem area.**

In our work, we study a game-theoretic model in which individuals take control over participation in data analytics projects in two ways: 1) individuals can contribute data at a self-chosen level of precision, and 2) individuals can decide whether they want to contribute at all (or not). From the analyst's perspective, we investigate to which degree the research analyst has flexibility to set requirements for data precision, so that individuals are still willing to contribute to the project, and the quality of the estimation improves.

We study this tradeoff scenario for populations of homogeneous and heterogeneous individuals, and determine Nash equilibria that reflect the optimal level of participation and precision of contributions. We further prove that the analyst can substantially increase the accuracy of the analysis by imposing a lower bound on the precision of the data that users can reveal.

***Keywords*—Non-cooperative game, public good, privacy, population estimate, data analytics, non-monetary incentives**

I. Introduction

A. Background

The seminal "How much Information Project?" report published in 2000 concluded that between 1 and 2 exabytes of unique information were produced worldwide per year which translated into about 250 megabytes of information for every human being [1], [2]. While those figures were (and are still) largely driven by commercial production of information, in recent years the amount of personal information produced by individuals has grown substantially. Now, Facebook alone absorbs about 220 petabytes of user-contributed data each year [3]. Recognizing the opportunities to economically benefit from this growth, personal data has been heralded as the "New Oil" of the 21st Century [4]. Similarly, opportunities are increasingly taken advantage of to utilize the data for research. From the individual's perspective the latter trend results in a tradeoff calculus.

On the one hand, individuals recognize that many complex challenges with societal importance, such as public health considerations, market-research or political decision-making [5], may benefit from a more rigorous analytic treatment, thanks to data analytics research and the newly-won abundance of personal information. From this perspective, many analytic results that are based on individuals' personal data can be interpreted as *public goods* with societal importance. For example, advancements to better understand certain illnesses do not only potentially benefit the contributors of personal data, but are often made accessible to people in a particular domain (e.g., citizen of a country, individuals in a certain social status or demographic category, or everybody).

On the other hand, the same individuals have justified *privacy concerns* about the release of their personal data. The reasons for privacy concerns can be quite diverse as outlined in Solove's privacy taxonomy [6]. Individuals may perceive the release and use of their data as an intrusion of their personal sphere [7], [8], or as a violation of their dignity [9]. In addition, they may fear this data can be abused for unsolicited advertisements, or social and economic discrimination (e.g., [10], [11]).

The published studies demonstrate the need to organize the collection of personal data when facing this users' tradeoff scenario, by implementing effective control and participation mechanisms. It has been shown that a majority of individuals consider it as important to be able to exercise *control* over the release of their personal data [12]. For example, a number of empirical studies have provided evidence for such desires for control in the medical domain [13]–[15]. Moreover, even if data privacy provisions are met, many respondents would still require notice and consent over their medical data release [14]–[16]. Finally, several studies show a high overall concern for certain data releases. For example, a meta-review of published surveys showed that in some contexts a majority of respondents were entirely uncomfortable with health research if effective notice and consent practices were absent [17]. Similar findings can be shown for other problem domains.

DOI 10.1109/CSF.2015.14

B. Problem Statement and Approach

Our research addresses the problem area identified in the above section. In this paper, we propose individuals' incentives to participate in data analysis projects. These individuals face a tradeoff between having privacy cost associated with their data release, but also deriving benefits from the analysis' results.

We are particularly motivated by the scenario when data about individuals is already stored in a secure database for a different primary purpose (e.g., social networking or medical services). An analyst can then request the participation of individuals in a data analysis project (via a notice and consent process with negligible cost) that provides a public good. More precisely, individuals make decisions about the release of a private value given a population-relevant metric. The analyst has the objective of accurately estimating the associated population average for all individuals.

Our main focus is on understanding the incentives of individuals to participate, and of the analyst to shape this decision-making process. From each individual's perspective, control over participation takes two forms: 1) individuals can contribute data at a self-chosen level of precision, and 2) individuals can decide whether they want to contribute at all (or not). From the analyst's perspective, we investigate to which degree the research analyst has flexibility to set requirements for data precision, so that individuals are still willing to contribute to the project, and the quality of the estimation improves.

Our work assumes that incentives for participation are non-monetary; that is, the main driver for data contributions is the interest in the derived public good. We base this assumption on the observation that direct monetary compensation for personal information has so-far received very little traction in the market for personal information, and that it meets little acceptance in consumer surveys.[1]

We follow a game-theoretic approach to investigate the outlined trade-off calculus. We iteratively develop a model, where the starting point is a simplified version of the work by [18], that captures the interaction between an analyst and a set of individuals who have control over the release of information to the analyst. We conduct a rigorous analysis and derive concrete results about the precision of contributions, the quality of the population estimate, and the overall willingness to contribute to the project.

C. Contributions

In this paper, we consider critical facets of realistic privacy decision-making, striking a good balance between model complexity and potential impact. We rigorously analyze a general model where users optimize a cost composed of an individual privacy cost and an estimation cost that captures the public good component of the analyst's estimation, both given by arbitrary functions satisfying relatively mild assumptions. In particular, we consider a general case with a continuous privacy cost function which allows users to choose a privacy level in a continuum of choices (and not simply a 0-1 choice). We first analyze the homogeneous agents case, and then we extend our results to the case of heterogeneous agents, providing in detail the actions the analyst should take in order to improve the estimation. Evidence that privacy concerns are heterogeneous is a particularly central cornerstone of the privacy literature [19], and such an extension is fundamental for the applicability of the model.

For both the homogeneous and the heterogeneous case, we determine Nash equilibria indicating the number of contributors and the optimal contribution levels by the individuals. We further prove that the analyst can increase the population estimate's accuracy simply by imposing a lower bound on the precision of the data that users can reveal (i.e., by restricting the level of precision of data contributions). While, for a fixed population of users providing data, increasing the precision of each data point clearly improves the population estimate's precision, the surprising and important aspect of our result lies in that the scheme remains incentive compatible, i.e., users are still willing to provide data with a higher precision rather than dropping out. We also show how to tune the minimum precision level the analyst should set in order to optimize the population estimate's accuracy. In our numerical simulations, we find a maximum improvement of the population estimate's accuracy in the order of $20 - 40\%$.

We further provide extensions of our modeling framework. First, we discuss a two-stage game in which the analyst may first recruit participants that commit to provide private data with a minimum precision; and only in a second stage, these agents would be asked to disclose their information. This captures scenarios in which agents are recruited for specific studies. Second, we also address the issue of costly acquisition of agents and their data for analysis purposes. While the no-cost-per-agent assumption we make throughout the remainder of the paper is a standard approach in most of the literature on public goods, we believe that certain practical scenarios require the appreciation of cost considerations, and this extension further completes our framework.

Our results provide a widely applicable method to increase the provision of a public good above voluntary contributions, simply by restricting the agents' strategy spaces. This method is attractive by its simplicity compared for instance to other schemes that involve monetary transfers; and could find utilization in other public good contexts.

Understanding the trade-off between privacy, the quality of data analysis results, and willingness-to-participate in such projects is of current and growing importance. Analysts should not rely on overly broad or ineffective (take-it-or-leave-it) notice and consent procedures that do not accurately reflect individuals' preferences. In many privacy-sensitive scenarios such as involving medical data it is particularly unethical to deprive individuals of their opportunities to make decisions about their data, and whether they want to be involved in certain analysis projects. However, better insights about the involved incentive structures are needed to guide public policy and advancements of privacy-aware data analysis.

Preliminary versions of some of the results presented in this paper appeared in our short paper [20], in the context of a simplified model with monomial privacy cost, linear estimation

[1] While related empirical data is sparse, a survey reported that only about 25% of the surveyed population would accept monetary compensation for personal information [12]. In contrast, offering discounts or free products/services for personal information is a common practice.

cost and homogeneous agents. Here, we provide results for the general framework introduced above that relaxes such assumptions, we provide detailed results of practical importance on how the analyst should optimally selected the minimum precision level, and we provide several further extensions. In Section IV-C, we also provide more detailed results in the simplified setting of [20], to qualitatively illustrate the results of the present paper.

D. Roadmap

Our paper is structured as follows. In Section II, we review related work. We develop and describe our model in Section III. We conduct our analysis in detail in Section IV on a canonical case of homogeneous agents. We extend the results to heterogeneous agents in Section V. We discuss extensions to our model in Section VI, and conclude in Section VII. Due to space constraints, the proofs of our results are included in the companion technical report [21].

II. Related Work

Our model draws on different lines of research including work on privacy in the context of data analytics, and game theoretic and public goods models. We also briefly review technical and cryptographic approaches, and behavioral research on control and data sharing.

Research on the optimal design of experiments assumes that already the stage of data collection can be influenced by the analyst in order to improve the learning of a linear model [22], [23]. In this paper, we allow the analyst to require data contributions at a certain level of precision to improve the computation of a population estimate, which is a related concept. Optimal design of experiments has been studied from the perspective of incentives [24], or with the scope of obtaining an unbiased estimator [25]. We propose to improve the design of experiments focusing on the privacy concerns of the agents.

Privacy-preserving techniques in the context of data analytics have a long history. Some recent papers propose new approaches, which allow users to protect their privacy selling aggregates of their data [26], [27]. The more classical framework of ϵ-differential privacy [28], [29], assumes that data are perturbed after an analysis has been conducted on unmodified inputs. That is, the analyst is considered trustworthy. In this framework, researchers have also studied the role of incentives [30]–[33]. Our work differs, as we assume agents to be releasing their data independently, and an untrusted data analyst which motivates perturbations of data before submission. The idea of affecting the level of precision of released personal data, adding noise in advance of data analysis has been studied in the context of privacy-preserving data-mining (see, e.g., [34], [35]) and specific application scenarios such as building decision trees [36], clustering [37], and association rule mining [38]. More recently, bounds have been derived on generic information-theoretic quantities and statistical estimation rates under a *local privacy* model which preserves the privacy of agents even from the learner (similarly to adding noise before revealing data) [39].

Recent work has also studied the combinatorial optimization problem when an analyst may buy unbiased samples of data from different providers with given but potentially heterogeneous variance-price combinations [40]. In another recent working paper, analysts can access unbiased samples of private data by compensating data subjects for their data release according to their preferences [41]. Those studies are complementary to our work in which data subjects individually decide in a game-theoretic framework on the degree of data accuracy given a trade-off between their privacy and the determination of a socially valuable population estimate.

From a mechanism design perspective, scenarios have been studied where survey subjects are assumed to potentially misreport their private values [42], [43], however, these behaviors are not studied in the context of a non-cooperative scenario. A mechanism design perspective is taken in [44] where the authors introduce monetary payments to create incentives for agents to give high quality data. Here, we do not consider monetary payments. A strategic approach is followed in [18], where an analyst performs a linear regression based on users' perturbed data. The authors in [18] treat the estimation accuracy as a public good and study the equilibrium accuracy achieved without introducing monetary payments and the resulting price of anarchy. Our starting point is a simplified version of the model in [18]. We continue this line of research by studying the benefits of restricting potential perturbation on the population estimate accuracy, and the incentives for participation in a game-theoretic framework.

Our research is also relevant to the context of the provisioning of public goods [45]. Our results show a new way of increasing the public good provision by restricting the agents' possible actions, as opposed to using monetary incentives. In addition, studies on interdependent privacy which capture the idea that data sharing by one agent impacts the privacy of other connected agents is complementary to our work [46], [47]. We model the scenario when sharing creates privacy risks for individuals, but positive benefits for all agents.

The aforementioned theoretical works are complemented by technical approaches (which do not utilize insights from game-theory) such as secure hardware-based private information retrieval which can be applied, for example, in the context of online behavioral advertisement [48]; see also other approaches for privacy-preserving online targeted advertisements [49]. Similarly, multi-party secure computation has been used to facilitate the fitting of logistic regression when data are held by separate parties [50], and homomorphic encryption has been applied to the scenario of linear regression [51]. Secure-computation notions of privacy have also been used in combination with game theory for privacy-preserving mechanism design [52], [53].

To facilitate the privacy negotiation process between a data subject and an analyst, different technical protocols have been proposed. Several works are connected to the Platform for Privacy Preferences Project (P3P) which offers a protocol allowing data collectors (e.g., websites) to declare their intended use of information they collect about data subjects [54], and also provides agent tools for the user to manage those data requests [55], [56]. More recent work, for example, addresses specific problem areas such as personalization [57]. Those mechanisms allow for user-specified policies regarding participation, but also minimum requirements for (not necessarily truthful) data sharing as specified by the analyst.

Research on user preferences and behaviors with respect to privacy has produced several results relevant to the context of our work. A survey study has shown that over 90% of the respondents agreed with the definition of privacy as control of personal information [12] which presumably would include an interest to decide over the participation in data analysis projects. In hypothetical scenarios, individuals typically report high attitudinal valuations for their private data [58]. However, in experiments with actual private data transfers researchers have observed low thresholds for the release of such data in exchange for free services/goods or discounts [19], [59], [60]. A root cause for this privacy dichotomy is the complexity of understanding personal information exchanges and their consequences [12].

The intricacies of human decision making have also been studied specifically focusing on the notion of control over information exchanges. Laboratory and online experiments have shown that control options have to be added with care to practically relevant scenarios [61]–[63]. For example, such options can elevate individuals' propensity to engage in riskier disclosures because their mere presence can contribute to a lowering of concerns over privacy [61]. Another experimental study found that allowing individuals to customize personal data exchanges does not increase the number of transactions even though individuals were able to exclude unwanted aspects of those transactions [64]. Overall the understanding of the involved attitudes and behaviors is still work in progress. In our paper, we propose a process that is relatively straightforward to implement and to understand from a user perspective. However, approaches that fully accommodate the stated behavioral hurdles remain the subject of future work for behavioral as well as theoretical scientists.

III. The Model

In this section, we present our model in detail. We describe the strategic interaction between the individuals (which we also refer to as agents), whose information is contained in a data repository, and how the analyst, wishing to observe the data and to perform a statistical analysis, may modify the estimation by varying selected parameters. The linear model approach we take here builds on the work of [18].

A. The Data Repository of Personal Data

Let $N = \{1, \dots, n\}$ denote the set of agents, whose personal data are contained in the data repository. In particular, we suppose that each agent $i \in N$ is associated with a *private variable* $y_i \in \mathbb{R}$, which contains sensitive information. Throughout our analysis, we suppose that there exists $y_M \in \mathbb{R}$, s.t., the private variables are of the form

$$y_i = y_M + \epsilon_i, \quad \forall i \in N, \tag{1}$$

where ϵ_i are i.i.d., zero-mean random variables with finite variance $\sigma^2 < \infty$, which capture the inherent noise. We stress that we make no further assumptions on the noise; in particular, we do not assume it is Gaussian. As a result, our model applies to a wide range of statistical inference problems, even cases where the distribution of variables is not known.

Paramter y_M represents the *mean* of the private variables y_i, and its knowledge is valuable to the analyst, for example as it allows him to predict the private variable of any agent whose data cannot be known (because it is not contained in the repository at that given moment, kept private by its owner, is not accessible due to limited computing resources, etc.). The analyst wishes to observe the available private variables y_i and to compute their average as an estimation of y_M. In our model, we suppose that the analyst does not know the mean y_M, that he wishes to estimate, but he knows the variance σ^2. We argue that observing the variability of an attribute in a population is easier than estimating the mean, both for the analyst and for the population (in [65], for example, the authors show how individuals value their age and weight information according to the relative variability).

B. The Precision and the Analyst's Estimation

We suppose that the analyst cannot directly access the private variables, rather she needs to ask the agents for their consent to be able to retrieve the information. As such, the agents have full control over their own private variables, and they have the choice to authorize or to deny the analyst's request. In particular, if wishing to contribute, but concerned about privacy, an agent can authorize the access to a perturbed value of the private variable. The *perturbed variable* has the form $\tilde{y}_i = y_i + z_i$, where z_i is a zero-mean random variable with variance σ_i^2. We assume that the $\{z_i\}_{i \in N}$ are independent and are also independent of the inherent noise variables $\{\epsilon_i\}_{i \in N}$. In practice, the agent chooses a given *precision* λ_i which corresponds to the inverse of the aggregate variance (inherent noise, plus artificially added noise) of the perturbed variable $\tilde{y}_i$, i.e.,

$$\lambda_i = 1/(\sigma^2 + \sigma_i^2) \in [0, 1/\sigma^2], \quad \forall i \in N.$$

In the choice of the precision level, we have the following two extreme cases:

- *(i)* when $\lambda_i = 0$, agent i has very high privacy concerns. This corresponds to adding noise of infinite variance or, equivalently, this represents the fact that agent i denies the access to her data;
- *(ii)* when $\lambda_i = 1/\sigma^2$, agent i has very low privacy concerns. This corresponds to authorizing the access to the real private variable y_i, without adding any additional noise to the data.

The strategy set $[0, 1/\sigma^2]$ contains all the possible choices for agent i: *denying*, *authorizing*, or any *intermediate level* of precision (which captures a wide range of privacy concerns as documented in behavioral studies [19]). We denote by $\boldsymbol{\lambda} = [\lambda_i]_{i \in N}$ the vector of the precisions.

Once each agent $i \in N$ has made her choice about the level of precision λ_i and, consequently, the perturbed variable $\tilde{y}_i$ has been computed, the analyst has access to both the set of precisions and the set of perturbed variables. Then, the analyst estimates the mean as

$$\hat{y}_M(\boldsymbol{\lambda}) = \frac{\sum_{i \in N} \lambda_i \tilde{y}_i}{\sum_{i \in N} \lambda_i}, \tag{2}$$

where perturbed variables with higher precision (i.e., smaller variance) receive a larger weight. This estimator is the standard *generalized least squares estimator*. It minimizes a weighted

square error in which the i-th term is weighted by the precision of the perturbed variable $\tilde{y}_i$. This estimator is unbiased, i.e., $\mathbb{E}[\hat{y}_M] = y_M$, and has variance

$$\sigma_M^2(\boldsymbol{\lambda}) = \mathbb{E}[(\hat{y}_M(\boldsymbol{\lambda}) - y_M)^2] = \frac{1}{\sum_{i\in N}\lambda_i} \in [\sigma^2/n, +\infty]. \tag{3}$$

In our model, the analyst aims at estimating the mean y_M, e.g., to be able to predict some additional private variables. Then, it is reasonable to assume that the analyst would use this estimator, as it is "good" for several reasons. In particular, it coincides with the *maximum-likelihood estimator* for Gaussian noise and, most importantly, it has minimal variance amongst the linear unbiased estimators for arbitrary noise distributions.

In the estimation, we have the following two extreme cases:

(i) when $\lambda_i = 0$ for each $i \in N$, the variance (3) is infinite. This corresponds to the situation in which each agent denies the access to her data, and then the analyst cannot estimate y_M;

(ii) when $\lambda_i = 1/\sigma^2$ for each $i \in N$, the analyst estimates y_M with variance σ^2/n, resulting only from the inherent noise. This corresponds to the situation in which each agent is authorizing the access to her data with maximum precision, i.e., no agent is perturbing her private variable.

For any level of precision in $[0, 1/\sigma^2]^n$, the estimated variance will be in $[\sigma^2/n, +\infty]$. The set of precision vectors for which the estimator has a finite variance is $[0, 1/\sigma^2]^n \setminus \{(0, \dots, 0)\}$.

C. The Estimation Game Γ

We next describe the interaction between the agents that results in their choices of precisions. We assume that each agent $i \in N$ wishes to minimize a cost function $J_i : [0, 1/\sigma^2]^n \to \bar{\mathbb{R}}_+$, s.t., for each $\boldsymbol{\lambda} \in [0, 1/\sigma^2]^n$,

$$J_i(\lambda, \boldsymbol{\lambda}_{-i}) = c_i(\lambda_i) + f(\boldsymbol{\lambda}), \tag{4}$$

where we use the standard notation $\boldsymbol{\lambda}_{-i}$ to denote the collection of actions of all agents but i. The cost function J_i of agent $i \in N$ comprises two non-negative components. The first component $c_i : [0, 1/\sigma^2] \to \mathbb{R}_+$ represents the privacy attitude of agent i, and we refer to it as the *privacy cost*: it is the (perceived or actual) cost that the individual incurs on account of the privacy violation sustained by revealing the private variable perturbed with a given precision. The second component $f : [0, 1/\sigma^2]^n \to \bar{\mathbb{R}}_+$ is the *estimation cost*, and we assume that it takes the form $f(\boldsymbol{\lambda}) = F(\sigma_M^2(\boldsymbol{\lambda}))$ where $F : [\sigma^2/n, +\infty) \to \mathbb{R}_+$ if the variance is finite, and $+\infty$ otherwise. It represents how well the analyst can estimate the mean y_M and it captures the idea that it is not only in the interest of the analyst, but also of the agents, that the analyst can determine an accurate estimate of the population average y_M.

In our model, the accuracy of the estimate can be understood as a public good, to which each user contributes with her choice of precision λ_i, at a given privacy cost. From this perspective, the assumption that the estimation cost is the same for all agents mirrors the usual standard assumption in the public good literature. Throughout our analysis, we make two additional assumptions:

Assumption 1: The privacy costs $c_i : [0, 1/\sigma^2] \to \mathbb{R}_+$, $i \in N$, are twice continuously differentiable, non-negative, non-decreasing, strictly convex and s.t. $c_i(0) = c_i'(0) = 0$.

Assumption 2: Function $F : [\sigma^2/n, +\infty) \to \mathbb{R}_+$ is twice continuously differentiable, non-negative, non-decreasing and strictly convex.

To describe the strategic interaction between the agents, we define the *estimation game* $\Gamma = \langle N, [0, 1/\sigma^2]^n, (J_i)_{i\in N}\rangle$ with set of agents N, strategy space $[0, 1/\sigma^2]$ for each agent $i \in N$ and cost function J_i given by (4).

D. The Modified Estimation Game $\Gamma(S, \eta)$

As we shall see (Section IV-A), game Γ has a unique Nash equilibrium for which the variance of the estimation is larger than the optimal one (σ^2/n) due to the excess noise added by agents to protect their privacy. We further investigate the situation in which the analyst can modify the game and try to mitigate the effect of agents' privacy concerns in order to reduce the estimation cost (i.e., to improve the accuracy of the estimation obtained). Specifically, the analyst can implement the following two variations of the model. First, she can choose a *minimum precision level* $\eta \in [0, 1/\sigma^2]$, which is equivalent to fixing a maximum variance for the noise that agents can add to perturb their data. As it is not practically possible to force agents to authorize the access to their data with a given precision, we still assume that the agents can choose to deny the authorization, which is equivalent to selecting a precision level equal to zero. Second, the analyst can request the access to the personal data to only a subset $S \subseteq N$ of agents, with $s = |S|$ (for example, excluding those agents who are the most concerned about privacy).

In the modified game, the agents are informed of the subset of individuals who are asked to reveal their personal data, and of the minimum precision level η. They choose their precision λ_i in the range imposed by the analyst $[\eta, 1/\sigma^2]$ or decide to deny the access, i.e., select their precision equal to 0. To analyze the strategic interaction between the agents in this variation, we define the game $\Gamma(S, \eta) = \langle S, [\{0\} \cup [\eta, 1/\sigma^2]]^s, (J_i)_{i\in S}\rangle$ (where the cost function J_i is still given by (4)), which is identical to Γ, except for the restricted set of agents and the restricted strategy space.

Observe that the original game Γ is a special case of this modified game $\Gamma(S, \eta)$, when $S = N$ and $\eta = 0$. We analyze the games Γ and $\Gamma(S, \eta)$ as *complete information games* between the agents, i.e., we assume that the set of agents, the action sets (in particular, when present, the value of the parameter η) and the costs are known by all the agents.

IV. The Homogeneous Agent Case

In this section, we detail the analysis in the symmetric case where all the agents have identical privacy concerns, i.e., we assume that the privacy cost functions of all agents are the same: $c_i(\cdot) = c(\cdot)$ for each $i \in N$. This special case highlights the key aspects of our approach and provides some interesting preliminary results that yield intuitive interpretations. We will generalize our results to the heterogeneous case in Section V.

A. The Estimation Game in the Homogeneous Case

We first analyze the estimation game Γ, in which all the agents in N are playing and the analyst allows them to choose any precision level between 0 and $1/\sigma^2$. A *Nash equilibrium* (in pure strategy) of this game is a strategy profile $\boldsymbol{\lambda}^* \in [0, 1/\sigma^2]^n$ satisfying

$$\lambda_i^* \in \underset{\lambda_i \in [0,1/\sigma^2]}{\arg\min}\ J_i(\lambda_i, \boldsymbol{\lambda}_{-i}^*), \quad \forall i \in N. \tag{5}$$

The game Γ with strategy space $[0, 1/\sigma^2]$ is a special case of the game in [18], where the existence of a unique Nash equilibrium is established. However, our specific assumptions allow us to characterize the equilibrium in more detail:

Theorem 1: The game Γ has a unique Nash equilibrium $\boldsymbol{\lambda}^*$ s.t. $\lambda_i^* = \lambda^* > 0$ for each $i \in N$.

The proof of this result exploits the fact that game Γ is a potential game to characterize the Nash equilibrium. Interestingly, we observe that non-participation by everybody, i.e., $\boldsymbol{\lambda} = (0, \dots, 0)$, cannot be an equilibrium. Indeed, as the estimation cost diverges at $\boldsymbol{\lambda} = (0, \dots, 0)$, every agent has a profitable deviation from this point since contributing any positive λ_i brings the estimation cost down to a finite cost. Note, however, that this is not an artifact of the model, as it remains true if we assume that the estimation cost is bounded but large enough to exceed the privacy cost.

We observe that, as a consequence of the symmetry of the game in the homogeneous case, all the agents at equilibrium choose the same precision level, which is a function $\lambda^* = \lambda^*(n)$ of the total number of agents n. Then, from the discussion above, it is clear that λ^* cannot be zero, so that all agents contribute a positive precision.

Due to the arbitrariness of the functions $F(\cdot)$ and $c(\cdot)$, the unique Nash equilibrium cannot be written in closed form. However, it is easily computable in practice either as the minimum of the potential function (which is convex) or as the unique solution of the following fixed point problem:

$$\lambda = g(n, \lambda),$$

where function $g : \mathbb{N}^* \times [0, 1/\sigma^2] \to [0, +\infty]$ is defined for each $\lambda \in (0, 1/\sigma^2]$ and for each $n \in \mathbb{N}^*$ as

$$g(n, \lambda) = \min\left\{ \sqrt{F'\left(\frac{1}{n\lambda}\right)\frac{1}{n^2 c'(\lambda)}}, 1/\sigma^2 \right\}$$

and is defined by continuity as $\lim_{\lambda \to 0^+} g(n, \lambda)$ for $\lambda = 0$ and for each $n \in \mathbb{N}^*$.

Given the unique Nash equilibrium $\boldsymbol{\lambda}^*(n)$, the variance (in Equation (3)) of the estimate of y_M obtained by the analyst at equilibrium is also a function of n, and given by the following expression:

$$\sigma_M^2(\boldsymbol{\lambda}^*(n)) = \frac{1}{n\lambda^*(n)}. \tag{6}$$

In Propositions 1 and 2 below, we derive the properties of the equilibrium precision and of the corresponding variance, when the number of agents varies.

Proposition 1: The equilibrium precision level $\lambda^*(n)$ satisfies:

- *(i)* $\lambda^*(n)$ is a non-increasing function of the number n of agents, and
- *(ii)* $\lim_{n \to +\infty} \lambda^*(n) = 0$.

Proposition 1 states that the equilibrium contribution of each agent decreases as the number of agents increases (Part *(i)*). This is a standard property in public good problems as agents choose their equilibrium contribution such that the marginal increase in the contribution cost equates the marginal decrease in the estimation cost, and the marginal effect of a single agent decreases when the number of agent increases. Proposition 1-*(ii)* shows that, in the limit when n becomes very large, the contribution of each agents tends to zero (i.e., each agent adds a variance tending to infinity to her data). It is interesting to notice that, given that the equilibrium prevision level $\lambda^*(n)$ goes to zero as n goes to infinity, the variance (6) cannot decrease in $1/n$ as in the standard case of the empirical mean of iid random variables of equal variance. This is because, here, the variance of each data point (or random variable) increases as the number of points increases. Yet, as the next proposition shows, the variance of the mean's estimate is still non-increasing.

Proposition 2: The equilibrium variance of the estimate of y_M satisfies:

- *(i)* $\sigma_M^2(\boldsymbol{\lambda}^*(n))$ is a non-increasing function of the number of agents n, and
- *(ii)* $\lim_{n \to +\infty} \sigma_M^2(\boldsymbol{\lambda}^*(n)) = 0$.

Proposition 2-*(i)* shows that, for the analyst, it is always better to have a larger number of agents giving data despite the fact that, when the number of agents increases, each agent gives data with smaller precision (see Proposition 1). Proposition 2-*(ii)* analyzes the case of a large number of agents n. Interestingly, when n gets large, the variance goes to zero, though at a rate smaller than $1/n$ as mentioned above. (We give an expression of the rate in Section IV-C for special functions F and c).

B. The Modified Estimation in the Homogeneous Case

We now move to the case where the analyst can restrict the set of agents, thereby asking to access the data of only a subgroup of them, and potentially introducing a minimum precision level $\eta \in [0, 1/\sigma^2]$. The final goal is to improve the estimation accuracy; formally, to estimate the mean y_M with a variance strictly smaller than $\sigma_M^2(\boldsymbol{\lambda}^*(n))$. We assume that the set $S \subseteq N$ of agents who can authorize access to their data (i.e., who are solicited by the analyst) is fixed, and we analyze how the estimation varies while moving only the parameter η. This variant is modeled by the game $\Gamma(S, \eta)$ defined in Section III-D, where η is now the only variable of the model. We suppose that the equilibrium precision level for the game $\Gamma(S, 0)$ is s.t. $\lambda^*(s) \neq 1/\sigma^2$ since, otherwise, the estimation would already be optimal with variance σ^2/s for $\eta = 0$.

A Nash equilibrium (in pure strategy) of the game $\Gamma(S, \eta)$ is a strategy profile $\boldsymbol{\lambda}^* \in \left[\{0\} \cup [\eta, 1/\sigma^2]\right]^s$ satisfying

$$\lambda_i^* \in \underset{\lambda_i \in \{0\} \cup [\eta, 1/\sigma^2]}{\arg\min}\ J_i(\lambda_i, \boldsymbol{\lambda}_{-i}^*), \quad \forall i \in S. \tag{7}$$

In the following theorem, we show that, if the analyst chooses a minimum precision level that is not "too big", the agents are still wishing to authorize access to their data at equilibrium. Recall that $S \subseteq N$ denotes the set of agents solicited by the analyst (who are the players of the game $\Gamma(S,\eta)$) and that $s = |S|$ denotes its cardinal.

Theorem 2: If $s = 1$, then for any $\eta \in [0, 1/\sigma^2]$, $\Gamma(S,\eta)$ has a unique Nash equilibrium $\lambda^*(s,\eta) = \max\{\lambda^*(1), \eta\}$. If $s > 1$, then there exists a unique parameter $\eta^*(s) \in [0, 1/\sigma^2]$ s.t.:

(i) for any $\eta \in [0, \eta^*(s)]$, $\Gamma(S,\eta)$ has a unique Nash equilibrium $\boldsymbol{\lambda}^*(s,\eta)$, s.t., $\lambda_i^*(s,\eta) = \lambda^*(s,\eta)$ for each $i \in S$, with

$$\lambda^*(s,\eta) = \begin{cases} \lambda^*(s) & \text{if } 0 \le \eta \le \lambda^*(s) \\ \eta & \text{if } \lambda^*(s) < \eta \le \eta^*(s); \end{cases} \quad (8)$$

(ii) for any $\eta \in (\eta^*(s), 1/\sigma^2]$, there does not exist a Nash equilibrium $\boldsymbol{\lambda}^*(s,\eta)$ s.t. $\lambda_i^*(s,\eta) \neq 0$ for each $i \in S$.

Theorem 2 introduces the quantity $\eta^*(s)$ which, as we will see, is crucial for the analyst. Similarly to $\lambda^*(s)$, the value of $\eta^*(s)$ cannot be written in closed form, but it can be computed as the unique solution of the following fixed point problem:

$$\eta = \tilde{g}(s,\eta),$$

where function $\tilde{g} : \mathbb{N}^* \times [0, 1/\sigma^2] \to [0, +\infty]$ is defined for each $\eta \in (0, 1/\sigma^2]$ and for each $n \in \mathbb{N}^*$ as

$$\tilde{g}(s,\eta) = \min\left\{ \frac{F\left(\frac{1}{(s-1)\eta}\right) - F\left(\frac{1}{s\eta}\right)}{c(\eta)} \cdot \eta, 1/\sigma^2 \right\}$$

and is defined by continuity as $\lim_{\eta \to 0^+} \tilde{g}(s,\eta)$ in $\eta = 0$ for each $n \in \mathbb{N}^*$. We can also show that $\lambda^*(s) < \eta^*(s)$ for all s (we obtain this result inside the proof of Theorem 3).

Theorem 2 characterizes the Nash equilibrium for different values of the parameter η. We observe that, as a consequence of the symmetry of the game, when $\eta \in [0, \eta^*(s)]$, the unique equilibrium of $\Gamma(S,\eta)$ is still symmetric, as it was for the unique equilibrium of the original game Γ. More specifically, if the analyst sets a minimum precision level η smaller than the unique equilibrium precision level $\lambda^*(s)$ of game Γ, the restriction of the strategy set does not have any effect on the outcome of the game. On the other hand, if the analyst sets a minimum precision level η in the interval $(\lambda^*(s), \eta^*(s)]$, all agents are still willing to participate with a precision $\eta > \lambda^*(s)$. This result matches with intuition, because even though agents' marginal costs are higher than the marginal benefits (the equilibrium choice is on the border of the strategy space $[\eta, 1/\sigma^2]$), their costs are still lower than if they choose a precision level zero. Therefore, agents do not have incentives to deviate. In the remaining range $(\eta^*(s), 1/\sigma^2]$, there does not exist an equilibrium such that each agent chooses a non-zero precision level. If there exist Nash equilibria, they are such that a subset $S' \subset S$ of agents choose the non-zero precision level $\lambda^*(s',\eta)$, while the others choose zero. The possible existence of these equilibria is not relevant for our analysis. In fact, such an equilibrium would provide the same estimation that the analyst can obtain by implementing the game $\Gamma(S',\eta)$ and, as we see in the following theorem, the estimation improves by maximizing the number of agents in the game.

The previous theorem is an important stepping stone allowing us to establish the main result of this section:

Theorem 3: The estimation variance at equilibrium is minimal for $S = N$ and $\eta = \eta^*(n)$. Moreover, we have

$$\sigma_M^2(\boldsymbol{\lambda}^*(n, \eta^*(n))) < \sigma_M^2(\boldsymbol{\lambda}^*(n)),$$

that is, setting a minimum precision level $\eta = \eta^*(n)$ strictly improves the estimation.

Theorem 3 shows that the analyst can indeed improve the quality of the estimation by setting a minimum precision level. It establishes that it is optimal, for the analyst, to solicit access to the private variable of all the agents whose data is contained in the data repository; and it provides the optimal minimum precision level $\eta = \eta^*(n)$ that the analyst should set to maximize the estimation precision. (Recall that $\eta^*(n)$ can be easily computed from the model's parameters by solving a fixed point problem.) Overall, Theorem 3 provides an implementable mechanism through which the analyst can improve the quality of the data provided by each user by imposing restrictions on the variance that users can add. In the next section, we study a special case with simple functions $F(\cdot)$ and $c(\cdot)$ in order to quantify precisely the improvement achieved.

C. The Special Case with Monomial Privacy Costs and Linear Estimation Cost

In this section, we illustrate the results of the previous sections on the special case where the privacy cost is monomial and the estimation cost is linear; i.e., we assume that the cost function in (4) has the form

$$J_i(\lambda_i, \boldsymbol{\lambda}_{-i}) = c\lambda_i^k + \sigma_M^2(\boldsymbol{\lambda}), \quad (9)$$

where $c \in (0, \infty)$ and $k \ge 2$ are constants. Note that, without loss of generality, in the linear estimation cost, we omit the constant factor (adding a constant to the cost does not modify the game solutions) as well as the slope factor (adding it would give an equivalent game with constant c rescaled). For this special case, we can determine both the equilibrium precision (without a minimum precision level) and the optimal minimum precision level in closed form. We can then graphically depict how the quantities vary while moving the model parameters, and explicitly compute the estimation improvement. A preliminary analysis of the simplified model with costs as in (9) was provided in our previous work [20]; we provide an extended analysis of this special case here thanks to the results of the previous section.

In the special case of costs given by (9), the equilibrium precision chosen by the agents in the game Γ simplifies to:

$$\lambda^*(n) = \begin{cases} \left(\frac{1}{ckn^2}\right)^{\frac{1}{k+1}} & \text{if } \left(\frac{1}{ckn^2}\right)^{\frac{1}{k+1}} \le 1/\sigma^2 \\ 1/\sigma^2 & \text{if } \left(\frac{1}{ckn^2}\right)^{\frac{1}{k+1}} > 1/\sigma^2. \end{cases} \quad (10)$$

As we have seen in the previous section (Theorem 3), it is optimal for the analyst to request access to the data of all

agents in N. In this special case, the corresponding optimal minimum precision level becomes

$$\eta^*(n) = \begin{cases} \left(\frac{1}{cn(n-1)}\right)^{\frac{1}{k+1}} & \text{if } \left(\frac{1}{cn(n-1)}\right)^{\frac{1}{k+1}} \leq 1/\sigma^2 \\ 1/\sigma^2 & \text{if } \left(\frac{1}{cn(n-1)}\right)^{\frac{1}{k+1}} > 1/\sigma^2. \end{cases}$$

Writing explicitly these two key quantities, we can immediately notice that, when c increases, i.e., when the agents are more concerned about privacy, they choose at equilibrium a smaller precision level $\lambda^*(n)$. Further, the minimum precision level $\eta^*(n)$ proposed by the analyst becomes smaller, if the agents are more sensitive about the protection of their data. In this special case, the properties of the results for the generic case are easy to spot. For instance, we have $\lambda^*(n) < \eta^*(n)$ for each $n \in \mathbb{N}^*$, and both of these quantities decrease and go to zero when n increases and goes to $+\infty$.

Most interestingly, the closed-form expressions that we have for this special case allow us to analyze the rate of decrease of the variance, and to quantify the improvement that can be achieved by imposing a minimum precision level. For n large enough (such that both $\lambda^*(n)$ and $\eta^*(n)$ are strictly smaller than $1/\sigma^2$), the variance at equilibrium level $\lambda^*(n)$ of game Γ is given by

$$\sigma_M^2(\boldsymbol{\lambda}^*(n)) = \frac{1}{n\left(\frac{1}{ckn^2}\right)^{\frac{1}{k+1}}},$$

while the variance at equilibrium level $\lambda^*(n, \eta^*(n))$ of game $\Gamma(N, \eta^*(n)))$ where the optimal minimum precision level is set is given by

$$\sigma_M^2(\boldsymbol{\lambda}^*(n, \eta^*(n))) = \frac{1}{n\left(\frac{1}{cn(n-1)}\right)^{\frac{1}{k+1}}}.$$

Both appear to have the same rate of decrease in $n^{\frac{-k+1}{k+1}}$ which is smaller than n^{-1} but becomes closer to n^{-1} as k tends to infinity. Intuitively, as the privacy cost becomes closer to a step function, the equilibrium precision level becomes less dependent on the number of agents so that we get closer to the case of averaging iid random variables of fixed variance. Consequently, for n large enough, the improvement is given by a factor:

$$\frac{\sigma_M^2(\boldsymbol{\lambda}^*(n))}{\sigma_M^2(\boldsymbol{\lambda}^*(n, \eta^*(n)))} = \left(\frac{kn}{n-1}\right)^{\frac{1}{k+1}} > 1, \tag{11}$$

which asymptotically becomes constant:

$$\frac{\sigma_M^2(\boldsymbol{\lambda}^*(n))}{\sigma_M^2(\boldsymbol{\lambda}^*(n, \eta^*(n)))} \sim_{n\to\infty} k^{\frac{1}{k+1}}. \tag{12}$$

Interestingly, we notice that this ratio of variances (characterizing the improvement when setting the optimal minimum precision level) depends on k, but not on c. (This holds even before the asymptotic regime, as long as n is large enough such that both $\lambda^*(n)$ and $\eta^*(n)$ are strictly smaller than $1/\sigma^2$.)

Figure 1 illustrates the asymptotic improvement ratio (12) for different values of k. We observe that it is bounded, it goes to 1 for large k's and it is in the range of $25-30\%$ improvement

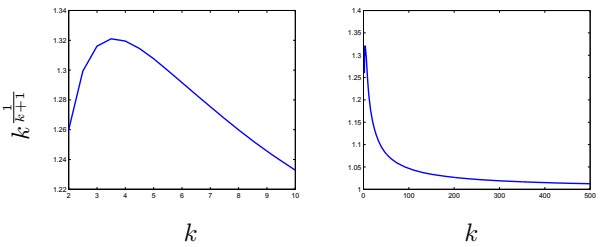

Fig. 1: Asymptotic improvement of the estimation choosing the optimum precision level η^* for values of $k = 2, \ldots, 10$ and for values of $k = 2, \ldots, 500$.

for values of k around $2 - 10$. Given that the ratio (11) converges towards its asymptote from above, this asymptotic improvement represents a lower bound of the improvement the analyst can achieve by implementing our mechanism with any finite number of agents n.

V. The Heterogeneous Agent Case

The previous section presents an exhaustive analysis of our model in the homogeneous case, i.e., when the agents exhibit the same privacy concerns. This simplified approach enables us to derive a first set of concrete results, intuition and qualitative understanding of the model and of the minimum contribution level mechanism. The results directly apply to homogeneous populations, and can serve as a first approximation by the analyst in other cases, i.e., whenever she does not have specific information about the agents. Indeed, the results are functions only of the total number of agents, and in practice this could represent the only available detail about the agents whose data is stored in the data repository. However, not all populations are homogeneous in their privacy concerns and having more details about the different privacy concerns of the agents allows for a customized analysis. Measuring how individuals value their private information is non-trivial, but researchers have conducted direct measurement surveys [58], [66] and various laboratory/field experiments [59], [60] allowing for an approximate ranking of users' privacy concerns, and context-specific valuations.

With this scope, we now extend our approach to the case in which the analyst faces a heterogeneous population. In this section, we remove the restricting hypothesis of homogeneity of the agents, and we allow them to exhibit different privacy concerns. Formally, the privacy cost function of an agent $i \in N$ is equal to $c_i(\cdot)$, where all the c_i's satisfy Assumption 1, but may be different from each other.

In order to model this situation, we follow the same approach that we used for the homogeneous case, i.e., we first analyze the situation in which the analyst implements the game Γ, without restricting the set of agents and without introducing a minimum precision level. Thereafter, we show how the analyst can improve the estimation by implementing a modified game $\Gamma(S, \eta)$.

A. The Estimation Game in the Heterogeneous Case

We start by analyzing the game Γ where each agent's action set is $[0, 1/\sigma^2]$. As for the homogeneous case, also in the

heterogeneous case we know that the equilibrium of the game Γ exists and is unique because we are considering a special case of the game in [18]. However, we can now characterize the equilibrium in more detail. The first result of the section, is presented in the following theorem.

Theorem 4: Assume that the privacy costs satisfy $c_1'(\lambda) \leq \cdots \leq c_n'(\lambda)$, for all $\lambda \in [0, 1/\sigma^2]$. Then, game Γ has a unique Nash equilibrium $\boldsymbol{\lambda}^*$ s.t., $0 < \lambda_n^* \leq \cdots \leq \lambda_1^*$.

Theorem 4 assumes that the agents can be ordered in such a way that, for any precision level $\lambda \in [0, 1/\sigma^2]$, an agent choosing precision λ has higher marginal privacy cost (and hence higher privacy cost since $c_i(0) = 0$ for all agents) than the previous agents if they choose the same precision level. This may require some re-ordering from the initial ordering, which comes without loss of generality. We believe that this assumption will often be reasonable in practice since agents who are more reluctant to increase the precision of their revealed data from a small precision (i.e., have higher marginal privacy cost for a small λ) will likely be more reluctant to increase the precision of their revealed data from a large precision (i.e., have higher marginal cost for a large λ too).

The proof of Theorem 4 exploits the potential nature of the game to characterize the Nash equilibrium. The unique Nash equilibrium, which cannot be written in closed form, can be easily computed as the minimum of the (convex) potential function of the game Γ, which is the function $\Phi : [0, 1/\sigma^2]^n \to \bar{\mathbb{R}}_+$, s.t., for each $\boldsymbol{\lambda} \in [0, 1/\sigma^2]^n$,

$$\Phi(\boldsymbol{\lambda}) = \sum_{j \in N} c_j(\lambda_j) + f(\boldsymbol{\lambda}). \tag{13}$$

We observe that, in the heterogeneous case, due to the asymmetry of the model, we no longer have a symmetric equilibrium. Moreover, the equilibrium strategy cannot be written as a function of the total number of agents n, as it depends on their privacy cost functions. We will use the notation $\boldsymbol{\lambda}^* = \boldsymbol{\lambda}^*(N)$ to denote that the equilibrium depends on the specific identity of the agents in the set of agents N. As expected, at equilibrium, agents with higher privacy concerns select lower precisions and, as for the homogeneous case, no agent decides to deny the access to her data. The fact that every agent contributes positively at Nash equilibrium stems from our assumption that giving a small amount of data implies very little cost since the marginal cost at zero is zero ($c'(0) = 0$). (Note, though, that some agents may contribute arbitrarily close to zero.) This assumption, although realistic, is not strictly necessary; but it greatly simplifies the presentation of our model and results.

Given the unique Nash equilibrium $\boldsymbol{\lambda}^*(N)$, the variance (3) of the estimate of y_M obtained by the analyst at equilibrium is given by the following expression:

$$\sigma_M^2(\boldsymbol{\lambda}^*(N)) = \frac{1}{\sum_{j \in N} \lambda_j^*(N)}. \tag{14}$$

Even if the equilibrium precisions chosen by the agents (and the corresponding variances) are not functions of only n, we can still generalize Propositions 1 and 2 to the heterogeneous case. In Propositions 3 and 4, we analyze how the equilibrium precision and the variance of the estimate at equilibrium vary when a new additional agent enters the game. Note that the following two propositions do not use the ordering assumption of Theorem 4.

Proposition 3: Given the game Γ, suppose that an additional $(n+1)$-th agent enters the game, and denote by $\boldsymbol{\lambda}^*(N \cup \{n+1\})$ the new equilibrium precision level. Then, for each $i \in N$, $\lambda_i^*(N \cup \{n+1\}) \leq \lambda_i^*(N)$.

Proposition 3 states that the equilibrium contribution of each agent decreases, as soon as a new agent enters the game.

Proposition 4: Given the game Γ, suppose that an additional $(n+1)$-th agent enters the game. Then, $\sigma_M^2(\boldsymbol{\lambda}^*(N \cup \{n+1\})) \leq \sigma_M^2(\boldsymbol{\lambda}^*(N))$.

Proposition 4 shows that, for the analyst, it is always better to let new agents enter the game despite the fact that, doing so, each other agent is giving data with a lower precision. Surprisingly, this is true even if the agent who enters has higher privacy concerns than any other agent in the game, and then would accordingly contribute the lowest quality data.

B. The Modified Estimation in the Heterogeneous Case

We now move to the case where the analyst can restrict the set of agents by introducing a minimum precision level $\eta \in [0, 1/\sigma^2]$. Again, her final goal is to improve the estimation accuracy. We consider at first the set of agents $S \subseteq N$ to be fixed, and we analyze how the estimation varies while moving only the parameter η. This variant is modeled by the game $\Gamma(S, \eta)$ defined in Section III-D, where η is now the only variable of the model. We denote by $\boldsymbol{\lambda}^*(S)$ the equilibrium precision level for the game $\Gamma(S, 0)$, and we suppose that it is such that there exists at least one agent $i \in S$ s.t. $\lambda_i^*(S) \neq 1/\sigma^2$; otherwise the estimation is already optimal with variance σ^2/s for $\eta = 0$.

The next result extends Theorem 2 to the heterogeneous case. We show that, if the analyst selects a minimum precision level which is not "too high", at equilibrium, all the agents (even the most concerned about privacy) are still willing to authorize access to their data (with perturbation).

Theorem 5: As in Theorem 4, assume that the privacy costs satisfy $c_1'(\lambda) \leq \cdots \leq c_n'(\lambda)$, for each $\lambda \in [0, 1/\sigma^2]$. Given the set of agents $S \subseteq N$, with cardinality $s \geq 1$:

- *(i)* if $s = 1$, then for any $\eta \in [0, 1/\sigma^2]$, $\Gamma(S, \eta)$ has a unique Nash equilibrium $\lambda_1^*(S, \eta) = \max\{\lambda_1^*(S), \eta\}$;
- *(ii)* if $s > 1$, then there exists a parameter $\eta^*(S) \in (\lambda^*(S), 1/\sigma^2]$ such that, for any $\eta \in [0, \eta^*(S)]$, $\Gamma(S, \eta)$ has a unique Nash equilibrium $\boldsymbol{\lambda}^*(S, \eta)$ with $\lambda_i^*(S, \eta) > 0$ for all $i \in S$.

Theorem 5 introduces a parameter $\eta^*(S)$ such that if the analyst sets a minimum precision level in $[0, \eta^*(S)]$, even the most privacy-concerned of the agents in S does not have an incentive to deviate to a zero precision level. As the theorem is stated, $\eta^*(S)$ is not unique (any value smaller than a valid $\eta^*(S)$ but still larger than $\lambda^*(S)$ will be suitable). However, let $\eta^*(S)$ be s.t.

$$c_n(\lambda_n^*(S, \eta^*(S))) = \tag{15}$$
$$F\left(\frac{1}{\sum_{j \in N, j \neq n} \lambda_j^*(S, \eta^*(S))}\right) - F\left(\frac{1}{\sum_{j \in N} \lambda_j^*(S, \eta^*(S))}\right),$$

where $\boldsymbol{\lambda}^*(S, \eta^*(S))$ is the local minimum of the potential function Φ defined as in (13), but on the domain $[\eta^*(S), 1/\sigma^2]^s$. We can prove that this $\eta^*(S)$ is unique, that it satisfies Theorem 5-*(ii)* and we conjecture that this definition gives the largest possible parameter satisfying Theorem 5-*(ii)*.

The result of Theorem 5 allows us to establish the main result of this section:

Theorem 6: As in Theorem 4, assume that the privacy costs satisfy $c_1'(\lambda) \leq \cdots \leq c_n'(\lambda)$, for each $\lambda \in [0, 1/\sigma^2]$. Let $\eta^*(N)$ be as in Theorem 5-*(ii)* for $S = N$. The analyst can improve the estimation by implementing the game $\Gamma(N, \eta^*(N))$ with minimum precision level $\eta^*(N)$, i.e.,

$$\sigma_M^2(\boldsymbol{\lambda}^*(N, \eta^*(N))) < \sigma_M^2(\boldsymbol{\lambda}^*(N)).$$

Theorem 6 shows that the analyst can improve the precision of the estimation of the mean y_M simply by setting a minimum precision level and soliciting access to the data from all the agents in N. This is true for any minimum precision level $\eta^*(N)$ such that Theorem 5-*(ii)* is satisfied and shows that, even in the heterogeneous case, it is possible to strictly improve the estimation by applying the minimum precision level mechanism. Here too, however, we conjecture that the parameter $\eta^*(N)$ solving (15) yields the highest possible improvement.

C. The Special Heterogeneous Case with Monomial Privacy Costs and Linear Estimation Cost

As for the homogeneous case, we now illustrate the results of the previous sections on the heterogeneous model in the special case of monomial privacy cost and linear estimation cost. In this simplified model, the cost function in (4) has the form

$$J_i(\lambda_i, \boldsymbol{\lambda}_{-i}) = c_i \lambda_i^k + \sigma_M^2(\boldsymbol{\lambda}), \tag{16}$$

with $c_i \in (0, \infty)$ for each $i \in N$ and $k \geq 2$. The assumption of Theorem 4 that agents can be ordered s.t. $c_1'(\lambda) \leq \ldots \leq c_n'(\lambda)$ for each $\lambda \in [0, 1/\sigma^2]$, translates now to requiring that $0 < c_1 \leq \ldots \leq c_n$ (which, in the case of monomial costs, is completely without loss of generality).

Even with such a simplified model, having heterogeneous agents does not allow us to write the key quantity in closed form as we did in the simplified homogeneous model in Section IV-C. However, it is still possible to provide clearer expressions and to quantify the variance improvement by setting a minimum precision level.

When the agents play the estimation game Γ, at equilibrium they choose a precision level that, if interior, can be written as

$$\lambda_i^*(N) = \left(\frac{1}{c_i k \left(\sum_{j \in N} \lambda_j^*(N) \right)^2} \right)^{\frac{1}{k-1}}.$$

The analyst can improve the estimation by setting a minimum precision level $\eta^*(N)$. In this simplified case, it takes the form

$$\eta^*(N) = \left(\frac{1}{c_n \left(\sum_{j \in N} \lambda_j^*(\eta^*(N)) \right) \left(\sum_{j \in N \setminus \{n\}} \lambda_j^*(\eta^*(N)) \right)} \right)^{\frac{1}{k-1}}.$$

Note that the two expressions above are in the form of fixed-point equations. It is interesting to note that when $k > c_n/c_1$ though, i.e., when the privacy cost of the agents are not too dispersed, this minimum precision level can be written in closed form as

$$\eta^*(N) = \left(\frac{1}{c_n n(n-1)} \right)^{\frac{1}{k+1}}. \tag{17}$$

It is then equal to the optimal precision level, when all the agents have the same privacy cost as the most privacy-concerned individual.

Figure 2 illustrates on an example the estimation improvement in the heterogeneous case when choosing $\eta^*(N)$ as above (which we conjectured is the optimal choice). We compare it with the improvement in the analogous homogeneous case when choosing the optimal $\eta^*(n)$ (see Theorem 3 which does not depend on c).

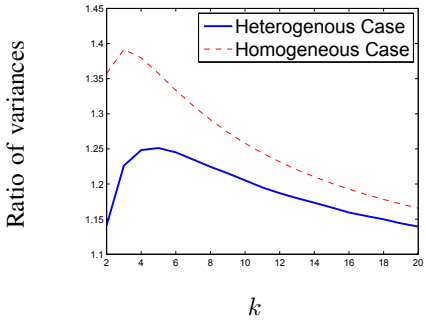

Fig. 2: Improvement of the estimation in $\Gamma(\eta)$ in the heterogeneous case choosing the optimum precision level $\eta^*(N)$, compared to the homogeneous case choosing the optimum precision level $\eta^*(n)$; for values of $k = 2, \ldots, 20$. In this example, $\mathbf{c} = (1, 1.5, 2, 2.5, 3)$, $1/\sigma^2 = 2$.

VI. Extensions of the Model

In this section, we extend our model in two directions. In Section VI-A, we propose an alternative modified estimation game, and we compare it with the one proposed in Section III-D. The main difference with the previous one is that it is a two-stage game. In Section VI-B, we add an important variable to our model by introducing a per-agent cost of collecting data. Both proposed extensions are included to derive qualitative insights about the practical applicability of the model, however, we defer an in-depth analysis to future work.

A. The Modified Two-Stage Game

In $\Gamma(N, \eta)$, both the decision to authorize the access (or to deny it) and the selection of a precision level (in case of authorization) are simultaneous. This variant captures cases in a realistic fashion where the analyst requests access to data already present in a repository. In different applications, however, the analyst may first recruit participants that commit to provide private data with a minimum precision; and only in a second stage (for example, as soon as the data becomes

available), these agents would be asked to disclose their information. This scenario applies, for example, to medical research studies or consumer decisions, and it motivates the study of a model where agents first decide to participate or not, and only then decide on the precision of the data released. Another motivation to study such a model is that it could lead to a higher estimation accuracy, in which case the analyst would want to implement it even if it does not naturally arise from the application at stake.

In this section, we investigate this extension of our original model in the simplified case with homogeneous monomial privacy costs and linear estimation cost as it is sufficient to understand and illustrate the qualitative differences between the two models. We leave the development of the more general model to future work. We also point out the possibility, for future work, of a similar extension, in which the agents asynchronously make decisions on whether or not to share their data (i.e., they make their sharing decisions based on actions taken by agents who were contacted earlier by the analyst). However, in absence of observability of the contribution decisions (as it is often the case in the medical domain due to confidentiality restrictions) even asynchronous decision-making can be approximated well with a simultaneous move model.

To investigate our variant of the model and to compare its outcome with the one of the game $\Gamma(N,\eta)$, we define a two-stage variant of the game. We assume that the agents are initially informed of the minimum precision level η. In a first stage, they have to decide if they want to deny access to their data, and exit the game, or if they wish to accept to authorize access. The set of agents who accepted to participate is revealed to all agents. In a second stage, the agents who decided to participate choose their precision in the imposed range $[\eta, 1/\sigma^2]$. Formally, this situation is modeled through the following two-stage game $\Gamma^2(\eta)$:

(i) In the *first stage*, the agents make a binary choice $p_i \in \{0,1\}$

$$\forall i \in N \quad p_i = \begin{cases} 0 \text{ if } i \text{ denies the access} \\ 1 \text{ if } i \text{ accepts to authorize.} \end{cases}$$

We denote by $\boldsymbol{p} \in \{0,1\}^n$ a strategy profile, $P = \{i \in N : p_i = 1\}$ the set of agents who accept, and $p = |P|$ its cardinality.

(ii) In the *second stage*, given $\boldsymbol{p} \in \{0,1\}^n$, the agents play a game $\Gamma^P(\eta) = \left\langle N, [\eta, 1/\sigma^2]^p \times \{0\}^{n-p}, (J_i)_{i\in N} \right\rangle$, where each agent $i \in P$ has strategy space $[\eta, 1/\sigma^2]$ whereas each agent $i \in N \setminus P$ has strategy space $\{0\}$ (i.e., the agents in $N \setminus P$ can only choose $\lambda_i = 0$, which, we reiterate, is equivalent to no participation).

We have already seen that the analyst can improve the estimation by modifying the original game, and that the optimal choice, in that previous setting (in the homogeneous case), is to implement the game $\Gamma(N, \eta^*(n))$. We now investigate whether the analyst can improve the estimation even more, while implementing the game $\Gamma^2(\eta)$ for an optimal choice of the minimum precision level η.

The games $\Gamma(N,\eta)$ and $\Gamma^2(\eta)$ differ in the information available to agents when choosing their precision (observe for instance that $\Gamma(N,0) = \Gamma$, while $\Gamma^2(0) \neq \Gamma$). In $\Gamma(N,\eta)$, both the decision to authorize the access or to deny it and the decision of the precision (in case of authorization) are simultaneous. In contrast, in $\Gamma^2(\eta)$, the set of agents who will authorize with precision of at least η is known at the time of choosing the exact precision.

As we did for the previous games, we study $\Gamma^2(\eta)$ as a *complete information game* between the agents, i.e., we assume that the set of agents, the action sets (in particular, when present, the value of the parameter η) and the costs are known by all the agents.

We study the pure strategy Nash equilibria of the game using backward induction. Given $\boldsymbol{p} \in \{0,1\}^n$ the outcome of the first stage, a Nash equilibrium for the second stage is a strategy profile $\boldsymbol{\lambda}^* \in [\eta, 1/\sigma^2]^p \times \{0\}^{n-p}$ s.t., for each $i \in N \setminus P$, $\lambda_i^* = 0$, while for each $i \in P$, λ_i^* is s.t.

$$\lambda_i^* \in \underset{\lambda_i \in [\eta, 1/\sigma^2]}{\arg\min} \; J_i(\lambda_i, \boldsymbol{\lambda}^*_{-i}). \tag{18}$$

If for each $\boldsymbol{p} \in \{0,1\}^n$ the second stage game has a unique solution $\boldsymbol{\lambda}^*(\boldsymbol{p},\eta)$ (as we will see, it is the case in our model), the choice that the agents make in the first stage determines univocally the outcome of the two-stage game. Then, $\Gamma^2(\eta)$ reduces to a one-stage game $\left\langle N, \{0,1\}^n, (J_i^1)_{i\in N} \right\rangle$, where the cost function $J_i^1 : \{0,1\}^n \to \mathbb{R}$, for each $\boldsymbol{p} \in \{0,1\}$, is given for all $i \in N$ by

$$J_i^1(\boldsymbol{p}) = J_i(\boldsymbol{\lambda}^*(\boldsymbol{p},\eta)) = c\lambda_i^*(\boldsymbol{p},\eta)^k + \sigma_M^2(\boldsymbol{\lambda}^*(\boldsymbol{p},\eta)). \tag{19}$$

Then, an equilibrium of the game $\Gamma^2(\eta)$ is a strategy profile $\boldsymbol{p}^* \in \{0,1\}$ s.t.

$$p_i^* \in \underset{p_i \in \{0,1\}}{\arg\min} J_i^1(p_i, \boldsymbol{p}^*_{-i}), \quad \forall i \in N. \tag{20}$$

We apply backward induction, by starting to analyze the second stage game. In the following lemma, we show the existence and the uniqueness of a Nash equilibrium for the game $\Gamma^P(\eta)$ when $\boldsymbol{p} \neq (0,\ldots,0)$.

Lemma 1: For each $\boldsymbol{p} \in \{0,1\}^n \setminus \{(0,\ldots,0)\}$, the game $\Gamma^P(\eta)$ has a unique Nash equilibrium $\boldsymbol{\lambda}^*(\boldsymbol{p},\eta)$, s.t., $\lambda_i^*(\boldsymbol{p},\eta) = \lambda^*(p,\eta)$ for each $i \in P$, with

$$\lambda^*(p,\eta) = \begin{cases} \lambda^*(p) & \text{if } 0 \leq \eta \leq \lambda^*(p) \\ \eta & \text{if } \lambda^*(p) < \eta \leq 1/\sigma^2, \end{cases} \tag{21}$$

(where $\lambda^*(p)$ is defined as in (10)) and $\lambda_i^*(\boldsymbol{p},\eta) = 0$, for each $i \in N \setminus P$.

We observe again that the equilibrium of the game restricted to agents in P is symmetric (i.e., each participating agent chooses the same precision level at equilibrium). We call $\lambda^*(p,\eta)$ the equilibrium precision for agents in P to emphasize the dependence on the cardinality p of the set P and on the parameter η. In fact, due to the symmetry, the optimal choice for an agent who decided to participate depends only on the number of agents who made the same choice as her in the first stage and not on the identity of these agents. Further, given P and η, the equilibrium of $\Gamma^P(\eta)$ is the same as the equilibrium of $\Gamma(N,\eta)$ given in Theorem 2, when replacing n by p. The only difference is that, even for large η the agents in P will

choose precision level η in $\Gamma^P(\eta)$ since they are committed to participate with precision of at least η. Consequently, the equilibrium of $\Gamma^P(\eta)$ always exists and it is s.t. each agent choosing a non-zero precision level.

As the second stage game has always a unique solution, we can apply backward induction, and the two-stage game $\Gamma^2(\eta)$ reduces to a one-stage game. The following lemma establishes the existence and uniqueness of its equilibrium for a minimum precision level.

Lemma 2: For any $\eta \in [\lambda^*(n-1), \eta^*(n)]$, the two-stage game $\Gamma^2(\eta)$ has a unique equilibrium given by $\boldsymbol{p}^* = (1, \ldots, 1)$.

Lemma 2 states that, if the analyst chooses a minimum precision level in the range $[\lambda^*(n-1), \eta^*(n)]$ and implements the two-stage game $\Gamma^2(\eta)$, then each agent will participate at equilibrium. The equilibrium contributions, given by Lemma 1, equal η for each agent since $\eta \geq \lambda^*(n-1) \geq \lambda^*(n)$. For η in the range $[\lambda^*(n-1), \eta^*(n)]$, the outcome of the two-stage game $\Gamma^2(\eta)$ is therefore the same as for the one-stage game $\Gamma(N, \eta)$. This is not the case, however, for other ranges of parameters. In particular, for $\eta < \lambda^*(n-1)$, all agents participate in $\Gamma(N, \eta)$ whereas they may not participate in $\Gamma^2(\eta)$. This is because, in $\Gamma^2(\eta)$, agents react in the second stage to the participation decisions of the first stage (typically if an agent chooses not to participate, the others increase their precisions in the second stage). As a consequence, agents can strategically choose their participation in the first stage to influence the precisions chosen in the second stage. Analysis of the existence and uniqueness of the Nash equilibrium in $\Gamma^2(\eta)$ in ranges of η outside $[\lambda^*(n-1), \eta^*(n)]$ is therefore more intricate. Nevertheless, we can establish our main result, namely that choosing $\eta = \eta^*(n)$ yields an optimal estimation variance for the analyst:

Theorem 7: For the game $\Gamma^2(\eta)$, with $\eta \in [0, 1/\sigma^2]$, the estimate's variance at equilibrium $\sigma_M^2(\boldsymbol{\lambda}^*(\boldsymbol{p}^*, \eta))$ is minimized for $\eta = \eta^*(n)$. The improvement obtained by setting the minimum precision level $\eta = \eta^*(n)$ is characterized, for n large enough, by the ratio

$$\frac{\sigma_M^2(\boldsymbol{\lambda}^*(n))}{\sigma_M^2(\boldsymbol{\lambda}^*(\boldsymbol{p}^*, \eta^*(n)))} = \left(\frac{kn}{n-1}\right)^{\frac{1}{k+1}} > 1, \quad (k \geq 2).$$

Theorem 7 shows that the optimal η is the same for the one-stage game $\Gamma(N, \eta)$ and the two-stage game $\Gamma^2(\eta)$, and both yield the same improvement for the analyst. As such, the discussion given in Section IV-B about the asymptotic behavior of this gain still holds. However, as mentioned, the two games $\Gamma(\eta)$ and $\Gamma^2(\eta)$ are not equivalent for each choice of the parameter η. In particular, we can infer from the proof of Theorem 7 that there is still a range of minimum precision levels for which the estimation is strictly improved, but this range is smaller than it was for $\Gamma(N, \eta)$.

B. The Estimation Game in the Presence of Per-Agent Costs

In this section, we propose an extension of our model to include the cost of collecting data. Indeed, in Section III and throughout this paper, we assumed that data is collected at no cost, and that the analyst aims at minimizing the variance of the mean estimation. The absence of per-agent cost (to solicit contributions) is a standard assumption in most of the public good literature. However, it could limit the appeal of our model in some applications. Here, we present preliminary results with arbitrary per-agent cost, restricted to the homogeneous case. We then introduce a simplified case with linear per-agent cost, to illustrate the qualitative difference to the zero per-agent cost case, in particular, the existence of an optimal number of agents n. The derivation of the optimal n would be slightly different when assuming, for example, a concave cost function. This is left as a possible future work suggestion (see Section VII).

When facing a per-agent cost, we can no longer rely on the fact that the analyst will always prefer to have the largest possible set of agents. Rather, she has to select the optimal subset of agents to include in the game. In the homogeneous case, selecting the optimal subset of agents reduces to selecting the optimal number of agents $n^* \in N$. To address this problem, we assume that, instead of aiming at minimizing the variance, the analyst aims at minimizing a cost function $J_A : \mathbb{N}^* \to \mathbb{R}$ defined as

$$J_A(n) = f(\boldsymbol{\eta}^*(n)) + Cn, \tag{22}$$

where f is the estimation cost defined in Section III, while C represents the *per-agent cost* of collecting personal data. We assume that the estimation cost is evaluated at equilibrium, when the analyst chooses the optimal minimum precision level. In fact, for a fixed n, $\boldsymbol{\eta}^*(n)$ provides the minimum variance and, consequently, the minimum estimation cost. The problem of the analyst now reduces to setting an optimal number of agents n^*.

Theorem 8: The function $J_A(n)$ has a minimum in $\mathbb{N}^*$. The optimal n^* is given by $n^* = \max\{m \in N^* | c(\eta^*(m)) \geq C\}$, if this set is non-empty, and by 1 otherwise.

Theorem 8 shows how the analyst can optimize the balance between the minimization of the estimation cost and the per-agent recruitment cost. In this situation, it is typically not optimal anymore to contact as many agents as possible. Of course, if the theoretically determined optimal number of agents equals or exceeds the size of the potential participant pool ($n^* \geq n$), then the analyst will contact all available agents. As $c(\eta^*(m))$ is non-increasing in m, n^* can be easily computed by the analyst, for example by implementing a bisection method on $[1, n]$, where n is the total number of agents whose data is contained in the repository.

VII. Concluding remarks

In this paper, we investigate the problem of estimating population averages from data provided by privacy-sensitive users. We assume that users can perturb their data before revealing it (e.g., by adding zero-mean noise) in order to protect their privacy. Users, however, benefit from a more accurate population estimate. Therefore, each user strategically selects the precision of her revealed data to balance her privacy cost and the cost incurred by a lower precision of the population estimate. We find that the resulting game has a unique Nash equilibrium and carefully study its properties.

We further prove that the analyst can increase the population estimate's accuracy simply by imposing a minimum precision level for the data which users can reveal (e.g., by restricting the variance of the noise users can add). The

surprising and important aspect of this result is that the scheme remains incentive-compatible, i.e., users are willing to provide data with a higher precision rather than dropping out. We also show how to tune the minimum precision level the analyst should set in order to optimize the population estimate's accuracy. In our numerical simulations, the maximum improvement of the population estimate's accuracy is in the order of $20 - 40\%$.

Our model treats the population estimate's accuracy as a public good (e.g., if one agent increases the precision of the data she gives, it benefits all users). Then, our results offer a novel method to increase the provision of a public good above voluntary contributions, simply by restricting the agents' strategy spaces. This method is attractive through its simplicity compared for instance to other schemes that involve monetary transfers, and could find application in other public good problem domains.

The results are derived for arbitrary functions for the privacy cost experienced by each user and for the estimation cost (satisfying relatively mild assumptions). This increases the robustness of our main results and allows for application to various situations. Further, we study the cases of homogeneous and heterogeneous agents. Indeed, for practical utilization of our work it is important to be able to accommodate different types of privacy preferences as evidenced by the literature on the value of privacy (which includes direct measurement surveys [58], [66] and laboratory/field experiments [59], [60]).

We also consider extensions of our basic model such as variations in the structure of decision-making. Introducing a two-stage structure impacts the available information to individuals, i.e., whether or not the set of contributing agents is determined before agents choose their precision levels. Surprisingly, we find that providing this information to users can never improve the estimation's accuracy. In future work, we plan to analyze other decision-making structures, such as when agents make decisions asynchronously and can utilize information about the previous contribution levels by other agents.

In our basic model, we assume that the analyst can collect data from n users at negligible cost. This assumption can be reasonable in scenarios where the data is already available in a repository, and the analyst merely has to inquire with individuals to contribute their data for secondary analysis. In this scenario, we showed that the population estimate's accuracy increases with n (although each individual then lowers the precision of her contribution). We further extend the model to handle applications where there could be a more substantial cost of collecting data per user (e.g., cost of sending a survey). In that case, it is no longer optimal for the analyst to collect data from all users but we show, in the homogeneous case, how the analyst can then select the optimal number of users. The method outlined for the homogeneous case also provides a trajectory to approach the task of selecting the optimal set of agents to solicit data from in the heterogeneous case, utilizing the ordering assumption of Theorem 4. Further, our results regarding the benefits of a minimum precision level apply also to costly data acquisition. In future work, we will consider non-linear cost (e.g., concave) to further generalize our results.

A unique Nash equilibrium exists for all considered cases and extensions. Computing the exact equilibrium strategies may be non-trivial for agents in practice. However, knowledge about the uniqueness of the optimal strategies suggests the possibility of reaching the equilibrium via *tacit coordination* when agents gain experience with comparable data contribution decisions [67]. In addition, providing a minimum precision level will further guide agents in their decision-making.

In this paper, we consider the problem of estimating the population average of a single scalar quantity. However, the results also serve as building blocks to tackle more complex scenarios. For example, an analyst may need to estimate averages of several quantities which are not independent (if the quantities are independent, our results readily apply by considering several independent instances of the model, possibly with different privacy costs). Further, the analyst may want to estimate the parameter of a linear model as in [18]. In both cases, the problem of selecting the users to solicit data from will become combinatorial and requires further study to find a suitable approximation. However, our techniques to characterize the equilibrium of the modified game will extend and will be instrumental in establishing the optimal strategy space to impose for a given set of users.

ACKNOWLEDGMENTS

This work was partially funded by the French Government (National Research Agency, ANR) through the "Investments for the Future" Program reference # ANR-11-LABX-0031-01; and by the Institut Mines-Télécom through the "Futur & Rupture" program. Jens Grossklags gratefully acknowledges the hospitality and support received as a Visiting Scientist at EURECOM during the earlier stages of this work. In addition, the authors would like to thank the anonymous reviewers for their detailed and helpful comments.

REFERENCES

[1] P. Lyman and H. Varian, "How much information?" 2000, available at: http://www2.sims.berkeley.edu/research/projects/how-much-info/.

[2] ——, "How much information 2003?" 2003, available at: http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/.

[3] J. Constine, "How big is Facebook's data? 2.5 billion pieces of content and 500+ terabytes ingested every day," *Techcrunch*, 2012.

[4] World Economic Forum, and Bain & Company, *Personal Data: The Emergence of a New Asset Class*. World Economic Forum, 2011.

[5] H. Varian, "Beyond big data," *Business Economics*, vol. 49, no. 1, pp. 27–31, 2014.

[6] D. Solove, "A taxonomy of privacy," *University of Pennsylvania Law Review*, vol. 154, no. 3, pp. 477–560, Jan. 2006.

[7] I. Altman, *The Environment and Social Behavior*. Belmont, 1975.

[8] S. Warren and L. Brandeis, "The Right to Privacy," *Harvard Law Review*, pp. 193–220, 1890.

[9] A. Westin, *Privacy and freedom*. New York: Atheneum, 1970.

[10] A. Acquisti and C. Fong, "An experiment in hiring discrimination via online social networks," Carnegie Mellon University, Tech. Rep., 2013, available at SSRN: http://ssrn.com/abstract=2031979.

[11] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris, "Crowd-assisted search for price discrimination in e-commerce: First results," in *Proceedings of the Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, 2013, pp. 1–6.

[12] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Security & Privacy*, vol. 3, no. 1, pp. 26–33, 2005.

[13] N. Kass, M. Natowicz, S. Hull, R. Faden, L. Plantinga, L. Gostin, and J. Slutsman, "The use of medical records in research: What do patients want?" *Journal of Law, Medicine & Ethics*, vol. 31, pp. 429–433, 2007.

[14] L. Damschroder, J. Pritts, M. Neblo, R. Kalarickal, J. Creswell, and R. Hayward, "Patients, privacy and trust: Patients' willingness to allow researchers to access their medical records," *Social Science & Medicine*, vol. 64, no. 1, pp. 223–235, 2007.

[15] M. Robling, K. Hood, H. Houston, R. Pill, J. Fay, and H. Evans, "Public attitudes towards the use of primary care patient record data in medical research without consent: A qualitative study," *Journal of Medical Ethics*, vol. 30, no. 1, pp. 104–109, 2004.

[16] D. Willison, L. Schwartz, J. Abelson, C. Charles, M. Swinton, D. Northrup, and L. Thabane, "Alternatives to project-specific consent for access to personal information for health research. What do canadians think?" in *Presentation at the 29th International Conference of Data Protection and Privacy Commissioners*, 2007.

[17] A. Westin, "How the public views privacy and health research," 2007, Institute of Medicine.

[18] S. Ioannidis and P. Loiseau, "Linear regression as a non-cooperative game," in *Web and Internet Economics*, ser. Lecture Notes in Computer Science, Y. Chen and N. Immorlica, Eds. Springer Berlin Heidelberg, 2013, vol. 8289, pp. 277–290.

[19] S. Spiekermann, J. Grossklags, and B. Berendt, "E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior," in *Proceedings of the 3rd ACM Conference on Electronic Commerce*, 2001, pp. 38–47.

[20] M. Chessa, J. Grossklags, and P. Loiseau, "A short paper on incentives to share private information for population estimates," in *Proceedings of the 19th International Conference on Financial Cryptography and Data Security (FC)*, 2015.

[21] ——, "A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications," 2015, technical report, available at arXiv:1505.02414.

[22] F. Pukelsheim, *Optimal design of experiments*. New York: Wiley, 1993.

[23] A. Atkinson, A. Donev, and R. Tobias, *Optimum experimental designs, with SAS*. Oxford University Press New York, 2007.

[24] T. Horel, S. Ioannidis, and S. Muthukrishnan, "Budget feasible mechanisms for experimental design," in *LATIN 2014: Theoretical Informatics*, ser. Lecture Notes in Computer Science, A. Pardo and A. Viola, Eds. Springer Berlin Heidelberg, 2014, vol. 8392, pp. 719–730.

[25] A. Roth and G. Schoenebeck, "Conducting truthful surveys, cheaply," in *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, 2012, pp. 826–843.

[26] C. Riederer, V. Erramilli, A. Chaintreau, B. Krishnamurthy, and P. Rodriguez, "For sale : Your data: By : You," in *Proceedings of the 10th ACM Workshop on Hot Topics in Networks*, 2011, pp. 13:1–13:6.

[27] I. Bilogrevic, J. Freudiger, E. De Cristofaro, and E. Uzun, "What's the gist? Privacy-preserving aggregation of user profiles," in *Computer Security - ESORICS 2014*, ser. Lecture Notes in Computer Science, M. Kutylowski and J. Vaidya, Eds. Springer International Publishing, 2014, vol. 8713, pp. 128–145.

[28] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*, ser. Lecture Notes in Computer Science, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds. Springer Berlin Heidelberg, 2006, vol. 4052, pp. 1–12.

[29] D. Kifer, A. Smith, and A. Thakurta, "Private convex empirical risk minimization and high-dimensional regression," *JMLR W&CP (Proceedings of COLT 2012)*, vol. 23, pp. 25.1–25.40, 2012.

[30] A. Ghosh and A. Roth, "Selling privacy at auction," in *Proceedings of the 12th ACM Conference on Electronic Commerce*, 2011, pp. 199–208.

[31] K. Nissim, R. Smorodinsky, and M. Tennenholtz, "Approximately optimal mechanism design via differential privacy," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 203–213.

[32] K. Ligett and A. Roth, "Take It or Leave It: Running a Survey When Privacy Comes at a Cost," in *Internet and Network Economics*, ser. Lecture Notes in Computer Science, P. Goldberg, Ed. Springer Berlin Heidelberg, 2012, vol. 7695, pp. 378–391.

[33] Y. Chen, S. Chong, I. Kash, T. Moran, and S. Vadhan, "Truthful mechanisms for agents that value privacy," in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC)*, 2013, pp. 215–232.

[34] J. Vaidya, C. Clifton, and Y. Zhu, *Privacy Preserving Data Mining*. Springer, 2006.

[35] J. Domingo-Ferrer, "A survey of inference control methods for privacy-preserving data mining," in *Privacy-Preserving Data Mining*, ser. Advances in Database Systems, C. Aggarwal and P. Yu, Eds. Springer, 2008, vol. 34, pp. 53–80.

[36] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 439–450.

[37] S. Oliveira and O. Zaiane, "Privacy preserving clustering by data transformation," in *Proceedings of the XVIII Simposio Brasileiro de Bancos de Dados*, 2003, pp. 304–318.

[38] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, "Disclosure limitation of sensitive rules," in *Proceedings of the Workshop on Knowledge and Data Engineering Exchange (KDEX'99)*, 1999, pp. 45–52.

[39] J. Duchi, M. Jordan, and M. Wainwright, "Local privacy and statistical minimax rates," in *Proceedings of the 54th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2013, pp. 429–438.

[40] R. Cummings, K. Ligett, A. Roth, Z. Wu, and J. Ziani, "Accuracy for sale: Aggregating data with a variance constraint," in *Proceedings of the Conference on Innovations in Theoretical Computer Science (ITCS)*, 2015, pp. 317–324.

[41] C. Aperjis, V. Gkatzelis, and B. Huberman, "Pricing private data," *Electronic Markets*, forthcoming.

[42] O. Dekel, F. Fischer, and A. D. Procaccia, "Incentive compatible regression learning," *Journal of Computer and System Sciences*, vol. 76, no. 8, pp. 759–777, 2010.

[43] J. Perote and J. Perote-Pena, "Strategy-proof estimators for simple regression," *Mathematical Social Sciences*, vol. 47, no. 2, pp. 153–176, 2004.

[44] Y. Cai, C. Daskalakis, and C. Papadimitriou, "Optimum statistical estimation with strategic data sources," 2014, preprint, available as arXiv:1408.2539.

[45] J. Morgan, "Financing public goods by means of lotteries," *Review of Economic Studies*, vol. 67, no. 4, pp. 761–84, 2000.

[46] G. Biczók and P. Chia, "Interdependent privacy: Let me share your data," in *Financial Cryptography and Data Security*, ser. Lecture Notes in Computer Science, A.-R. Sadeghi, Ed. Springer, 2013, vol. 7859, pp. 338–353.

[47] Y. Pu and J. Grossklags, "An economic model and simulation results of app adoption decisions on networks with interdependent privacy consequences," in *Decision and Game Theory for Security*, R. Poovendran and W. Saad, Eds. Springer, 2014, vol. 8840, pp. 246–265.

[48] M. Backes, A. Kate, M. Maffei, and K. Pecina, "Obliviad: Provably secure and practical online behavioral advertising," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2012, pp. 257–271.

[49] S. Guha, B. Cheng, and P. Francis, "Privad: Practical privacy in online advertising," in *Proceedings of the 8th USENIX Conference on Networked Systems Design and Implementation (NSDI)*, 2011, pp. 169–182.

[50] Y. Nardi, S. Fienberg, and R. Hall, "Achieving both valid and secure logistic regression analysis on aggregated data from different private sources," *Journal of Privacy and Confidentiality*, vol. 4, no. 1, 2012.

[51] R. Hall, S. Fienberg, and Y. Nardi, "Secure multiple linear regression based on homomorphic encryption," *Journal of Official Statistics*, vol. 27, no. 4, pp. 669–691, 2011.

[52] M. Naor, B. Pinkas, and R. Sumner, "Privacy preserving auctions and mechanism design," in *Proceedings of the 1st ACM Conference on Electronic Commerce (EC)*, 1999, pp. 129–139.

[53] S. Izmalkov, S. Micali, and M. Lepinski, "Rational secure computation and ideal mechanism design," in *Proceedings of the 46th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2005, pp. 585–594.

[54] L. Cranor, *Web privacy with P3P - The platform for privacy preferences*. O'Reilly, 2002.

[55] O. Berthold and M. Köhntopp, "Identity management based on P3P," in *Designing Privacy Enhancing Technologies*, ser. Lecture Notes in Computer Science, H. Federrath, Ed. Springer, Berlin Heidelberg, 2001, vol. 2009, pp. 141–160.

[56] L. Cranor, M. Arjula, and P. Guduru, "Use of a P3P user agent by early adopters," in *Proceedings of the ACM Workshop on Privacy in the Electronic Society*, 2002, pp. 1–10.

[57] Y. Wang and A. Kobsa, "Respecting users' individual privacy constraints in web personalization," in *User Modeling 2007*, ser. Lecture Notes in Computer Science, C. Conati, K. McCoy, and G. Paliouras, Eds. Springer Berlin Heidelberg, 2007, vol. 4511, pp. 157–166.

[58] A. Acquisti and J. Grossklags, "An online survey experiment on ambiguity and privacy," *Communications & Strategies*, vol. 49, no. 4, pp. 19–39, 2012.

[59] A. Acquisti, L. John, and G. Loewenstein, "What is privacy worth?" *Journal of Legal Studies*, vol. 42, no. 2, 2013.

[60] J. Grossklags and A. Acquisti, "When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information," in *Proceedings of the Workshop on the Economics of Information Security*, 2007.

[61] A. Acquisti, I. Adjerid, and L. Brandimarte, "Gone in 15 seconds: The limits of privacy transparency and control," *IEEE Security & Privacy*, vol. 11, no. 4, pp. 72–74, 2013.

[62] L. Brandimarte, A. Acquisti, and G. Loewenstein, "Misplaced confidences: Privacy and the control paradox," *Social Psychological and Personality Science*, vol. 4, no. 3, pp. 340–347, 2013.

[63] N. Wang, H. Xu, and J. Grossklags, "Third-party apps on Facebook: Privacy and the illusion of control," in *Proceedings of the 5th ACM Symposium on Computer Human Interaction for Management of Information Technology*, 2011.

[64] N. Wang, J. Grossklags, and H. Xu, "An online experiment of privacy authorization dialogues for social applications," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*, 2013, pp. 261–272.

[65] B. Huberman, E. Adar, and L. Fine, "Valuating privacy," *IEEE Security & Privacy*, vol. 3, no. 5, pp. 22–25, 2005.

[66] C. Bauer, J. Korunovska, and S. Spiekermann, "On the value of information: What Facebook users are willing to pay," in *Proceedings of the 33rd International Conference on Information Systems*, 2012.

[67] J. Van Huyck, R. Battalio, and R. Beil, "Tacit coordination games, strategic uncertainty, and coordination failure," *The American Economic Review*, vol. 80, no. 1, pp. 234–248, 1990.