

Linear Regression from Strategic Data Sources

NICOLAS GAST, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, France

STRATIS IOANNIDIS, Northeastern University, USA

PATRICK LOISEAU, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG France
and MPI-SWS, Germany

BENJAMIN ROUSSILLON, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, France

Linear regression is a fundamental building block of statistical data analysis. It amounts to estimating the parameters of a linear model that maps input features to corresponding outputs. In the classical setting where the precision of each data point is fixed, the famous Aitken/Gauss-Markov theorem in statistics states that generalized least squares (GLS) is a so-called “Best Linear Unbiased Estimator” (BLUE). In modern data science, however, one often faces *strategic data sources*; namely, individuals who incur a cost for providing high-precision data. For instance, this is the case for personal data, whose revelation may affect an individual’s privacy—which can be modeled as a cost—or in applications such as recommender systems, where producing an accurate estimate entails effort.

In this article, we study a setting in which features are public but individuals choose the precision of the outputs they reveal to an analyst. We assume that the analyst performs linear regression on this dataset, and individuals benefit from the outcome of this estimation. We model this scenario as a game where individuals minimize a cost composed of two components: (a) an (agent-specific) disclosure cost for providing high-precision data; and (b) a (global) estimation cost representing the inaccuracy in the linear model estimate. In this game, the linear model estimate is a public good that benefits all individuals. We establish that this game has a unique non-trivial Nash equilibrium. We study the efficiency of this equilibrium and we prove tight bounds on the price of stability for a large class of disclosure and estimation costs. Finally, we study the estimator accuracy achieved at equilibrium. We show that, in general, Aitken’s theorem does not hold under strategic data sources, though it does hold if individuals have identical disclosure costs (up to a multiplicative factor). When individuals have non-identical costs, we derive a bound on the improvement of the equilibrium estimation cost that can be achieved by deviating from GLS, under mild assumptions on the disclosure cost functions.

CCS Concepts: • **Mathematics of computing** → **Regression analysis**; • **Theory of computation** → **Quality of equilibria**; • **Security and privacy** → **Economics of security and privacy**; • **Theory of computation** → *Algorithmic game theory and mechanism design; Models of learning*;

This article is an extended version of “Linear regression as a non-cooperative game,” by Ioannidis and Loiseau [29].

This work was supported by the French National Research Agency (ANR) through the “Investissements d’avenir” program (ANR-15-IDEX-02), through MIAI @ Grenoble-Alpes (ANR-19-P3IA-0003), and through grant ANR-16-TERC0012; by the DGA; and by the Alexander von Humboldt Foundation. Stratis Ioannidis acknowledges support from NSF grants CCF-1750539 and CNS-1717213.

Authors’ addresses: N. Gast, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000, Grenoble, France; email: nicolas.gast@inria.fr; S. Ioannidis, Northeastern University, Boston, MA 02120; email: ioannidis@ece.neu.edu; P. Loiseau, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000, Grenoble, France, MPI-SWS, D-66123, Saarbrücken, Germany; email: patrick.loiseau@inria.fr; B. Roussillon, Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000, Grenoble, France; email: benjamin.roussillon@inria.fr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2167-8375/2020/05-ART10 \$15.00

<https://doi.org/10.1145/3391436>

Additional Key Words and Phrases: Linear regression, Aitken theorem, Gauss-Markov theorem, strategic data sources, potential game, price of stability

ACM Reference format:

Nicolas Gast, Stratis Ioannidis, Patrick Loiseau, and Benjamin Roussillon. 2020. Linear Regression from Strategic Data Sources. *ACM Trans. Econ. Comput.* 8, 2, Article 10 (May 2020), 24 pages.
<https://doi.org/10.1145/3391436>

1 INTRODUCTION

The statistical analysis of data is a cornerstone of many scientific disciplines. The core problem of estimating the parameters of a model is classic and well understood in the standard setting in which any noise or distortions present in the data are exogenous. Naturally, the quality of the data, as captured by such noise or distortions, is key to an estimator’s accuracy. In many instances, however, obtaining high quality data may be associated with a cost incurred by the data source. For example, this is the case when the data is of a personal nature and provided by privacy-conscious individuals. The quality of the data provided in this case can come at a cost of a violation of privacy [15, 24, 33]. The desire for privacy incentivizes individuals to obfuscate their private information, or, in the extreme, altogether refrain from any disclosure. An additional setting in which high-quality data may come at a cost is when the data quality depends on effort exerted [8, 48], i.e., improved quality is the result of increased effort expended by the data source. This setting naturally arises in, e.g., open collaboration projects, such as Wikipedia, but also in online recommender systems, where individuals need to exert effort (complete surveys, click “like” buttons, etc.) to disclose their preferences. Just as in the privacy case, a data source may choose to not exert the effort required to produce high-quality, low-noise responses or, in the extreme, altogether refrain from reporting anything meaningful.

In this setting, where providing high-quality data comes at a cost, it makes sense to consider strategic behavior among data sources. In particular, one should ask: Why would strategic data sources provide any data at all? The existing literature focuses either on the case where individuals receive a monetary compensation to provide data [1, 8, 15, 24, 35, 48] or on the case where individuals care about the quality of the estimation but only w.r.t. predictions over their own features [11, 18, 41]. We consider another possibility—namely, that *globally successful data analysis may also provide a utility to the individuals from which the data is collected*. This is evident in medical studies: An experiment may lead to the discovery of a treatment for a disease, from which an experiment subject may benefit. In the case of recommender systems, users may indirectly benefit from overall service improvements, as data disclosed may lead to, e.g., improved product recommendations or better-targeted advertising. Similarly, open collaboration projects, by their nature, implicitly assume a common underlying utility, linked to the success of the collaboration. If such benefits outweigh associated privacy or effort costs, then individuals may consent to the collection and analysis of high-quality data, e.g., by participating in a clinical trial, completing a survey, or disclosing their preferences in a recommender service.

In this article, we approach the above issues through a non-cooperative game, focusing on the basic statistical analysis task of *linear regression*. We consider the following formal setting: A set of individuals $i \in \{1, \dots, n\}$ participates in an experiment in which they are asked to provide data to an analyst. Each individual i is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$, capturing public information such as age, gender, and so on, and possesses a private variable $y_i \in \mathbb{R}$ —e.g., the true answer to a survey, the outcome of a medical test, or how much they like a product. The analyst wishes to perform linear regression over the data, i.e., compute a vector $\boldsymbol{\beta} \in \mathbb{R}^d$ such that:

$$y_i \approx \boldsymbol{\beta}^T \mathbf{x}_i, \quad \text{for all } i \in \{1, \dots, n\}.$$

We assume that individuals benefit from the correct estimation of β : for example, if the analyst learns β , an individual may benefit due to, e.g., better medical treatment, improved recommendations. However, individuals *do not* disclose their true private variables to the analyst. Instead, they provide a perturbed version \tilde{y}_i , constructed by *adding noise* to the private variable y_i . This is because there is a cost associated with the disclosure of the private variable: The higher the variance of the noise an individual adds, the lower the cost (e.g., due to privacy violation or effort exerted) she incurs. However, high noise variance lowers the accuracy of the analyst's estimate of β , the linear model computed in aggregate across multiple individuals. As such, the individuals need to strike a balance between the cost they incur through disclosure and the utility they accrue from accurate model prediction.

We make the following contributions:

- (i) We model interactions between data sources as a non-cooperative game, in which each data source selects the precision (i.e., the inverse of the noise variance) of the private variable she strategically discloses. A data source's decision minimizes a cost function made up of two components: (a) a *disclosure cost*, which is an increasing function of the chosen precision, and (b) an *estimation cost*, which decreases as the accuracy of the analyst's estimation of β increases. Formally, the estimation cost is a function of the covariance matrix of the estimate of β .
- (ii) We characterize the Nash equilibria of the above game. In particular, we show that it is a potential game and that, under appropriate assumptions on the disclosure and estimation costs, there exists a unique pure Nash equilibrium at which individual costs are finite.
- (iii) Armed with this result, we determine the game's efficiency, providing bounds for the price of stability for several cases of disclosure and estimation costs.
- (iv) Finally, we turn our attention to the analyst's estimation algorithm. In the presence of non-strategic data sources, the Aitken theorem¹ states that generalized least squares estimation yields minimal covariance among linear unbiased estimators. We challenge this theorem in the context of strategic data sources and obtain both positive and negative results:
 - a. We show that, in general, an equivalent of Aitken theorem no longer holds under strategic data sources. We exhibit a series of counter-examples that shows that, when data sources have non-identical disclosure cost functions, there exist linear estimators that lead to more efficient equilibria (i.e., that attain more accurate estimates at equilibrium) than generalized least squares.
 - b. We show that, when agents have monomial cost functions with identical exponents (but with possibly non-identical multiplicative factors), then generalized least squares is optimal among the class of unbiased linear estimators even in the strategic setting: It indeed yields the most accurate estimate at equilibrium.
 - c. Finally, even when generalized least squares is suboptimal, under mild assumptions on the disclosure cost functions, we show that the improvement of the equilibrium estimation cost that can be achieved by deviating from generalized least squares is bounded by a factor that depends on the heterogeneity of data source disclosure costs.

Our results imply that the optimality of the generalized least squares estimator does not persist if data sources strategically choose the variance of their data.

More broadly, we model the outcome of a statistical data analysis—the estimator's accuracy—as a *public good*: Data sources contribute to the public good by providing high-precision data at a disclosure cost and benefit in return from the global estimator's accuracy. As is natural in such public

¹ Also known as the Gauss-Markov theorem in the special case of ordinary least squares.

good games, we find that data sources typically contribute a level of data precision at equilibrium that is suboptimal from the social welfare perspective (i.e., there is partial free-riding). More surprisingly, though, we also find that under such strategic data sources, standard statistics results are challenged: It is sometimes possible to deviate from the standard estimator to increase the public good provision at equilibrium without involving any monetary compensation.

The remainder of this article is organized as follows: We present related work in Section 2. Section 3 contains a review of linear regression and the definition of our non-cooperative game. We characterize Nash equilibria in Section 4 and discuss their efficiency in Section 5. Our results on the optimality (or non-optimality) of generalized least squares are in Section 6, and our conclusions in Section 7. All proofs are relegated to appendices.

2 RELATED WORK

Data Perturbation for Privacy. Perturbing a dataset before submitting it as input to a data mining algorithm has a long history in privacy-preserving data-mining (see, e.g., References [19, 47]). Independent of an algorithm, early research focused on perturbing a dataset prior to its public release [21, 46]. Perturbations tailored to specific data mining tasks have also been studied in the context of, e.g., reconstructing the original distribution of the underlying data [2], building decision trees [2], clustering [40], and association rule mining [4]. We approach such perturbation techniques via a non-cooperative setting, where individuals strategically choose the perturbation to their data.

The above setting differs from the framework of ϵ -differential privacy [22, 32], which has also been studied from the perspective of mechanism design [39]. In differential privacy, noise is added to the *output* of a computation, which is subsequently publicly released. The analyst performing the computation is *a priori* trusted; as such, individuals submit unadulterated inputs. Several works study mechanisms incentivizing data disclosure under costs quantified by differential privacy [15, 16, 24, 33], whereby individuals are compensated for the privacy cost they incur. In contrast, we do not assume that the analyst is trusted, which motivates input perturbation. Such input perturbations also correspond to the more recently studied notion of local differential privacy [20, 31], though such studies focus on the privacy/utility tradeoff, ignoring the strategic aspect of the input perturbation.

Strategic Data Sources and Data Elicitation. A few recent works consider settings where sources may choose their effort when generating data [8, 35, 48] or have heterogeneous costs due to disclosure [1]. In all these works, data sources are assumed to maximize the payment received (minus cost of effort). We note that in Cai et al. [8] and Westenbroek et al. [48], which are the closest to our work, the disclosure costs of data sources are linear in the exerted effort, whereas we use more general convex costs. The data elicitation literature also includes related problems, in which one tries to incentivize an expert to truthfully reveal her prediction of an event, typically using scoring rules [23] (see also the literature on incentives in crowdsourcing [17]).

A number of papers also consider data acquisition in sequential settings [1, 10, 34]. All this literature, however, considers agents that aim to maximize the payment received but are insensitive to the quality of the learning result. Moreover, agents aim to optimize payments while the learning algorithm is fixed; the only exceptions to the latter are References [9, 14], which are restricted to the case of averaging and do not consider learning tasks such as regression. In contrast, in this work, we do not involve payments but assume that data sources benefit from the result of the learning algorithm.

Strategy-proof Statistical Inference. Several papers study regression from the perspective of mechanism design, whereby private variables are directly reported by strategic agents. In

particular, Dekel et al. [18] consider a broad class of regression problems in which data sources may misreport their private values and determine loss functions under which empirical risk minimization is group strategyproof. The special case of linear regression is also treated, albeit in a more restricted setting, by Perote and Perote-Peña [41], who identify more general strategyproof mechanisms for the 2-dimensional case. More recently, Chen et al. [11] consider a similar setting and propose a family of group strategyproof regression mechanisms for any dimension, extending the results of both References [18] and [41]. As in our article, those works assume that the independent variables (the \mathbf{x}_i 's) are public information and mostly look at mechanism design without money. Several papers also analyze similar problems in the case of classification [25, 36] (see also a recent variant in Reference [6]).

In contrast to this line of research, our work assumes that the analyst uses a fixed algorithm (GLS or a linear unbiased estimator). We also assume that individuals choose the precision of the data reported (and not directly the reported value), and no design is required, as the chosen precision is assumed known. The main difference, however, is conceptual: In References [11, 18, 41], agents care about the estimation error on their instance only, whereas we assume that agents benefit equally from the global downstream effects of an accurate predictor. We consider noise addition as a non-cooperative game, focusing on pure Nash equilibrium as a solution concept and studying its efficiency.

Non-cooperative Regression Games. Closer to our setting, Hossain and Shah [28] also consider the pure Nash equilibrium as a solution concept in regression games and investigate its efficiency, albeit in a model closer to References [11, 18] than to ours. Interestingly, this work considers the mean squared error, a standard quantity to measure a model's quality in linear regression, instead of our estimation cost based on the covariance matrix. Our estimation cost, however, includes a somewhat broader family of functions satisfying mild assumptions (see Assumptions 2 and 4).

Our article is an extended version of "Linear regression as a non-cooperative game," by Ioannidis and Loiseau [29]. We generalize and tighten the price of stability results and correct Theorem 6 of Reference [29], which stated that GLS is optimal for all cost functions. We show that this result is not true in general but that (i) it holds when agents have identical cost functions (up to a multiplicative constant) and (ii) the sub-optimality of GLS in the case of non-identical disclosure cost functions can be bounded under mild assumptions.

Experimental Design. In classic experimental design [5, 7, 42], an analyst observes the public features of a set of experiments and determines which experiments to conduct with the objective of learning a linear model (from non-strategic sources). The quality of an estimated model is quantified through a scalarization of its variance [7]. As discussed in Section 3.3, many such scalarizations are used in the literature, including the so-called A-optimality, C-optimality, and D-optimality criteria we define in Equation (4). We focus on non-negative scalarizations to ensure meaningful notions of efficiency (as determined by the price of stability in Section 5). Among these classic scalarizations, A-optimality and E-optimality satisfy both our technical assumptions (Assumptions 2 and 4), while D-optimality satisfies only our convexity assumption (Assumption 2). As we note in Section 3.3, convexity implies that the information gain (i.e., the cost reduction) due to new experiments is a submodular function. This has implications about mechanism design as well. For example, Horel et al. [27] exploit this to produce a polytime mechanism with approximation guarantees for a version of the experimental design problem in which subjects report their private values truthfully but may lie about the costs they require for their participation.

Public Good Provision Problems. We finally note that our model has analogies to models used in *public good provision problems* (see, e.g., Reference [38] and references therein). Indeed, the estimate variance reduction can be seen as a public good in that, when a source contributes data,

all other sources in the game benefit. As is standard in such literature, our model assumes that the disclosure costs (corresponding to provision costs in public good problems) and the estimation cost (mapping to the public good benefit) are fully separable. This analogy is pushed further in References [12, 13], where the authors propose a simple mechanism to increase the provision of the public good at equilibrium in the simple case of averaging (corresponding to the standard public good framework).

3 MODEL DESCRIPTION

In this section, we give a detailed description of our linear regression game and the agents involved. Before discussing strategic considerations, we give a brief technical review of linear models, as well as key properties of least squares estimators; all related results presented here are classic (see, e.g., Reference [26]).

Notational conventions. We use boldface type (e.g., \mathbf{x} , \mathbf{y} , $\boldsymbol{\beta}$) to denote vectors (all vectors are column vectors) and capital letters (e.g., A , B , V) to denote matrices. As usual, we denote by $S_+^d, S_{++}^d \subset \mathbb{R}^{d \times d}$ the sets of (symmetric) positive semidefinite (PSD) and positive definite matrices of size $d \times d$, respectively. For two positive semidefinite matrices $A, B \in S_+^d$, we write that $A \geq B$ if $A - B \in S_+^d$; recall that \geq defines a partial order over S_+^d . We say that $F : S_+^d \rightarrow \mathbb{R}$ is non-decreasing in the positive semidefinite order if $F(A) \geq F(A')$ for any two $A, A' \in S_+^d$ such that $A \geq A'$. Moreover, we say that a matrix-valued function $F : \mathbb{R}^n \rightarrow S_+^d$ is *matrix convex* if $\alpha F(\boldsymbol{\lambda}) + (1 - \alpha)F(\boldsymbol{\lambda}') \geq F(\alpha\boldsymbol{\lambda} + (1 - \alpha)\boldsymbol{\lambda}')$ for all $\alpha \in [0, 1]$ and $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \mathbb{R}^n$.

3.1 Linear Models

Consider a set of n data sources, henceforth referred to as *agents*, denoted by $N \equiv \{1, \dots, n\}$. Each agent $i \in N$ is associated with a vector $\mathbf{x}_i \in \mathbb{R}^d$, the *feature vector*, which is public; for example, this vector may correspond to publicly available demographic information about the agent, such as age, gender, and so on. Each $i \in N$ is also associated with a private variable $y_i \in \mathbb{R}$; for example, this may express the likelihood that this agent contracts a disease, the concentration of a substance in her blood, or a true answer to a survey by that agent.

We assume that the agent's private variable y_i is a linear function of her public features \mathbf{x}_i . In particular, there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^d$, the *model*, such that the private variables are given by

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i, \quad \text{for all } i \in N, \quad (1)$$

where the “inherent noise” variables $\{\epsilon_i\}_{i \in N}$ are i.i.d.² zero-mean random variables in \mathbb{R} with finite variance σ^2 . We make no further assumptions on the noise; in particular, we *do not* assume it is Gaussian.

An *analyst* wishes to observe the y_i 's and infer the model $\boldsymbol{\beta} \in \mathbb{R}^d$. This type of inference is ubiquitous in experimental sciences and has a variety of applications. For example, the magnitude of $\boldsymbol{\beta}$'s coordinates captures the effect that features (e.g., age or weight) have on y_i (e.g., the propensity to get a disease), while the sign of a coordinate captures positive or negative correlation. Knowing $\boldsymbol{\beta}$ can also aid in prediction: An estimate of private variable $y \in \mathbb{R}$ of a new sample with features $\mathbf{x} \in \mathbb{R}^d$ is given by the inner product $\boldsymbol{\beta}^T \mathbf{x}$. We note that the linear relationship between y_i and \mathbf{x}_i expressed in Equation (1) is in fact quite general. For example, the case where $y_i = f(\mathbf{x}) + \epsilon_i$, where f is a polynomial function of degree 2, reduces to a linear model by considering the transformed feature space whose features compose the monomials $x_{ik}x_{ik'}$, for $1 \leq k, k' \leq d$. More generally,

²To ease notation, we assume that the variance of the inherent noise σ^2 is identical for all agents, but all results of the article remain valid if we allow this variance (or equivalently the upper bound on the precision $1/\sigma^2$, see below) to depend on the identity of Agent i .

the same principle can be applied to reduce to Equation (1) any function class spanned by a finite set of basis functions over \mathbb{R}^d [26].

3.2 Generalized Least Squares Estimation

We consider a setup in which the agents choose the precision of the data that they provide. That is, they do not directly provide y_i but rather a perturbed variable \tilde{y}_i , which we assume is an unbiased estimate of y_i with variance σ_i^2 . For example, in the case of privacy, agents *distort* their private variable by adding excess noise: Each $i \in N$ computes $\tilde{y}_i = y_i + z_i$ where z_i is a zero-mean random variable with variance σ_i^2 ; we assume that $\{z_i\}_{i \in N}$ are independent and are also independent of the inherent noise variables $\{\epsilon_i\}_{i \in N}$. In the case of effort, the variance σ_i^2 captures the effort exerted by the agent in generating the label \tilde{y}_i . Each agent reveals to the analyst (a) the perturbed variable \tilde{y}_i and (b) the variance σ_i^2 . As a result, the aggregate variance of the reported value is $\sigma^2 + \sigma_i^2$ and its precision (the inverse of the aggregate variance) is $\lambda_i \equiv \frac{1}{\sigma^2 + \sigma_i^2}$.

Note that, as a consequence of the above description, our model assumes that the analyst can observe the (true) precision of the private data revealed by the analyst. This is reasonable in settings where the data are stored in a trusted database and the agent grants access to it under a given precision, and the noise is added by a third party (e.g., the database itself). In medical research, for instance, one can imagine that the data are stored in a hospital database; a patient would then grant access to it with a given precision and the analyst would receive the perturbed data directly from the hospital. In other applications, such as surveys, one can also imagine that, rather than providing a specific value, agents would provide an interval whose size naturally translates to precision.

In turn, having access to the perturbed variables \tilde{y}_i , $i \in N$, and the corresponding precisions, the analyst estimates β through *generalized least squares* (GLS) estimation. Denote by $\lambda = [\lambda_i]_{i \in N}$ the vector of precisions and by $\Lambda = \text{diag}(\lambda)$ the diagonal matrix whose diagonal is given by vector λ . Then, the generalized least squares estimator is given by:

$$\hat{\beta}_{\text{GLS}} = \arg \min_{\beta \in \mathbb{R}^d} \left(\sum_{i \in N} \lambda_i (\tilde{y}_i - \beta^T \mathbf{x}_i)^2 \right) = (X^T \Lambda X)^{-1} X^T \Lambda \tilde{\mathbf{y}}, \quad (2)$$

where $\tilde{\mathbf{y}} = [\tilde{y}_i]_{i \in N}$ is the n -dimensional vector of perturbed variables, and $X = [\mathbf{x}_i^T]_{i \in N} \in \mathbb{R}^{n \times d}$ the $n \times d$ matrix whose rows compose the transposed feature vectors. Throughout our analysis, we assume that $n \geq d$ and that X has rank d .

Note that $\tilde{\mathbf{y}} \in \mathbb{R}^n$ is a random variable and as such, by Equation (2), so is $\hat{\beta}_{\text{GLS}}$. It can be shown that $\mathbb{E}(\hat{\beta}_{\text{GLS}}) = \beta$ (i.e., $\hat{\beta}_{\text{GLS}}$ is unbiased), and

$$V(\lambda) \equiv \text{Cov}(\hat{\beta}_{\text{GLS}}) = \mathbb{E} \left[(\hat{\beta}_{\text{GLS}} - \beta)^T (\hat{\beta}_{\text{GLS}} - \beta) \right] = (X^T \Lambda X)^{-1}.$$

The covariance V captures the uncertainty of the estimation of β . The matrix

$$A(\lambda) \equiv X^T \Lambda X = \sum_{i \in N} \lambda_i \mathbf{x}_i \mathbf{x}_i^T$$

is known as the *precision* matrix. It is positive semidefinite, i.e., $A(\lambda) \in S_+^d$, but it may not be invertible: This is the case when $\text{rank}(X^T \Lambda) < d$, i.e., the vectors \mathbf{x}_i , $i \in N$, for which $\lambda_i > 0$, do not span \mathbb{R}^d . Put differently, if the set of agents providing useful information does not include d linearly independent vectors, there exists a direction $\mathbf{x} \in \mathbb{R}^d$ that is a “blind spot” to the analyst: The analyst has no way of predicting the value $\beta^T \mathbf{x}$. In this degenerate case the number of solutions to the least squares estimation problem (2) is infinite, and the covariance is not well-defined (it is infinite in all such directions \mathbf{x}). Note, however, that, since X has rank d (and hence $X^T X$ is

invertible), the set of λ for which the precision matrix is invertible is non-empty. In particular, it contains $(0, 1/\sigma^2]^n$, since $A(\lambda) \in S_{++}^d$ if $\lambda_i > 0$ for all $i \in N$.

3.3 Non-cooperative Game Model of Strategic Data Sources

The perturbed variables \tilde{y}_i are motivated by the fact that strategic data sources incur a cost to provide high-precision data—for instance, due to privacy concerns, an agent may be reluctant to grant unfettered access to her private variable or release it in the clear. However, it may be to the agent's advantage that the analyst learns the model β . In our running medical example, learning that, e.g., a disease is correlated to an agent's weight or her cholesterol level may lead to a cure, which in turn may be beneficial to the agent.

We model the above considerations through cost functions. Recall that the action of each agent $i \in N$ amounts to choosing the noise level of the perturbation, captured by the variance $\sigma_i^2 \in [0, \infty]$. For notational convenience, we use the equivalent representation $\lambda_i = 1/(\sigma^2 + \sigma_i^2) \in [0, 1/\sigma^2]$ for the action of an agent. Note that $\lambda_i = 0$ (or, equivalently, infinite variance σ_i^2) corresponds to no participation: In terms of estimation through Equation (2), it is as if this perturbed value is not reported.

Each agent $i \in N$ chooses her action $\lambda_i \in [0, 1/\sigma^2]$ to minimize her cost

$$J_i(\lambda_i, \lambda_{-i}) = c_i(\lambda_i) + f(\lambda), \quad (3)$$

where we use the standard notation λ_{-i} to denote the collection of actions of all agents but i . The cost function $J_i : \mathbb{R}_+^n \rightarrow \mathbb{R}_+$ of agent $i \in N$ composes two non-negative components. We refer to the first component $c_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ as the *disclosure cost*: It is the cost that the agent incurs for providing the perturbed variable. The second component is the *estimation cost*, and we assume that it takes the form $f(\lambda) = F(V(\lambda))$, if $A(\lambda) \in S_{++}^d$, and $f(\lambda) = \infty$ otherwise. The mapping $F : S_{++}^d \rightarrow \mathbb{R}_+$ is known as a *scalarization* [7]. It maps the covariance matrix $V(\lambda)$ to a scalar value $F(V(\lambda))$ and captures how well the analyst can estimate the model β . The estimation cost $f : \mathbb{R}_+^n \rightarrow \mathbb{R}_+ = \mathbb{R}_+ \cup \{\infty\}$ is the so-called *extended-value extension* of $F(V(\lambda))$: It equals $F(V(\lambda))$ in its domain and $+\infty$ outside its domain.

Main Assumptions. Throughout our analysis, we make the following two assumptions:

ASSUMPTION 1. The disclosure costs $c_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, $i \in N$, are non-negative, continuous, non-decreasing, and convex.

ASSUMPTION 2. The scalarization $F : S_{++}^d \rightarrow \mathbb{R}_+$ is non-negative, continuous, increasing in the positive semidefinite order, and convex.

The monotonicity assumptions in Assumptions 1 and 2 are standard and natural. Increasing the precision λ_i leads to a higher disclosure cost. In contrast, increasing λ_i can only decrease the estimation cost: This is because decreasing the variance of an agent's provided perturbed variable also decreases the variance in the positive semidefinite sense (as the matrix inverse is a PSD-decreasing function).

The convexity assumption in Assumption 2 is also standard and natural. Intuitively, the naturalness of Assumption 2 stems from the following observation: The convexity of F implies that the so-called *information gain*, i.e., the relative reduction in F as a new label is collected, exhibits a diminishing returns property, as additional labels affect estimation quality less and less. Scalarizations of positive semidefinite matrices and, in particular, of the covariance matrix $V(\lambda)$, are abundant in statistical inference literature in the context of experimental design [5, 7, 42] (also known as batch active learning). Similar to our setting, in experimental design an analyst has access to samples with known feature vectors $x_i \in \mathbb{R}^d$, $i \in N$ and wishes to conduct a limited number of

k experiments, where $k \ll N$, to collect labels $y_i \in \mathbb{R}$ for a subset of these samples. Given budget k , the experimental design problem amounts to determining which labels to collect. The standard approach is to accomplish this by minimizing a scalarization function of the covariance of the estimator applied to the labels selected [5, 7, 42]. Three examples of such estimators encountered often in practice are the so-called A-optimality, E-optimality, and D-optimality criteria:

$$F_1(V) = \text{trace}(V), \quad F_2(V) = \|V\|_F^2, \quad F_3(V) = \log \det(I + V), \quad (4)$$

where $\|\cdot\|_F$ is the Frobenius norm and I is the identity matrix. All three scalarizations satisfy Assumption 2. The convexity of these scalarizations implies that, if repetitions are allowed (i.e., an experiment can be conducted multiple times), the analyst can determine which fraction of her experiments should be performed on a given sample by solving a convex optimization problem (see, e.g., Reference [7]). However, if repetitions are not allowed, convexity implies that experimental design can be cast as a submodular maximization problem subject to cardinality constraints (see, e.g., Reference [27]), which is NP-hard for the objectives in Equation (4) but admits a poly-time approximation. Submodular maximization arises precisely due to the aforementioned diminishing returns property of the information gain under new labels; this, in turn, is a direct consequence of Assumption 2.

Note that, as a further consequence of Assumption 2, the extended-value extension f is convex (in λ). The convexity of $F(V(\cdot))$ follows from the fact that it is the composition of the increasing convex function $F(\cdot)$ with the matrix convex function $V(\cdot)$; the latter is convex, because the matrix inverse is matrix convex and the precision matrix $A(\lambda)$ is an affine function of λ .

Additional Assumptions. Our result on Nash equilibrium existence and uniqueness (Theorem 1) relies on Assumptions 1 and 2. Our bounds on the price of stability (Theorem 3) and our Aitken-type results (Theorem 5) use two additional assumptions that further constrain the shape of the disclosure costs and the scalarization function:

ASSUMPTION 3. *There exist $1 \leq p_{\min} \leq p_{\max} \in \mathbb{R}_+ \cup \{+\infty\}$ such that, for all $i \in N$, the disclosure cost $c_i : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ satisfies:*

$$a^{p_{\min}} c_i(\lambda) \leq c_i(a\lambda) \leq a^{p_{\max}} c_i(\lambda), \quad \text{for all } \lambda \in \mathbb{R}_+ \text{ and } a \geq 1. \quad (5)$$

ASSUMPTION 4. *There exists $q \geq 1$ such that the scalarization $F : S_{++}^d \rightarrow \mathbb{R}_+$ is q -homogeneous, i.e., it satisfies:*

$$F(aM) = a^q F(M), \quad \text{for all } M \in S_{++}^d \text{ and } a \geq 1. \quad (6)$$

Intuitively, Assumption 3 captures “near-homogeneity” of the disclosure cost functions. It is, for example, satisfied when all agents have monomial disclosure costs $c_i(\lambda) = r_i \lambda^{p_i}$, where r_i is a constant, with different exponents $p_i \in [p_{\min}, p_{\max}]$. Assumption 4 is also a homogeneity assumption. It holds for a broad class of interesting scalarizations, such as any norm taken to any power. In particular, it holds for F_1 and F_2 in Equation (4), i.e., the trace and the squared Frobenius norm (with $q = 1$ and $q = 2$, respectively), which are classical scalarizations in the statistical inference literature in the context of experimental design [5, 7, 42]. Note that Assumption 4 also implies that $f(a\lambda) = a^{-q} f(\lambda)$, for all $\lambda \in [0, 1/\sigma^2]^n$ and $a \geq 1$.

Game notation. We denote by $\Gamma = \langle N, [0, 1/\sigma^2]^n, (J_i)_{i \in N} \rangle$ the game with set of agents $N = \{1, \dots, n\}$, where each agent $i \in N$ chooses her action λ_i in her action set $[0, 1/\sigma^2]$ to minimize her cost $J_i : [0, 1/\sigma^2]^n \rightarrow \mathbb{R}_+$, given by Equation (3). We refer to a $\lambda \in [0, 1/\sigma^2]^n$ as a *strategy profile* of the game Γ . We analyze the game as a *complete information game*, i.e., we assume that the set of agents, the action sets, and utilities are known by all agents.

4 NASH EQUILIBRIA

We begin our analysis by characterizing the Nash equilibria of the game Γ . In the game Γ , each agent chooses her contribution λ_i to minimize her cost. A Nash equilibrium (in pure strategy) is a strategy profile λ^* satisfying

$$\lambda_i^* \in \arg \min_{\lambda_i} J_i(\lambda_i, \lambda_{-i}^*), \quad \text{for all } i \in N.$$

Observe first that Γ is a potential game [37]. Indeed, define the function $\Phi : [0, 1/\sigma^2]^n \rightarrow \bar{\mathbb{R}}$ such that

$$\Phi(\lambda) = f(\lambda) + \sum_{i \in N} c_i(\lambda_i), \quad (\lambda \in [0, 1/\sigma^2]^n). \quad (7)$$

Then for every $i \in N$ and for every $\lambda_{-i} \in [0, 1/\sigma^2]^{n-1}$, we have

$$J_i(\lambda_i, \lambda_{-i}) - J_i(\lambda'_i, \lambda_{-i}) = \Phi(\lambda_i, \lambda_{-i}) - \Phi(\lambda'_i, \lambda_{-i}), \quad \forall \lambda_i, \lambda'_i \in [0, 1/\sigma^2]. \quad (8)$$

Therefore, Γ is a potential game with potential function Φ . From Equation (8), we see that (as for any convex potential game) the set of Nash equilibria coincides with the set of local minima of function Φ .

Note that there may exist Nash equilibria λ^* for which $f(\lambda^*) = \infty$. For instance, if $d \geq 2$, $\lambda^* = 0$ is a Nash equilibrium. Indeed, in that case, no agent has an incentive to deviate, since a single $\lambda_i > 0$ still yields a non-invertible precision matrix $A(\lambda)$. In fact, any profile λ for which $A(\lambda)$ is non-invertible, and remains so under unilateral deviations, is an equilibrium. We call such Nash equilibria (at which the estimation cost is infinite) *trivial*. Existence of trivial equilibria can be avoided using slight model adjustments: For instance, one can alter the game definition to disallow infinite variances. Alternatively, the existence of d non-strategic agents whose feature vectors span \mathbb{R}^d is also sufficient to enforce a finite covariance at all λ across strategic agents.

In the remainder, we focus on the more interesting *non-trivial* equilibria. Using the potential game structure of Γ , we derive the following result:

THEOREM 1. *Under Assumptions 1 and 2, there exists a unique non-trivial equilibrium of the game Γ .*

This result is proved in Appendix A. The potential game structure of Γ has another interesting implication: If agents start from an initial strategy profile λ such that $f(\lambda) < \infty$, the so-called *best-response dynamics* converge towards the unique non-trivial equilibrium (see, e.g., Reference [44]). This implies that the non-trivial equilibrium is the only equilibrium reached when, e.g., all agents start with non-infinite noise variance.

5 PRICE OF STABILITY

Having established the uniqueness of a non-trivial equilibrium in our game, we turn our attention to issues of efficiency. We define the *social cost* function $C : \mathbb{R}^n \rightarrow \mathbb{R}_+$ as the sum of all agent costs and say that a strategy profile λ^{opt} is *socially optimal* if it minimizes the social cost, i.e.,

$$C(\lambda) = \sum_{i \in N} c_i(\lambda_i) + nf(\lambda), \quad \text{and} \quad \lambda^{\text{opt}} \in \arg \min_{\lambda \in [0, 1/\sigma^2]^n} C(\lambda).$$

Let $\text{opt} = C(\lambda^{\text{opt}})$ be the minimal social cost. We define the *price of stability* (*price of anarchy*) as the ratio of the social cost of the best (worst) Nash equilibrium in Γ to opt , i.e.,

$$\text{PoS} = \min_{\lambda \in \text{NE}} \frac{C(\lambda)}{\text{opt}}, \quad \text{and} \quad \text{PoA} = \max_{\lambda \in \text{NE}} \frac{C(\lambda)}{\text{opt}},$$

where $NE \subseteq [0, 1/\sigma^2]^n$ is the set of Nash equilibria of Γ . Clearly, in the presence of trivial equilibria, the price of anarchy is infinity. We thus turn our attention to determining the price of stability. Note, however, that since the non-trivial equilibrium is unique (Theorem 1), the price of stability and the price of anarchy coincide under the slight model adjustments discussed in Section 4 that eliminate trivial equilibria.

The fact that our game admits a potential function has the following immediate consequence (see, e.g., References [44, 45]):

THEOREM 2. *Under Assumptions 1 and 2, $\text{PoS} \leq n$.*

The proof of Theorem 2 can be found in Appendix B. Improved bounds can be obtained for specific estimation and disclosure cost functions. The following result provides tighter bounds when the disclosure costs and the scalarization satisfy Assumptions 3 and 4.

THEOREM 3. *In addition to Assumptions 1 and 2, assume that the disclosure cost functions satisfy Assumption 3 with $p_{\min} \geq 1$ and $p_{\max} \in \mathbb{R} \cup \{\infty\}$ and that the scalarization F satisfies Assumption 4 for some $q \geq 1$. Then, the price of stability satisfies $\text{PoS} \leq n^{\frac{q}{p_{\min}+q}}$. Additionally, for all $p_{\min}, q \geq 1$, and all $\varepsilon > 0$, there exists a game in which the estimation cost and the disclosure costs satisfy Assumptions 3 and 4, respectively, such that $\text{PoS} \geq n^{\frac{q}{p_{\min}+q}}(1 - \varepsilon)$.*

The proof of Theorem 3 can be found in Appendix C. Note that, as the bound does not depend on p_{\max} , we can set $p_{\max} = \infty$ in Assumption 3, which is equivalent to replacing this assumption by

$$a^{p_{\min}} c_i(\lambda) \leq c_i(a\lambda), \quad \text{for all } \lambda \in \mathbb{R}_+ \text{ and } a \geq 1.$$

The proof of the upper bound relies on deriving a “good” solution from the social optimum and showing that, if the PoS is too high, this “good” solution attains a lower potential than a Nash equilibrium (a contradiction). The proof of the lower bound in Theorem 3 relies on explicitly characterizing the socially optimal profile in a certain game class and showing it equals the Nash equilibrium λ^* multiplied by a scalar. We note that the theorem states that, among monomial disclosure costs and for any estimation cost satisfying Assumption 4, the largest PoS is $n^{\frac{q}{1+q}}$ and is attained for linear disclosure costs. Similarly, among all estimation costs satisfying Assumption 4 and all disclosure costs satisfying the assumptions presented in Theorem 3, the largest PoS is n ; this is approached as q tends to infinity.

We note that a similar worst-case efficiency of linear functions among convex cost families has also been observed in the context of other games, including routing [43] and resource allocation games [30]. As such, Theorem 3 indicates that this behavior emerges in our linear regression game as well but only concerning the disclosure cost: We observe a worst-case efficiency of linear functions in this game for the disclosure cost but a worst-case efficiency of highly convex functions for the estimation cost.

6 AN AITKEN-TYPE THEOREM FOR NASH EQUILIBRIA

Until this point, we have assumed that the analyst uses the generalized least-square estimator (2) to estimate model β . In the non-strategic case, where λ (and, equivalently, the added noise variance) is fixed, the generalized least-square estimator is known to satisfy a strong optimality property: The so-called Aitken/Gauss-Markov theorem, which we briefly review below, states that it is a “Best Linear Unbiased Estimator,” a property commonly referred to as BLUE. In this section, we investigate how this result extends in the strategic case, i.e., when λ^* is not *a priori* fixed but is the equilibrium reached by agents; crucially, the latter depends on the estimator used by the analyst.

6.1 Linear Unbiased Estimators and the Aitken Theorem

A linear estimator $\hat{\beta}_L$ of the model β is a linear map of the perturbed variables \tilde{y} ; i.e., it is an estimator that can be written as $\hat{\beta}_L = L\tilde{y}$ for some matrix $L \in \mathbb{R}^{d \times n}$. A linear estimator is called *unbiased* if $\mathbb{E}[L\tilde{y}] = \beta$ (the expectation taken over the inherent and extra noise). Recall by Equation (2) that the generalized least-square estimator $\hat{\beta}_{\text{GLS}}$ is an unbiased linear estimator with $L = (X^T \Lambda X)^{-1} X^T \Lambda$ and covariance $\text{Cov}(\hat{\beta}_{\text{GLS}}) = (X^T \Lambda X)^{-1}$. Any linear estimator $\hat{\beta}_L = L\tilde{y}$ can be written without loss of generality as

$$L = (X^T \Lambda X)^{-1} X^T \Lambda + D^T, \quad (9)$$

where

$$D = D(X) \in \mathbb{R}^{d \times n}, \quad (10)$$

is a matrix that may depend on X but does not depend on Λ . It is easy to verify that $\hat{\beta}_L$ is unbiased if and only if

$$D^T X = 0; \quad (11)$$

in turn, using this result, the covariance of any linear unbiased estimator can be shown to be

$$\text{Cov}(\hat{\beta}_L) = (X^T \Lambda X)^{-1} + D^T \Lambda^{-1} D \succeq \text{Cov}(\hat{\beta}_{\text{GLS}}). \quad (12)$$

In other words, the covariance of the generalized least-square estimator is minimal in the positive-semidefinite order among the covariances of *all linear unbiased estimators*. This optimality result is known as the Aitken theorem [3]. Applied specifically to homoscedastic noise (i.e., when all noise variances are identical), it is known as the Gauss-Markov theorem [26], which establishes the optimality of the ordinary least squares estimator. Both theorems provide a strong argument in favor of using least squares to estimate β in the presence of fixed noise variance (i.e., non-strategic agents).

6.2 Extension of the Non-cooperative Game to Linear Unbiased Estimators

Suppose now that the data analyst uses a linear unbiased estimator $\hat{\beta}_L$ with a given matrix $L \in \mathbb{R}^{n \times d}$ that may depend on X . Similarly to the model introduced in Section 3.3, we define a game Γ_L in which each agent i chooses her λ_i to minimize her cost; this time, however, the estimation cost depends on the variance of $\hat{\beta}_L$. A natural question to ask is the following: Is it possible that, despite the fact that the analyst uses an estimator that is “inferior” to $\hat{\beta}_{\text{GLS}}$ in the BLUE sense, an equilibrium reached under $\hat{\beta}_L$ is *better* than the equilibrium reached under $\hat{\beta}_{\text{GLS}}$ in terms of equilibrium estimation cost? If so, despite the Aitken theorem, the data analyst would have an incentive to use $\hat{\beta}_L$ instead and to inform the agents that she will use $\hat{\beta}_L$ and not $\hat{\beta}_{\text{GLS}}$.

In this section, we provide both a positive and a negative answer to this question, depending on specific assumptions on the disclosure costs. Formally, we consider the game $\Gamma_L = \langle N, [0, 1/\sigma^2]^n, (J_i)_{i \in N} \rangle$ defined as in Section 3.3, except that the estimation cost is the extended-value extension of $F(V_L(\lambda))$ with

$$V_L(\lambda) \equiv (X^T \Lambda X)^{-1} + D^T \Lambda^{-1} D, \quad (\Lambda = \text{diag } \lambda), \quad (13)$$

where D is defined as in Equation (11).

Recall that D may depend on X but does not depend on the precision Λ . Then, Γ_L is still a potential game with potential function $\Phi(\lambda) = f_L(\lambda) + \sum_{i \in N} c_i(\lambda_i)$ where $f_L(\lambda) = F(V_L(\lambda))$. This potential function has the same form as the potential of the original game, given by Equation (7). Moreover, as the function $V_L(\cdot)$ given by Equation (13) is a matrix convex function, the extended-value extension $f_L(\cdot)$ is convex. This shows that the potential function is convex. Since the proof

of Theorem 1 mostly relies on the convexity of the potential, a straightforward adaptation yields the following result:

THEOREM 4. *Under Assumptions 1 and 2, for any linear estimator $L \in \mathbb{R}^{n \times d}$, there exists a unique non-trivial equilibrium of the game Γ_L .*

As for the case of GLS, this result follows from the uniqueness of a minimizer of the potential function attained in the effective domain. In what follows, we denote the unique non-trivial equilibrium of Γ_L by λ_L^* and we denote by λ_{GLS}^* the equilibrium of the game Γ with the same parameters except for the estimator.

6.3 Optimality of GLS

6.3.1 Theoretical Bound and Optimality Condition. For a given linear unbiased estimator L , the estimation cost at equilibrium is $f_L(\lambda_L^*)$. We say that a linear estimator is efficient if it provides a small estimation cost at equilibrium. In the following theorem, we provide both a negative and a positive result about the efficiency of GLS: On the one hand, GLS is not always the most efficient estimator; on the other hand, under Assumptions 3 and 4, the ratio between the estimation cost at equilibrium of GLS and any other estimator is bounded by $\frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})}$; in other words, GLS is never too far from the most efficient estimator.

THEOREM 5. *Assume that the disclosure cost and scalarization functions satisfy Assumptions 1 and 2. Then:*

- (i) *There exists a game Γ such that GLS is not the most efficient estimator; i.e., there exists an unbiased linear estimator L such that, for these game parameters,*

$$f_L(\lambda_L^*) < f_{\text{GLS}}(\lambda_{\text{GLS}}^*).$$

- (ii) *For all games that additionally satisfy Assumptions 3 and 4, GLS is $\frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})}$ -optimal, i.e., for all unbiased estimators L*

$$f_{\text{GLS}}(\lambda_{\text{GLS}}^*) \leq \frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})} f_L(\lambda_L^*).$$

The proof is provided in Appendix D. Note that the bound in Theorem 5(ii) is clearly smaller than or equal to p_{\max}/p_{\min} . By remarking that it can be written as $(1 + \frac{q}{p_{\min}})/(1 + \frac{q}{p_{\max}})$, it is easy to see that it is also smaller than or equal to $1 + q$. This shows that GLS is p_{\max}/p_{\min} -optimal for any q and $(1 + q)$ -optimal for any p_{\min}, p_{\max} . Note also that Theorem 5(ii) trivially implies the following:

COROLLARY 1. *Under Assumptions 1 and 2, Assumption 3 with $p_{\min} = p_{\max} = p$, and Assumption 4, GLS is the most efficient estimator.*

Note that $p_{\min} = p_{\max} = p$, which literally translates to $c_i(a\lambda) = a^p c_i(\lambda)$ for all $i \in N, \lambda \in \mathbb{R}_+$ and $a \geq 1$, means that all agents have monomial cost functions with the same exponent. Put differently, for all i , there exists a constant $r_i > 0$ such that $c_i(\lambda_i) = r_i \lambda_i^p$. Theorem 5(i) may seem counter-intuitive, as GLS is optimal in the case of non-strategic agents: By Aitken's theorem, if precisions are fixed and known, then the best linear unbiased estimator is GLS, i.e., for all λ : $f_L(\lambda) > f_{\text{GLS}}(\lambda)$. Our result demonstrates that this is not the case with strategic agents.

6.3.2 Numerical Illustration of the Non-optimality of GLS. The proof of the non-optimality of GLS (Theorem 5(i)) is a constructive proof that uses a counter-example with two agents in a one-dimensional model ($d = 1$) where both agents have the same public data. This raises the question

of whether the suboptimality of GLS arises in higher dimensions or, more generally, in more complicated scenarios. Although extending our analytical proof to more general cases appears to be difficult, in this section, we provide three numerical counter-examples that illustrate the gap of sub-optimality of GLS. In particular, our numerical counter-examples suggest that the sub-optimality of GLS is not limited to the simple counter-example of our analytical proof.

These counter-examples are constructed by using an estimator $L(\delta)$ equal to GLS plus a small perturbation term of the form δ times D^T , i.e.,

$$L(\delta) = \text{GLS} + \delta D^T \equiv (X^T \Lambda X)^{-1} X^T \Lambda + \delta D^T,$$

for an appropriately selected D . The idea behind our counter-examples is that when using a perturbed estimator (with perturbation $\delta > 0$) that is less accurate than GLS under non-strategic agents, some agents will tend to choose a higher precision than under GLS at equilibrium. In all of our numerical examples, a small enough δ leads to an estimation cost at equilibrium smaller than the one of GLS, because some agents will use a higher precision. When δ increases too much, the gain brought by the higher precision of agents is canceled by the loss of precision that is caused by using the estimator $L(\delta)$ that is less precise than GLS.

In all of our examples, the equilibrium costs of the estimators are very close to that of GLS and our examples are far from attaining the bound $\frac{p_{\max}(q+p_{\min})}{p_{\min}(q+p_{\max})}$ provided by Theorem 5. We believe that this bound is loose and can probably be refined.

We present three examples, because each example is of independent interest. The first two involve 1-dimensional models ($d = 1$). In the first example, we use a perturbation term that affects all agents. For this example, we believe that GLS is sub-optimal only when the two exponents p_{\min} and p_{\max} are significantly different. In the second example, we use a perturbation that only affects two “less generous” agents. This allows us to build counter-examples with similar disclosure costs (with exponents $p_{\min} = 1.01$ and $p_{\max} = 1.1$). Our third example includes several counter examples in settings for different values $d \geq 2$. This setting has d symmetrical agents and a single $(d + 1)$ -th agent whose public vector \mathbf{x}_{d+1} is significantly different.

To ensure the reproducibility of these results, we make the code used to compute the equilibria and to produce the figures in this section publicly available.³

Example 1 (1-dimensional model with two agents). We consider a 1-dimensional model ($d = 1$) with two agents ($n = 2$) in which the public data of each agent is $x_i = 1$. For such a game, the estimator GLS is $(X^T \Lambda X)^{-1} X^T \Lambda \tilde{\mathbf{y}} = (\lambda_1 + \lambda_2)^{-1} \lambda^T \tilde{\mathbf{y}}$ and its covariance is $1/(\lambda_1 + \lambda_2)$. We consider a linear estimator $L(\delta)$ of the form

$$\text{GLS} + \begin{bmatrix} \sqrt{\delta} \\ -\sqrt{\delta} \end{bmatrix} = \begin{bmatrix} \lambda_1/(\lambda_1 + \lambda_2) + \sqrt{\delta} \\ \lambda_2/(\lambda_1 + \lambda_2) - \sqrt{\delta} \end{bmatrix}.$$

According to Equation (12), its covariance is $1/(\lambda_1 + \lambda_2) + \delta/\lambda_1 + \delta/\lambda_2$, where $\delta/\lambda_1 + \delta/\lambda_2$ is the loss of precision due to using a linear estimator that is less precise than GLS. We assume that the disclosure cost of Agent 1 is $c_1(\lambda) = \lambda^{1.01}$ ($p_{\min} = 1.01$) while the disclosure cost of Agent 2 is $c_2(\lambda) = \lambda^{20}$ ($p_{\max} = 20$). The scalarization function is the identity, which means that $f_{L(\delta)}(\lambda) = 1/(\lambda_1 + \lambda_2) + \delta/\lambda_1 + \delta/\lambda_2$. We set the maximal precision to $1/\sigma^2 = 1$.

In Figure 1(a), we plot the estimation cost at equilibrium $f_{L(\delta)}(\lambda_{L(\delta)}^*)$ as a function of δ . We observe that with GLS we get an estimation cost of approximately 0.99. When δ increases, the estimation cost at equilibrium decreases up to $\delta = 0.012$ for which it reaches approximately 0.96. This decrease is explained by the fact that for small δ , the gain due to a higher precision used by

³<https://github.com/ngast/strategicLinearRegression>.

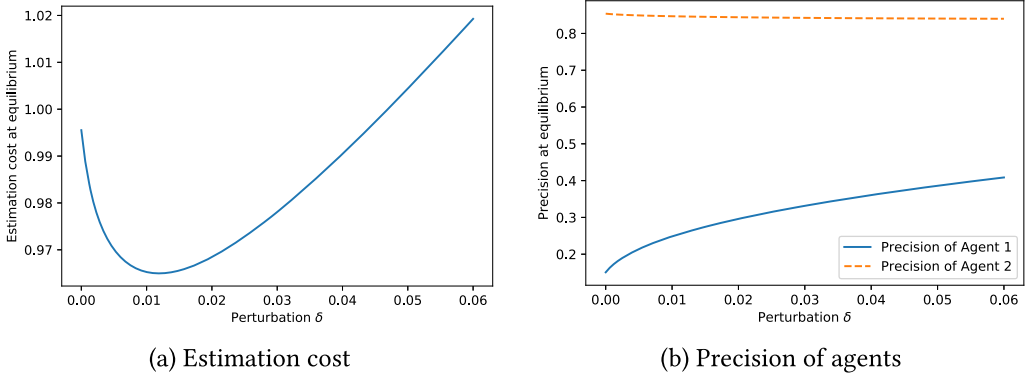


Fig. 1. Counter-example 1: Estimation cost and precision of agents as a function of the perturbation δ .

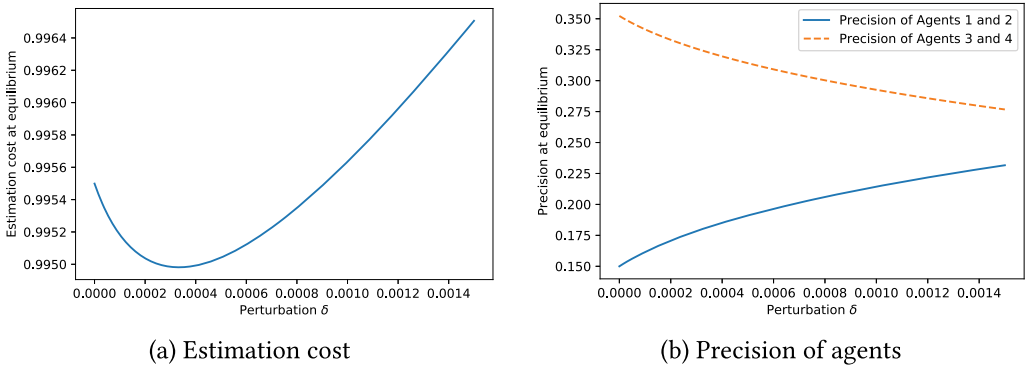


Fig. 2. Counter-example 2: Estimation cost and precision of agents as a function of the perturbation δ .

Agent 1 is larger than the loss of precision $\delta/\lambda_1 + \delta/\lambda_2$. When δ exceeds 0.012, this loss of precision is more important than the gain due to higher precision. This behavior is further illustrated in Figure 1(b), where we plot the precision released by the two agents. We observe that the precision of Agent 1 increases with δ while the precision of Agent 2 decreases (slightly).

Example 2 (1-dimensional model with four agents). We consider a 1-dimensional game with four agents in which the public data of each agent equal $x_i = 1$. Agents 1 and 2 have disclosure costs $c_i(\lambda) = \lambda^{1.01}$ while Agents 3 and 4 have disclosure costs $c_i(\lambda) = \lambda^{1.1}$. We consider a linear unbiased estimator that is equal to GLS plus a perturbation cost that only affects the first two agents: $D = [\sqrt{\delta}, -\sqrt{\delta}, 0, 0]$. Note that this perturbation is only applied to the most selfish agents, as they are the ones we must incentivize to give more. We set the maximal precision to $1/\sigma^2 = 1$.

In Figure 2, we plot the estimation cost at equilibrium $f_{L(\delta)}(\lambda_{L(\delta)}^*)$ as a function of δ . With GLS ($\delta = 0$), we get an estimation cost of 0.9955, which is larger than the value 0.9950 that we obtain for $\delta = 3 \cdot 10^{-4}$. As for Example 1, when δ increases, the precisions used by the least generous agents (Agents 1 and 2) increase while the precisions of the most generous agents decrease.

While the previous two counter-examples are in dimension 1 and with agents that all have $x_i = 1$, the sub-optimality of GLS is not limited to that case. To illustrate that, we consider in the next counter-example models in dimension d with $d \geq 2$. Note that, as we assume that matrix X has rank d , we need at least d players whose feature vectors x_i 's span the d dimensions. Note also

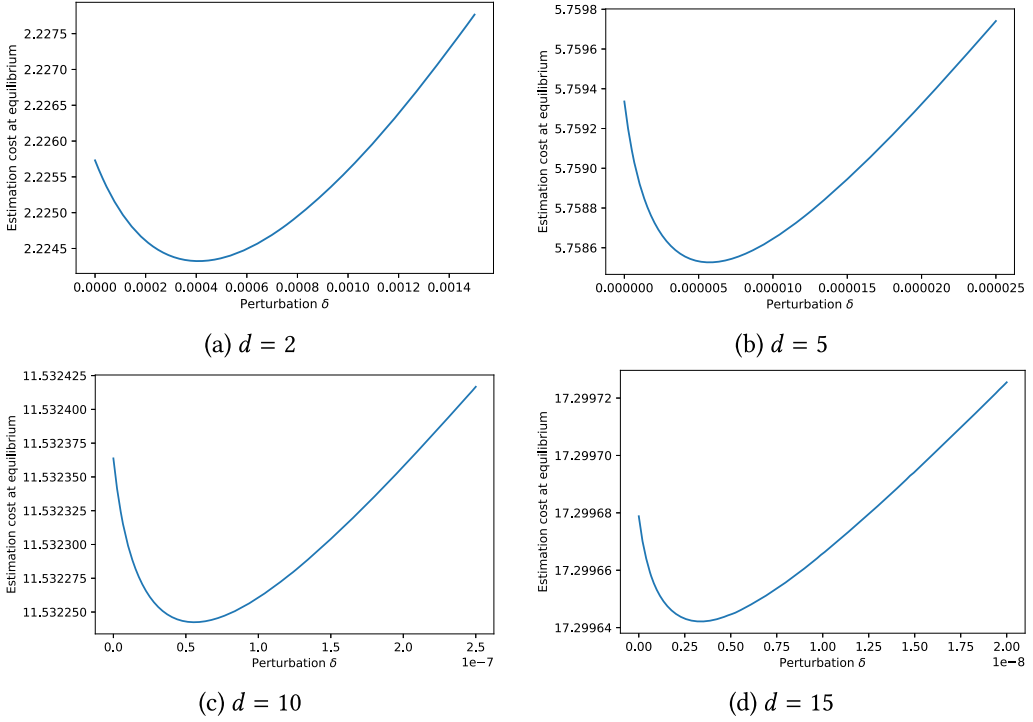


Fig. 3. Counter-example 3: Estimation cost and precision of agents as a function of the perturbation δ for models in dimension $d \geq 2$.

that, with d players in d dimensions, GLS is the only linear unbiased estimator. Indeed, as matrix X would then be invertible, the condition in Equation (11) leads to $D^T = 0$. In Example 3 below, we consider the simplest case of models with $d + 1$ agents, though it is clear that one could construct similar counter examples with any number of agents larger than or equal to $d + 1$.

Example 3 (d -dimensional models with $d + 1$ agents). We consider a d -dimensional game with $d + 1$ agents. The public data of the first d agents span the d dimensions: \mathbf{x}_i is a vector where all components equal 0 except the i th one that is equal to 1. All components of the public data of Agent $d + 1$ are equal to $1/d$: $\mathbf{x}_{d+1} = [1/d, \dots, 1/d]^T$. We assume that the disclosure costs of the first d agents are $c_i(\lambda) = \lambda^{20}$ (for $i \in \{1, \dots, d\}$), and the disclosure cost of the last agent is $c_{d+1}(\lambda) = \lambda^{1.5}$. We set the maximal precision to $1/\sigma^2 = 1$.

The perturbation matrix D is a $(d + 1) \times d$ matrix whose first column is $\sqrt{\delta}[1, \dots, 1, -d]$, all other entries being 0. Hence, the public feature matrix X and the perturbation matrix D are the following $(d + 1) \times d$ matrices:

$$X = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \\ 1/d & \dots & 1/d \end{bmatrix}, \quad D = \begin{bmatrix} \sqrt{\delta} & 0 & 0 & \dots \\ \vdots & 0 & 0 & \dots \\ \sqrt{\delta} & 0 & 0 & \dots \\ -d\sqrt{\delta} & 0 & 0 & \dots \end{bmatrix}. \quad (14)$$

It is easy to verify that $D^T X = 0$, which implies $L(\delta) = \text{GLS} + D^T$ is an unbiased estimator.

In Figure 3, we report the estimation cost at equilibrium $f_{L(\delta)}(\lambda_{L(\delta)}^*)$ as a function of δ . We consider models of dimension $d \in \{2, 5, 10, 15\}$. We observe that for all dimensions d , the behavior is

similar to the one observed in Figure 1(a) and 2(a): When δ is small enough, using the estimator $L(\delta)$ provides a higher precision at equilibrium (i.e., a lower equilibrium estimation cost as seen on the graphs). This comes from the fact that when δ increases, the precision at equilibrium provided by Agent $d + 1$ increases with δ , whereas the precision provided by Agents 1 to d is almost independent of δ . When δ increases too much, the estimation cost increases again because of the non-optimality of the estimator $L(\delta)$ (for given individual precisions). We also observe that the maximal gain that can be obtained by using an estimator other than GLS (and the perturbation δ for which it is achieved with our particular perturbation matrix D) seems to decrease when the dimension d increases.

Finally, although the public feature matrix X in Equation (14) has a particular form, many d -dimensional models with $d + 1$ agents can be cast in this model via an appropriate change of basis. In fact, we conjecture that for any matrix of public features X with at least $d + 1$ agents, there exist disclosure costs such that GLS is not optimal.

7 CONCLUDING REMARKS

This article studies linear regression in the presence of strategic data sources, modeling the precision choice as a non-cooperative game with a public good component. We establish existence of a unique non-trivial Nash equilibrium and study its efficiency for a large class of disclosure and estimation cost functions. We also show an extension of the Aitken/Gauss-Markov theorem to this non-cooperative setup under certain conditions and examples in which the generalized least squares estimator is not optimal.

Our Aitken/Gauss-Markov-type theorem is weaker than these two classical results in three ways. First, we proved the generalized least squares estimator is only approximately optimal in the case of a homogeneous estimation cost and of near-homogeneous disclosure cost and is not always optimal. Second, the optimality of the generalized least squares estimator in the case of monomial disclosure cost functions of same degree is shown w.r.t. the homogeneous scalarization chosen, rather than the positive semidefinite order. Finally, Theorem 5 applies to linear estimators whose difference from GLS does not depend on the actions λ . In the presence of arbitrary dependence on λ , the non-trivial equilibrium need not be unique (or even exist). Understanding when this occurs, and proving optimality results in this context, also remains open.

Our model assumes that the precision chosen by each agent is known to the analyst. Amending this assumption brings issues of truthfulness into consideration: in particular, an important open question is whether there exists an estimator (viewed as a mechanism) that induces truthful precision reporting among agents, at least in equilibrium. An Aitken-type theorem seems instrumental in establishing such a result.

Our analysis of the Nash equilibrium assumes complete information—that is, that agents know the costs and features of other agents. Extending it to a Bayesian setting is an interesting open direction. Still, even when the complete information assumption does not hold, our present results are indicative in at least two ways. First, as our game is a potential game, we know that many natural dynamic evolutions of the game will converge to the Nash equilibrium. Second, if the number of agents grows large, we expect that the Nash equilibrium and Bayesian Nash equilibrium would be close, since the empirical distribution of costs/features would then be close to the true underlying distribution.

APPENDICES

A PROOF OF THEOREM 1

In this proof, we show that the potential function Φ is strictly convex on its effective domain, which implies that the set of Nash equilibria that lies in its effective domain coincides with the set

of local minima of Φ . By strict convexity, Φ has at most one such local equilibrium. To conclude the proof, we then show that this minimum is attained.

The potential function $\Phi(\lambda) = f(\lambda) + \sum_i c_i(\lambda_i)$ takes values in the extended positive real numbers line $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{+\infty\}$. By Assumption 1, the disclosure costs $c_i(\cdot)$ are finite on $[0, 1/\sigma^2]$, since they are continuous on a compact set. Therefore, $\Phi(\cdot)$ is finite if and only if $f(\cdot)$ is finite, i.e., $\text{dom } \Phi \equiv \{\lambda : \Phi(\lambda) < \infty\} = \text{dom } f$, where dom is the effective domain. Recall that, since X has rank d , $(0, 1/\sigma^2]^n \subseteq \text{dom } \Phi$, and $\text{dom } \Phi$ is non-empty.

Recall that $V(\lambda) = (X^T \Lambda X)^{-1}$. This implies that V is strictly convex and goes to infinity when $A(\lambda) = X^T \Lambda X$ goes to a non-invertible matrix (i.e., the largest eigenvalue of V goes to infinity for any sequence λ_n that converges to a λ such that $A(\lambda)$ is non-invertible). As F is convex and increasing, this shows that $f(\lambda) = F(V(\lambda))$ is strictly convex and goes to $+\infty$ when $A(\lambda)$ goes to a non-invertible matrix, which then implies that $f(\lambda) : [0, 1/\sigma^2]^n \rightarrow \bar{\mathbb{R}}_+$ is continuous. As the functions c_i are convex, we conclude that the potential function Φ is strictly convex and continuous on $\bar{\mathbb{R}}_+$.

Let B be the subset of λ such that $\Phi(\lambda) \leq \Phi(1/\sigma^2 \dots 1/\sigma^2)$. By continuity and convexity of Φ , B is a non-empty convex and compact subset of $[0, 1/\sigma^2]^n$ on which $\Phi(\lambda) < \infty$. This implies that the unique minimum of Φ is attained in $B \subseteq \text{dom}(\Phi)$. \square

B PROOF OF THEOREM 2

Under Assumptions 1 and 2, the unique non-trivial equilibrium λ^* minimizes the potential function $\Phi(\lambda) = \sum_{i \in N} c_i(\lambda_i) + f(\lambda)$. Then, for λ^{opt} a minimizer of the social cost:

$$\Phi(\lambda^*) \leq \Phi(\lambda^{\text{opt}}) = \sum_{i \in N} c_i(\lambda_i^{\text{opt}}) + f(\lambda^{\text{opt}}) \leq \sum_{i \in N} c_i(\lambda_i^{\text{opt}}) + n f(\lambda^{\text{opt}}) = \text{opt}$$

by the positivity of f . However, $C(\lambda^*) \leq n\Phi(\lambda^*)$, by the positivity of c_i , and the theorem follows. \square

C PROOF OF THEOREM 3

To simplify the notation, in this proof, we write p instead of p_{\min} ; hence, we show that $\text{PoS} \leq n^{\frac{q}{p+q}}$.

Upper Bound. Recall that Assumption 3 implies that $\forall \lambda \in \mathbb{R}_+, \forall a \geq 1, a^p c_i(\lambda) \leq c_i(a\lambda)$. This implies, by rewriting the assumption with $\lambda' = a\lambda$, that $c_i(\frac{\lambda'}{a}) \leq a^{-p} c_i(\lambda')$ for all $a \geq 1$ and for all λ' .

Recall that we denote by λ^* the unique non-trivial Nash equilibrium. Suppose that $\text{PoS} > n^{\frac{q}{p+q}}$; that is,

$$\sum_{i \in N} c_i(\lambda_i^*) + n f(\lambda^*) > n^{\frac{q}{q+p}} \left(\sum_{i \in N} c_i(\lambda_i^{\text{opt}}) + n f(\lambda^{\text{opt}}) \right).$$

We will show that this implies that λ^* is not an equilibrium, which is a contradiction.

By using that $c_i(\lambda_i^*) \geq 0$ and dividing the above inequality by n , we obtain:

$$\begin{aligned} \sum_{i \in N} c_i(\lambda_i^*) + f(\lambda^*) &\geq \frac{1}{n} \left(\sum_{i \in N} c_i(\lambda_i^*) + n f(\lambda^*) \right) \\ &> n^{-\frac{p}{q+p}} \sum_{i \in N} c_i(\lambda_i^{\text{opt}}) + n^{\frac{q}{p+q}} f(\lambda^{\text{opt}}) \\ &\geq \sum_{i \in N} c_i \left(\frac{\lambda_i^{\text{opt}}}{n^{\frac{1}{p+q}}} \right) + f \left(\frac{\lambda^{\text{opt}}}{n^{\frac{1}{p+q}}} \right), \end{aligned}$$

where for the last inequality, we used Assumption 3 and Assumption 4 with $a = n^{1/(p+q)}$.

To conclude the proof, we remark that $\frac{\lambda^{\text{opt}}}{n^{1/(p+q)}} \leq \lambda^{\text{opt}}$, which implies that $\frac{\lambda^{\text{opt}}}{n^{1/(p+q)}}$ is a valid strategy profile. This would imply that λ^* is not the minimum of the potential function, which is a contradiction. Thus, we have $\text{PoS} \leq n^{\frac{q}{p+q}}$. \square

Lower Bound. Fix $p \geq 1$ and $q \geq 1$. We consider a 1-dimensional model ($d = 1$) with $x_1 = 1$ and $\sigma^2 = (q/p)^{1/(p+q)}$. Let $c_i(\lambda_i) = \lambda_i^p$ for all i and $F(V) = \text{trace}(V)^q = V^q$ (the last equality holds, because when $d = 1$, the co-variance matrix is a scalar). Hence, the co-variance matrix is $V(\lambda) = (\sum_{i \in N} \lambda_i)^{-1}$.

As all agents are identical, and by uniqueness of the Nash equilibrium, the Nash equilibrium is a symmetric Nash equilibrium where all agents will give the same value λ^* where λ^* is the unique minimizer of the potential function:

$$n\lambda^p + (n\lambda)^{-q}.$$

The minimum of this function is attained when its derivative is equal to 0. This implies that $np\lambda^{p-1} = nq(n\lambda)^{-q-1}$, which implies that $\lambda^{p+q} = (q/p)n^{-1-q}$. This shows that $\lambda^* = ((q/p)n^{-1-q})^{1/(p+q)}$.

Similarly, the socially optimal λ^{opt} is also symmetric and is attained when all agents give λ^{opt} the unique minimizer of the social cost:

$$n\lambda^p + n(n\lambda)^{-q}.$$

This implies that

$$\lambda^{\text{opt}} = (n(q/p)n^{-1-q})^{1/(p+q)} = n^{1/(p+q)}\lambda^*. \quad (15)$$

Hence, we get:

$$\begin{aligned} \text{PoS} &= \frac{C(\lambda^*)}{C(\lambda^{\text{opt}})} = \frac{n(\lambda^*)^p + n(n\lambda^*)^{-q}}{n(\lambda^{\text{opt}})^p + n(n\lambda^{\text{opt}})^{-q}} \\ &= \frac{(\lambda^*)^p + (n\lambda^*)^{-q}}{(\lambda^{\text{opt}})^p + (n\lambda^{\text{opt}})^{-q}} \\ &= \frac{(n\lambda^*)^{-q}}{(n\lambda^{\text{opt}})^{-q}} \frac{(\lambda^*)^{p+q} + 1}{(\lambda^{\text{opt}})^{p+q} + 1} \\ &= \left(\frac{\lambda^{\text{opt}}}{\lambda^*}\right)^q \frac{1 + (\lambda^*)^{p+q}}{1 + (\lambda^{\text{opt}})^{p+q}} \\ &= n^{q/(p+q)} \frac{1 + (q/p)n^{-1-q}}{1 + (q/p)n^{-q}}, \end{aligned}$$

where we use the expression in Equation (15) for λ^* and λ^{opt} in the last line. This shows that, for any ϵ , for large enough n , the price of stability is at least $n^{q/(p+q)}(1 - \epsilon)$. \square

D PROOF OF THEOREM 5

D.1 Proof of (i)

We consider the same setting as Example 1, i.e., a 1-dimensional model ($d = 1$) with two agents in which the public data of each agent is $x_i = 1$. For such a game, the GLS estimator is $(X^T \Lambda X)^{-1} X^T \Lambda \tilde{\mathbf{y}} = (\lambda_1 + \lambda_2)^{-1} \lambda^T \tilde{\mathbf{y}}$ and its covariance is $1/(\lambda_1 + \lambda_2)$. We consider a linear estimator $\hat{\beta}(\delta)$ with $\delta \geq 0$ of the form $\hat{\beta}_{\text{GLS}} + \delta^T \tilde{\mathbf{y}}$ where $\delta \in \mathbb{R}^2$ is a vector with coefficients $\delta_1 = -\delta_2 = \sqrt{\delta}$. Note that $\delta_1 = -\delta_2$ guarantees that this linear estimator is unbiased. We assume that the disclosure cost of Agent 1 is $c_1(\lambda) = \lambda^{p_1}$ while the disclosure cost of Agent 2 is $c_2(\lambda) = \lambda^{p_2}$. For a given δ , we denote the equilibrium of the game by $\lambda^*(\delta)$.

Overall, this proof is decomposed in two steps:

Step 1: We compute the derivative of the estimation cost at $\delta = 0$ to show that it is negative if and only if $\lambda_1^*(0)(2p_1 - p_2 - p_1p_2) + \lambda_2^*(0)(2p_2 - p_1 - p_1p_2) > 0$.

Step 2: We show that there exists $x > 0$ such that the above inequality is satisfied for $p_1 = 1/x$ and $p_2 = 1 + x$.

We describe both steps in detail below.

Step 1. According to Equation (12), the covariance of the estimator is $1/(\lambda_1 + \lambda_2) + \delta/\lambda_1 + \delta/\lambda_2$, where $\delta/\lambda_1 + \delta/\lambda_2$ is the loss of precision due to using a linear estimator that is less precise than GLS. We assume that the scalarization function is the identity, which means that the estimation cost is

$$f_\delta(\lambda) = \frac{1}{\lambda_1 + \lambda_2} + \frac{\delta}{\lambda_1} + \frac{\delta}{\lambda_2}. \quad (16)$$

The equilibrium $\lambda^*(\delta)$ is the minimum of the potential function $\Phi_\delta(\lambda) = f_\delta(\lambda) + \lambda_1^{p_1} + \lambda_2^{p_2}$. The estimation cost at equilibrium is $f_\delta(\lambda^*(\delta))$. Our goal in this step is to compute the derivative of $f_\delta(\lambda^*(\delta))$ with respect to δ and to obtain a condition that ensures that it is negative at $\delta = 0$. Let us denote by $(\lambda^*)_i'(\delta) = d\lambda_i^*(\delta)/(d\delta)$ the derivative of $\lambda_i^*(\delta)$ with respect to δ . To simplify notation, we will omit the dependence on δ and simply denote $\lambda_i^* = \lambda_i^*(0)$ and $\lambda_i' = (\lambda^*)_i'(0)$ when it is not confusing. The derivative of the estimation cost evaluated at $\delta = 0$ is equal to

$$\left. \frac{d}{d\delta}(f_\delta(\lambda^*(\delta))) \right|_{\delta=0} = -\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^2} + \frac{1}{\lambda_1^*} + \frac{1}{\lambda_2^*} = -\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^2} + \frac{\lambda_1^* + \lambda_2^*}{\lambda_1^* \lambda_2^*}. \quad (17)$$

In particular, the above derivative is negative if and only if $\frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3} \lambda_1^* \lambda_2^* > 1$. In what follows, we compute the derivatives λ_i' as a function of the values of λ_i^* and p_i .

The equilibrium $\lambda^*(\delta)$ is the minimum of the potential function $\Phi_\delta(\lambda) = f_\delta(\lambda) + \lambda_1^{p_1} + \lambda_2^{p_2}$. By using the first order condition $\partial \Phi_\delta / \partial \lambda_i = 0$, this implies that for all $\delta \geq 0$:

$$-\frac{1}{(\lambda_1^*(\delta) + \lambda_2^*(\delta))^2} - \frac{\delta}{(\lambda_i^*(\delta))^2} + p_i(\lambda_i^*(\delta))^{p_i-1} = 0, \quad \text{for } i \in \{1, 2\}. \quad (18)$$

The derivative of $\lambda_i^*(\delta)$ with respect to δ exists thanks to the implicit function theorem. By differentiating Equation (18) with respect to δ , we obtain

$$\begin{aligned} 0 &= \frac{d}{d\delta} \left(-\frac{1}{(\lambda_1^*(\delta) + \lambda_2^*(\delta))^2} - \frac{\delta}{(\lambda_i^*(\delta))^2} + p_i(\lambda_i^*(\delta))^{p_i-1} \right) \\ &= 2 \frac{(\lambda^*)_1'(\delta) + (\lambda^*)_2'(\delta)}{(\lambda_1^*(\delta) + \lambda_2^*(\delta))^3} - \frac{1}{(\lambda_i^*(\delta))^2} + 2\delta \frac{(\lambda^*)_i'(\delta)}{(\lambda_i^*(\delta))^3} + p_i(p_i - 1)(\lambda_i^*(\delta))^{p_i-2} (\lambda^*)_i'(\delta). \end{aligned} \quad (19)$$

Equation (18), evaluated at $\delta = 0$, shows that $p_i(\lambda_i^*)^{p_i-1} = \frac{1}{(\lambda_1^* + \lambda_2^*)^2}$. Evaluating Equation (19) at $\delta = 0$ and plugging this equality gives

$$\begin{aligned} 0 &= 2 \frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3} - \frac{1}{(\lambda_i^*)^2} + p_i(p_i - 1)(\lambda_i^*)^{p_i-2} \lambda_i' \\ &= 2 \frac{\lambda_1' + \lambda_2'}{(\lambda_1^* + \lambda_2^*)^3} - \frac{1}{(\lambda_i^*)^2} + \frac{1}{(\lambda_1^* + \lambda_2^*)^2} \frac{p_i - 1}{\lambda_i^*} \lambda_i'. \end{aligned} \quad (20)$$

To isolate the term $\lambda'_1 + \lambda'_2$, we multiply the above equation by $\lambda_i^*/(p_i - 1)$ and we sum over $i \in \{1, 2\}$. This gives:

$$\begin{aligned} 0 &= 2 \frac{\lambda'_1 + \lambda'_2}{(\lambda_1^* + \lambda_2^*)^3} \left(\frac{\lambda_1^*}{p_1 - 1} + \frac{\lambda_2^*}{p_2 - 1} \right) - \frac{1}{\lambda_1^*(p_1 - 1)} - \frac{1}{\lambda_2^*(p_2 - 1)} + \frac{\lambda'_1 + \lambda'_2}{(\lambda_1^* + \lambda_2^*)^2} \\ &= \frac{\lambda'_1 + \lambda'_2}{(\lambda_1^* + \lambda_2^*)^3} \left(\frac{2\lambda_1^*}{p_1 - 1} + \frac{2\lambda_2^*}{p_2 - 1} + \lambda_1^* + \lambda_2^* \right) - \frac{1}{\lambda_1^*(p_1 - 1)} - \frac{1}{\lambda_2^*(p_2 - 1)}. \end{aligned}$$

This shows that

$$\begin{aligned} \frac{\lambda'_1 + \lambda'_2}{(\lambda_1^* + \lambda_2^*)^3} &= \frac{\frac{1}{\lambda_1^*(p_1 - 1)} + \frac{1}{\lambda_2^*(p_2 - 1)}}{\frac{2\lambda_1^*}{p_1 - 1} + \frac{2\lambda_2^*}{p_2 - 1} + \lambda_1^* + \lambda_2^*} \\ &= \frac{1}{\lambda_1^* \lambda_2^*} \frac{\lambda_2^*(p_2 - 1) + \lambda_1^*(p_1 - 1)}{2\lambda_1^*(p_1 - 1) + 2\lambda_2^*(p_2 - 1) + (\lambda_1^* + \lambda_2^*)(p_1 - 1)(p_2 - 1)}. \end{aligned}$$

In particular, this implies that the derivative Equation (17) is negative if and only if

$$\frac{\lambda_2^*(p_2 - 1) + \lambda_1^*(p_1 - 1)}{2\lambda_1^*(p_1 - 1) + 2\lambda_2^*(p_2 - 1) + (\lambda_1^* + \lambda_2^*)(p_1 - 1)(p_2 - 1)} > 1.$$

After some algebra, this gives

$$\lambda_1^*(2p_1 - p_2 - p_1 p_2) + \lambda_2^*(2p_2 - p_1 - p_1 p_2) > 0, \quad (21)$$

where, again, by abuse of notation, we denote $\lambda_1^* = \lambda_1^*(0)$ and $\lambda_2^* = \lambda_2^*(0)$.

Step 2. We now consider $p_1 = 1 + 1/x$ and $p_2 = 1 + x$ and $x \rightarrow \infty$. To emphasize the dependence in x , let us denote by $\lambda^*(x) = (\lambda_1^*(x), \lambda_2^*(x))$ the value of the precision at equilibrium (for GLS) and $\Phi_x(\cdot)$ the potential of the game. By definition, $\lambda^*(x)$ minimizes $\Phi_x(\lambda) = 1/(\lambda_1 + \lambda_2) + \lambda_1^{1+1/x} + \lambda_2^{1+x}$. This implies that for all $\epsilon > 0$, $\Phi_x(\lambda^*(x)) \leq \Phi_x(0, 1 - \epsilon)$. As $\lim_{x \rightarrow \infty} \Phi_x(0, 1 - \epsilon) = 1/(1 - \epsilon)$ and because this is true for all ϵ , this implies that

$$\lim_{x \rightarrow \infty} \Phi_x(\lambda^*(x)) = \lim_{x \rightarrow \infty} \left(\frac{1}{\lambda_1^*(x) + \lambda_2^*(x)} + (\lambda_1^*(x))^{1+1/x} + (\lambda_2^*(x))^{1+x} \right) \leq 1.$$

This implies that $\lim_{x \rightarrow \infty} \lambda_1^*(x) = 0$ and $\lim_{x \rightarrow \infty} \lambda_2^*(x) = 1$.

For our values of $p_1 = 1 + 1/x$ and $p_2 = 1 + x$, the left-hand side of Equation (21) equals $\lambda_1(x)(1/x - 2x - 1) + \lambda_2(x)(x - 2/x - 1)$. As $\lim_{x \rightarrow \infty} \lambda_2(x) = 1$ and $\lim_{x \rightarrow \infty} \lambda_1(x) = 0$, this term is asymptotically equivalent to x and is therefore positive for x large enough. This implies that there exists a value x such that $d/(d\delta)f_\delta(\lambda^*(\delta)) < 0$. Hence, for this x , there exists a perturbation value $\delta > 0$ such that $\hat{\beta}(\delta)$ is an estimator that is more efficient than GLS. \square

D.2 Proof of (ii)

We will start by proving Lemma 1. This lemma can easily be explained if we recall Assumption 3 and Assumption 4. Indeed, they dictate how the different components of the potential function behave when all agents multiply or divide the amount of information they give. If the sum of individual costs is too great compared to the common cost, then dividing the amount that all agents give greatly reduces the individual costs while slightly augmenting the common cost, which is beneficial. The same goes the other way around where agents multiply the amount they give. This formalizes an intuition that one can have about this model: There is a balance between the individual costs paid to achieve the objective of reducing the common cost and the objective itself.

LEMMA 1. Under Assumptions 1, 2, 3, and 4, the ratio between the sum of individual costs and the common cost is bounded. Formally, the equilibrium λ^* satisfies:

$$\sum_{i \in N} c_i(\lambda_i^*) \leq \frac{q}{p_{\min}} f(\lambda^*) \quad \text{and} \quad f(\lambda^*) \leq \frac{p_{\max}}{q} \sum_{i \in N} c_i(\lambda_i^*).$$

PROOF. This proof mainly relies on the fact that λ^* is the minimum of the potential function. Let λ^* be the unique non-trivial equilibrium. Let $\kappa \in (0, 1)$ be a multiplicative factor applied to the equilibrium profile. As λ^* is the minimum of the potential function, we have $\Phi(\lambda^*) \leq \Phi(\kappa\lambda^*)$ and $\Phi(\lambda^*) \leq \Phi(\lambda^*/\kappa)$. This implies that:

$$\begin{aligned} \sum_{i \in N} c_i(\lambda_i^*) + f(\lambda^*) &\leq \sum_{i \in N} c_i(\kappa\lambda_i^*) + f(\kappa\lambda^*) \leq \kappa^{p_{\min}} \sum_{i \in N} c_i(\lambda_i^*) + \kappa^{-q} f(\lambda^*), \quad \text{and} \\ \sum_{i \in N} c_i(\lambda_i^*) + f(\lambda^*) &\leq \sum_{i \in N} c_i(\lambda_i^*/\kappa) + f(\lambda^*/\kappa) \leq \kappa^{-p_{\max}} \sum_{i \in N} c_i(\lambda_i^*) + \kappa^q f(\lambda^*). \end{aligned}$$

The above equations imply that:

$$\begin{aligned} (1 - \kappa^{p_{\min}}) \sum_{i \in N} c_i(\lambda_i^*) &\leq (\kappa^{-q} - 1) f(\lambda^*), \quad \text{and} \\ (1 - \kappa^q) f(\lambda^*) &\leq (\kappa^{-p_{\max}} - 1) \sum_{i \in N} c_i(\lambda_i^*). \end{aligned}$$

As $\kappa \in (0, 1)$, we have $1 - \kappa^{p_{\min}} > 0$ and $\kappa^{-p_{\max}} - 1 > 0$. Hence, the above equations imply that

$$\frac{1 - \kappa^q}{\kappa^{-p_{\max}} - 1} f(\lambda^*) \leq \sum_{i \in N} c_i(\lambda_i^*) \leq \frac{\kappa^{-q} - 1}{1 - \kappa^{p_{\min}}} f(\lambda^*).$$

This inequality is valid for every $\kappa \in (0, 1)$. As $\lim_{\kappa \rightarrow 1} \frac{1 - \kappa^q}{\kappa^{-p_{\max}} - 1} = q/p_{\max}$ and $\lim_{\kappa \rightarrow 1} \frac{1 - \kappa^{-q}}{1 - \kappa^{p_{\min}}} = q/p_{\min}$, this gives

$$\frac{q}{p_{\max}} f(\lambda^*) \leq \sum_{i \in N} c_i(\lambda_i^*) \leq \frac{q}{p_{\min}} f(\lambda^*). \quad \square$$

We are now ready to prove Theorem 5(ii). Let $\Phi_L(\lambda) = \sum_{i \in N} c_i(\lambda_i) + f_L(\lambda)$ be the potential function for any linear unbiased estimator and $\Phi_{\text{GLS}}(\lambda) = \sum_{i \in N} c_i(\lambda_i) + f_{\text{GLS}}(\lambda)$ be the potential function for GLS. Recall that λ_L^* and λ_{GLS}^* denote the non-trivial equilibria for the linear unbiased estimator and for GLS, respectively. By optimality of GLS, for all λ , we have $f_L(\lambda) \geq f_{\text{GLS}}(\lambda)$. This implies that for all λ , we have $\Phi_L(\lambda) \geq \Phi_{\text{GLS}}(\lambda)$. Therefore,

$$\Phi_L(\lambda_L^*) = \min_{\lambda} \Phi_L(\lambda) \geq \Phi_{\text{GLS}}(\lambda_{\text{GLS}}^*) = \min_{\lambda} \Phi_{\text{GLS}}(\lambda). \quad (22)$$

By applying the inequalities of Lemma 1, we obtain:

$$\Phi_{\text{GLS}}(\lambda_{\text{GLS}}^*) = \sum_{i \in N} c_i((\lambda_{\text{GLS}}^*)_i) + f_{\text{GLS}}(\lambda_{\text{GLS}}^*) \geq \frac{q}{p_{\max}} f_{\text{GLS}}(\lambda_{\text{GLS}}^*) + f_{\text{GLS}}(\lambda_{\text{GLS}}^*),$$

and

$$\Phi_L(\lambda_L^*) = \sum_{i \in N} c_i((\lambda_L^*)_i) + f_L(\lambda_L^*) \leq \frac{q}{p_{\min}} f_L(\lambda_L^*) + f_L(\lambda_L^*).$$

Combining the above two inequalities with Equation (22), we conclude that

$$f_{\text{GLS}}(\lambda_{\text{GLS}}^*) \leq \frac{\frac{q}{p_{\min}} + 1}{\frac{q}{p_{\max}} + 1} f_L(\lambda_L^*) = \frac{p_{\max}(q + p_{\min})}{p_{\min}(q + p_{\max})} f_L(\lambda_L^*).$$

ACKNOWLEDGMENTS

We thank the editor and the three anonymous reviewers for their particularly thoughtful comments and feedback that significantly improved the article.

REFERENCES

- [1] Jacob Abernethy, Yiling Chen, Chien-Ju Ho, and Bo Waggoner. 2015. Low-cost learning via active data procurement. In *Proceedings of the 16th ACM Conference on Economics and Computation (EC'15)*. 619–636.
- [2] Rakesh Agrawal and Ramakrishnan Srikant. 2000. Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 439–450.
- [3] A. C. Aitken. 1935. On least squares and linear combinations of observations. *Proc. Roy. Soc. Edinburgh* 55 (1935), 42–48.
- [4] Mike Atallah, Elisa Bertino, Ahmed Elmagarmid, Mohamed Ibrahim, and Vassilios Verykios. 1999. Disclosure limitation of sensitive rules. In *Proceedings of the Workshop on Knowledge and Data Engineering Exchange (KDEX'99)*. 45–52.
- [5] A. C. Atkinson, A. N. Donev, and R. D. Tobias. 2007. *Optimum Experimental Designs, with SAS*. Oxford University Press, New York.
- [6] Omer Ben-Porat and Moshe Tennenholtz. 2019. Regression equilibrium. In *Proceedings of the ACM Conference on Economics and Computation (EC'19)*. 173–191.
- [7] S. Boyd and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.
- [8] Y. Cai, C. Daskalakis, and C. H. Papadimitriou. 2015. Optimum statistical estimation with strategic data sources. In *Proceedings of the 28th Conference on Learning Theory (COLT'15)*. 40.1–40.40.
- [9] Ioannis Caragiannis, Ariel D. Procaccia, and Nisarg Shah. 2016. Truthful univariate estimators. In *Proceedings of the 33rd International Conference on Machine Learning (ICML'16)*.
- [10] Yiling Chen, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. 2018. Optimal data acquisition for statistical estimation. In *Proceedings of the ACM Conference on Economics and Computation (EC'18)*. 27–44.
- [11] Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. 2018. Strategyproof linear regression in high dimensions. In *Proceedings of the ACM Conference on Economics and Computation (EC'18)*. 9–26.
- [12] Michela Chessa, Jens Grossklags, and Patrick Loiseau. 2015. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *Proceedings of the 28th IEEE Computer Security Foundations Symposium (CSF'15)*.
- [13] Michela Chessa and Patrick Loiseau. 2017. On non-monetary incentives for the provision of public goods. Retrieved from <https://ideas.repec.org/p/gre/wpaper/2017-24.html>.
- [14] Anil Kumar Chorppe and Tansu Alpcan. 2013. Trading privacy with incentives in mobile commerce: A game theoretic approach. *Perv. Mob. Comput.* 9, 4 (2013), 598–612.
- [15] Rachel Cummings, Stratis Ioannidis, and Katrina Ligett. 2015. Truthful linear regression. In *Proceedings of the 28th Conference on Learning Theory (COLT'15)*, Vol. 40. 1–36.
- [16] Pranav Dandekar, Nadia Fawaz, and Stratis Ioannidis. 2014. Privacy auctions for recommender systems. *ACM Trans. Econ. Comput.* 2, 3 (July 2014), 12:1–12:22.
- [17] Anirban Dasgupta and Arpita Ghosh. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web (WWW'13)*. 319–330.
- [18] Ofer Dekel, Felix Fischer, and Ariel D. Procaccia. 2010. Incentive compatible regression learning. *J. Comput. Syst. Sci.* 76, 8 (2010), 759–777.
- [19] Josep Domingo-Ferrer. 2008. A survey of inference control methods for privacy-preserving data mining. In *Privacy-preserving Data Mining*. Springer, 53–80.
- [20] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. 2013. Local privacy and statistical minimax rates. In *Proceedings of the 54th IEEE Symposium on Foundations of Computer Science (FOCS'13)*. 429–438.
- [21] George T. Duncan and Sumitra Mukherjee. 2000. Optimal disclosure limitation strategy in statistical databases: Detering tracker attacks through additive noise. *J. Amer. Statist. Assoc.* 95, 451 (2000), 720–729.
- [22] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP'06)*. 1–12.
- [23] Rafael M. Frongillo, Yiling Chen, and Ian A. Kash. 2015. Elicitation for aggregation. In *Proceedings of the 29th Conference on Artificial Intelligence (AAAI'15)*.
- [24] Arpita Ghosh and Aaron Roth. 2011. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce (EC'11)*. 199–208.
- [25] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. 2016. Strategic classification. In *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science (ITCS'16)*. 111–122.

- [26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2 ed.). Springer.
- [27] Thibaut Horel, Stratis Ioannidis, and S. Muthukrishnan. 2014. Budget feasible mechanisms for experimental design. In *Proceedings of the 11th Latin American Theoretical INformatics Symposium (LATIN'14)*. 719–730.
- [28] Safwan Hossain and Nisarg Shah. 2020. Pure nash equilibria in linear regression. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'20)*.
- [29] Stratis Ioannidis and Patrick Loiseau. 2013. Linear regression as a non-cooperative game. In *Proceedings of the 9th International Conference on Web and Internet Economics (WINE'13)*. 277–290.
- [30] Ramesh Johari and John N. Tsitsiklis. 2004. Efficiency loss in a network resource allocation game. *Math. Oper. Res.* 29, 3 (2004), 407–435.
- [31] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2016. Extremal mechanisms for local differential privacy. *J. Mach. Learn. Res.* 17, 17 (2016), 1–51.
- [32] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. 2012. Private convex empirical risk minimization and high-dimensional regression. In *Proceedings of the 25th Conference on Learning Theory (COLT'12)*. 25.1–25.40.
- [33] Katrina Ligett and Aaron Roth. 2012. Take it or Leave it: Running a survey when privacy comes at a cost. In *Proceedings of the 8th International Conference on Web and Internet Economics (WINE'12)*. 378–391.
- [34] Yang Liu and Yiling Chen. 2016. A bandit framework for strategic regression. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'16)*. 1821–1829.
- [35] Yuan Luo, Nihar B. Shah, Jianwei Huang, and Jean Walrand. 2015. Parametric prediction from parametric agents. In *Proceedings of the 10th Workshop on the Economics of Networks, Systems and Computation (NetEcon'15)*. 57–57.
- [36] Reshef Meir, Ariel D. Procaccia, and Jeffrey S. Rosenschein. 2012. Algorithms for strategyproof classification. *Artif. Intell.* 186 (2012), 123–156.
- [37] Dov Monderer and Lloyd S. Shapley. 1996. Potential games. *Games Econ. Behav.* 14, 1 (1996), 124–143.
- [38] John Morgan. 2000. Financing public goods by means of lotteries. *Rev. Econ. Stud.* 67, 4 (Oct. 2000), 761–84.
- [39] Kobbi Nissim, Rann Smorodinsky, and Moshe Tennenholtz. 2012. Approximately optimal mechanism design via differential privacy. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS'12)*. 203–213.
- [40] Stanley R. M. Oliveira and Osmar R. Zaiane. 2003. Privacy preserving clustering by data transformation. In *Proceedings of the Brazilian Symposium on Databases (SBBD'03)*. 304–318.
- [41] Javier Perote and Juan Perote-Pena. 2004. Strategy-proof estimators for simple regression. *Math. Soc. Sci.* 47, 2 (2004), 153–176.
- [42] F. Pukelsheim. 2006. *Opt. Des. Exper.* Vol. 50. Society for Industrial Mathematics.
- [43] Tim Roughgarden and Éva Tardos. 2002. How bad is selfish routing? *J. ACM* 49, 2 (Mar. 2002), 236–259.
- [44] William H. Sandholm. 2010. *Population Games and Evolutionary Dynamics*. The MIT Press.
- [45] Guido Schäfer. 2011. Online Social Networks and Network Economics. Lecture notes, Sapienza University of Rome.
- [46] Joseph F. Traub, Yechiam Yemini, and H. Woźniakowski. 1984. The statistical security of a statistical database. *ACM Trans. Datab. Syst.* 9, 4 (1984), 672–679.
- [47] Jaideep Vaidya, Christopher W. Clifton, and Yu Michael Zhu. 2006. *Privacy Preserving Data Mining*. Springer.
- [48] Tyler Westenbroek, Roy Dong, Lillian J. Ratliff, and S. Shankar Sastry. 2020. Competitive statistical estimation with strategic data sources. *IEEE Trans. Automat. Control* 65, 4 (2020), 1537–1551.

Received July 2019; revised December 2019; accepted February 2020