Accuracy for Sale: Aggregating Data with a Variance Constraint

Rachel Cummings* Katrina Ligett* Aaron Roth† Zhiwei Steven Wu† Juba Ziani*

August 8, 2014

Abstract

We consider the problem of a data analyst who may purchase an unbiased estimate of some statistic from multiple data providers. From each provider i, the analyst has a choice: she may purchase an estimate from that provider that has variance chosen from a finite menu of options. Each level of variance has a cost associated with it, reported (possibly strategically) by the data provider. The analyst wants to choose the *minimum cost* set of variance levels, one from each provider, that will let him combine his purchased estimators into an aggregate estimator that has variance at most some fixed desired level. Moreover, he wants to do so in such a way that incentivizes the data providers to truthfully report their costs to the mechanism.

We give a dominant strategy truthful solution to this problem that yields an estimator that has optimal expected cost, and violates the variance constraint by at most an additive term that tends to zero as the number of data providers grows large.

*Computing and Mathematical Sciences, California Institute of Technology; {rachelc, katrina, jziani}@caltech.edu

†Computer and Information Science, University of Pennsylvania; {aaroth,wuzhiwei}@cis.upenn.edu

1 Introduction

We consider a *data analyst* who wishes to compute an unbiased estimate of some underlying population statistic, by buying and aggregating data from multiple strategic data providers. Each data provider may experience different costs for different levels of data accuracy (variance), and may strategically price access to his data if doing so would benefit him. The analyst must design a mechanism for choosing which level of accuracy to purchase from each provider, and for combining the purchased data into a single aggregate quantity that forms an unbiased estimator of the statistic of interest. Her goal is to do so at minimum cost, given some target level of overall accuracy.

This model captures a number of interesting scenarios. For example:

- Each "data provider" might in fact be a single individual, who is selling a (possibly perturbed) bit signifying some property of interest to the data analyst (e.g., the cancer status of the individual). Here, the variance of each of these estimates comes from two sources: each individual's bit is the realization of an independent sample from some underlying population distribution, with an inherent variance. A data provider may also add his own perturbation (e.g., noise from a Gaussian or Laplace distribution) in order to guarantee a certain level of (differential) privacy. He can therefore potentially offer the data analyst access to his data at a menu of different variance levels (and costs), corresponding to differing levels of privacy protection. Intuitively, the different costs the individual experiences at different levels of accuracy may correspond to his (potentially arbitrary) preferences for privacy. Since we model data providers as strategic agents, they will report the cost that maximizes their utility, and their reported costs need not necessarily match their true costs. We allow each individual to report an arbitrary cost separately for each variance level, so this approach does not require assuming that agent preferences for privacy respect any fixed functional form.

- Each data provider might be an organization (such as a university) that has the ability to collect a random sample of varying size from a sub-population that it controls (e.g. students, professors, etc). Under the assumption that the individuals in the data provider's populations are sampled i.i.d. from some underlying distribution, the variance of the estimate that they offer is inversely proportional to the number of individuals that they sample. Here, the costs for different levels of variance correspond to the costs required to recruit different numbers of participants to a study. These costs may differ between organizations, and behave in complicated ways: for example, the marginal cost for each additional sample might be decreasing (if there are economies of scale – for example by advertising on a campus TV station), or might be increasing (for example, after exhausting the undergraduate population at a university, obtaining additional samples may require recruiting faculty, which is more difficult). Again, because we allow data providers to report arbitrary cost schedules corresponding to different variance levels, we need make no assumptions about the form that these costs take.

1.1 Our Results and Techniques

We model the data analyst's problem as a combinatorial optimization problem: From each of the data providers, the analyst buys an unbiased estimator of the population statistic of interest, for which she must choose a variance from a fixed, finite menu of options. Given these purchased estimators, the data analyst may then take any convex combination to obtain her final unbiased estimator of the underlying population statistic. The choices made by the data analyst affect both

the variance of the final estimator that she derives, as well as the total payment that she must make. We consider the problem of finding the *cheapest* way of constructing an estimator that has variance below some fixed desired level, specified in advance by the data analyst.

Our main tool in solving this problem is linear programming. However, the solution is not straightforward. First, our problem actually consists of two nested optimization problems: we must choose a variance level for each of the estimators, and then we must find the optimal weighted linear combination of these estimators. Rather than solving these problems separately, we use the KKT conditions to derive a closed form for the optimal weights to use in the linear combination of each of the estimators *as a function of their variance*. This allows us to express the problem as a one-shot optimization problem, with decision variables only for the choices of variance for each estimator. Unfortunately, the natural fractional relaxation of this optimization problem (in which the data analyst may fractionally choose different variance levels) is non-convex. Instead, we consider a further (linear) relaxation of the constraint in our problem, which matches the original constraint only for integer solutions. We show that all optimal extreme points of the linear program that result from this relaxation do in fact yield integer choices for all but *at most one* data provider, and then show that if the number of data providers is sufficiently large, then rounding the one fractional assignment to an integer assignment only marginally violates our target variance constraint.

We note that our algorithm chooses the *minimum expected cost* lottery over purchase decisions from among a pre-specified feasible set of lotteries, and hence is *maximum-in-distributional-range*. This means it can be paired with payments such that truthful reporting of costs is a dominant strategy for each of the data providers. (We recall that although we allow data providers to misreport their costs, they cannot lie about their data or its accuracy.)

In summary, we show the following theorem:

Theorem 1 (Informal)**.** *Given any finite menu of variance levels, and any feasible aggregate variance level for the data analyst, there exists a dominant strategy truthful mechanism that selects the minimum cost assignment of variance levels to providers, and generates an unbiased linear estimator that satisfies the analyst's variance constraint up to an additive term that tends to 0 as the number of data providers grow large.*

Finally, we observe that VCG payments (although always truthful) do *not* guarantee individual rationality in our setting. We prove an upper bound on the degree to which individual rationality can be violated for any player, and hence can add a fixed amount to the payment given to each player, to guarantee individual rationality for all providers with sufficiently low minimum costs.

1.2 Related Work

There is a growing body of work [10, 15, 19, 7, 8, 17, 9] related to our first motivating example: buying sensitive data from individuals. This line of work considers the problem of incentivizing individuals to provide their data to an analyst, when they experience a cost — usually due to privacy loss — from sharing their data. These papers have used differential privacy, defined in [6], to combat this privacy loss, but have generally offered only a single privacy level to participants, or have made assumptions about the functional form of this privacy loss in terms of the differential privacy parameter. Our construction, on the other hand, allows the analyst to offer each data provider a menu of different variance levels, corresponding to different levels of differential privacy, and allows the agents to express *arbitrary* costs for each level independently. This requires no assumptions at all about the functional form of agent costs. As in our setting, these papers all

consider the individual data providers to be selfish agents, and thus allow agents to strategically misreport their costs to secure them a higher payment. Recently, [9] considered a setting where individuals have unverifiable data, and can also misreport their data. We restrict to a setting where data is verifiable (as the other papers in this literature do), but allow individuals to lie about their costs for providing data. In this setting, we are making the implicit assumption that the sensitive data held by individuals is independent of their privacy costs. This is motivated by an impossibility result of [10], later strengthened by [17], that when privacy costs are correlated with data, no mechanism can satisfy individual rationality and estimate the statistic of interest with non-trivial accuracy, while making finite payments.

Our paper is also related to the vast body of work on optimal experiment design, in which an analyst wishes to learn parameters of an underlying distribution by optimizing a multi-set of samples to observe from the population, each at some cost. (For a survey of results see [18] or [2]; for a textbook treatment see Section 7.5 of [3].) Each data sample is an "experiment" with observable attributes. The analyst assumes a linear relationship between experiment attributes and outcomes, and wishes to accurately learn the linear parameter by performing a collection of experiments subject to a budget constraint. Although the problem we consider seems to be a special case of experiment design (i.e. with attribute vector of dimension 0), these two problems differ in a few key aspects. First, optimal experiment design allows the analyst to *repeat experiments*, and performing the same experiment multiple times may result in different outcomes. We do not allow this, and we further constrain the problem to require the analyst to buy exactly one "experiment" (i.e. observation of data) from each data provider. The techniques used in this line of work do not generically extend to the more constrained setting that we consider. Additionally, optimal experiment design is a problem in the *full information* setting, so the cost of each experiment is a priori known by the analyst. That is, data providers cannot misreport their costs. The experimental design literature generally gives approximation algorithms which are not maximal in range, and hence do not yield truthful mechanisms. We consider a setting in which data providers are strategic agents, and we must additionally ensure that our optimization process is incentive compatible. Recent work of [13] considered experimental design for strategic agents, but their work considered a different accuracy objective (other than minimizing variance), and their techniques fail under the additional constraint that the analyst can buy at most one estimate from each data provider.

The truthfulness of our mechanism depends on a property called *maximal-in-distributional-range* (MIDR), defined in [4]. MIDR mechanisms are guaranteed to select a distribution over outputs that maximizes expected welfare. Similar properties were also used in [1], [5] and [14]. [4] showed that MIDR mechanisms are *truthful-in-expectation* when paired with VCG payments. That is, with an MIDR mechanism, no player can increase her expected utility by lying about her private information. We first show that our proposed mechanism is MIDR, and then use this result to show that data providers do not have an incentive to misreport their costs.

2 Preliminaries and Model

We consider an analyst who wishes to estimate the expected value μ of some statistic on the underlying population. She has access to a set of n data providers, each of which is capable of providing some unbiased estimate μ_i of the statistic of interest with different levels of variance $\mathbb{E}\left[(\mu_i - \mu)^2\right]$. The provider may also experience some cost for computing the estimate at each

variance level. The analyst's goal is to obtain an accurate unbiased estimate for μ, using the estimates from the providers, while minimizing the social cost for computing such data.

We equip the analyst with a mechanism that offers a menu specifying a discrete range of possible variance levels $0 < v_1 < v_2 < \ldots < v_m < \infty$, and asks each provider i report back a set of costs $\{c_{ij}\}_{j=1}^m$ for computing the estimates at all levels. The mechanism then selects a variance level to purchase from each provider, and generates an estimate for μ that is a weighted sum of the providers' reported estimates μ_i's: $\widehat{\mu} = \sum_i w_i \mu_i$. Note that the expectation $\mathbb{E}[\widehat{\mu}] = \sum_i w_i \mathbb{E}[\mu_i] = \sum_i w_i \mu$, so $\widehat{\mu}$ will be an unbiased estimate as long as $\sum_i w_i = 1$. The following proposition, often called the Bienaymé formula, allows us to express the variance of $\widehat{\mu}$ as a linear combination of the variances of μ_i.

Proposition 1. *Let $X_1, \ldots, X_n$ be uncorrelated real-valued random variables, and $w_1, \ldots, w_n$ be any real numbers, then*

$$Var\left(\sum_i w_i X_i\right) = \sum_i w_i^2 \, Var(X_i).$$

The goal of the analyst is to minimize the total cost among all providers, while maintaining a guarantee that the variance of $\widehat{\mu}$ is below some threshold α. This can be expressed in the following program, where each x_{ij} indicates whether we assign provider i to variance level j:

$$\min \sum_{i,j} x_{ij} c_{ij} \tag{1}$$

$$\text{subject to } \sum_i w_i^2 \left(\sum_j x_{ij} v_j\right) \leq \alpha \tag{2}$$

$$\sum_j x_{ij} = 1 \text{ for all } i \tag{3}$$

$$x_{ij} \in \{0, 1\} \text{ for all } (i, j) \tag{4}$$

$$\sum_i w_i = 1 \text{ and for all } i, w_i \geq 0 \tag{5}$$

2.1 Mechanism Design Basics

We study our optimization problem in the setting of mechanism design, with n players, and a set Ω of possible outcomes. In particular, this set Ω corresponds to the set of possible assignments of players to variance levels. Each player also has a cost function $c_i \colon \Omega \to \mathbb{R}$, where $c_i(\omega)$ is player i's cost for outcome ω. Let $c = (c_1, \ldots, c_n)$ denote the profile of cost functions for all players. We want to minimize total cost, so our objective is $\sum_i^n c_i(\omega)$. A *(direct-revelation) mechanism* $\mathcal{M}$ consists of an *allocation rule* $\mathcal{A}$, a function mapping reported cost profiles to outcomes, and a *payment rule* p, a function mapping cost profiles to a payments to each player. Such a mechanism takes as input reported cost functions from the players, and outputs (possibly randomly) an allocation ω and payments to all the players. Two important desiderata in mechanism design are *truthfulness* and *individual rationality*.

Definition 1 (Truthful-in-Expectation)**.** *A mechanism $\mathcal{M} = (\mathcal{A}, p)$ on n players is* (dominant strategy) truthful-in-expectation *if for any reported cost profile c and for all $i \in [n]$:*

$$\mathop{\mathbb{E}}_{\mathcal{M}}[p_i(c) - c_i(\mathcal{A}(c))] \geq \mathop{\mathbb{E}}_{\mathcal{M}}[p_i((c_i, c_{-i})) - c_i(\mathcal{A}(c_i, c_{-i}))].$$

Definition 2 (Individually Rational)**.** *A mechanism $\mathcal{M} = (\mathcal{A}, p)$ is* individually rational *(IR) if for any reported cost profile c and for all $i \in [n]$:*

$$\mathop{\mathbb{E}}_{\mathcal{M}}[p_i(c) - c_i(\mathcal{A}(c))] \geq 0.$$

We will use *VCG-based mechanisms* to minimize total cost while achieving truthfulness. A *VCG mechanism* is defined by the allocation rule that selects the cost-minimizing outcome $\omega^* \in \arg\min_{\omega\in\Omega} \sum_i c_i(\omega)$ for any reported cost functions, and the payment rule p that rewards each player his "externality":

$$p_i(c) = \min_{\omega\in\Omega_{-i}} \sum_{i'\neq i} c_{i'}(\omega) - \sum_{i'\neq i} c_{i'}(\omega^*). \tag{6}$$

Let $dist(\Omega)$ be the set of all probability distributions over the set of outcomes Ω, and let $\mathcal{R} \subseteq dist(\Omega)$ be a compact subset. Then a *maximal-in-distributional-range* (MIDR) allocation rule is defined as sampling an outcome ω from distribution $D^* \in \mathcal{R}$, where D^* minimizes the expected total cost $\mathbb{E}_{\omega\sim D^*}[\sum_i c_i(\omega)]$ over all distributions in $\mathcal{R}$. A VCG payment rule can be defined accordingly:

$$p_i(c) = \min_{D'\in dist(\Omega_{-i})} \mathop{\mathbb{E}}_{\omega\sim D'}\left[\sum_{i'\neq i} c_{i'}(\omega)\right] - \mathop{\mathbb{E}}_{\omega\sim D^*}\left[\sum_{i'\neq i} c_{i'}(\omega)\right].$$

It is known from [4] that when an MIDR allocation rule is paired with a VCG payment rule, the resulting mechanism is truthful-in-expectation.

To guarantee individual rationality, we pay each player some entrance reward R before running the MIDR mechanism so that $R + \mathbb{E}[p_i(c) - c_i(\mathcal{A}(c))] \geq 0$ for all player i. It suffices to set $R \geq \max_i \mathbb{E}[p_i(c) - c_i(\mathcal{A}(c))]$, and in Section 4.2 we derive a more refined bound for R to get individual rationality.

3 Rewriting the Program

The optimization problem introduced in Section 2 is non-convex because the variance constraint (2) contains the product of decision variables x_{ij} and w_i. To achieve convexity, we will transform the program in three steps:

1. First, we will eliminate the decision variables w_i by deriving a closed form solution for the weights w_i that minimize variance, once the variables x_{ij} are fixed. However, this will still leave us with a non-convex optimization problem.

2. Next, we will replace the non-convex constraint derived above with a linear constraint, that is identical whenever the x_{ij} variables take on integral values.

3. Finally in Section 4, we relax the integrality constraint. Because our linear variance constraint is no longer identical to the original "correct" non-convex variance constraint, we must in the end argue that a rounded solution does not substantially violate the original constraint.

First, to simplify notation, for any assignment $\{x_{ij}\}$, let $\widehat{v}_i$ denote the variance level assigned to provider i. We want to write w_i as a function of $\widehat{v}_i$'s. In particular, given the variance assignments, we want to choose the weights w_i so that the variance of the aggregate statistic $\widehat{\mu}$ is minimized.

Lemma 1. *Given a variance level assignment $\{\widehat{v_i}\}$, the weight vector w^* that minimizes the variance of $\widehat{\mu} = \sum_i w_i \mu_i$ satisfies*

$$w_i^* = \frac{1/\widehat{v_i}}{\sum_i 1/\widehat{v_i}} \qquad \textit{for all } i.$$

Proof. The problem can be written as a convex program

$$\min \sum_i w_i^2 \widehat{v_i} \qquad \text{subject to } \sum_i w_i = 1 \text{ and } w_i \geq 0 \text{ for all } i \tag{7}$$

We know that strong duality holds because the program satisfies Slater's condition, and the Lagrangian is given by

$$\mathcal{L}(w, \lambda) = \sum_i \widehat{v_i} \cdot w_i^2 - \lambda \left(1 - \sum_i w_i\right) = w^T V w - \lambda(1 - \mathbb{1}^T w),$$

where $V = diag(v^1, \ldots, v^n)$. Note that $\nabla_w \mathcal{L}(w, \lambda)^T = 2Vw + \lambda \mathbb{1}$. By KKT conditions, $\nabla_w \mathcal{L}(w^*, \lambda)^T = 0$, and so $w^* = -\frac{\lambda}{2} V^{-1} \mathbb{1}$, which gives $\min_w \mathcal{L}(w, \lambda)^T = -\frac{\lambda^2}{4} \sum_i 1/\widehat{v_i} - \lambda$. Now the dual problem can be written as

$$\begin{aligned} \max_{\lambda} \min_{w \geq 0} \mathcal{L}(w, \lambda) &= \max_{\lambda} \left[-\frac{\lambda^2}{4} \sum_i 1/\widehat{v_i} - \lambda \right] \\ &= \max_{\lambda} \left[-\left(\sum_i 1/\widehat{v_i} \right) \left(\lambda/2 + \frac{1}{\sum_i 1/\widehat{v_i}} \right)^2 + \frac{1}{\sum_i 1/\widehat{v_i}} \right]. \end{aligned}$$

It is easy to see that the maximum is reached at $\lambda^* = \frac{-2}{\sum_i 1/\widehat{v_i}}$. It follows that

$$w^* = \frac{-\lambda^*}{2} V^{-1} \mathbb{1} = \frac{V^{-1}\mathbb{1}}{\sum_i 1/\widehat{v_i}}, \quad \text{and so,} \quad w_i^* = \frac{1/\widehat{v_i}}{\sum_i 1/\widehat{v_i}} \text{ for all } i.$$

□

Lemma 1 shows that we can rewrite the variance constraint of $\widehat{\mu}$ as

$$\sum_i \left(\frac{1/\widehat{v_i}}{\sum_i 1/\widehat{v_i}} \right)^2 \widehat{v_i} = \sum_i \frac{1/\widehat{v_i}}{(\sum_i 1/\widehat{v_i})^2} = \frac{1}{\sum_i 1/\widehat{v_i}} \leq \alpha.$$

Plugging in $\widehat{v_i} = \sum_j x_{ij} v_j$ and taking the inverse on both sides, constraint (2) becomes

$$1/\alpha \leq \sum_i \frac{1}{\sum_j x_{ij} v_j} \tag{8}$$

Note that the constraints are not linear, but since each $x_{ij} \in \{0,1\}$, and only one $x_{ij} = 1$ for each i, we have $1/\sum_j x_{ij} v_j = \sum_j x_{ij}/v_j$. Thus, we can write our whole program as the following ILP.

$$\min_{x_{ij}} \sum_{i,j} x_{ij} c_{ij} \tag{9}$$

$$\text{subject to} \quad 1/\alpha \leq \sum_i \sum_j x_{ij}/v_j \tag{10}$$

$$\sum_j x_{ij} = 1 \text{ for all } i \tag{11}$$

$$x_{ij} \in \{0,1\} \text{ for all } (i,j) \tag{12}$$

Remark 1. Note that our problem is only interesting if the target variance α is in the range of $[v_1/n, v_m/n]$. This is due to the following observation based on constraint (10): if $1/\alpha < n/v_m$, then the problem is trivial since the variance constraint is satisfied by any assignment; if $1/\alpha > n/v_1$, then the problem is infeasible, i.e. even if we assign the lowest variance level to all providers, the variance constraint is still violated.

4 An MIDR Mechanism via a Linear Programming Relaxation

In order to obtain a computationally efficient mechanism, we consider the LP relaxation of the integer linear program we derived in the previous section, by replacing constraint (12) by $x_{ij} \geq 0$ for all (i,j). We interpret a fractional solution $x_i = (x_{i1}, \ldots, x_{im})$ as a lottery over assignments for player i, i.e. the probabilities of getting assigned to different variance levels. Since the objective is to minimize the total cost, the LP gives a maximum in distributional range allocation rule, where the restricted distributional range is,

$$S_\alpha = \{x \geq 0 \mid \sum_{ij} x_j = 1 \text{ for all } i, \text{ and } \sum_{ij} x_{ij}/v_j \geq 1/\alpha\}.$$

Given a collection of reported costs, our mechanism first computes a distribution x over assignments, based on the MIDR allocation rule defined by the LP. We then pay each provider based on the VCG payment rule, in addition to some entrance reward R. Given the realized variance assignment sampled from x, we ask the providers to compute their estimates μ_i at the corresponding variance levels. Finally, we re-weight the estimates to obtain the linear combination estimator $\widehat{\mu}$ with the minimum variance based on the optimal re-weighting rule in Lemma 1. The formal description of our mechanism is presented in Algorithm 1.

Theorem 2. *Given n data providers with reported costs $\{c_{ij}\}$ for variance levels $\{v_j\}$ and a feasible target variance level α, Algorithm 1 selects an minimum expected cost assignment with a truthful-in-expectation mechanism, and,*

1. *for any $\varepsilon > 0$, computes an estimate $\widehat{\mu}$ with variance $Var(\widehat{\mu}) \leq (1+\varepsilon)\alpha$ as long as*

$$n \geq \left(\frac{v_m}{v_1} - 1\right)\left(\frac{1}{\varepsilon} + 1\right),$$

Algorithm 1 MIDR Mechanism for Buying Estimates

Input: Data providers' reported costs $\{c_{ij}\}$ for different variance levels $\{v_1, \ldots, v_m\}$, target variance α, initial payment R

Compute assignment and payment based on MIDR allocation rule and VCG payment rule:

$$x^* \in \arg\min_{x \in S_\alpha} \sum_i c_{ij} x_{ij} \qquad p_i = \min_{x_{-i} \in S_\alpha} \left[\sum_{i' \neq i} c_{i'}(x) \right] - \sum_{i' \neq i} c_{i'}(x^*) + R$$

Let $\widehat{v} = (\widehat{v}_1, \ldots, v_n)$ be the realized variance assignments sampled from x^* and

$$w_i = \frac{1/\widehat{v}_i}{\sum_i 1/\widehat{v}_i} \qquad \text{for all } i.$$

Collect the estimates from providers $\{\mu_i\}$ based on $\widehat{v}$

Output: $\sum_i w_i \mu_i$ as our estimate $\widehat{\mu}$

2. the mechanism is individually rational if the entrance reward $R \geq \max_i \min_j c_{ij}$.

The properties of cost minimization and truthfulness follow from the MIDR allocation rule and VCG payments. We show the other two properties in the following subsections.

Remark 2. To achieve a 2-approximation for the variance (i.e. $\varepsilon = 1$), it will suffice to have $n = 2v_m/v_1$ providers. Plugging in the bound in Remark 1, the meaningful range of target variance should be $v_1^2/2v_m \leq \alpha \leq v_1/2$. Note that $v_1/v_m < 1$, so this range is always non-empty.

4.1 Variance Violation

The fractional solution we obtain could violate the variance constraint (8), and so could the final assignment sampled from the fractional solution. Let x be an optimal solution to the LP, then x violates the variance constraint (8) by at most

$$\Delta(x) = \sum_i \sum_j x_{ij}/v_j - \sum_i 1/\sum_j x_{ij} v_j = \sum_i (\sum_j x_{ij}/v_j - 1/\sum_j x_{ij} v_j).$$

The quantity $\Delta(x)$ represents the distance between the "real" desired variance constraint and our linear relaxation. Note that for any agent who happens to receive an integral allocation, the corresponding terms in the two constraints are equal, but they may diverge for agents who have fractional allocations. To simplify and bound this quantity, we show that at any optimal fractional solution, all but at most one agent receives an integral allocation:

Lemma 2. *At any extreme point x^* of the feasible region for the LP, there are at least $n-1$ indices i such that $x_{ij} \in \{0, 1\}$ for all j.*

Proof. Suppose not. Then let x be a point in the feasible set S_α such that at least two players (say players 1 and 2) are assigned to lotteries. In other words, each of these two players are assigned

nonzero weight on at least two different variance levels. Let $a < b, k < l$ be the indices such that $x_{1a}, x_{1b}, x_{2k}, x_{2l} \notin \{0, 1\}$. Let $\varepsilon > 0$ be a small enough number such that

$$x_{1a} \pm \varepsilon, x_{1b} \pm \varepsilon, x_{2k} \pm \varepsilon, x_{2l} \pm \varepsilon \in [0, 1]$$

and

$$x_{1a} \pm \varepsilon', x_{1b} \pm \varepsilon', x_{2k} \pm \varepsilon', x_{2l} \pm \varepsilon' \in [0, 1],$$

where $\varepsilon' = \varepsilon \left(\frac{1/v_a - 1/v_b}{1/v_k - 1/v_l} \right)$. Now consider the following two points that differ from x only in four coordinates:

$$\begin{aligned} y: \; & y_{1a} = x_{1a} + \varepsilon, \; y_{1b} = x_{1b} - \varepsilon, \; y_{2k} = x_{2k} - \varepsilon', \text{ and } y_{2l} = x_{2l} + \varepsilon' \\ z: \; & z_{1a} = x_{1a} - \varepsilon, \; z_{1b} = x_{1b} + \varepsilon, \; z_{2k} = x_{2k} + \varepsilon', \text{ and } z_{2l} = x_{2l} - \varepsilon' \end{aligned}$$

Note that $x = 1/2(y + z)$, and recall that $1/\alpha \leq \sum_i \sum_j x_{ij}/v_j$ because $x \in S_\alpha$. Furthermore,

$$\begin{aligned} \sum_i \sum_j y_{ij}/v_j &= \sum_i \sum_j x_{ij}/v_j + \varepsilon(1/v_a - 1/v_b) + \varepsilon'(1/v_l - 1/v_k) \\ &= \sum_i \sum_j x_{ij}/v_j + \varepsilon \left[1/v_a - 1/v_b + (1/v_l - 1/v_k) \frac{1/v_a - 1/v_b}{1/v_k - 1/v_l} \right] \\ &= \sum_i \sum_j x_{ij}/v_j \geq 1/\alpha. \end{aligned}$$

Similarly, $\sum_{i,j} z_{ij}/v_j = \sum_{i,j} x_{ij}/v_j \geq 1/\alpha$, so both y and z are in the feasible region S_α. Since x is a convex combination of y and z that are both in S_α, we know that x cannot be an extreme point of the feasible region.

□

Lemma 2 says that at any extreme point x, at least $n - 1$ players have an integral assignment in x. To use this property, we will compute the solution using an (ellipsoid-based) polynomial-time LP solver from [16] that always returns an optimal extreme point solution.[1] Now we can bound the variance of our aggregate estimate $\widehat{\mu}$.

Lemma 3. *For any $\varepsilon > 0$, the variance of our estimate $Var(\widehat{\mu}) \leq (1 + \varepsilon)\alpha$, as long as*

$$n \geq \left(\frac{v_m}{v_1} - 1 \right) \left(\frac{1}{\varepsilon} + 1 \right).$$

Proof. Suppose that n satisfies the bound above. If the solution x is fully integral, then the variance is no more than α. Otherwise let a be the data provider receiving a lottery in x. Since for every player i with an integral assignment $\sum_j x_{ij}/v_j = \sum_j 1/\sum_j x_{ij} v_j$, we can further simplify,

$$\Delta(x) = \sum_j x_{aj}/v_j - 1/\sum_j x_{aj} v_j.$$

[1] The algorithm consists of two steps: first compute a sufficiently near optimal solution $\widehat{x}$ using the ellipsoid algorithm; then round the solution $\widehat{x}$ to an optimal extreme point solution x^* using the method of continued fractions. For more details, see [16].

Then we can bound the violation of (8) by the final assignment $\widehat{v}$:

$$\sum_j x_{aj}/v_j - 1/v_m \le 1/v_1 - 1/v_m.$$

In other words, the resulting variance $\mathrm{Var}(\widehat{\mu})$ satisfies

$$\frac{1}{\mathrm{Var}(\widehat{\mu})} \ge \frac{1}{\alpha} - (\frac{1}{v_1} - \frac{1}{v_m}).$$

Since we assume $n > v_m/v_1 - 1$, we have $n/v_m - (1/v_1 - 1/v_m) > 0$. As stated earlier in Remark 1, the only interesting range of α is $v_1/n \le \alpha \le v_m/n$. (Recall that if $\alpha < v_1/n$, then the problem is infeasible; if $\alpha > v_m/n$, then the problem is trivial.) For the remainder of the proof, we assume $\alpha \in [v_1/n, v_m/n]$. By this assumption, $1/\alpha - (1/v_1 - 1/v_m) > 0$, and so,

$$\begin{aligned}\mathrm{Var}(\widehat{\mu}) \le \frac{1}{\frac{1}{\alpha} - \frac{1}{v_1} + \frac{1}{v_m}} &= \alpha \left(\frac{1}{1 - \frac{\alpha}{v_1 v_m}(v_m - v_1)} \right) \\ &\le \alpha \left(\frac{n}{n - (\frac{v_m}{v_1} - 1)} \right) \le (1+\varepsilon)\alpha.\end{aligned}$$

□

We also give an example in Appendix A showing that this analysis cannot be improved, and we do need $n = \Omega(v_m/v_1)$ to approximately satisfy the target variance constraint.

4.2 Individual Rationality

In order to ensure individual rationality, we need to set the entrance reward R large enough, so that for each player i , $R + p_i - c_i \ge 0$, where c_i denotes the cost for player i to provide its assigned estimate. To reason about the payment player i gets, we need to compute the following two costs C_1 and C_2, for all players except i. Let x^* be the optimal (fractional) solution for our LP, and $\widehat{v_i}$ be the expected variance level assigned to player i: $\sum_j x^*_{ij} v_j$. Let OPT denote the optimal min-cost value in the LP, and C_1 denote the total cost for all players except i in x^*:

$$C_1 = \min \sum_{a \ne i,j} x_{aj} c_{aj} \tag{13}$$

$$\text{subject to } \sum_{a \ne i,j} x_{aj}/v_j \ge 1/\alpha - 1/\widehat{v_i} \tag{14}$$

$$\sum_j x_{aj} = 1 \text{ for all } a \tag{15}$$

$$x_{aj} \ge 0 \text{ for all } (a,j) \tag{16}$$

Let C_2 be the minimum cost had we removed agent i from the input:

$$C_2 = \min \sum_{a \neq i,j} x_{aj} c_{aj} \tag{17}$$

$$\text{subject to } \sum_{a \neq i,j} x_{aj}/v_j \geq 1/\alpha \tag{18}$$

$$\sum_j x_{aj} = 1 \text{ for all } a \tag{19}$$

$$x_{aj} \geq 0 \text{ for all } (a, j) \tag{20}$$

The VCG payment given to player i in Algorithm 1 is $p_i = C_2 - C_1$. Note that since the second LP is more constrained than the first, we know $C_2 \geq C_1$ and the payment is always non-negative. We can write down the expected utility of player i:

$$R + p_i - c_i = R + C_2 - C_1 - c_i = R + C_2 - \text{OPT}\,.$$

Lemma 4. *The mechanism in Algorithm 1 is individually rational if the entrance reward satisfies*

$$R \geq \max_i \min_j c_{ij}.$$

Proof. Let x_{-i} be the optimal assignment for the second program (with optimal objective value at C_2). Now let's add back player i to the problem, and construct an assignment x such that $x = (x_i, x_{-i})$, where x_i assigns player i to the assignment with minimum cost $(\min_j c_{ij})$.

Note that x is a feasible solution to our original problem since x_{-i} already satisfies the variance constraint. It follows that the cost given by x is at least as large as OPT, the optimal solution:

$$\text{OPT} \leq C_2 + \min_j c_{ij}.$$

Therefore, as long as $R \geq \min_j c_{ij}$ for each player i, we have individual rationality. □

We give an example in Appendix A to show that this bound is tight. In particular, our example shows that without the entrance reward, individual rationality constraint could be violated by up to $\min_j c_{ij}$ for each player i.

Remark 3. Let $c_{min} = \max_i \min_j c_{ij}$. If costs are drawn from a known distribution, the analyst can set R to ensure that with high probability, all players have $c_{min} \leq R$. If c_{min} is unbounded, it is clear that no Groves mechanism[2] can be individually rational for all players in this setting. The Green-Laffont-Holmström theorem [11, 12] shows that under certain technical conditions, any mechanism which is dominant strategy incentive compatible and maximizes welfare must be a Groves mechanism. Thus without additional information on the players' costs, we should not hope to satisfy individual rationality for all players while still achieving our other desiderata.

[2] A *Groves mechanism* is one which selects the welfare maximizing outcome, and each player's payment is his externality plus an amount that is independent of his report. In particular, the payments induced by any Groves mechanism to a player i are shifts of the payments induced by our mechanism, by an amount that is independent of player i's report. Hence by reporting a large enough value of c_{min}, individual rationality can always be violated by a Groves mechanism.

References

[1] Aaron Archer, Christos Papadimitriou, Kunal Talwar, and Éva Tardos. An approximate truthful mechanism for combinatorial auctions with single parameter agents. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pages 205–214, 2003.

[2] Anthony Atkinson, Alexander Donev, and Randall Tobias. *Optimum Experimental Designs, with SAS*. Oxford Statistical Science, 2007.

[3] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] Shahar Dobzinski and Shaddin Dughmi. On the power of randomization in algorithmic mechanism design. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '09, pages 505–514, 2009.

[5] Shahar Dobzinski and Noam Nisan. Limitations of VCG-based mechanisms. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, STOC '07, pages 338–344, 2007.

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC'06, pages 265–284, 2006.

[7] Lisa K. Fleischer and Yu-Han Lyu. Approximately optimal auctions for selling privacy when costs are correlated with data. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 568–585, 2012.

[8] Arpita Ghosh and Katrina Ligett. Privacy and coordination: computing on databases with endogenous participation. In *Proceedings of the 14th ACM conference on Electronic Commerce*, pages 543–560. ACM, 2013.

[9] Arpita Ghosh, Katrina Ligett, Aaron Roth, and Grant Schoenebeck. Buying private data without verification. In *Proceedings of the 15th ACM Conference on Economics and Computation*, EC '14, pages 931–948, 2014.

[10] Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 199–208, 2011.

[11] Jerry R. Green and Jean-Jacques Laffont. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica*, 45(2):427–438, 1977.

[12] Bengt Holmström. Groves' scheme on restricted domains. *Econometrica*, 47(5):1137–1144, 1979.

[13] Thibaut Horel, Stratis Ioannidis, and S. Muthukrishnan. Budget feasible mechanisms for experimental design. In Alberto Pardo and Alfredo Viola, editors, *LATIN 2014: Theoretical Informatics*, Lecture Notes in Computer Science, pages 719–730. 2014.

[14] Ron Lavi and Chaitanya Swamy. Truthful and near-optimal mechanism design via linear programming. *J. ACM*, 58(6):1–24, December 2011.

[15] Katrina Ligett and Aaron Roth. Take it or leave it: Running a survey when privacy comes at a cost. In Paul W. Goldberg, editor, *Internet and Network Economics*, volume 7695 of *Lecture Notes in Computer Science*, pages 378–391. 2012.

[16] George L. Nemhauser and Laurence A. Wolsey. *Integer and combinatorial optimization*. Wiley interscience series in discrete mathematics and optimization. Wiley, 1988.

[17] Kobbi Nissim, Salil Vadhan, and David Xiao. Redrawing the boundaries on purchasing data from privacy-sensitive individuals. In *Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, ITCS '14, pages 411–422, 2014.

[18] Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50. Society for Industrial and Applied Mathematics, 2006.

[19] Aaron Roth and Grant Schoenebeck. Conducting truthful surveys, cheaply. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 826–843, 2012.

A Tightness of Our Bounds

A.1 Example for Variance Violation Bound

Consider an example where there are only two options of variance levels, v_1 and v_2, and we set the target variance $\alpha = \frac{v_1 v_2}{nv_1 + \delta(v_2 - v_1)}$. Suppose the reported costs $c_{i1} = t_1$ and $c_{i2} = t_2$ for each player $i \in [n-1]$, and $c_{n1} < t_1$ and $c_{n2} = t_2$ for player n. We also assume that $t_2 < t_1$. Let x denote the assignment such that $x_{i1} = 0$ and $x_{i2} = 1$ for each $i \in [n-1]$, and $x_{n1} = \delta \in (0,1)$ and $x_{n2} = 1 - \delta$. That is, the assignment gives v_2 to the first $(n-1)$ players, and give a lottery between the two levels to player n. Note that

$$\frac{1}{\alpha} = \frac{n-\delta}{v_2} + \frac{\delta}{v_1}.$$

We know that the fractional solution x exactly satisfies the variance constraint 8, and is also the optimal min-cost solution. Therefore, with probability $(1-\delta)$, the realized variance is

$$\frac{1}{\mathrm{Var}(\widehat{\mu})} = \frac{n}{v_2} = \frac{n-\delta}{v_2} + \frac{\delta}{v_1} + \frac{\delta}{v_2} - \frac{\delta}{v_1} = \frac{1}{\alpha} - \delta\left(\frac{1}{v_1} - \frac{1}{v_2}\right) > 0.$$

It follows that

$$\mathrm{Var}(\widehat{\mu}) = \frac{\alpha}{1 - \alpha\delta\left(\frac{1}{v_1} - \frac{1}{v_2}\right)} = \frac{\alpha}{1 - \frac{\delta\left(\frac{1}{v_1} - \frac{1}{v_2}\right)}{\frac{n}{v_2} + \delta\left(\frac{1}{v_1} - \frac{1}{v_2}\right)}} = \left(1 + \frac{\delta\left(\frac{v_2}{v_1} - 1\right)}{n}\right)\alpha$$

If we want $\frac{\delta(v_2/v_1 - 1)}{n} \le \varepsilon$, we would need to have the number of providers

$$n \ge \left(\frac{v_2}{v_1} - 1\right)\frac{\delta}{\varepsilon}.$$

For δ close to 1 and constant ε, the number of providers we need does scale with v_2/v_1, which shows that the $\Omega(v_m/v_1)$ for n is essentially tight.

A.2 Example for Entrance Reward Bound

Consider an example with two providers, two possible variance levels v_1, v_2 such that $v_2 = 2v_1$, and target variance $\alpha = v_1$. Suppose the costs satisfy $c_{11} = c_{21} = t$ and $c_{12} = c_{22} = t - \varepsilon$ for some $\varepsilon > 0$.

Since we need to an estimate from each provider, the optimal solution is to assign v_2 to both players, which yields cost $\mathrm{OPT} = 2t - 2\varepsilon$. Now suppose we remove any provider from the mechanism. Then we would assign the remaining provider to v_1, which yield cost $C_2 = t$. Therefore, the utility for each provider is

$$R + C_2 - \mathrm{OPT} = R + t - 2(t + \varepsilon) = R + 2\varepsilon - t = R + \varepsilon - t.$$

In order to ensure non-negative utility, we need $R \geq t - \varepsilon$. Note that the right hand side tends to $\max_i \min_j c_{ij}$ when ε tends to 0. Therefore, the bound in Lemma 4 is tight.