Incentivizing Truthful Submissions in a Data Marketplace for Mean Estimation

Keran Chen
University of Wisconsin-Madison

Alex Clinton
University of Wisconsin-Madison

Kirthevasan Kandasamy
University of Wisconsin-Madison

Abstract

We study a data marketplace where a broker intermediates between buyers, who seek to estimate the mean μ of an unknown normal distribution $\mathcal{N}(\mu, \sigma^2)$, and contributors, who can collect data from this distribution at a cost. The broker delegates data collection work to contributors, aggregates reported datasets, sells it to buyers, and redistributes revenue as payments to contributors. We aim to maximize welfare or profit under key constraints: individual rationality for buyers and contributors, incentive compatibility (contributors are incentivized to comply with data collection instructions and truthfully report the collected data), and budget balance (total contributor payments equals total revenue). We first compute welfare/profit-optimal prices under truthful reporting; however, to incentivize data collection and truthful data reporting, we adjust them based on discrepancies in contributors' reported data. This yields a Nash equilibrium (NE) where the two lowest-cost contributors collect all data. We complement this with two impossibility results: *(i)* no nontrivial dominant-strategy incentive-compatible mechanism exists in this problem, and *(ii)* no mechanism outperforms ours in a NE.

1 INTRODUCTION

With the ubiquity of AI, data has become a critical economic resource, driving operations and innovation across many domains. However, not everyone has the means to collect data on their own and thus must rely on *data marketplaces* to acquire the data they need. In recent years, data marketplaces have emerged across various domains, including materials science cit (2024), advertising goo (2024), computer systems Microsoft (2023), insurance dat (2024), freight fre (2024), logistics blo (2024), and others (aws, 2024; ibm, 2024; sno, 2024).

In data marketplaces, strategic data contributors aim to minimize their data collection costs while maximizing compensation from buyers. This can lead to behaviors such as under-collecting data to reduce costs, or misreporting their collected datasets to increase payments. Such actions undermine trust in buyers, who seek high-quality data for their learning tasks. These considerations motivate the central goal of this paper: design data marketplaces that *disincentivize strategic contributor behavior*.

Most prior work on data marketplaces assume that data contributors will always truthfully report data, and focus on setting prices for buyers and payments to contributors based on the *quantity of data*. This overlooks the fact that buyers ultimately care about performance on a given learning task, and pricing based on quantity alone is meaningless when strategic contributors can report data untruthfully. For instance, naively paying contributors based on the size of the dataset they contribute, incentivizes them to report large fabricated (fake) datasets to dishonestly inflate earnings. Prior work on incentivizing truthful data reporting do not consider broader market constraints (e.g. contributor payments must originate from buyers who derive value from the data they purchase).

1.1 Summary of Contributions

In this work, we study these fundamental challenges in a normal mean estimation setting.

Problem Formalism. In §2, we formalize the problem set up. A set of contributors $\mathcal{C}$ act as the *agents*, each able to collect i.i.d. samples from an *unknown* normal distribution $\mathcal{N}(\mu, \sigma^2)$, each incurring a per-sample cost c_i. We assume these costs are public information.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

A set of buyers $\mathcal{B}$ seek to estimate μ by purchasing data from the contributors. A broker serves as the *principal*, who facilitates these transactions by delegating data collection work to contributors (which may not be followed), and receiving data submissions (which may be reported untruthfully). The broker then sells subsets of the received data to buyers and redistributes the revenue to contributors.

Mechanism design objectives. We consider two objectives: welfare and contributors' profit. Welfare is defined as the total value that buyers derive from the allocated data minus contributors' total data-collection costs. Revenue is the total payment collected from buyers. Contributors' profit is then defined as revenue minus contributors' total data-collection costs. The broker itself is not profit-seeking and only serves as the mechanism operator, its objective is to maximize either welfare or profit, while satisfying key market constraints: *(i) Budget balance (BB):* total contributor payments equal total revenue from buyers. *(ii) Individually rational for contributors (IRC):* contributor payments cover data collection costs. *(iii) Incentive-compatible for contributors (ICC):* contributors are incentivized to behave well, *i.e.*, follow the broker's instructions on data collection, and report the collected data truthfully; specifically, being well-behaved constitutes a Nash equilibrium (NE) for all contributors. *(iv) Individually rational for buyers (IRB):* buyers' prices should not exceed their valuation.

Impossibility Results. In §3, we establish two key impossibility results. First, there exists no nontrivial dominant-strategy incentive-compatible (DSIC) mechanism for this problem, i.e. collecting the specified amount and reporting it truthfully is the best strategy regardless of others' strategies. Second, in any NE of any mechanism, the maximum achievable profit is OPT $-(c_2 - c_1)$ where c_1, c_2 are the costs of the cheapest and second cheapest agents, and OPT is a welfare/profit-optimal baseline in which contributors are not strategic. Both results stem from a key insight: when others collect no data, an agent can fabricate data at no cost from an any normal distribution.

Mechanism Design. Next, we design mechanisms that satisfy the four requirements outlined above while maximizing either the welfare (§4) or profit (§5, Appendix E).

Key algorithmic insights and analysis. To minimize costs, one might expect the broker to delegate all data collection to the contributor with the lowest cost. However, this approach is not ICC, as a sole contributor could fabricate data, and the broker—having no other data source—would be unable to detect it. We show, however, that two agents suffice, as the broker can assess each agent's submission by comparing it against the other's. Thus, we assign data collection to the two cheapest contributors, allocating nearly all points to the cheapest, and just one point to the second-cheapest, whose primary role is to enable the broker to verify the submission of the cheapest.

Our mechanisms first determine the total data collection amount, buyers' dataset sizes and prices, and contributor payments to maximize welfare/profit, assuming well-behaved contributors. However, buyers' actual prices include an additional term based on the difference between the means of the two contributors' reported datasets, and this variation is reflected in the contributor payments. This design ensures ICC, as each contributor is incentivized to report truthfully when the other does, in order to minimize the discrepancy. This mechanism achieves a welfare (profit) which matches (nearly matches) the OPT $-(c_2 - c_1)$ upper bound proved in our impossibility result in §3.

1.2 Related Work

Incentivizing Data Collection. Prior work has studied methods to incentivize data collection, both via payments in principal-agent models (Cai et al., 2015; Cummings et al., 2015; Fallah et al., 2024), and via reciprocal data sharing in collaborative learning settings (Karimireddy et al., 2022; Huang et al., 2023; Sim et al., 2020; Blum et al., 2021). However, they assume agents will always truthfully report the collected data. Indeed, in many cases, they are able to design DSIC mechanisms for contributors, which, as we demonstrate, is impossible in our problem. Moreover, some methods assume specific forms for the valuation of the data for the principal. In contrast, our approach is more general: we only assume that buyers' valuation decreases with their estimation error.

Incentivizing Truthful Data Reporting. Previous work on incentivizing truthful reporting of already collected data (Zheng et al., 2024; Chen et al., 2020; Ghosh et al., 2014) do not study many of the market constraints we study here, such as efficiency, IRB, and IRC. In some of these models, agents do not incur costs to collect data. Moreover, they assume that the principal has access to a large budget for incentivization, but do not address the source of these funds. In contrast, our setting requires that payments originate from buyers who derive value from the data contributed by contributors. Other related work Cummings et al. (2015) assumes that agents submit unbiased data and only add noise under a variance constraint; however, this restriction on agents' strategic behavior is too strong for practical settings. Chen et al. (2018) investigate truthful mechanisms for obtaining data from strategic agents who have different costs; however, these agents can only misreport their costs and not the actual data. Miller et al. (2005) introduced peer prediction for eliciting truthful reports without ground truth, focusing purely on report incentives; our mechanism additionally ensures market feasibility, accounting for contributors' data collection costs and other key market constraints.

Truthful Contributions in Collaborative Mean Estimation. Recent work has explored collaborative mean estimation, where a group of strategic agents collaborate to estimate the mean of a distribution (Chen et al., 2023; Clinton et al., 2024; Dorner et al., 2023). A common technique in our method and these works is comparing an agent's reported data mean against the mean of other agents to incentivize truthful reporting. However, our analysis techniques are different from these methods, in part because we must adhere to market constraints that do not arise in payment-less collaborative learning settings, and in part because all three methods assume specific forms for the agents' valuation of data.

Moral Hazard. The issue of contributors submitting fabricated data can also be analyzed under the framework of *moral hazard*, where agents (contributors) deviate from agreed-upon tasks (data collection) with a principal (broker) (Laffont and Martimort, 2009; Holmström, 1979; Mirrlees, 1999; Helpman and Laffont, 1975). As in classical moral hazard settings, the outcome of a contributor's work (the collected dataset) is stochastic. However, unlike typical models, our broker lacks knowledge of the signal distribution given contributors' effort (the learning problem would be trivial if the data distribution were known) and cannot directly observe the outcomes of these efforts, relying instead on potentially dishonest reports from the contributors themselves.

Data Pricing. Our approach to profit maximization leverages an extensive body of work on data pricing (Agarwal et al., 2020, 2019; Bergemann and Bonatti, 2019; Chen and Hossain, 2023; Bergemann et al., 2018; Babaioff et al., 2012; Mehta et al., 2021; Pei, 2020). Of these, pertinent to us, is a line of work on *pricing ordered items*, where unit-demand buyers share a common preference ranking over goods (Chawla et al., 2022; Hartline and Koltun, 2005; Chen et al., 2024). We will use algorithms from these works when studying envy-free profit maximization.

2 PROBLEM SET UP

We now present the problem set up. We describe the marketplace environment in §2.1, followed by outlining the mechanism design problem in §2.2. A table of notations is provided in Appendix F.

2.1 Description of the Marketplace

A data marketplace involves three key actors: *(i)* a finite set of *strategic* data contributors $\mathcal{C} = \{1, \ldots, |\mathcal{C}|\}$, *(ii)* a finite set of *non-strategic* buyers $\mathcal{B}$, and *(iii)* a trusted broker who facilitates transactions between contributors and buyers. Contributors collect samples from a *common* normal distribution $\mathcal{N}(\mu, \sigma^2)$. Here, σ^2 is known. Neither the contributors, buyers, or the broker know μ, nor do they have any ancillary information, such as a prior. Buyers wish to estimate μ by purchasing data from the contributors, but have different valuations based on estimation errors. Contributor i incurs a cost c_i to collect each sample, where c_i is *known to the broker*.

Interaction Protocol. The marketplace operates in the following sequence:

1. *Mechanism design (broker)*: The broker designs and publishes a mechanism M to delegate data collection work to contributors, elicit the collected data, set prices, sell data to buyers, and redistribute revenue back to contributors.
2. *Delegation of work (broker)*: The broker requests each contributor i to collect $N'_i = N_i(\{c_i\}_{i\in\mathcal{C}}) \in \mathbb{N}$ samples, where N_i maps the costs to requested quantities.
3. *Data collection (contributors)*: Each contributor i collects $n'_i = n_i(N'_i, c_i) \in \mathbb{N}$ i.i.d. samples at cost $c_i n'_i$, yielding a dataset $Z_i = \{z_{i,1}, \ldots, z_{i,n'_i}\} \in \mathbb{R}^{n'_i}$. The function n_i maps the broker's request and the contributor's cost to the actual number of samples she will collect (which may not be the amount requested).
4. *Data submission (contributors)*: Each contributor submits a dataset $X'_i = \{x'_{i,1}, x'_{i,2}, \ldots\} = X_i(Z_i, N'_i, c_i) \in \bigcup_{\ell=0}^{\infty} \mathbb{R}^{\ell}$. The submission function X_i may modify the collected dataset Z_i, enabling strategic behaviors (*e.g.*, fabrication) for personal gain.
5. *Data allocation and purchases (broker)*: The broker allocates to each buyer $j \in \mathcal{B}$ a random subset of size $m'_j = m_j(\{X'_i\}_{i\in\mathcal{C}}, \{c_i\}_{i\in\mathcal{C}}) \in \mathbb{N}$ drawn from $\bigcup_{i\in\mathcal{C}} X'_i$, and charges a price $p'_j = p_j(\{X'_i\}_{i\in\mathcal{C}}, \{c_i\}_{i\in\mathcal{C}}) \in \mathbb{R}$. Let Y'_j be the dataset received by buyer j, obtained by sampling a random subset of size m'_j from the union of all contributors' datasets $\bigcup_{i\in\mathcal{C}} X'_i$.
6. *Revenue redistribution (broker)*: The broker redistributes total revenue $\sum_{j\in\mathcal{B}} p'_j$ to contributors via payments $\{\pi'_i\}_{i\in\mathcal{C}}$, where $\pi'_i = \pi_i(\{X'_i\}_{i\in\mathcal{C}}, \{c_i\}_{i\in\mathcal{C}}) \in \mathbb{R}$, and each payment function π_i maps available information to contributor i's payment.

Notation. Primed symbols denote realizations of quantities that are functions of information available to a contributor or broker, *e.g.*, in step 3, the function n_i maps the requested amount and cost to the actual amount contributor i will collect, denoted n'_i.

Mechanism. Let $\mathcal{X} = \bigcup_{\ell=0}^{\infty} \mathbb{R}^\ell$ denote the space of datasets. A broker's mechanism is specified by the tuple of (possibly randomized) functions $M = (\{N_i\}_{i\in\mathcal{C}}, \{m_j\}_{j\in\mathcal{B}}, \{p_j\}_{j\in\mathcal{B}}, \{\pi_i\}_{i\in\mathcal{C}})$. Here, $N_i : \mathbb{R}_+^{|\mathcal{C}|} \to \mathbb{N}$ determines the amount of data to request from contributor i based on their costs. Next, $m_j : \mathcal{X}^{|\mathcal{C}|} \times \mathbb{R}_+^{|\mathcal{C}|} \to \mathbb{N}$ determines the size of the dataset sold to buyer j; it should satisfy $m_j(\{X'_i\}_{i\in\mathcal{C}}, \cdot) \leq \sum_{i\in\mathcal{C}} |X'_i|$ for all $j \in \mathcal{B}$, as the broker cannot sell more data to any buyer than she has received from contributors. Next, $p_j : \mathcal{X}^{|\mathcal{C}|} \times \mathbb{R}_+^{|\mathcal{C}|} \to \mathbb{R}$ specifies the price charged to buyer j. Finally, $\pi_i : \mathcal{X}^{|\mathcal{C}|} \times \mathbb{R}_+^{|\mathcal{C}|} \to \mathbb{R}$ determines the payment to contributor i.

Contributors. Without loss of generality, we assume contributors are ordered by increasing cost: $c_1 \leq c_2 \leq \cdots \leq c_{|\mathcal{C}|}$. After the mechanism is published, each contributor i selects a strategy, which is a tuple of (possibly randomized) functions $s_i = (n_i, X_i)$, where: $n_i : \mathbb{N} \times \mathbb{R}_+ \to \mathbb{N}$ determines how much data to collect, based on the broker's request and the cost; and $X_i : \mathcal{X} \times \mathbb{N} \times \mathbb{R}_+ \to \mathcal{X}$ determines what dataset to submit, possibly modifying or fabricating the collected dataset. This formalism accommodates a wide range of strategic behavior, including under-collecting data to reduce data collection costs (*i.e.*, $n'_i = n_i(N'_i, c_i) < N'_i$), and fabricating data to appear compliant (*i.e.*, $|X'_i| = N_i > |Z_i| = n'_i$, where $X'_i = X_i(Z_i, N_i, c_i)$).

Well-behaved contributor. Our goal will be to design mechanisms that incentivize contributors to be "well-behaved", *i.e.*, truthfully report their costs, collect the requested amount of data, and submit it without alteration. To formalize this, let $\mathbb{I}$ denote the identity function, which returns its first argument: $\mathbb{I}(x, y, z, \dots) = x$. Define the well-behaved strategy for contributor i as $s_i^{\mathbb{I}} = (\mathbb{I}, \mathbb{I})$. Let $s^{\mathbb{I}}$ ($s_{-i}^{\mathbb{I}}$) denote the strategy profile in which all contributors (all except i) are well-behaved.

Contributor utility. If a contributor samples n'_i points and receives payment $\pi'_i \in \mathbb{R}$, her *ex post* utility is $\pi'_i - c_i n'_i$. As these quantities depend on the mechanism M and all contributors' strategies s, her *ex ante* utility u_i^{C} can be defined as[1],

$$u_i^{\mathrm{C}}(M, s) := \inf_{\mu\in\mathbb{R}} \mathbb{E}\left[\pi'_i - c_i n'_i\right]. \tag{1}$$

Here, the expectation is over stochasticity in data generation, and any randomness in the mechanism and contributor strategies. We consider the infimum (worst-case) value over μ as μ is unknown, and a process should yield reliable estimates across all $\mathcal{N}(\mu, \sigma^2)$. To illustrate further, suppose the true mean is $\mu = \widetilde{\mu}$. Suppose a contributor chooses not to collect any data, $n'_i = 0$ (incurring zero cost), and reports a large vector of repeated copies of $\widetilde{\mu}$, i.e., $X'_i = (\widetilde{\mu}, \dots, \widetilde{\mu})$. This benefits buyers, as they may obtain arbitrarily accurate estimates of μ using the sample mean. As a result, buyers would be willing to pay more, which increases the contributor's payment without incurring any cost. However, this strategy is viable only if the contributor knew *a priori* that $\mu = \widetilde{\mu}$. Taking the infimum over μ accounts for the fact that μ is unknown and makes u_i^{C} well-defined[2].

Buyers. Buyer j receives a dataset Y'_j from the mechanism, pays the specified price p'_j to the broker, and estimates μ via the sample mean $\widehat{\mu}(Y'_j)$. While prior work on data marketplaces has often modeled buyer valuation as a function of dataset size (Zhang et al., 2023), such a model is inadequate when contributors may misreport data. A more principled approach, rooted in classical statistics, measures performance by the estimation error $|\widehat{\mu}(Y'_j) - \mu|$ (Stein, 1981). Motivated by this, we adopt a more appropriate model where a buyer's valuation depends on the estimation error $|\widehat{\mu}(Y'_j) - \mu|$, which directly reflects the buyer's ultimate objective. We provide additional background on normal mean estimation in Appendix A.

Buyer valuations. We will find it useful to define three related valuation functions for buyers:

1. *Error-based valuation v_j^{e}:* Different buyers may have differing valuations for varying errors, which we model using buyer-specific valuation functions $v_j^{\mathrm{e}} : \mathbb{R}_+ \to [0, 1]$. Here, $v_j^{\mathrm{e}}(e)$ is buyer j's *ex-post* value when her estimation error is e. We adopt general error-based valuation functions v_j^{e} to capture heterogeneity across buyers. In real-world markets, buyers often differ in their sensitivity to estimation errors: some may place high value on small accuracy improvements, while others may tolerate larger errors. Allowing v_j^{e} to be buyer-specific better reflects diverse buyer objectives. The next two definitions build on v_j^{e}.
2. *Valuation under a mechanism and strategy profile v_j^{B}:* Recall that buyer j's dataset Y'_j is a random subset of size m'_j drawn from all contributors' submissions (step 5 in interaction protocol). It depends

[1] Making the dependencies on the mechanism and strategies explicit, we can write this as $u_i^{\mathrm{C}}(M, s) := \inf_{\mu\in\mathbb{R}} \mathbb{E}\left[\pi_i\big(\{X_j(Z_j)\}_{j\in\mathcal{C}}, \{c_j\}_{j\in\mathcal{C}}\big) - c_i n_i\left(N_i(\{c_j\}_{j\in\mathcal{C}}), c_i\right)\right]$. To reduce notational clutter, we will use the primed notation going forward.

[2] An alternative approach to address uncertainty about μ is to place a prior over μ and take an expectation with respect to that prior. While we do not study this Bayesian setting, our high-level ideas can be adapted to it.

on the mechanism M, the contributor strategy profile s, and the stochastic datasets $\{Z_i\}_{i\in\mathcal{C}}$ collected by the contributors. Thus, buyer j's value under mechanism M and strategy profile s is:

$$v_j^{\mathrm{B}}(M,s) := \inf_{\mu\in\mathbb{R}} \mathbb{E}\left[v_j^{\mathrm{e}}\left(\left|\,\widehat{\mu}(Y_j') - \mu\,\right|\right)\right]. \quad (2)$$

3. *Valuation under clean data* v_j^{iid}*:* As a benchmark, we also define a size-based valuation v_j^{iid}, where $v_j^{\mathrm{iid}}(m)$ is buyer j's expected value upon receiving m i.i.d. (clean) samples:

$$v_j^{\mathrm{iid}}(m) = \inf_{\mu\in\mathbb{R}} \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)^m}\left[v_j^{\mathrm{e}}\left(\left|\,\widehat{\mu}(X) - \mu\,\right|\right)\right]. \quad (3)$$

We will assume that v_j^{e} is such that $v_j^{\mathrm{iid}}(m)$ is increasing in m, *i.e.*, if a buyer receives more clean data, her value increases in expectation. As in (1), the infimum in (2) and (3) ensures v_j^{B} and v_j^{iid} are well-defined as μ is unknown.

Buyer utility. A buyer's utility is her valuation minus her payment. Defining this is somewhat subtle, as infima and expectations do not commute. We consider two candidate definitions:

$$u_j^{\mathrm{B}}(M,s) := v_j^{\mathrm{B}}(M,s) - \sup_{\mu\in\mathbb{R}} \mathbb{E}\left[p_j'\right]. \quad (4)$$

$$u_j^{\mathrm{B}}(M,s) := \inf_{\mu\in\mathbb{R}} \mathbb{E}\left[v_j^{\mathrm{e}}\left(\left|\widehat{\mu}(Y_j') - \mu\,\right|\right) - p_j'\right]. \quad (5)$$

Here, (4) measures the gap between the buyer's worst-case expected value (2) and her worst-case expected price, whereas (5) computes the buyer's *ex-post* utility $v_j^{\mathrm{e}}(e) - p_j$ for realized error e, and price p_j' followed by a worst-case expectation. Clearly, (4) $\le$ (5), so guarantees established for (4) also hold for (5). Accordingly, we adopt (4) as our measure of buyer utility. That said, as we will see, in our mechanism, both expressions will be equal, as the price is independent of μ.

Non-strategic buyers, known buyer valuations. As we focus on contributor-side challenges, we assume buyers are non-strategic and that the broker knows buyer valuations $\{v_j^{\mathrm{e}}\}_{j\in\mathcal{B}}$. Extensions to unknown valuations—elicited via bids or learned in repeated markets—is left for future work.

Public Information. We assume that the set of contributors $\mathcal{C}$, buyers $\mathcal{B}$, buyer valuations $\{v_j^{\mathrm{e}}\}_{j\in\mathcal{B}}$, and the variance σ^2 are public information known to all parties. Contributors' costs $\{c_i\}_{i\in\mathcal{C}}$ are known only to the contributors themselves and the broker. A contributor's strategy and the broker's mechanism may depend on this information, but we suppress this dependence for simplicity.

2.2 The Mechanism Design Problem

Market Constraints. We wish to design a mechanism in which buyers and contributors do not lose by participation, contributors are incentivized to behave well, and payments to contributors equal the revenue. We now formalize these desiderata:

1. *Budget balance (BB):* M satisfies budget balance if the total payments to contributors is equal to the total revenue, i.e., $\sum_{i\in\mathcal{C}} \pi_i' = \sum_{j\in\mathcal{B}} p_i'$.
2. *Individually rational for contributors (IRC):* M is *ex-ante* individually rational for the contributors if $u_i^{\mathrm{C}}(M,s^{\mathbb{I}}) \ge 0$ for all contributors $i\in\mathcal{C}$.
3. *Incentive-compatible for contributors (ICC):* M is incentive-compatible for contributors if $s^{\mathbb{I}}$ is a Nash equilibrium. That is, for all contributors $i\in\mathcal{C}$, and for every alternative strategy s_i for i, we have $u_i^{\mathrm{C}}(M,s^{\mathbb{I}}) \ge u_i^{\mathrm{C}}(M,(s_i,s_{-i}^{\mathbb{I}}))$. (In §3, we show that no DSIC mechanisms are possible).
4. *Individually rational for buyers (IRB):* M is *ex ante* individually rational for the buyers if $u_j^{\mathrm{B}}(M,s^{\mathbb{I}}) \ge 0$ for all buyers $j\in\mathcal{B}$.

Efficiency. A mechanism should satisfy some notion of efficiency while meeting the above criteria, where common objectives include welfare and profit maximization. Our contributor-side mechanism supports both. Here, we focus on welfare maximization, which is simpler to implement on the buyer side. In §5 and Appendix E, we extend these ideas to profit maximization.

First, let us define the welfare $W(M,s;\mu)$ of a mechanism M under a strategy profile s:

$$W(M,s;\mu) = \mathbb{E}\left[\sum_{j\in\mathcal{B}} v_j^{\mathrm{B}}(M,s) - \sum_{i\in\mathcal{C}} c_i n_i'\right]. \quad (6)$$

For an arbitrary mechanism, the welfare may depend on reported data, which itself depends on the actual distribution $\mathcal{N}(\mu,\sigma^2)$. However, we will see that for our mechanism, under $s^{\mathbb{I}}$, the welfare $W(M,s^{\mathbb{I}};\mu)$ is independent of μ.

A baseline for welfare maximization. Our approach is to define a welfare-optimal benchmark OPT and design a mechanism that satisfies the four desiderata while achieving welfare close to OPT. A natural baseline is the maximum welfare achievable when contributors are non-strategic and follow $s^{\mathbb{I}}$. In this case, since buyers receive i.i.d. data, their valuations depend only on the data amounts $\{m_j'\}_{j\in\mathcal{B}}$ (see (3)). Additionally, the data allocated to any buyer cannot exceed the total collected: $m_j' \le \sum_{i\in\mathcal{C}} n_i' = \sum_{i\in\mathcal{C}} N_i'$. These observations define the baseline OPT:

$$\mathrm{OPT} = \max_{M} \mathbb{E}\left[\left(\sum_{j\in\mathcal{B}} v_j(M,s^{\mathbb{I}}) \;-\; \sum_{i\in\mathcal{C}} c_i N_i'\right)\right]$$

$$\stackrel{(a)}{=} \max_{\{N_i'\}_i,\{m_j'\}_j} \left(\sum_{j\in\mathcal{B}} v_j^{\mathrm{iid}}(m_j') - \sum_{i\in\mathcal{C}} c_i N_i'\right)$$

$$= \max_{\{N'_i\}_i} \left(\max_{\{m'_j\}_j} \sum_{j \in \mathcal{B}} v_j^{\text{iid}}(m'_j) - \sum_{i \in \mathcal{C}} c_i N'_i \right)$$

$$\stackrel{(b)}{=} \max_{\{N'_i\}_i} \left(\sum_{j \in \mathcal{B}} v_j^{\text{iid}} \left(\sum_{i \in \mathcal{C}} N'_i \right) - \sum_{i \in \mathcal{C}} c_i N'_i \right)$$

$$\stackrel{(c)}{=} \sum_{j \in \mathcal{B}} v_j^{\text{iid}} \left(N^{\text{OPT}} \right) - c_1 N^{\text{OPT}}, \tag{7}$$

where $N^{\text{OPT}} = \text{argmax}_{N' \in \mathbb{N}} \sum_{j \in \mathcal{B}} v_j^{\text{iid}}(N') - c_1 N'$. Here, (a) follows from (3), the fact that welfare depends only on the quantities $\{N'_i\}_i, \{m'_j\}_j$, not on payments or prices, and that N'_i does not depend on the reported data. As v_j^{iid} is non-decreasing, step (b) sets $m'_j = \sum_{i \in \mathcal{C}} N'_i$ to maximize $v_j^{\text{iid}}(m'_j)$ under the constraint $m_j \leq \sum_{i \in \mathcal{C}} N'_i$, reflecting that welfare is maximized when all collected data is allocated to all buyers. Step (c) assigns all data collection to contributor 1, the lowest-cost contributor, to minimize total cost.

Some Remarks. We conclude this section with three observations. *(1) Buyer valuation for untruthful data:* We do not explicitly model the fact that truthful data maximizes buyer valuations. Doing so would require additional assumptions on v_i and an analysis of optimal estimators. Instead, we define a baseline where *non-strategic* contributors follow the rules, ensuring buyers receive truthful data. We aim to design a mechanism that incentivizes strategic contributors to also follow the rules while approximating the baseline. *(2) Negative prices/payments:* Since data is random, data-dependent prices and payments are also random. We allow *ex post* negative prices (broker pays buyers) and payments (contributors pay broker). Although unconventional, this enables strong ICC guarantees. Notably, in our mechanism, the *ex-ante* (i.e. expected) prices and payments are always non-negative. *(3) On BB:* Payments cannot exceed revenue to ensure market feasibility. We additionally require equality to ensure that trade benefits are fully distributed to contributors, leaving no unallocated excess revenue. Unlike the other requirements, BB must hold always and not just in expectation, as the market must be feasible for every realization of data.

3 IMPOSSIBILITY RESULTS

We first present two impossibility results for this problem. Theorem 3.1 states that no nontrivial dominant-strategy incentive-compatible mechanism exists, motivating our focus on NE in the ICC guarantee. Theorem 3.2 establishes an upper bound on the maximum achievable welfare in any NE of any mechanism that ensures truthful data reporting. The proofs of both theorems are given in Appendix D.

Theorem 3.1. *For any mechanism M, if $s = \{(n_i, X_i)\}_{i \in \mathcal{C}}$ is a dominant strategy profile, then no contributor collects any data, i.e., $\forall i \in \mathcal{C}$, $n'_i = n_i(N'_i, c_i) = 0$.*

Theorem 3.2. *Let* OPT *be as defined in* (7)*. For any mechanism M, if strategy profile $s = \{(n_i, \mathbb{I})\}_{i \in \mathcal{C}}$ is a Nash equilibrium, then $W(M, s; \mu) \leq \text{OPT} - (c_2 - c_1)$.*

Next, we will design a mechanism that achieves the welfare upper bound established here, while at the same time satisfying BB, IRC, ICC, and IRB.

4 METHODS AND ANALYSIS

We have outlined our mechanism in Algorithm 1. For the purpose of extending this algorithm for profit maximization, we use four sets of input parameters: $\widetilde{\text{OPT}}$ is the optimal baseline welfare or profit, $N^{\widetilde{\text{OPT}}}$ specifies the total amount of data that needs to be collected, $\{\widetilde{m}_j\}_{j \in \mathcal{B}}$ specifies the amounts of data to sell to each buyer, and $\{\widetilde{p}_j\}_{j \in \mathcal{B}}$ is the expected price to charge each buyer. For welfare maximization, we will set $\widetilde{\text{OPT}} = \text{OPT}$, $N^{\widetilde{\text{OPT}}} = N^{\text{OPT}}$, where OPT and N^{OPT} are as defined in (7), and set $\widetilde{m}_j = N^{\text{OPT}}$, $\widetilde{p}_j = v_j^{\text{iid}}(N^{\text{OPT}})$ for all j. We will choose these values differently for profit maximization in Appendix E.

We will first describe our mechanism and then motivate our design choices. In line 2, our mechanism instructs the cheapest contributor to collect $N^{\widetilde{\text{OPT}}} - 1$ points, the second cheapest contributor to collect 1 point, and the remaining to collect no points. It then aggregates all the received datasets and chooses $m'_j = \widetilde{m}_j$ random points to sell to buyer j (for welfare maximization, the broker will sell all the data to all buyers as $\widetilde{m}_j = N^{\text{OPT}}$). He then charges buyer j p'_j as shown below in (8), and issues payments π'_i to the contributors as shown below in (9).

To state prices and payments, for $i \in \{1, 2\}$, let X'_{-i} be the dataset submitted by the other contributor, $N'_{-i} = N^{\text{OPT}} - N'_i$ be the amount of data the other contributors should collect, and $d_i = c_i (N'_i)^2 / \sigma^2$. Recall that the input $\widetilde{p}_j$ specifies buyer j's expected price. We can now express p'_j as:

$$p'_j = \sum_{i \in \{1,2\}} \mathbb{I}\left(|X'_i| = N'_i\right) \left(\widetilde{p}_j \frac{N'_i}{N^{\widetilde{\text{OPT}}}} + \frac{d_i}{|\mathcal{B}|} \frac{\sigma^2}{N'_{-i}} + \frac{d_i}{|\mathcal{B}|} \frac{\sigma^2}{N'_i} \right)$$
$$- \sum_{i \in \{1,2\}} \frac{d_i}{|\mathcal{B}|} \left(\widehat{\mu}(X'_1) - \widehat{\mu}(X'_2) \right)^2. \tag{8}$$

Next, we set payments $\pi'_i = 0$ for contributors $i \in \{3, \ldots, |\mathcal{C}|\}$. For contributors $i \in \{1, 2\}$ we have,

Algorithm 1 Mechanism

1: **Input:** Target welfare or profit $\widetilde{\mathrm{OPT}}$, Total amount of data to collect $N^{\widetilde{\mathrm{OPT}}} \in \mathbb{N}$, Amounts to sell to buyers $\{\widetilde{m}_j\}_{j\in\mathcal{B}} \in \mathbb{N}^{|\mathcal{B}|}$, Expected prices to charge buyers $\{\widetilde{p}_j\}_{j\in\mathcal{B}} \in \mathbb{R}^{|\mathcal{B}|}$.
2: Instruct contributors to collect $\{N_i'\}_{i\in\mathcal{C}}$ where,
$$N_1' = N^{\widetilde{\mathrm{OPT}}} - 1,\ N_2' = 1,\ N_j' = 0 \text{ for all } j \geq 3.$$
3: Receive datasets $\{X_i'\}_{i\in\mathcal{C}}$ from contributors.
4: Set $m_j' = \widetilde{m}_j$. Buyer j receives m_j' randomly chosen points from $\bigcup_{i\in\mathcal{C}} X_i'$.
5: Charge p_j' from each buyer $j \in \mathcal{B}$ (see (8)). # p_j' uses $\widetilde{p}_j$
6: Pay π_i' to each contributor $i \in \mathcal{C}$ (see (9)).

$$\pi_i' = \mathbb{I}\left(|X_i'| = N_i'\right)\left((\widetilde{\mathrm{OPT}} + c_1 - c_2)\frac{N_i'}{N^{\widetilde{\mathrm{OPT}}}} + c_i N_i' + d_i\frac{\sigma^2}{N_{-i}'} + d_i\frac{\sigma^2}{N_i'}\right) - d_i\big(\widehat{\mu}(X_1') - \widehat{\mu}(X_2')\big)^2. \tag{9}$$

Design Choices. As indicated in (7), collecting N^{OPT} data maximizes welfare. To minimize total cost, one would expect all the work to be assigned to the cheapest contributor. However, if only one contributor were to collect all the data, she could fabricate it; the broker—having no other data source—will not be able to detect it. Fortunately, we show that using just one data point from a second contributor is sufficient to incentivize contributor 1 to follow the rules. Hence, we ask the second cheapest contributor to collect one point, and the rest is collected by contributor 1.

Contributor payments distribute the profit proportional to contributions, with contributor i's *expected payment* given by $c_i N_i' + (\mathrm{OPT} + c_1 - c_2)N_i'/N^{\mathrm{OPT}}$. Actual payments are adjusted based on the difference between the means of contributors' datasets, with a larger difference lowering payments. We prove that each contributor should report truthfully when the other does so, to minimize this difference, thus achieving ICC. This difference is reflected in buyer prices to ensure BB and IRB.

The following theorem summarizes the key properties of our mechanism. It satisfies all the required constraints while achieving a welfare that matches the upper bound in Theorem 3.2, demonstrating that the mechanism is essentially unimprovable. The proof is given in Appendix B.

Theorem 4.1. *Let M be the mechanism in Algorithm 1 executed with inputs $\widetilde{\mathrm{OPT}} = \mathrm{OPT}$, $N^{\widetilde{\mathrm{OPT}}} = N^{\mathrm{OPT}}$, $\widetilde{m}_j = N^{\mathrm{OPT}}$, and $\widetilde{p}_j = v_j^{\mathrm{iid}}(N^{\mathrm{OPT}})$ for all $j \in \mathcal{B}$, where OPT and N^{OPT} are as defined in (7). Then, M satisfies BB, IRC, ICC, and IRB. Moreover, its expected welfare at the well-behaved NE is $W(M, s^{\mathbb{I}}; \mu) = \mathrm{OPT} - (c_2 - c_1)$.*

Discussion. We highlight three observations. *1) Multiple minimum-Cost Contributors:* If multiple contributors share the minimum cost, evenly distributing data collection among them yields a welfare of OPT and reduces the variance in the comparison terms of equations (8) and (9). This lowers the variance of prices and payments which may be desirable in practice. *2) Beyond Mean Estimation:* Our insights likely generalize to learning tasks beyond mean estimation, such as the Gaussian family, where similar discrepancy-based incentives may be constructed with minor modifications. We anticipate the absence of nontrivial DSIC mechanisms and the sufficiency of two contributors for an optimal NE. However, the requirement of a single sample from the second contributor is specific to normal mean estimation as it has a single learnable parameter. Generally, the number of samples needed from the second contributor may scale with the model's degrees of freedom. *3) Envy-free pricing:* The pricing in (8) may seem "unfair" as buyers may pay different prices for identical datasets. A natural fairness requirement in this context is envy-freeness (EF) (Varian, 1973; Guruswami et al., 2005). While EF is not our primary focus here for simplicity, §5 and Appendix E show that our mechanism can be adapted for EF profit maximization.

4.1 Proof Sketch of ICC in Theorem 4.1

We now, outline a proof sketch for incentive-compatibility. To prove ICC, we must show that $s^{\mathbb{I}}$ is a Nash equilibrium of M. That is, we will show, $\forall i \in \mathcal{C}$, $\forall (n_i, X_i)$,
$$u_i^{\mathrm{C}}\left(M, ((n_i, X_i), s_{-i}^{\mathbb{I}})\right) \leq u_i^{\mathrm{C}}\left(M, (s_i^{\mathbb{I}}, s_{-i}^{\mathbb{I}})\right).$$

Note that because we are only asking the two contributors with the lowest collection costs to collect data, we can focus on $\{1, 2\} \in \mathcal{C}$ as the above inequality is immediately satisfied for all other contributors. We now prove the inequality for these contributors in three steps, each with a corresponding lemma.

Step 1. Let (n_i, X_i) be given. First we show that a contributor can improve her utility by switching their strategy from (n_i, X_i) to (n_i, X_a), where X_a is a function that always submits exactly the requested amount of points N_i', while maintaining the same sample mean as X_i. This is true as our mechanism determines a contributors penalty based on the sample mean of their submitted data and whether or not they submitted exactly N_i' points. This is stated formally in the lemma below.

(Lemma 1): Fix any (n_i, X_i). Let $X_a : \mathcal{X} \to \mathbb{R}^{N_i'}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}(X_a(Z_i)) = \widehat{\mu}(X_i(Z_i))$. Then,

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_i), s_{-i}^{\mathbb{I}}\right)\right) \leq u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right).$$

Step 2. The second step is showing that a contributor can improve their utility by switching their strategy from (n_i, X_a) to (n_i, X_b), where X_b is a function that always submits exactly N_i' points and whose sample mean agrees with the sample mean of the original dataset she collected Z_i. This step uses the term in (9) which penalizes contributors according to the difference in reported means. Its proof uses techniques from minimax optimality for normal mean estimation (Lehmann and Casella, 2006). This is stated formally in the lemma below.

(Lemma 2): Fix any n_i. Consider any $X_a : \mathcal{X} \to \mathbb{R}^{N_i'}$ and let $X_b : \mathcal{X} \to \mathbb{R}^{N_i'}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}(X_b(Z_i)) = \widehat{\mu}(Z_i)$. Then,

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right) \leq u_i^{\mathrm{C}}\left(M, \left((n_i, X_b), s_{-i}^{\mathbb{I}}\right)\right).$$

Step 3. Finally, we show a contributor can improve their utility by switching their strategy from (n_i, X_b) to $(\mathbb{I}, X_b)$. We prove that when a contributor is using X_b their utility is concave in n_i', which is maximized when $n_i' = N_i'$. We then show the utility under $(\mathbb{I}, X_b)$ is the same as under $s^{\mathbb{I}}$.

(Lemma 3): Fix any n_i. Let $X_b : \mathcal{X} \to \mathbb{R}^{N_i}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}(X_b(Z_i)) = \widehat{\mu}(Z_i)$. Then,

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_b), s_{-i}^{\mathbb{I}}\right)\right) \leq u_i^{\mathrm{C}}\left(M, \left(s_i^{\mathbb{I}}, s_{-i}^{\mathbb{I}}\right)\right).$$

These three steps now combine to prove ICC. For any (n_i, X_i), we then have that

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_i), s_{-i}^{\mathbb{I}}\right)\right) \leq u_i^{\mathrm{C}}\left(M, s^{\mathbb{I}}\right).$$

5 ENVY-FREE PROFIT MAXIMIZATION

Our mechanism in Algorithm 1 naturally extends to other buyer-side objectives. We illustrate this with a profit-maximization use case in Appendix E. We summarize the key ideas here.

For this, let us first define the profit of a mechanism M under a strategy profile s: the total revenue—i.e., the sum of all payments received from buyers—minus the total cost incurred for data collection: $\sum_{j\in\mathcal{B}} p_j' - \sum_{i\in\mathcal{C}} c_i n_i'$. Hence, the expected profit is given by,

$$\mathrm{profit}(M, s; \mu) = \mathbb{E}\left[\sum_{j\in\mathcal{B}} p_j' - \sum_{i\in\mathcal{C}} c_i n_i'\right]. \tag{10}$$

A Baseline for Envy-free Profit Maximization. Similar to (7), we first establish a baseline for profit maximization when contributors are non-strategic. Here, the broker seeks to maximize profit subject to two buyer-side incentive constraints: individual rationality for the buyer (IRB) and envy-freeness (EF). Intuitively, EF requires that no buyer prefers another's dataset and price over their own (Varian, 1973; Guruswami et al., 2005). In this baseline, as in (7), the lowest-cost contributor collects all the data. Unfortunately, unlike in (7), no closed-form expressions for the realizations of $\{N_i'\}_{i\in\mathcal{C}}$, $\{m_j'\}_{j\in\mathcal{B}}$, or $\{p_j'\}_{j\in\mathcal{B}}$ are available in the envy-free profit-maximization problem.

However, to apply Algorithm 1, we must specify its four inputs. We can approximate the required quantities algorithmically using the following methods.

Algorithms for Envy-free Profit Maximization. A key technical insight, proved in Lemma 6, is that *envy-free revenue maximization* reduces to the *pricing ordered items* (POI) problem introduced by Chawla et al. (2007). Although POI is NP-hard, prior work provides methods for obtaining $\mathcal{O}(\epsilon)$-additive approximations (Chawla et al., 2007; Hartline and Karlin, 2007; Chen et al., 2024). Algorithm 2 builds on these methods to compute a profit-optimal solution under truthful contributors. It searches over data collection amounts; for each, it uses POI techniques to compute an approximately optimal revenue, and the corresponding dataset sizes and prices for buyers. By selecting the data collection amount that maximizes revenue minus cost (assuming only the cheapest contributor collects data), we obtain $N^{\widetilde{\mathrm{OPT}}}$.

Envy-free Profit Maximization with Strategic Contributors. Finally, we invoke Algorithm 1, setting $N^{\widetilde{\mathrm{OPT}}}$ as the data collection amount from the above step, and $\widetilde{\mathrm{OPT}}$, $\{\widetilde{m}_j\}_{j\in\mathcal{B}}$, $\{\widetilde{p}_j\}_{j\in\mathcal{B}}$ as the corresponding profit, dataset sizes, and prices. This yields a mechanism with profit $\mathrm{OPT} - c_2 + c_1 - \mathcal{O}(\epsilon)$, where OPT denotes the optimal baseline profit, while satisfying IRB, IRC, BB, ICC and EF.

6 CONCLUSION AND DISCUSSION

We design a data marketplace where a broker mediates transactions between buyers who aim to estimate the mean of an unknown normal distribution, and strategic data contributors, who collect data at a cost. Our mechanism admits a Nash equilibrium (NE) in which the two lowest-cost contributors collect and truthfully report all the data. This equilibrium maximizes welfare or profit subject to key market constraints. Matching impossibility results establish that our mechanism is unimprovable.

Future directions include extending the framework beyond mean estimation to other learning tasks. A limitation of our current model is that contributors' costs are assumed to be known. Relaxing this assumption to unknown costs, where contributors may strategically report both costs and data, introduces substantial additional challenges and is left to future work. It would also be interesting to allow buyers to behave strategically, for example by misreporting their valuations to obtain more favorable allocations, which would lead to a different mechanism-design problem beyond our current scope.

Acknowledgments

This work was supported in part by the National Science Foundation Award IIS-2441796.

References

(2024). Ads Data Hub. developers.google.com/ads-data-hub/guides/intro. Accessed: 2022-05-10.

(2024). Aws data exchange overview. https://aws.amazon.com/data-exchange/.

(2024). Bloomberg Enterprise Access Point. https://www.bloomberg.com/professional/product/enterprise-access-point/.

(2024). Citrine Informatics – Accelerating Materials Innovation. URL: https://citrine.io/. Accessed: March 9, 2024.

(2024). DAT Freight and Analytics. URL: www.dat.com/sales-inquiry/freight-market-intelligence-consortium. Accessed: July 9, 2024.

(2024). Datarade. https://datarade.ai/.

(2024). IBM Data Fabric. https://www.ibm.com/data-fabric.

(2024). Snowflake Data Marketplace. URL: https://www.snowflake.com/en/. Accessed: December 26, 2024.

Agarwal, A., Dahleh, M., Horel, T., and Rui, M. (2020). Towards data auctions with externalities. *arXiv preprint arXiv:2003.08345*.

Agarwal, A., Dahleh, M., and Sarkar, T. (2019). A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726.

Babaioff, M., Kleinberg, R., and Paes Leme, R. (2012). Optimal mechanisms for selling information. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 92–109.

Bergemann, D. and Bonatti, A. (2019). Markets for information: An introduction. *Annual Review of Economics*, 11(1):85–107.

Bergemann, D., Bonatti, A., and Smolin, A. (2018). The design and price of information. *American economic review*, 108(1):1–48.

Blum, A., Haghtalab, N., Phillips, R. L., and Shao, H. (2021). One for one, or all for all: Equilibria and optimality of collaboration in federated learning. In *International Conference on Machine Learning*, pages 1005–1014. PMLR.

Cai, Y., Daskalakis, C., and Papadimitriou, C. (2015). Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory*, pages 280–296. PMLR.

Chawla, S., Hartline, J. D., and Kleinberg, R. (2007). Algorithmic pricing via virtual valuations. In *Proceedings of the 8th ACM Conference on Electronic Commerce*, pages 243–251.

Chawla, S., Rezvan, R., Teng, Y., and Tzamos, C. (2022). Pricing ordered items. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 722–735.

Chen, K., Huh, J. S., and Kandasamy, K. (2024). Learning to Price Homogeneous Data. In *Advances in Neural Information Processing Systems*.

Chen, Y. and Hossain, S. (2023). Equilibrium and learning in fixed-price data markets with externality. *arXiv preprint arXiv:2302.08012*.

Chen, Y., Immorlica, N., Lucier, B., Syrgkanis, V., and Ziani, J. (2018). Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 27–44.

Chen, Y., Shen, Y., and Zheng, S. (2020). Truthful data acquisition via peer prediction. *Advances in Neural Information Processing Systems*, 33:18194–18204.

Chen, Y., Zhu, X., and Kandasamy, K. (2023). Mechanism Design for Collaborative Normal Mean Estimation. In *Advances in Neural Information Processing Systems*.

Clinton, A., Chen, Y., Zhu, X., and Kandasamy, K. (2024). Data Sharing for Mean Estimation Among Heterogeneous Strategic Agents.

Cummings, R., Ligett, K., Roth, A., Wu, Z. S., and Ziani, J. (2015). Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 conference on innovations in theoretical computer science*, pages 317–324.

Dorner, F. E., Konstantinov, N., Pashaliev, G., and Vechev, M. (2023). Incentivizing Honesty among

Competitors in Collaborative Learning and Optimization. *arXiv preprint arXiv:2305.16272*.

Fallah, A., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2024). Optimal and differentially private data acquisition: Central and local mechanisms. *Operations Research*, 72(3):1105–1123.

Ghosh, A., Ligett, K., Roth, A., and Schoenebeck, G. (2014). Buying private data without verification. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 931–948.

Guruswami, V., Hartline, J. D., Karlin, A. R., Kempe, D., Kenyon, C., and McSherry, F. (2005). On profit-maximizing envy-free pricing. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2005)*, pages 1164–1173. Society for Industrial and Applied Mathematics.

Hartline, J. and Karlin, A. (2007). Profit maximization in mechanism design. *Algorithmic Game Theory*, 331.

Hartline, J. D. and Koltun, V. (2005). Near-optimal Pricing in Near-linear Time. In *Proceedings of the 9th International Conference on Algorithms and Data Structures*.

Helpman, E. and Laffont, J.-J. (1975). On moral hazard in general equilibrium theory. *Journal of Economic Theory*, 10(1):8–23.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, pages 74–91.

Huang, B., Karimireddy, S. P., and Jordan, M. I. (2023). Evaluating and incentivizing diverse data contributions in collaborative learning. *arXiv preprint arXiv:2306.05592*.

James, W. and Stein, C. (1992). Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory*, pages 443–460. Springer.

Karimireddy, S. P., Guo, W., and Jordan, M. I. (2022). Mechanisms that incentivize data sharing in federated learning. *arXiv preprint arXiv:2207.04557*.

Laffont, J.-J. and Martimort, D. (2009). The theory of incentives: the principal-agent model. In *The theory of incentives*. Princeton university press.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Mehta, S., Dawande, M., Janakiraman, G., and Mookerjee, V. (2021). How to sell a data set? pricing policies for data monetization. *Information Systems Research*, 32(4):1281–1297.

Microsoft (2023). Azure data share overview. *Microsoft Documentation*.

Miller, N., Resnick, P., and Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, 51(9):1359–1373.

Mirrlees, J. A. (1999). The theory of moral hazard and unobservable behaviour: Part i. *The Review of Economic Studies*, 66(1):3–21.

Pei, J. (2020). A survey on data pricing: from economics to data science. *IEEE Transactions on knowledge and Data Engineering*, 34(10):4586–4608.

Sim, R. H. L., Zhang, Y., Chan, M. C., and Low, B. K. H. (2020). Collaborative machine learning with incentive-aware model rewards. In *International conference on machine learning*, pages 8927–8936. PMLR.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151.

Varian, H. R. (1973). Equity, envy, and efficiency.

Zhang, M., Beltrán, F., and Liu, J. (2023). A survey of data pricing for data marketplaces. *IEEE Transactions on Big Data*, 9(4):1038–1056.

Zheng, S., Kwon, Y., Qi, X., and Zou, J. (2024). Truthful dataset valuation by pointwise mutual information. *arXiv preprint arXiv:2405.18253*.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A REVIEW OF NORMAL MEAN ESTIMATION

We first briefly review normal mean estimation (Stein, 1981; James and Stein, 1992). A process[3] M collects a finite dataset $Z = \{z_\ell\}_\ell$ of i.i.d. points from a normal distribution $\mathcal{N}(\mu, \sigma^2)$, and reports a finite dataset $Y = \{y_\ell\}_\ell$ to a learner. The learner knows the variance σ^2, but not the mean μ. She estimates μ via the sample mean $\widehat{\mu}(Y) = \frac{1}{|Y|}\sum_{y\in Y} y$. Conventionally, one studies the *loss* as a function of the estimation error $|\widehat{\mu}(Y) - \mu|$ (Stein, 1981). Here, we instead model the learner's *value* via a decreasing function $v^{\mathrm{e}} : \mathbb{R}_+ \to [0, 1]$, where $v^{\mathrm{e}}(e)$ is her value for achieving error e. The learner's valuation v for a process M, which depends on her error-based valuation v^{e}, is:

$$v(M) = \inf_{\mu\in\mathbb{R}} \mathbb{E}\left[v^{\mathrm{e}}\left(|\,\widehat{\mu}(Y) - \mu\,|\right)\right]. \tag{11}$$

The expectation is over any randomness in M and in generating the original dataset Z. The infimum over μ captures the worst-case performance, ensuring reliability across all possible $\mathcal{N}(\mu, \sigma^2)$ since μ is unknown. To illustrate, let $\widetilde{\mu} \in \mathbb{R}$ be arbitrary. A (pathological) process which always reports repeated $\widetilde{\mu}$ values $(\widetilde{\mu}, \ldots, \widetilde{\mu})$ regardless of the collected dataset, achieves maximal value when $\mu = \widetilde{\mu}$ but fails elsewhere. The infimum ensures $v(M)$ is well-defined, despite μ being unknown.

B PROOF OF THEOREM 4.1

(BB): We will first show that the mechanism satisfies the budget balance property, that is the total contributor payments is equal to the total revenue.

Proof. From the definition of the payment rule in M and a sequence of algebraic manipulations we have

$$\begin{aligned}
\sum_{i\in\mathcal{C}} \pi'_i &= \sum_{i\in\mathcal{C}} \left(\left((\mathrm{OPT}{+}c_1{-}c_2)\frac{N'_i}{N^{\mathrm{OPT}}} + c_i N'_i + \frac{d_i\sigma^2}{N'_i} + \frac{d_i\sigma^2}{N'_{-i}}\right) - d_i\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2\right) \\
&= \frac{(\mathrm{OPT}{+}c_1{-}c_2)}{N^{\mathrm{OPT}}}\sum_{i\in\mathcal{C}} N'_i + \sum_{i\in\mathcal{C}} c_i N'_i + \sum_{i\in\mathcal{C}}\left(\frac{d_i\sigma^2}{N'_{-i}} + \frac{d_i\sigma^2}{N'_i} - d_i\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2\right) \\
&= (\mathrm{OPT}{+}c_1{-}c_2) + \sum_{i\in\mathcal{C}} c_i N'_i + \sum_{j\in\mathcal{B}}\sum_{i\in\mathcal{C}}\left(\frac{d_i\sigma^2}{|\mathcal{B}|\,N'_{-i}} + \frac{d_i\sigma^2}{|\mathcal{B}|\,N'_i} - \frac{d_i}{|\mathcal{B}|}\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2\right) \\
&= \sum_{j\in\mathcal{B}} \widetilde{p}_j + \sum_{j\in\mathcal{B}}\sum_{i\in\mathcal{C}}\left(\frac{d_i\sigma^2}{|\mathcal{B}|\,N'_{-i}} + \frac{d_i\sigma^2}{|\mathcal{B}|\,N'_i} - \frac{d_i}{|\mathcal{B}|}\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2\right) \\
&= \sum_{j\in\mathcal{B}}\left(\sum_{i\in\mathcal{C}}\left(\frac{\widetilde{p}_j N'_i}{N^{\mathrm{OPT}}} + \frac{d_i\sigma^2}{|\mathcal{B}|\,N'_{-i}} + \frac{d_i\sigma^2}{|\mathcal{B}|\,N'_i} - \frac{d_i}{|\mathcal{B}|}\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2\right)\right) = \sum_{j\in\mathcal{B}} p'_j.
\end{aligned}$$

Therefore, M satisfies the budget balance property. □

(IRC): Next, we will show that under M, $s^{\mathbb{I}}$ is *ex-ante* individually rational for contributors.

[3] Typically, one need not model such a process explicitly: often, the learner receives Z as is, or a random subset, keeping Y i.i.d. In a marketplace, although data is sampled i.i.d., strategic contributors may misreport data to maximize personal gain, in which case Y is no longer an i.i.d. dataset from $\mathcal{N}(\mu, \sigma^2)$. In what follows, we will find it useful to explicitly model such a process induced by a mechanism and contributor strategies.

Proof. Under the recommended strategy, we have $n_i' = N_i'$ and $X_i' = Z_i$, the utility for contributor i is

$$
\begin{aligned}
u_i^{\mathrm{C}}(M, s^{\mathbb{I}}) &= \inf_{\mu \in \mathbb{R}} \mathbb{E}\left[\mathbb{I}\left(|X_i'| = N_i'\right)\left((\mathrm{OPT}{+}c_1{-}c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - d_i\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2 - c_i N_i'\right] \\
&= \left((\mathrm{OPT}{+}c_1{-}c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - c_i N_i' - d_i \sup_{\mu \in \mathbb{R}} \mathbb{E}\left[\left(\widehat{\mu}(Z_i) - \widehat{\mu}(Z_{-i})\right)^2\right] \\
&= \left((\mathrm{OPT}{+}c_1{-}c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - c_i N_i' - \left(\frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) \\
&= (\mathrm{OPT}{+}c_1{-}c_2)\frac{N_i'}{N^{\mathrm{OPT}}} \geq 0.
\end{aligned}
$$

□

Remark 1. *Our mechanism, as presented in Algorithm 1, charges each buyer their* ex-ante *value for the data (under truthful reporting), which is why the buyer's* ex-ante *utility is zero. However, it is straightforward to see that by slightly reducing the price and proportionally decreasing contributor payments, one can also satisfy individual rationality for the contributors (IRC) while ensuring that buyers receive strictly positive utility.*

(ICC): We defer the proof of the ICC property to Appendix C.

(IRB): Finally, we will show that under M, $s^{\mathbb{I}}$ is individually rational for buyers.

Proof. Under the recommended strategy, the expected payment for buyer i is

$$
\begin{aligned}
\mathbb{E}\left[p_j'\right] &= \sum_{i \in \{1,2\}} \mathbb{I}\left(|X_i'| = N_i'\right)\left(\widetilde{p}_j \frac{N_i'}{N^{\mathrm{OPT}}} + \frac{d_i}{|\mathcal{B}|}\frac{\sigma^2}{N_{-i}'} + \frac{d_i}{|\mathcal{B}|}\frac{\sigma^2}{N_i'}\right) - \sum_{i \in \{1,2\}} \frac{d_i}{|\mathcal{B}|}\mathbb{E}\left[\left(\widehat{\mu}(X_i') - \widehat{\mu}(X_{-i}')\right)^2\right] \\
&= \left(\widetilde{p}_j + \sum_{i \in \{1,2\}} \frac{d_i}{|\mathcal{B}|}\frac{\sigma^2}{N_{-i}'} + \sum_{i \in \mathcal{C}} \frac{d_i}{|\mathcal{B}|}\frac{\sigma^2}{N_i'}\right) - \sum_{i \in \{1,2\}} \frac{d_i}{|\mathcal{B}|}\left(\frac{\sigma^2}{N_i'} + \frac{\sigma^2}{N_{-i}'}\right) \\
&= \widetilde{p}_j.
\end{aligned}
\tag{12}
$$

Therefore, combing with definition of buyer's utility (4), and buyers's value for i.i.d. data (3), buyer j's utility under recommended strategy is

$$
u_j^{\mathrm{B}}(M, s^{\mathbb{I}}) := v_j^{\mathrm{B}}(M, s^{\mathbb{I}}) - \sup_{\mu \in \mathbb{R}} \mathbb{E}\left[p_j'\right] = v_j^{\mathrm{iid}}(N^{\mathrm{OPT}}) - \sup_{\mu \in \mathbb{R}} \mathbb{E}\left[p_j'\right] = v_j^{\mathrm{iid}}(N^{\mathrm{OPT}}) - \widetilde{p}_j = 0,
$$

where the third step is by (12) and the last step is by definition $\widetilde{p}_j = v_j^{\mathrm{iid}}(N^{\mathrm{OPT}})$. □

Efficiency: Finally, we will show that under M, when the contributors follow $s^{\mathbb{I}}$, the expected welfare achieves the optimal welfare upper bound in Theorem 3.2.

Proof. Recall that contributor 1 collects $N_1' = N^{\mathrm{OPT}} - 1$ points, and contributor 2 collects $N_2' = 1$ point. Under truthful reporting, as contributors have i.i.d data (3), the expected welfare (6) is

$$
W(M, s^{\mathbb{I}}; \mu) = \sum_{j \in \mathcal{B}} v_j^{\mathrm{iid}}\left(N^{\mathrm{OPT}}\right) - c_1(N^{\mathrm{OPT}} - 1) - c_2 = \mathrm{OPT} + c_1 - c_2,
$$

the last step is by definition of welfare-optimal benchmark $\mathrm{OPT} = \sum_{j \in \mathcal{B}} v_j^{\mathrm{iid}}\left(N^{\mathrm{OPT}}\right) - c_1 N^{\mathrm{OPT}}$ in (7).

Extension to Capacity Constraints. Our framework can also accommodate contributor-side capacity constraints. In particular, each contributor may have an upper limit on how much data she can collect. In this case, the broker simply needs to respect these limits when assigning data-collection tasks. The mechanism itself remains unchanged; only the feasible set of allocations is restricted accordingly. Under this modification, the same reasoning and proof continue to apply, so our main guarantees still hold.

□

C PROOF OF ICC

We will now show the ICC property. We recommend that the readers review the proof sketch in §4.1.

We need to show that $s^{\mathbb{I}}$ is a Nash equilibrium of M. That is, we will show, $\forall i \in \mathcal{C}$, $\forall (n_i, X_i)$,

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_i), s_{-i}^{\mathbb{I}}\right)\right) \le u_i^{\mathrm{C}}\left(M, \left(s_i^{\mathbb{I}}, s_{-i}^{\mathbb{I}}\right)\right).$$

This result is proved via key three steps, formalized by the lemmas below. Fix any (n_i, X_i). Let $X_a : \mathcal{X} \to \mathbb{R}^{N_i'}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}\left(X_a(Z_i)\right) = \widehat{\mu}\left(X_i(Z_i)\right)$. Let $X_b : \mathcal{X} \to \mathbb{R}^{N_i'}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}\left(X_b(Z_i)\right) = \widehat{\mu}\left(Z_i\right)$. We then have that

$$\begin{aligned}
u_i^{\mathrm{C}}\left(M, \left((n_i, X_i), s_{-i}^{\mathbb{I}}\right)\right) &\le u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right) && \text{(Lemma 1)} \\
&\le u_i^{\mathrm{C}}\left(M, \left((n_i, X_b), s_{-i}^{\mathbb{I}}\right)\right) && \text{(Lemma 2)} \\
&\le u_i^{\mathrm{C}}\left(M, \left(s_i^{\mathbb{I}}, s_{-i}^{\mathbb{I}}\right)\right). && \text{(Lemma 3)}
\end{aligned}$$

Lemma 1. *Fix any (n_i, X_i). Let $X_a : \mathcal{X} \to \mathbb{R}^{N_i'}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}\left(X_a(Z_i)\right) = \widehat{\mu}\left(X_i(Z_i)\right)$. Then,*

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_i), s_{-i}^{\mathbb{I}}\right)\right) \le u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right).$$

Proof. Under M we have that

$$\begin{aligned}
&u_i^{\mathrm{C}}\left(M, \left((n_i, X_i), s_{-i}^{\mathbb{I}}\right)\right) \\
&= \inf_{\mu \in \mathbb{R}} \mathbb{E}\left[\mathbb{I}\left(|Y_i| = N_i'\right)\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - d_i\left(\widehat{\mu}(X_i(Z_i)) - \widehat{\mu}(Z_{-i})\right)^2 - c_i n_i'\right] \\
&\le \inf_{\mu \in \mathbb{R}} \mathbb{E}\left[\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - d_i\left(\widehat{\mu}(X_a(Z_i)) - \widehat{\mu}(Z_{-i})\right)^2 - c_i n_i'\right] \\
&= u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right).
\end{aligned}$$

The inequality follows from the definition of X_a. □

Lemma 2. *Fix any n_i. Consider any $X_a : \mathcal{X} \to \mathbb{R}^{N_i'}$ and let $X_b : \mathcal{X} \to \mathbb{R}^{N_i'}$ be a function such that $\forall Z_i \in \mathcal{X}$, $\widehat{\mu}\left(X_b(Z_i)\right) = \widehat{\mu}\left(Z_i\right)$. Then,*

$$u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right) \le u_i^{\mathrm{C}}\left(M, \left((n_i, X_b), s_{-i}^{\mathbb{I}}\right)\right).$$

Proof. From the definition of X_a we have

$$\begin{aligned}
&u_i^{\mathrm{C}}\left(M, \left((n_i, X_a), s_{-i}^{\mathbb{I}}\right)\right) \\
&= \inf_{\mu \in \mathbb{R}} \mathbb{E}\left[\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - d_i\left(\widehat{\mu}(X_a(Z_i)) - \widehat{\mu}(Z_{-i})\right)^2 - c_i n_i'\right] \\
&= \left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - c_i n_i' + d_i \inf_{\mu \in \mathbb{R}} \mathbb{E}\left[-\left(\widehat{\mu}(X_a(Z_i)) - \widehat{\mu}(Z_{-i})\right)^2\right] \\
&= \left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}} + c_i N_i' + \frac{d_i\sigma^2}{N_{-i}'} + \frac{d_i\sigma^2}{N_i'}\right) - c_i n_i' - d_i \sup_{\mu \in \mathbb{R}} \mathbb{E}\left[\left(\widehat{\mu}(X_a(Z_i)) - \widehat{\mu}(Z_{-i})\right)^2\right]. \qquad (13)
\end{aligned}$$

Now notice that because the other agents truthfully submit a combined N_{-i}' points under $s_{-i}^{\mathbb{I}}$, we have that

$$\begin{aligned}
\sup_{\mu \in \mathbb{R}} \mathbb{E}\left[\left(\widehat{\mu}(X_a(Z_i)) - \widehat{\mu}(Z_{-i})\right)^2\right] &= \sup_{\mu \in \mathbb{R}}\left(\mathbb{E}\left[\left(\widehat{\mu}(X_a(Z_i)) - \mu\right)^2\right] + \mathbb{E}\left[\left(\widehat{\mu}(Z_{-i}) - \mu\right)^2\right]\right) \\
&= \sup_{\mu \in \mathbb{R}}\left(\mathbb{E}\left[\left(\widehat{\mu}(X_a(Z_i)) - \mu\right)^2\right]\right) + \frac{\sigma^2}{N_{-i}'}.
\end{aligned}$$

Therefore, we can rewrite (13), using the fact that the sample mean is minimax optimal (Lehmann and Casella, 2006) and the definition of X_b, to get

$$
\begin{aligned}
&\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}}+c_iN_i'+\frac{d_i\sigma^2}{N_{-i}'}+\frac{d_i\sigma^2}{N_i'}\right)-c_in_i'-\frac{d_i\sigma^2}{N_{-i}'}-d_i\sup_{\mu\in\mathbb{R}}\mathbb{E}\left[\left(\widehat{\mu}(X_a(Z_i))-\mu\right)^2\right]\\
\leq&\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}}+c_iN_i'+\frac{d_i\sigma^2}{N_{-i}'}+\frac{d_i\sigma^2}{N_i'}\right)-c_in_i'-\frac{d_i\sigma^2}{N_{-i}'}-d_i\sup_{\mu\in\mathbb{R}}\mathbb{E}\left[\left(\widehat{\mu}(Z_i)-\mu\right)^2\right]\\
=&\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}}+c_iN_i'+\frac{d_i\sigma^2}{N_{-i}'}+\frac{d_i\sigma^2}{N_i'}\right)-c_in_i'-d_i\sup_{\mu\in\mathbb{R}}\mathbb{E}\left[\left(\widehat{\mu}(Z_i)-\widehat{\mu}(Z_{-i})\right)^2\right]\\
=&\,u_i^{\mathrm{C}}\left(M,\left((n_i,X_b),s_{-i}^{\mathbb{I}}\right)\right).
\end{aligned}
$$

□

Lemma 3. *Fix any n_i. Let $X_b:\mathcal{X}\to\mathbb{R}^{N_i'}$ be a function such that $\forall Z_i\in\mathcal{X}$, $\widehat{\mu}\left(X_b(Z_i)\right)=\widehat{\mu}\left(Z_i\right)$. Then,*

$$u_i^{\mathrm{C}}\left(M,\left((n_i,X_b),s_{-i}^{\mathbb{I}}\right)\right)\leq u_i^{\mathrm{C}}\left(M,\left(s_i^{\mathbb{I}},s_{-i}^{\mathbb{I}}\right)\right).$$

Proof. From the definition of X_b we have that

$$
\begin{aligned}
&u_i^{\mathrm{C}}\left(M,\left((n_i,X_b),s_{-i}^{\mathbb{I}}\right)\right)\\
=&\inf_{\mu\in\mathbb{R}}\mathbb{E}\left[\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}}+c_iN_i'+\frac{d_i\sigma^2}{N_{-i}'}+\frac{d_i\sigma^2}{N_i'}\right)-d_i\left(\widehat{\mu}(X_b(Z_i))-\widehat{\mu}(Z_{-i})\right)^2-c_in_i'\right]\\
=&\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}}+c_iN_i'+\frac{d_i\sigma^2}{N_{-i}'}+\frac{d_i\sigma^2}{N_i'}\right)-c_in_i'-d_i\sup_{\mu\in\mathbb{R}}\mathbb{E}\left[\left(\widehat{\mu}(Z_i)-\widehat{\mu}(Z_{-i})\right)^2\right]\\
=&\left((\mathrm{OPT}+c_1-c_2)\frac{N_i'}{N^{\mathrm{OPT}}}+c_iN_i'+\frac{d_i\sigma^2}{N_{-i}'}+\frac{d_i\sigma^2}{N_i'}\right)-c_in_i'-\frac{d_i\sigma^2}{n_i'}-\frac{d_i\sigma^2}{N_{-i}'}.
\end{aligned}
$$

Define the concave function $l:\mathbb{R}_+\to\mathbb{R}$, given by $l(n_i')=-c_in_i'-\frac{d_i\sigma^2}{n_i'}$. Observe that $l'(n_i')=-c_i+\frac{d_i\sigma^2}{n_i'^2}$ so $l'(n_i')=0\iff n_i'=\sigma\sqrt{\frac{d_i}{c_i}}=N_i'$. Therefore the choice of n_i' which maximizes $u_i^{\mathrm{C}}\left(M,\left((n_i,X_b),s_{-i}^{\mathbb{I}}\right)\right)$ is N_i'. Together with the fact that $\widehat{\mu}(X_b(Z_i))=\widehat{\mu}(Z_i)=\widehat{\mu}(\mathbb{I}(Z_i))$ we conclude

$$u_i^{\mathrm{C}}\left(M,\left((n_i,X_b),s_{-i}^{\mathbb{I}}\right)\right)\leq u_i^{\mathrm{C}}\left(M,\left((\mathbb{I},X_b),s_{-i}^{\mathbb{I}}\right)\right)=u_i^{\mathrm{C}}\left(M,\left((\mathbb{I},\mathbb{I}),s_{-i}^{\mathbb{I}}\right)\right)=u_i^{\mathrm{C}}\left(M,\left(s_i^{\mathbb{I}},s_{-i}^{\mathbb{I}}\right)\right).$$

□

D PROOF OF IMPOSSIBILITY RESULT IN SECTION §3

We will now prove our impossibility results in Section §3. In this section, with a slight abuse of notation, we will use 0 to denote both the scalar zero, and the zero function.

The following lemma is the key technical ingredient in proving both Theorem 3.1 and Theorem 3.2. It states that when other agents are not collecting any data, then, regardless of their submission functions, the best response for an agent is to also not collect any data. Note that even when an agent has not collected any data, an untruthful submission function may actually report some data.

Lemma 4. *Let M be any mechanism and let $s_{-i}=\{(0,X_j)\}_{j\neq i}$ be a strategy profile for all agents except i, where $\{X_j\}_{j\neq i}$ are arbitrary submission functions. If $s_i=(n_i,X_i)$ is the best response of agent i to s_{-i}, then $n_i=0$ (i.e., $n_i'=n_i(N_i',c_i)=0$ for all N_i',c_i).*

We will prove Lemma 4 in §D.1. We first prove Theorem 3.1 and Theorem 3.2 using Lemma 4.

Theorem 3.1. *For any mechanism M, if $s=\{(n_i,X_i)\}_{i\in\mathcal{C}}$ is a dominant strategy profile, then no contributor collects any data, i.e., $\forall i\in\mathcal{C}$, $n_i'=n_i(N_i',c_i)=0$.*

Proof of Theorem 3.1. Let $s_{-i} = \{(0, X_j)\}_{j \neq i}$ be a strategy profile for all agents except i, where $\{X_j\}_{j \neq i}$ are arbitrary submission functions. Since s is a dominant strategy profile we know that $\forall s_i'$

$$u_i^{\mathrm{C}}(M, (s_i, s_{-i})) \geq u_i^{\mathrm{C}}(M, (s_i', s_{-i}))$$

i.e. s_i is the best response to s_{-i}. By Lemma 4 we conclude that $n_i' = 0$. □

Theorem 3.2. *Let* OPT *be as defined in* (7)*. For any mechanism M, if strategy profile $s = \{(n_i, \mathbb{I})\}_{i \in \mathcal{C}}$ is a Nash equilibrium, then $W(M, s; \mu) \leq \mathrm{OPT} - (c_2 - c_1)$.*

We will first present a technical result, which follows as a corollary of Lemma 4. We will then use this result to prove Theorem 3.2.

Corollary 5 (Corollary of Lemma 4)**.** *If $s = \{(n_i, X_i)\}_{i \in \mathcal{C}}$ is a NE under M and $\forall j \neq i$, $n_j = 0$, then $n_i = 0$.*

Proof. Since s is a NE under M, $\forall s_i'$ we have that $u_i^{\mathrm{C}}(M, (s_i, s_{-i})) \geq u_i^{\mathrm{C}}(M, (s_i', s_{-i}))$, i.e. s_i is the best response to s_{-i}. By the hypothesis of the corollary, we also know that $s_{-i} = \{(0, X_j)\}_{j \neq i}$. Therefore, we can apply Lemma 4 to conclude $n_i = 0$. □

Proof of Theorem 3.2. As defined in (6), the welfare $W(M, s; \mu)$ of a mechanism M under a strategy profile s is

$$W(M, s; \mu) = \mathbb{E}\left[\sum_{j \in \mathcal{B}} v_j^{\mathrm{B}}(M, s) - \sum_{i \in \mathcal{C}} c_i n_i'\right].$$

Under strategy s, the contributors are submitting a total of $N^{\mathrm{tot}} := \sum_i n_i'$ truthful points to the broker. By buyers's value for i.i.d. data (3) and (2), we have $v_j^{\mathrm{B}}(M, s) = v_j^{\mathrm{iid}}(m_j')$. Since $m_j' \leq N^{\mathrm{tot}}$, it follows that $v_j^{\mathrm{iid}}(m_j') \leq v_j^{\mathrm{iid}}(N^{\mathrm{tot}})$. Therefore,

$$W(M, s; \mu) = \mathbb{E}\left[\sum_{j \in \mathcal{B}} v_j^{\mathrm{iid}}(m_j') - \sum_{i \in \mathcal{C}} c_i n_i'\right] \leq \mathbb{E}\left[\sum_{j \in \mathcal{B}} v_j^{\mathrm{iid}}(N^{\mathrm{tot}}) - \sum_{i \in \mathcal{C}} c_i n_i'\right]$$

Corollary 5 tells us that under the Nash equilibrium $s = \{(n_i, \mathbb{I})\}_{i \in \mathcal{C}}$, at least two contributors must collect data. Assuming that $c_1 \leq c_2 \leq \cdots \leq c_{|\mathcal{B}|}$, we have that the cheapest way for at least two contributors to collect a total of $N^{\mathrm{tot}} := \sum_i n_i'$ points is to have agent 1 collect $N^{\mathrm{tot}} - 1$ points and agent 2 collect 1 point. Combining this with the fact that OPT is the best welfare achievable when contributors are non-strategic, we get

$$\begin{aligned} W(M, s; \mu) &\leq \mathbb{E}\left[\sum_{j \in \mathcal{B}} v_i^{\mathrm{iid}}\left(N^{\mathrm{tot}}\right) - \sum_{i \in \mathcal{C}} c_i n_i'\right] \leq \mathbb{E}\left[\sum_{j \in \mathcal{B}} v_i^{\mathrm{iid}}\left(N^{\mathrm{tot}}\right) - c_1\left(N^{\mathrm{tot}} - 1\right) - c_2\right] \\ &\leq \max_N \left\{\sum_{i \in \mathcal{C}} v_i(N) - c_1 N\right\} - (c_2 - c_1) = \mathrm{OPT} - (c_2 - c_1). \end{aligned}$$

The last step follows from equation (7), the definition of OPT. □

D.1 Proof of Lemma 4

At a high level, Lemma 4 claims that when the other agents do not collect data, it is best for agent i to not collect data. This is because the mechanism has no knowledge of the true value of $\mu \in \mathbb{R}$ and thus cannot penalize agent i for submitting fabricated data since the others to do not provide data to validate against. Therefore, agent i has the ability to forgo the cost of data collection without decreasing the payment they receive. Hence, agents i's best response involves collecting no data. The following proof formalizes this intuition.

Proof. Suppose for a contradiction that $n_i' = n_i(N_i', c_i) > 0$. Define the strategy $s_{i,\mu'} = (n_i, X_{i,\mu'})$ where $X_{i,\mu'}$ is the submission function which disregards Z_i and instead samples and submits n_i' points from $N(\mu', \sigma^2)$ using X_i, i.e. $X_{i,\mu'}(\cdot) := X_i(Z_i)$, $Z_i \sim N(\mu', \sigma^2)^{n_i'}$. Applying the definition of contributor i's utility we see that

$$\inf_{\mu \in \mathbb{R}} \mathop{\mathbb{E}}_{Z_i \sim N(\mu, \sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j \neq i}\right)\right] - c_i n_i' \geq \inf_{\mu \in \mathbb{R}} \mathop{\mathbb{E}}_{Z_i \sim N(\mu', \sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j \neq i}\right)\right] - c_i n_i'$$

$$\Rightarrow \inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu,\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] \geq \inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu',\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right].$$

For clarity we omit writing the dependence on M and s_{-i} under the expectation as it is not relevant to the proof. Now notice that since all of the other agents are not collecting data we have

$$\begin{aligned}
\inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu',\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] &= \mathop{\mathbb{E}}_{Z_i\sim N(\mu',\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] \\
&\geq \inf_{\mu'\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu',\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] \\
&= \inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu,\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right].
\end{aligned}$$

Because we have both matching upper and lower bounds, we conclude that $\forall \mu' \in \mathbb{R}$

$$\inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu,\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] = \mathop{\mathbb{E}}_{Z_i\sim N(\mu',\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right].$$

Intuitively this says that the expected payment an agent receives is independent of which normal distribution they submit data from. But this means that agent i can avoid the cost of data collection by submitting fabricated data with no reduction in payment. More formally, for any $\mu' \in \mathbb{R}$, consider the strategy $\tilde{s}_i := (0, X_{i,\mu'})$. Under this strategy agent i's utility is

$$\begin{aligned}
u_i^{\mathrm{C}}\left(M, (\tilde{s}_i, s_{-i})\right) &= \inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu',\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] \\
&= \inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu,\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] \\
&> \inf_{\mu\in\mathbb{R}} \mathop{\mathbb{E}}_{Z_i\sim N(\mu,\sigma^2)^{n_i'}} \left[\pi_i\left(X_i(Z_i), (X_j(\varnothing))_{j\neq i}\right)\right] - c_i n_i' = u_i^{\mathrm{C}}\left(M, (s_i, s_{-i})\right).
\end{aligned}$$

But this contradicts that s_i is the best response to s_{-i}. Therefore, $n_i' = 0$. □

E ENVY-FREE PROFIT MAXIMIZATION

In this section, we will demonstrate how our mechanism can also be applied for envy-free profit maximization. For this, let us define the profit of a mechanism M under a strategy profile s: the total revenue—i.e., the sum of all payments received from buyers—minus the total cost incurred for data collection: $\sum_{j\in\mathcal{B}} p_j' - \sum_{i\in\mathcal{C}} c_i n_i'$. Hence, the expected profit is given by,

$$\operatorname{profit}(M, s; \mu) = \mathbb{E}\left[\sum_{j\in\mathcal{B}} p_j' - \sum_{i\in\mathcal{C}} c_i n_i'\right]. \tag{14}$$

Envy-freeness. First, note that it is not very hard to see our mechanism in Algorithm 1 already maximizes profit without additional constraints – it maximizes the total welfare, while ensuring that the buyer's utility is 0. However, as noted in §4, this pricing scheme may be "unfair" to some buyers, as all buyers pay different amounts for the same dataset. A natural fairness requirement in this context is envy-freeness (EF) (Varian, 1973; Guruswami et al., 2005), which states that no buyer should prefer another's data allocation and price to theirs. We have formalized it as a fifth desideratum, in addition to the four given in §2.2:

5. *Envy-free for the buyers (EFB): M* is envy-free for the buyers if no buyer prefers another buyer's dataset size selection rule and pricing rule to her own, when contributors follow $s^{\mathbb{I}}$. To state this formally, note that we can write a buyer's utility as a function of her own dataset size selection rule and pricing rule (with a slight abuse of notation):

$$u_j^{\mathrm{B}}(M, s) = \underbrace{\inf_{\mu\in\mathbb{R}} \mathbb{E}\left[v_j^{\mathrm{e}}\left(\left|\,\widehat{\mu}(Y_j') - \mu\,\right|\right)\right]}_{:=f(m_j,s)} - \underbrace{\sup_{\mu\in\mathbb{R}} \mathbb{E}\left[p_j'\right]}_{:=g(p_j,s)} = u_j^{\mathrm{B}}(m_j, p_j, s). \tag{15}$$

With this definition, the envy-freeness requirement can be stated as: $u_j^{\mathrm{B}}(m_j, p_j, s^{\mathbb{I}}) \geq u_j^{\mathrm{B}}(m_k, p_k, s^{\mathbb{I}})$ for all buyers $j, k \in \mathcal{B}$.

We will show how Algorithm 1 can also be used for envy-free profit maximization (Our ideas can also be adapted for envy-free welfare maximization). First, in §E.1, we establish a profit-maximization baseline similar to (7) which assumes non-strategic contributors. Unfortunately, unlike in (7), there are no closed form solutions for this baseline; hence, in §E.2, we design algorithms for envy-free profit maximization with non-strategic contributors. Then, in §E.3, we will combine the algorithm in §E.2 with Algorithm 1 to derive mechanisms in the presence of strategic contributors.

E.1 A Baseline for Profit Maximization

We construct the profit maximization baseline similar to (7), assuming that contributors are non-strategic and follow $s^{\mathbb{I}}$. In this case, since buyers receive i.i.d. data, their valuations depend only on the data amounts $\{m'_j\}_{j\in\mathcal{B}}$ (see (3)). Therefore, buyer j's utility can be written as,

$$u_j^{\mathrm{B}}(M, s^{\mathbb{I}}) = u_j^{\mathrm{B}}(m_j, p_j, s^{\mathbb{I}}) = u_j^{\mathrm{B}}(m'_j, p'_j, s^{\mathbb{I}}) = v_j^{\mathrm{iid}}(m'_j) - p'_j.$$

Let us refer to all $\{m'_j, p'_j\}_{j\in\mathcal{B}} \in (\mathbb{R}\times\mathbb{N})^{|\mathcal{B}|}$ values which satisfy IRB and EFB under $s^{\mathbb{I}}$ as *envy-free pricing schemes*. Let $\mathbf{EF}(N)$ denote all envy-free pricing schemes where the maximum amount of data sold is at most N, *i.e.*, $\max_j m'_j \leq N$, and let $\mathbf{EF} = \bigcup_{N\geq 0}\mathbf{EF}(N)$.

Optimal baseline profit. The optimal baseline profit OPT, subject to the IRB and EFB constraints, can therefore be written as shown in (16). We have,

$$\mathrm{OPT} = \max_{M;M \text{ is envy-free}} \mathrm{profit}(M, s^{\mathbb{I}}; \mu) = \max_{\{m'_j,p'_j\}_j\in\mathbf{EF},\ \{N'_i\}_i\in\mathbb{N}^{|\mathcal{C}|}} \left(\sum_{j\in\mathcal{B}} p'_j - \sum_{i\in\mathcal{C}} c_i N'_i\right). \tag{16}$$

Above, the second step follows from the fact that since contributors will collect the requested amount and submit it truthfully, the broker's task reduces to finding *scalar* values for dataset sizes and prices for buyers, and data collection amounts for contributors so as to maximize profit. (Note that OPT in (16) for envy-free profit maximization is different to OPT in (7) for welfare maximization.)

Revenue-optimal envy-free pricing schemes. To better understand OPT, we first note that as buyer valuations only depend on the amount of data they receive (and not which contributors collected the data), the broker's decisions for buyers' dataset sizes $\{m'_j\}_{j\in\mathcal{C}}$ and prices $\{p'_j\}_{j\in\mathcal{C}}$ need to depend on $\{N'_i\}_{i\in\mathcal{C}}$ only through their sum $N^{\mathrm{tot}} = \sum_{i\in\mathcal{C}} N'_i$, i.e. the total amount of data collected. The optimal revenue $\mathrm{rev}^{\mathrm{OPT}}$ under IRB, EFB constraints when the broker has received N^{tot} total data points can be written as shown below:

$$\mathrm{rev}^{\mathrm{OPT}}(N^{\mathrm{tot}}) = \max_{\{m'_j,p'_j\}_{j\in\mathcal{B}}\in\mathbf{EF}(N^{\mathrm{tot}})} \left(\sum_{j\in\mathcal{B}} p'_j\right). \tag{17}$$

From optimal revenue to optimal profit. The following straightforward calculations lead to the following expression for the optimal envy-free profit OPT (16):

$$\begin{aligned}
\mathrm{OPT} &= \max_{\{m'_j,p'_j\}_{j\in\mathcal{B}}\in\mathbf{EF},\{N'_i\}_i} \left(\sum_{j\in\mathcal{B}} p'_j \; - \; \sum_{i\in\mathcal{C}} c_i N'_i\right) \\
&= \max_{\{N'_i\}_i} \left(\max_{\{m'_j,p'_j\}_{j\in\mathcal{B}}\in\mathbf{EF}(\sum_i N'_i)} \left(\sum_{j\in\mathcal{B}} p'_j\right) \; - \; \sum_{i\in\mathcal{C}} c_i N'_i\right) \\
&= \max_{\{N'_i\}_i} \left(\mathrm{rev}^{\mathrm{OPT}}\left(\sum_{i\in\mathcal{C}} N'_i\right) \; - \; \sum_{i\in\mathcal{C}} c_i N'_i\right) \; = \; \mathrm{rev}^{\mathrm{OPT}}\left(N_1^{\mathrm{OPT}}\right) - c_1 N_1^{\mathrm{OPT}}.
\end{aligned} \tag{18}$$

where $N_1^{\mathrm{OPT}} = \mathrm{argmax}_{N\in\mathbb{N}}\, \mathrm{rev}^{\mathrm{OPT}}(N) - c_1 N$. The last step uses the observation that agent 1 has the smallest cost. Intuitively, the broker should ask the cheapest agent to collect all the data.

Unfortunately, unlike in (7), there is no closed-form expressions for OPT, N_1^{OPT}, which can be used as inputs for Algorithm 1. Next, we will show how these quantities can be approximated algorithmically.

E.2 Designing Profit-optimal Envy-free Pricing Schemes with Non-strategic Contributors

Our first key insight here is that designing an optimal envy-free pricing scheme $\{(m'_j, p'_j)\}_{j\in\mathcal{B}}$ reduces to constructing a revenue-optimal pricing curve (Lemma 6). Specifically, the seller posts a pricing curve $q : \{0,\dots,N\} \to [0,1]$ with $q(0)=0$, and buyers choose their purchase quantities to maximize utility (in contrast, in our setting, the broker directly determines dataset sizes and prices for each buyer). When buyers have monotonic valuations, revenue-optimal pricing aligns with the ordered item pricing problem (Chawla et al., 2022), allowing us to leverage existing algorithms. We first define the ordered item pricing problem below. To simplify the exposition, we do so in the context of data pricing[4].

Definition 1. *(Ordered item pricing (Chawla et al., 2022)) Suppose the broker has $N^{\mathrm{tot}} \in \mathbb{N}$ data points and posts a pricing curve q. Buyers then decide how much to purchase based on q. A utility-maximizing buyer j selects $m \in \{0,\dots,N^{\mathrm{tot}}\}$ to maximize utility, $v_j(m) - q(m)$. If multiple values of m yield the same maximum utility, the buyer chooses the largest dataset. Thus, buyer j purchases $\overline{m}_j(N^{\mathrm{tot}}, q)$ data points, where*

$$\overline{m}_j(N^{\mathrm{tot}}, q) := \max\left\{ \operatorname*{argmax}_{m\le N^{\mathrm{tot}}} \big(v_j(m) - q(m) \big) \right\}. \tag{19}$$

It follows that the revenue from buyer j is $q(\overline{m}_j(N^{\mathrm{tot}}, q))$. Hence, the total revenue $\overline{\mathrm{rev}}(N^{\mathrm{tot}}, q)$ of a pricing curve q is as defined below. The broker wishes to find a pricing function q^{OPT} which maximizes the cumulative revenue. The optimal revenue $\overline{\mathrm{rev}}^{\mathrm{OPT}}$ can be defined as:

$$\overline{\mathrm{rev}}(N^{\mathrm{tot}}, q) := \textstyle\sum_{j\in\mathcal{B}} q\left(\overline{m}_j(N^{\mathrm{tot}}, q)\right), \qquad \overline{\mathrm{rev}}^{\mathrm{OPT}}(N^{\mathrm{tot}}) := \max_q\ \overline{\mathrm{rev}}(N^{\mathrm{tot}}, q). \tag{20}$$

From ordered item pricing to revenue-optimal envy-free mechanisms. Given a pricing curve q with N^{tot} points, it is straightforward to construct an envy-free pricing scheme $\{m'_j, p'_j\}_{j\in\mathcal{B}} \in \mathbf{EF}(N^{\mathrm{tot}})$ while achieving the same revenue as q, i.e. $\overline{\mathrm{rev}}(N^{\mathrm{tot}}, q) = \sum_{j\in\mathcal{B}} p'_j$.

To do so, we can set $m'_j = \overline{m}_j(N^{\mathrm{tot}}, q)$ and $p'_j = q(m'_j)$ (see (19)). Since each buyer selects data to maximize her utility (19), it follows that $u_j^{\mathrm{B}}(m'_j, p'_j) \ge u_j^{\mathrm{B}}(m, q(m))$ for every other $m \in [N]$, and in particular, for all other buyers' dataset sizes $\{m_k\}_{k\neq j}$. This ensures EFB. Moreover, as buying no data ($m'_j = 0$ at $p'_j = 0$) yields zero utility, IRB is also satisfied.

This also implies that the revenue of the optimal envy-free pricing scheme is at least as large as the revenue of the optimal pricing curve, i.e. $\mathrm{rev}^{\mathrm{OPT}}(N^{\mathrm{tot}}) \ge \overline{\mathrm{rev}}^{\mathrm{OPT}}(N^{\mathrm{tot}})$ (see (17) and (20)).

Our next lemma shows the converse, thus establishing an equivalence between optimizing ordered item pricing curves and envy-free pricing schemes. Thus, we have $\mathrm{rev}^{\mathrm{OPT}}(N^{\mathrm{tot}}) = \overline{\mathrm{rev}}^{\mathrm{OPT}}(N^{\mathrm{tot}})$. The proof is provided in §E.2.1.

Lemma 6. *Given the set of buyers $\mathcal{B}$, along with their valuations $\{v_i\}_{i\in\mathcal{B}}$. Let N^{tot} be the amount of data available to the broker. For every envy-free pricing scheme $(\{m'_j\}_{j\in\mathcal{B}}, \{p'_j\}_{j\in\mathcal{B}}) \in \mathbf{EF}(N^{\mathrm{tot}})$, there exists a pricing curve q which achieves a higher revenue, i.e., $\overline{\mathrm{rev}}(N^{\mathrm{tot}}, q) \ge \sum_{j\in\mathcal{B}} p'_j$.*

Algorithms for ordered item pricing. The ordered item pricing problem is known to be NP-hard (Chawla et al., 2007; Hartline and Koltun, 2005). As a result, prior work has developed approximation algorithms. An ordered item pricing algorithm, denoted by A, takes as input the buyer valuations $\{v_j\}_{j\in\mathcal{B}}$, the total amount of data N^{tot}, and an approximation parameter ϵ. It outputs a pricing curve q such that the cumulative revenue satisfies $\overline{\mathrm{rev}}(N^{\mathrm{tot}}, q) \ge \overline{\mathrm{rev}}^{\mathrm{OPT}}(N^{\mathrm{tot}}) - |\mathcal{B}|\mathcal{O}(\epsilon)$. Table 1 summarizes algorithms from prior work under various assumptions on buyer valuations.

[4] In Chawla et al. (2022), the seller has $N^{\mathrm{tot}} \in \mathbb{N}$ items, and a set of *unit-demand* buyers, i.e. wish to purchase only a single item, with non-decreasing valuations $v_i : [N^{\mathrm{tot}}] \to [0,1]$. While buyers may have different valuations for the same good, their preference ranking is the same. The seller should choose a pricing function $q : [N^{\mathrm{tot}}] \to \mathbb{R}_+$ to maximize revenue.

Table 1: Algorithms from Prior Work

Algorithm	Assumptions on buyer valuations	Time complexity
Hartline and Koltun (2005)	–	$\widetilde{\mathcal{O}}(2^N \epsilon^{-N})$
Chawla et al. (2022)	Monotonicity	$N^{\mathcal{O}\left(\epsilon^{-2}\log \epsilon^{-1}\right)}$
Chen et al. (2024)	Monotonicity, K types	$\widetilde{\mathcal{O}}(N^K \epsilon^{-K})$
	Monotonicity, K types, smoothness	$\widetilde{\mathcal{O}}\left(\epsilon^{-2K}\right)$
	Monotonicity, K types, diminishing return	$\widetilde{\mathcal{O}}(\epsilon^{-3K}\log^K(N))$

Table 1: Algorithms from prior work applicable for pricing ordered items, along with their assumptions on buyer valuations and time complexities to obtain an $|\mathcal{B}|\mathcal{O}(\epsilon)$-optimal solution when there are N total points. Chawla et al. (2022) were the first to study this problem, assuming monotonic buyer valuations. Chen et al. (2024) studied this problem in the context of data pricing, introducing additional assumptions on buyer valuations, such as finite buyer types, smoothness, and diminishing returns.

Algorithm 2

Input: Buyer valuations $\{v_j\}_{j\in\mathcal{B}}$, contributor costs $\{c_i\}_{i\in\mathcal{C}}$, an algorithm for ordered item pricing A, approximation parameter ϵ.
for $N \in \{1, 2, \dots \frac{|\mathcal{B}|}{c_1}\}$:
    $q_N \leftarrow A(\{v_j\}_{j\in\mathcal{B}}, N, \epsilon)$. # Execute A to obtain an $|\mathcal{B}|\mathcal{O}(\epsilon)$-optimal pricing curve.
    $m_{N,j} \leftarrow \overline{m}_j(N, q_N)$ for all $j \in \mathcal{B}$. # Compute buyer dataset sizes. See (19).
    $\text{profit}_N \leftarrow \sum_{i\in\mathcal{B}} q_N(m_{N,i}) - c_1 N$. # Compute the profit if collecting N data.
$N^+ \leftarrow \text{argmax}_{N\in\left\{1,2,\dots\frac{|\mathcal{B}|}{N^{\text{tot}}}\right\}} \text{profit}_N$. # Find the optimal data collection amount.
$m_j^+ \leftarrow m_{N^+,j}$, $q_j^+ \leftarrow q_{N^+}(m_j^+)$. # Optimal prices and dataset sizes for buyer.
Return: N^+, $\{m_j^+\}_{j\in\mathcal{B}}$, $\{q_j^+\}_{j\in\mathcal{B}}$

Designing Profit-optimal Envy-free Mechanisms. The above observations suggest the following procedure to design an approximately profit-optimal mechanism, by leveraging an algorithm for ordered item pricing. For any given $N^{\text{tot}} \in \mathbb{N}$, we apply A to determine an approximately optimal pricing curve. To maximize profit, we assign all data collection to contributor 1 and optimize over N^{tot} to maximize revenue minus $c_1 N^{\text{tot}}$.

Algorithm 2 outlines this procedure. It takes as input buyer valuations $\{v_j\}_{j\in\mathcal{B}}$, contributor costs $\{c_i\}_{i\in\mathcal{C}}$, an ordered item pricing algorithm A, and an approximation parameter ϵ. It returns a data quantity N^+, the corresponding pricing curve q_{N^+}, and dataset allocations $\{m_j^+\}_{j\in\mathcal{B}}$, all of which together guarantee an $\mathcal{O}(|\mathcal{B}|\epsilon)$-optimal revenue. Since buyer valuations are bounded in $[0, 1]$, the maximum total buyer value is $|\mathcal{B}|$; thus, it suffices to search for $N \leq |\mathcal{B}|/c_1$. While our brute-force approach is admittedly inefficient; we leave it to future work to develop more computationally efficient procedures to find the optimal N^+. The following lemma lower bounds the profit of Algorithm 2, with the proof given in Appendix C.

Lemma 7. *Suppose we execute Algorithm 2 with an algorithm A for ordered item pricing. Let N^+, $\{m_j^+\}_{j\in\mathcal{B}}$, $\{q_j^+\}_{j\in\mathcal{B}}$ be the returned values of Algorithm 2. Consider a mechanism M where we let the contributors collectively collect N^+ amount of data (i.e., $N_1' = N^+$, and $N_i' = 0$ for all $i \neq 1$). The broker then sell m_j^+ data to each buyer at price q_j^+, and issues payment $\pi_1' = \sum_j q_j^+$ to contributor 1 ($\pi_i' = 0$, for all $i \neq 1$). Assume that the contributors follow $s^{\mathbb{I}}$, we then have* $\text{profit}(M, s^{\mathbb{I}}; \mu) \geq \text{OPT} - |\mathcal{B}|\mathcal{O}(\epsilon)$.

E.2.1 Proof of Lemma 6

Proof. We need to show that for any envy-free pricing scheme $\left\{(m_j', p_j')\right\}_{j\in\mathcal{B}}$, there exists a non-decreasing price curve $q : \{0, 1, \dots, N^{\text{tot}}\} \to [0, 1]$ that yields a revenue of at least $\sum_{j\in\mathcal{B}} p_j'$. Without lose of generality, we assume

$m'_1 \leq m'_2 \leq \cdots \leq m'_{|\mathcal{B}|}$. We define q as follows:

$$q(m) := \begin{cases} p'_1, & \text{if } m \leq m'_1 \\ p'_2, & \text{if } m'_1 < m \leq m'_2 \\ \vdots & \\ p'_{|\mathcal{B}|}, & \text{if } m'_{|\mathcal{B}|-1} < m \leq N^{\text{tot}} \end{cases}$$

This price function q is non-decreasing and has at most $|\mathcal{B}|$ steps. By the purchase model of ordered item pricing (19), each buyer j would purchase $\overline{m}_j(N^{\text{tot}}, q)$ data points under price curve q, where

$$\overline{m}_j(N^{\text{tot}}, q) := \max \left\{ \underset{m \leq N^{\text{tot}}}{\operatorname{argmax}} \left(v_j(m) - q(m) \right) \right\}$$

It is sufficient to show that $\overline{m}_j(N^{\text{tot}}, q) \geq m'_j$ holds for every buyer j. Then, as the price curve is non-decreasing, the revenue from each buyer when using the pricing curve q would be larger than her price in the envy-free pricing scheme. Therefore,

$$\overline{\text{rev}}(N^{\text{tot}}, q) = \sum_{j \in \mathcal{B}} q(\overline{m}_j(N^{\text{tot}}, q)) \geq \sum_{j \in \mathcal{B}} q(m'_j) = \sum_{j \in \mathcal{B}} p'_j.$$

which will complete the proof.

To show that $\overline{m}_j(N^{\text{tot}}, q) \geq m'_j$, let us first consider an arbitrary $m \leq m'_{|\mathcal{B}|}$. For any $m \leq m'_{|\mathcal{B}|}$, let $k \in [|\mathcal{B}|]$ be such that $m'_{k-1} < m \leq m'_k$. As valuations are non-decreasing we have,

$$v_j(m) - q(m) \leq v_j(m'_k) - q(m) = v_j(m'_k) - q(m'_k) \leq v_j(m'_j) - q(m'_j).$$

This implies that $m'_j \in \underset{m \leq m'_{|\mathcal{B}|}}{\arg\max} \, (v_j(m) - q(m))$. Finally, as $m'_{|\mathcal{B}|} \leq N^{\text{tot}}$, we have

$$m'_j \leq \max \left(\underset{m \leq m'_{|\mathcal{B}|}}{\arg\max} \, (v_j(m) - q(m)) \right) \leq \max \left(\underset{m \leq N^{\text{tot}}}{\arg\max} \, (v_j(m) - q(m)) \right) = \overline{m}_j(N^{\text{tot}}, q).$$

□

E.2.2 Proof of Lemma 7

Proof. First, let $q' = A\left(\{v_j\}_{j \in \mathcal{B}}, N_1^{\text{OPT}}, \epsilon\right)$ denote the price curve obtained by applying A to the buyers' valuations $\{v_j\}_{j \in \mathcal{B}}$, total amount of data N_1^{OPT}, and the approximation parameter ϵ. Using the definition of OPT in equation (18) and the equivalence between envy-free pricing schemes and ordered item pricing curves (see Lemma 6 and the discussion above it), we have

$$\text{OPT} = \text{rev}^{\text{OPT}} \left(N_1^{\text{OPT}}\right) - c_1 N_1^{\text{OPT}} = \overline{\text{rev}}^{\text{OPT}}(N_1^{\text{OPT}}) - c_1 N_1^{\text{OPT}}.$$

Since algorithm A returns a $|\mathcal{B}|\mathcal{O}(\epsilon)$ solution we get

$$\overline{\text{rev}}^{\text{OPT}}(N_1^{\text{OPT}}) - c_1 N_1^{\text{OPT}} \leq \overline{\text{rev}}(N_1^{\text{OPT}}, q') + |\mathcal{B}|\mathcal{O}(\epsilon) - c_1 N_1^{\text{OPT}}.$$

Finally, from the definition of N^+: $N^+ = \arg\max_{N \in \left\{1, 2, \ldots, \frac{|\mathcal{B}|}{N^{\text{tot}}}\right\}} \text{profit}_N$, since $N^+ \leq \frac{|\mathcal{B}|}{N^{\text{tot}}}$, we have $\text{profit}_{N^+} = \overline{\text{rev}}(N^+, q_{N^+}) - c_1 N^+ \geq \overline{\text{rev}}(N_1^{\text{OPT}}, q') - c_1 N_1^{\text{OPT}}$. Therefore,

$$\begin{aligned} \overline{\text{rev}}(N_1^{\text{OPT}}, q') + |\mathcal{B}|\mathcal{O}(\epsilon) - c_1 N_1^{\text{OPT}} &\leq \overline{\text{rev}}(N^+, q_{N^+}) - c_1 N^+ + |\mathcal{B}|\mathcal{O}(\epsilon) \\ &= \text{profit}(M, s^{\mathbb{I}}; \mu) + |\mathcal{B}|\mathcal{O}(\epsilon) \end{aligned}$$

so we conclude that

$$\text{profit}(M, s^{\mathbb{I}}; \mu) \geq \text{OPT} - |\mathcal{B}|\mathcal{O}(\epsilon).$$

□

E.3 Designing Profit-optimal Envy-free Mechanisms with Strategic Contributors

Our mechanism with strategic contributors will execute Algorithm 1 with inputs obtained from the results of Algorithm 2. The following theorem describes the mechanism formally and states its theoretical properties.

Theorem E.1. *Let N^+, $\{m_j^+\}_{j\in\mathcal{B}}$, $\{q_j^+\}_{j\in\mathcal{B}}$ be the output of Algorithm 2. Denote $\mathrm{OPT}^+ = \sum_{j\in\mathcal{B}} q_j^+ - c_1 N^+$ be the profit of this envy-free pricing scheme. Let M be the mechanism in Algorithm 1 executed with inputs $\widetilde{\mathrm{OPT}} = \mathrm{OPT}^+$, $N^{\widetilde{\mathrm{OPT}}} = N^+$, $\widetilde{m}_j = m_j^+$, and $\widetilde{p}_j = q_j^+$ for all $j \in \mathcal{B}$. Then, M satisfies BB, IRC, ICC, IRB, and EFB. Moreover, its expected profit at the well-behaved NE satisfies*

$$\mathrm{profit}(M, s^{\mathbb{I}}; \mu) \geq \mathrm{OPT} - (c_2 - c_1) - |\mathcal{B}|\mathcal{O}(\epsilon),$$

Proof. The proofs of the BB, ICC, and IRC properties are identical to Theorem 4.1—we only change the input of Algorithm1 while maintaining the same payment format in equations (8) and (9). We will prove IRB, and EFB and then prove the lower bound on the profit.

(IRB): As shown in (12), $\mathbb{E}\left[p_j'\right] = \widetilde{p}_j$, since we let $\widetilde{p}_j = q_j^+$ be the output of Algorithm 2, we have $\mathbb{E}\left[p_j'\right] = q_j^+$. Then, combing with definition of buyer's utility (4), and buyers's value for i.i.d. data (3), buyer j's utility under recommended strategy is

$$\begin{aligned} u_j^{\mathrm{B}}(M, s^{\mathbb{I}}) &:= u_j^{\mathrm{B}}(m_j, p_j, s^{\mathbb{I}}) := v_j^{\mathrm{B}}(M, s^{\mathbb{I}}) - \sup_{\mu\in\mathbb{R}} \mathbb{E}_{M,s^{\mathbb{I}}}\left[p_j'\right] \\ &= v_j^{\mathrm{iid}}(m_j^+) - \sup_{\mu\in\mathbb{R}} \mathbb{E}_{M,s^{\mathbb{I}}}\left[p_j'\right] = \ v_j^{\mathrm{iid}}(m_j^+) - q_j^+ \geq 0. \end{aligned} \tag{21}$$

In the last step, recall that $(m_j^+, q_j^+)_{j\in\mathcal{B}}$ are generated by Algorithm 2, combing with the buyer's purchase model in (19), we have $v_j^{\mathrm{iid}}(m_j^+) - q_j^+ \geq 0$.

(EFB): Under M, following $s^{\mathbb{I}}$ is envy-free for buyers.

Proof. To show that when contributors follow $s^{\mathbb{I}}$, M is envy-free for buyers, we must prove that $u_j^{\mathrm{B}}(m_j, p_j, s^{\mathbb{I}}) \geq u_j^{\mathrm{B}}(m_k, p_k, s^{\mathbb{I}})$, $\forall j, k \in \mathcal{B}$. As demonstrated in (21), we know $u_j^{\mathrm{B}}(M, s^{\mathbb{I}}) := \ u_j^{\mathrm{B}}(m_j, p_j, s^{\mathbb{I}}) = v_j^{\mathrm{iid}}(m_j^+) - q_j^+ \geq 0$. Similarly, by the definition of buyers utility (4) and buyer's valuation for i.i.d. data (3), we have

$$u_j^{\mathrm{B}}(m_k, p_k, s^{\mathbb{I}}) := v_j^{\mathrm{iid}}(m_k^+) - \sup_{\mu\in\mathbb{R}} \mathbb{E}_{M,s^{\mathbb{I}}}\left\{p_k'\right\} = \ v_j^{\mathrm{iid}}(m_k^+) - q_k^+ \leq \ v_j^{\mathrm{iid}}(m_j^+) - q_j^+.$$

The last step is by the definition of $\{q_j^+\}_{j\in\mathcal{B}}$ in Algorithm 2, and buyer's purchase model (19). □

Approximately Optimal Profit: Under M, when the contributors follow following $s^{\mathbb{I}}$, the expected profit of buyers approximates the optimal profit upper bound (defined in Theorem 3.2) with additive error $|\mathcal{B}|\mathcal{O}(\epsilon)$.

Proof. By definition of contributors' cumulative profit (14),

$$\begin{aligned} \mathrm{profit}(M, s^{\mathbb{I}}; \mu) &= \mathbb{E}_{M,s^{\mathbb{I}}}\left[\sum_{j\in\mathcal{B}} p_j' - \sum_{i\in\mathcal{C}} c_i N_i'\right] = \mathbb{E}_{M,s^{\mathbb{I}}}\left[\sum_{j\in\mathcal{B}} p_j'\right] - \sum_{i\in\mathcal{C}} c_i N_i' \\ &= \sum_{j\in\mathcal{B}} \widetilde{p}_j - \sum_{i\in\mathcal{C}} c_i N_i' = \sum_{j\in\mathcal{B}} q_j^+ - \sum_{i\in\mathcal{C}} c_i N_i' \qquad \text{(By proof of } \textbf{IRB}\text{, equation (12))} \end{aligned}$$

Recall that $\overline{\mathrm{rev}}(N^+, q_{N^+}) = \sum_{j\in\mathcal{B}} q_j^+$ and that $N^+ = \sum_i N_i'$, $N_1' = N^+ - 1$, $N_2' = 1$, we have

$$\begin{aligned} \mathrm{profit}(M, s^{\mathbb{I}}; \mu) &= \overline{\mathrm{rev}}(N^+, q_{N^+}) - c_1\left(N^+\right) - (c_2 - c_1) \geq \ \overline{\mathrm{rev}}(N_1^{\mathrm{OPT}}, q_{N_1^{\mathrm{OPT}}}) - c_1\left(N_1^{\mathrm{OPT}}\right) - (c_2 - c_1) \\ &\geq \overline{\mathrm{rev}}^{\mathrm{OPT}}\left(N_1^{\mathrm{OPT}}\right) - c_1\left(N_1^{\mathrm{OPT}}\right) - |\mathcal{B}|\mathcal{O}(\epsilon) - (c_2 - c_1) = \ \mathrm{OPT} - (c_2 - c_1) - |\mathcal{B}|\mathcal{O}(\epsilon), \end{aligned}$$

where the first inequality is by the optimality of N^+ in Algorithm 2, the second inequality is because algorithm A returns a $|\mathcal{B}|\mathcal{O}(\epsilon)$ approximation, and the last step is by definition of OPT (18) and the fact that $\mathrm{rev}^{\mathrm{OPT}}(N_1^{\mathrm{OPT}}) = \overline{\mathrm{rev}}^{\mathrm{OPT}}(N_1^{\mathrm{OPT}})$. □

□

F Notations

The following table contains the notations used in this paper.

Table 2: Table of Notations.

Notation	Meaning
$\mathcal{B}$	The set of buyers.
$\mathcal{C}$	The set of contributors.
c_i	The cost for each contributor to collect one data point.
N_i'	Amount of data the broker asks each contributor to collect.
$N_i : \{c_i\}_{i\in\mathcal{C}} \to N_i'$	The function mapping the cost to N_i'.
n_i'	The amount of data each contributor actually collects.
$n_i : (N_i', c_i) \to n_i'$	The contributor i's strategy function to decide amount of data to collect.
Z_i	The set of data contributor i actually collects, containing n_i' i.i.d. points.
X_i'	The set of data each contributor reports.
X_i	The contributors' data reporting function, may enable strategic behaviors (e.g., fabrication).
$s_i = (n_i, X_i)$	Contributor i's strategy profile.
$s_i^{\mathbb{I}} = (\mathbb{I}, \mathbb{I})$	Contributor is well-behaved: follow the instruction to collect N_i' data and submit truthfully.
m_j'	The amount of data points the broker allocates to each buyer $j \in \mathcal{B}$.
m_j	The broker's data allocation function to decide m_j'.
p_j'	The broker charges buyer j a price p_j' for m_j' data points.
p_j	The broker's price function to decide p_j'.
π_i'	Payment that the broker assigns to each contributor.
π_i	The broker's payment function to decide how to reward contributor i for data collection.
M	Mechanism, $M = (\{N_i\}_{i\in\mathcal{C}}, \{m_j\}_{j\in\mathcal{B}}, \{p_j\}_{j\in\mathcal{B}}, \{\pi_i\}_{i\in\mathcal{C}})$.
v^{e}	Error-based buyer valuation (11).
v^{B}	Buyer valuation under a mechanism and a strategy profile (2).
v^{iid}	Buyer valuation under clean data (3).
$u_j^{\mathrm{B}}(M, s^{\mathbb{I}})$	Buyer utility under mechanism M and strategy s (4).
$W(M, s)$	Welfare of a mechanism M under a strategy profile s (6).
OPT	Welfare (7) or profit (16) maximization baseline.
N^{OPT}	The amount of data to be collected to achieve welfare/profit-optimal welfare baseline OPT.
$\widetilde{\mathrm{OPT}}$	Input parameter in Algorithm 1, denoting the welfare/profit-optimal baseline.
$N^{\widetilde{\mathrm{OPT}}}$	Input parameter in Algorithm 1, denoting the total amount of data to be collected.
$\widetilde{m}_j$	Input parameter in Algorithm 1, denoting the amount of data to sell to buyers.
$\widetilde{p}_j$	Input parameter in Algorithm 1, denoting the expected price to charge buyers.
$\mathrm{profit}(M, s)$	Profit of a mechanism M under a strategy profile s (14).
A	Ordered item pricing algorithm, introduced in Section E.2.