Signaling in Data Markets via Free Samples

author names withheld

Under Review for NExT-Game 2026

Abstract

We study a setting in which a data buyer seeks to estimate an unknown parameter by purchasing samples from one of K data sellers. Each seller has privately known data quality (e.g., high vs. low variance) and a private per-sample cost. We consider a multi-stage game in which the first stage is a free-trial stage in which the sellers have the option of signaling data quality by offering a few samples of data for free. Buyers update their beliefs based on the sample variance of the free data and then run a procurement auction to buy data in a second stage. For the auction stage, we characterize an approximately optimal Bayesian incentive compatible mechanism: the buyer selects a single seller by minimizing a belief-adjusted virtual cost and chooses the purchased sample size as a function of posterior quality and virtual cost. For the free-trial stage, we characterize the equilibrium, taking the above mechanism as the continuation game. Free trials may fail to emerge: for some parameters, all sellers reveal zero samples. However, under sufficiently strong competition (large K), there is an equilibrium in which sellers reveal the maximum allowable number of samples; in fact, it is the unique equilibrium.

1. Introduction

Data markets are increasingly central to economic and scientific activity: a decision maker can purchase datasets from competing providers to estimate a latent quantity of interest, yet the quality of these datasets is often known only to the sellers and is hard to verify ex ante. A natural response—mirroring "free trials" in software and consumer goods—is to let sellers reveal a small number of samples for free so that buyers can assess quality before purchasing. This paper studies when, and why, free trials emerge in competitive data markets, and when they fail to do so. See Section A for a discussion of related work on data markets, strategic information revelation, free samples, competition, and persuasion.

We model a market with K data sellers, each with privately known quality $\sigma_i \in \{\sigma_L, \sigma_H\}$ and per-sample cost c_i, and a buyer who wants to estimate an unknown parameter θ from i.i.d. normal samples. Sellers first choose how many samples $m_i \in \{0, \ldots, M\}$ to reveal for free; the buyer updates beliefs from the empirical variance of the revealed samples and then runs a procurement mechanism that decides from whom to buy, how many additional samples to buy, and what to pay. Free samples are restricted to evaluation, consistent with "evaluation-only" licensing.

This setup couples a multi-stage signaling problem with a downstream Bayesian mechanism-design problem. We first derive an approximately optimal Bayesian incentive-compatible mechanism for the procurement stage: under standard regularity conditions on the cost distribution, the optimal mechanism for a real-valued relaxation of the buyer's problem is *single-sourced*—the buyer purchases all samples from a single seller chosen by minimizing a belief-adjusted virtual cost index—and the integer mechanism obtained by rounding down the relaxed allocation retains both BIC and approximate optimality.

Taking this mechanism as the continuation game, we analyze approximate subgame-perfect equilibria in sellers' free-trial choices, and find two qualitatively different regimes. For some parameter ranges, no seller shares any free samples in equilibrium, even for arbitrarily large K: the prospect of having a poor draw revealed (which sharply lowers the selection probability when σ_H is realized) and of selling fewer samples upon winning (because tighter posteriors lower the buyer's chosen sample size) together dominate the benefit of signaling. Under sufficiently strong competition, however, every seller shares the maximum allowable number of samples in equilibrium, and this is the *unique* equilibrium given the buyer's mechanism: small differences in posterior quality translate into first-order advantages in winning probability, and any deviator's selection probability collapses exponentially in K. We complement these two extremes with simulations across $(\sigma_L, \sigma_H, K, M)$ that exhibit intermediate symmetric equilibria with $0 < m^* < M$, and show that multiple symmetric equilibria can coexist for the same parameters.

2. Model

We consider a market with K data sellers and a buyer who wants to estimate a parameter θ. Seller $i \in [K]$ can provide i.i.d. samples from $\mathcal{N}(\theta, \sigma_i^2)$, where $\sigma_i \in \{\sigma_L, \sigma_H\}$ (with $\sigma_L < \sigma_H$) is seller i's privately known data quality and c_i is its privately known per-sample cost (operational or privacy). Ex ante, $\sigma_i = \sigma_L$ with probability μ and c_i is drawn from a continuous distribution with pdf f and cdf F on $[c_{\min}, c_{\max}]$, $c_{\min} > 0$. These distributions are independent and identical across sellers.

Timing and information. The interaction is a multi-stage game. (i) Each seller i commits to a free-trial size $m_i \in \{0, 2, 3, \ldots, M\}$ before (σ_i, c_i) are realized; we exclude $m_i = 1$ so that the sample variance is well defined. (ii) The buyer arrives, (σ_i, c_i) are realized, and each seller submits a reported cost c_i' together with a free-sample dataset $\mathcal{S}_i^f$ of size m_i drawn i.i.d. from $\mathcal{N}(\theta, \sigma_i^2)$ (sellers cannot cherry-pick). (iii) The buyer forms a Bayesian posterior $\pi_i(\mathcal{S}_i^f)$ over σ_i^2 from the sample variance of $\mathcal{S}_i^f$ (a mean-invariant statistic, so no prior over θ is needed), and then chooses how many samples $n_i(\boldsymbol{c}'; \boldsymbol{\pi})$ to purchase from each seller and a payment $t_i(\boldsymbol{c}'; \boldsymbol{\pi})$. Free samples can be used only for evaluation (belief updating), consistent with evaluation-only licensing [see, e.g., 55].

Estimation and utilities. The buyer estimates θ by a weighted average $\hat{\theta} = \sum_{i=1}^{K} w_i \frac{1}{n_i} \sum_{j=1}^{n_i} x_j^i$ with $w_i \geq 0$ and $\sum_i w_i = 1$, which under the buyer's beliefs has variance $\sum_{i=1}^{K} w_i^2 \bar{\sigma}_i^2 / n_i$, where $\bar{\sigma}_i^2 := \pi_i(\sigma_L^2)\sigma_L^2 + \pi_i(\sigma_H^2)\sigma_H^2$ is the posterior mean variance. Seller i's utility is $U_i = t_i(\boldsymbol{c}'; \boldsymbol{\pi}) - n_i(\boldsymbol{c}'; \boldsymbol{\pi})\, c_i$, and the buyer trades off precision against payments:

$$U_b(\boldsymbol{c}', \boldsymbol{w}; \boldsymbol{\pi}; \boldsymbol{n}, \boldsymbol{t}) := -\sum_{i=1}^{K} w_i^2 \frac{\bar{\sigma}_i^2}{n_i} - \lambda \sum_{i=1}^{K} t_i(\boldsymbol{c}'; \boldsymbol{\pi}), \tag{1}$$

where λ is the buyer's relative weight on payments versus estimation error.

Solution concept. We use ζ-approximate subgame-perfect Bayesian Nash equilibrium. Fix beliefs $\boldsymbol{\pi}$. By the revelation principle the buyer's continuation $(\boldsymbol{n}, \boldsymbol{t})$ solves

$$\max_{\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{n}} \mathbb{E}_{\boldsymbol{c}}\big[U_b(\boldsymbol{c}, \boldsymbol{w}; \boldsymbol{\pi}; \boldsymbol{n}, \boldsymbol{t})\big] \quad \text{s.t. Bayesian IC and interim IR for every seller.} \tag{2}$$

A pair $(\boldsymbol{n}^*, \boldsymbol{t}^*)$ is a ζ-*approximate solution* to (2) if it satisfies BIC and IR exactly and the buyer's expected cost (negative utility) is within a factor $1+\zeta$ of optimal. The triple $(\boldsymbol{m}^*, \boldsymbol{n}^*, \boldsymbol{t}^*)$ is a ζ-*approximate subgame-perfect equilibrium* if $(\boldsymbol{n}^*, \boldsymbol{t}^*)$ is a ζ-approximate solution and no seller has a profitable unilateral deviation in $m_i' \in \{0, 1, \ldots, M\}$, evaluating ex-ante payoffs over $(\boldsymbol{c}, \mathcal{S}^f)$ under $(\boldsymbol{n}^*, \boldsymbol{t}^*)$. The approximation enters only through the buyer's continuation; sellers best-respond exactly.

3. The Mechanism Design Subproblem

We start by analyzing the buyer's choice of allocation and payment functions, given the free samples received from all data sellers and the induced belief profile $\boldsymbol{\pi}$. We first make the following customary regularity assumption on the distribution of sample costs, i.e., $F(\cdot)$, which ensures that the *virtual cost*, as defined below, is non-decreasing.

Assumption 1 (Monotonicity of virtual costs) *The virtual cost $\psi(c) := c + F(c)/f(c)$ is a non-decreasing function of c.*

This assumption is standard in mechanism design and holds for a variety of distributions including uniform, normal, and exponential distributions [see, e.g., 56].

To solve the mechanism design problem (2), we first relax it by allowing the per-seller sample size to be real-valued, then round the resulting allocation down to the nearest integer; we show this is approximately optimal.

Proposition 1 (Approximately Optimal Purchasing and Payment Rules) *Suppose Assumption 1 holds. Given the sequence of free samples shared by the sellers $\left(\mathcal{S}_i^f\right)_{i=1}^K$ inducing beliefs $(\pi)_{i=1}^K$ with means $(\bar{\sigma}_i^2)_{i=1}^K$, there is a sample purchasing and payment rule that forms a Bayesian Incentive Compatible mechanism with the following properties. When the sellers report costs $c_1^{'}, \ldots, c_K^{'}$,*

1. *The buyer buys only from one seller $i^* \in [K]$ with $i^* \in \text{argmin}_{i \in [K]} \bar{\sigma}_i^2 \psi(c_i')$.*
2. *The buyer purchases $\lfloor n^* \rfloor$ samples from seller i^*, where $n^* = \bar{\sigma_{i^*}} / \sqrt{\lambda \psi(c_{i^*}^{'})}$.*
3. *The expected payment each seller receives is the expectation of product of the number of samples purchased from the seller and the virtual cost of the seller $\mathbb{E}_{\boldsymbol{c}}[n_i(\boldsymbol{c}; \boldsymbol{\pi}) \psi(c_i)]$.*
4. *This mechanism yields a utility that is suboptimal by a factor of $1 - \frac{1}{n^*-1}$ compared to the optimal BIC mechanism.*

The remainder of this section is devoted to proving this result. Before doing so, however, we make two remarks. First, as the proof will establish, the optimal mechanism corresponding to the aforementioned relaxed problem is *single-sourced*, meaning that the buyer purchases all samples (i.e., n^* samples) from a single data seller, denoted by seller i^*. Our rounded allocation preserves this property as well.

Second, it is straightforward to see that, under the assumption

$$\sigma_L \geq \alpha \sqrt{\lambda \psi(c_{\max})}, \tag{3}$$

for some positive constant $\alpha \geq 3$, we can ensure that $n^* \geq \alpha$, which in turn guarantees a $(\alpha - 2)/(\alpha - 1)$-approximation factor. If we are satisfied with high-probability lower bounds on n^*, it is straightforward to see that $\psi(c_{\max})$ can be replaced in this condition by a smaller constant, since the winning data seller i^* is likely to have a low cost. While we do not formalize this point here, we emphasize that assumptions of this form are generally unavoidable; otherwise, if λ or the costs are too high, the cost of purchasing samples may induce the data buyer to purchase no samples at all. We next present the proof.

Proof [Proof sketch of Theorem 1] We first identify the variance-minimizing weights for any fixed sample allocation: by Cauchy–Schwarz, $w_i^* \propto n_i/\bar{\sigma}_i^2$ (Theorem 4 in Section C). Substituting these weights, applying Myerson's lemma to the procurement subproblem, and minimizing the resulting buyer objective pointwise in $\boldsymbol{c}$ shows that all samples should be purchased from the single seller minimizing $\bar{\sigma}_i^2\psi(c_i)$, with $n^* = \bar{\sigma}_{i^*}/\sqrt{\lambda\psi(c_{i^*})}$. Assumption 1 makes this allocation monotone in the reported cost, so the Myerson payment rule yields a BIC mechanism. Rounding n^* down preserves monotonicity (and hence BIC) and decreases utility by at most a factor $1/(n^* - 1)$. The full proof appears in Section C. ■

4. Free-Sample Equilibria

We now analyze sellers' free-sample choices, taking the mechanism of Theorem 1 as the buyer's continuation. The approximation in our equilibrium concept enters only through this continuation, since the integer-valued buyer problem need not admit a closed form; an exact-equilibrium analysis can be recovered by restricting the buyer ex ante to the mechanism of Theorem 1. Throughout this section we additionally assume:

Assumption 2 (Bounds on the cost density) *The cost density f satisfies $\ell \leq f(c) \leq L$ for every $c \in [c_{min}, c_{max}]$.*

4.1. The uninformative equilibrium

When sellers share a large number of samples at equilibrium, the buyer is more informed about the quality of data that they purchase. On the other hand, when fewer samples are shared at equilibrium, the buyer has a higher degree of uncertainty about the data quality purchased. Our first result shows that for some problem parameters, the equilibrium can be maximally uninformative. That is, there is an equilibrium where no seller shares any free samples.

Theorem 2 (Uninformative Equilibrium) *Suppose that Assumption 1 and Assumption 2 hold and that $\sigma_L > \sqrt{\lambda\psi(c_{max})}$. Then, for any number of sellers K and any $\zeta > 0$, there exist values of σ_L, σ_H, μ such that there is a ζ-approximate subgame perfect Nash equilibrium where all sellers share zero free samples.*

Note that there are two conflicting forces shaping data sellers' incentives to share data. On the one hand, competition among sellers may push each seller to reveal their quality; otherwise, they may lose even if they possess high-quality (low-variance) samples. On the other hand, revealing information has two potential drawbacks. First, if a seller's sample variance is low, they may indeed win, but the buyer will purchase fewer samples from them, as suggested by Theorem 1, since the buyer can achieve the desired precision with a smaller number of samples when quality is known to be high. Second, if the realized sample variance is high, the seller may regret revealing this information.

Interestingly, as the proof shows, for any number of sellers (that is, for any level of competition) there exist parameter regimes in which the second force dominates, to the point that sellers may prefer not to share any samples for free. In particular, as the proof shows, this outcome arises when the ratio σ_H/σ_L is sufficiently large, which amplifies the disincentive to share free samples.

Proof sketch of Theorem 2. We argue that the no-sharing profile, paired with the mechanism of Theorem 1, is an approximate subgame-perfect equilibrium by comparing each seller's utility before and after a unilateral deviation. *Before deviation*, a symmetry argument together with the Myerson payment from Theorem 1 produces a strictly positive $\Omega(1/(K(K+1)L^2))$ lower bound on per-seller utility, scaled by $\sqrt{(\mu\sigma_L^2 + (1-\mu)\sigma_H^2)/(\lambda\psi(c_{\max}))}$ in the $m = 0$ case (Theorem 5 in Section D). *After deviation*, the buyer's posterior over the seller's quality moves away from the prior: when σ_L^2 is realized it shifts toward σ_L^2, decreasing the payment upon selection; when σ_H^2 is realized it shifts toward σ_H^2, decreasing the chance of selection. We make this belief shift quantitative via a chi-squared tail bound on the likelihood ratio (Theorem 6) and tune $\Delta_L, \Delta_H, \delta$ together with a sufficiently large σ_H/σ_L ratio so that the post-deviation utility is strictly below $\bar{U}_{\text{lb}}(0)$. Choosing σ_L to satisfy (3) with $1/(\alpha-1) < \zeta$ delivers the ζ-approximate equilibrium. The full proof appears in Section D.

4.2. The maximally informative equilibrium

In this section, we ask the question the other way around. If we fix the problem's parameters and increase the number of sellers, does the incentive to share data dominate? It turns out that the answer is yes. In fact, as the number of sellers grows, sharing the maximum possible number of samples M becomes an equilibrium—and indeed the unique equilibrium. The following result formalizes this claim.

Theorem 3 (Equilibrium is maximally informative when there are many sellers) *Suppose that Assumption 1 and Assumption 2 hold. Let $\sigma_L \geq \alpha\sqrt{\lambda\psi(c_{\max})}$ for some $\alpha \geq 3$. Then:*

- *For every σ_L, σ_H, μ, there exists $\bar{K}$, such that, for every number of sellers $K \geq \bar{K}$, the strategy profile in which each seller shares the maximum possible number of free samples (M) and the buyer runs the mechanism in Theorem 1 is a $1/(\alpha-1)$-approximate subgame perfect equilibrium.*
- *Assuming the buyer runs the mechanism in Theorem 1, the above strategy profile is the unique equilibrium.*

Proof sketch of Theorem 3. The argument hinges on a fundamental floor for the posterior mean variance: when only $m < M$ samples are shared, the posterior mean is at least $\sigma_{lb}(m)$, and σ_{lb} is strictly decreasing in m (Theorem 7 in Section E). Hence any seller sharing M samples is dominated whenever both their realized cost $c_i \leq c_0$ and posterior mean $\bar{\sigma}_i \leq \sigma_0$ for suitable thresholds c_0, σ_0 tied to the gap between $\sigma_{lb}(M-1)$ and $\sigma_{lb}(M)$. Calling this event E_i^{select}, one shows it has probability $q > 0$ for any seller sharing M samples (Theorem 8). Letting J be the number of M-sharers, selection probability for any $m < M$-sharer is at most $(1-q)^J$ (Theorem 9), while expected utility of an M-sharer is at least $\Omega(1/J^3)$ (Theorem 10). The exponential-vs-polynomial

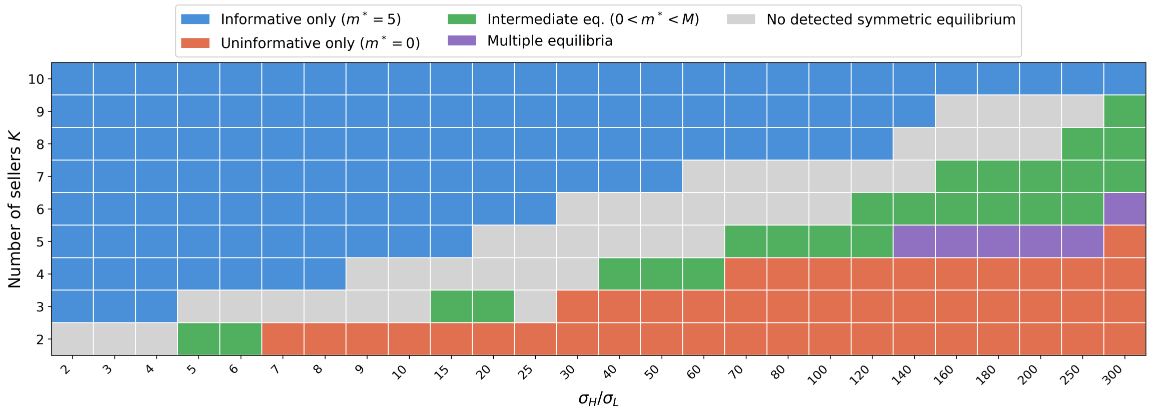

Figure 1: Phase diagram of symmetric equilibria across $(K, \sigma_H/\sigma_L)$. Each cell is colored based on the equilibrium detected in the corresponding regime of parameters. Blue: only the informative equilibrium ($m^* = M$). Red: only the uninformative ($m^* = 0$). Green: only an intermediate equilibrium ($m^* = 2$). Purple: coexistence of multiple symmetric equilibria. Gray: no symmetric equilibrium found.

contrast rules out unilateral deviations from the all-M profile and, by a two-case argument splitting on whether $J \geq K^{1/4}$, also rules out any other profile. The full proof appears in Section E.

5. Numerical Simulations

Theorems 2 and 3 characterize the uninformative and maximally informative equilibria in limiting regimes: the former when σ_H/σ_L is large, the latter when K is large. To map out the equilibrium structure across the full parameter space—and in particular to probe whether additional equilibria exist—we conduct Monte Carlo simulations.

We simulate the game with K symmetric sellers whose per-sample costs are drawn i.i.d. from $\text{Uniform}[c_{\min}, c_{\max}]$ and whose data variances are σ_L^2 with probability μ and σ_H^2 with probability $1 - \mu$. The buyer runs the approximately optimal mechanism of Theorem 1. We fix $\sigma_H = 50$ and vary $\sigma_L = \sigma_H/r$ to sweep the ratio $r = \sigma_H/\sigma_L$, with $\mu = 0.6$, $\lambda = 0.007$, $c_{\min} = 0.5$, $c_{\max} = 2.0$, and $M = 5$. For each candidate symmetric strategy $m^* \in \{0, 2, 3, \ldots, M\}$, we estimate seller 1's expected profit under every deviation using $N = 100{,}000$ Monte Carlo samples. We consider symmetric strategies and find the strategy where each seller shares m^* samples an equilibrium only if its estimated profit exceeds every alternative by at least two combined standard errors. While this approach ensures that the detected equilibria are indeed equilibria with significant certainty, it cannot detect equilibria where deviation does not decrease utility or decreases utility by a small amount.

Figure 1 maps the equilibrium structure across $K \in \{2, \ldots, 10\}$ and $\sigma_H/\sigma_L \in \{2, \ldots, 300\}$. Five regions emerge:

1. *Informative equilibrium* (blue). At low σ_H/σ_L or high K, the unique symmetric equilibrium is $m^* = M$. For $K \geq 10$, this is the only equilibrium at every ratio tested.

2. *Uninformative equilibrium* (red). At high σ_H/σ_L and small K, the equilibrium $m^* = 0$ exists.

3. *Intermediate equilibrium* (green). Between the maximally informative and maximally uninformative regions, we have an intermediate equilibrium with $m^* = 2$ emerges. This equilibrium appears for every K from 2 to 9 at appropriate ratios.

4. *Multiple equilibria* (purple). For a band of parameters (e.g., $K = 5$ at $\sigma_H/\sigma_L \in [140, 250]$; $K = 6$ at $\sigma_H/\sigma_L = 300$), both $m^* = 0$ and $m^* = 2$ are simultaneously equilibria.

5. *No detected symmetric equilibrium* (gray). In some regions, we do not detect any symmetric strategies to be equilibria. Note that this does not mean that there are no symmetric equilibria for those parameters. But rather, we are not able to detect it given the uncertainty in our Monte Carlo sampling.

References

[1] Laura Abrardi, Carlo Cambini, and Flavio Pino. Data brokers competition, synergic datasets, and endogenous information value. *International Journal of Industrial Organization*, page 103146, 2025.

[2] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14 (4):218–256, 2022.

[3] Daron Acemoglu, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Learning from reviews: The selection effect and the speed of learning. *Econometrica*, 90(6):2857–2899, 2022.

[4] Alessandro Acquisti and Hal R Varian. Conditioning prices on purchase history. *Marketing Science*, 24(3):367–381, 2005.

[5] Anish Agarwal, Munther Dahleh, and Tuhin Sarkar. A marketplace for data: An algorithmic solution. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 701–726, 2019.

[6] Anish Agarwal, Munther Dahleh, Thibaut Horel, and Maryann Rui. Towards data auctions with externalities. *Games and Economic Behavior*, 148:323–356, 2024.

[7] Nivasini Ananthakrishnan, Stephen Bates, Michael Jordan, and Nika Haghtalab. Delegating data collection in decentralized machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 478–486. PMLR, 2024.

[8] Pak Hung Au and Keiichi Kawai. Competitive information disclosure by multiple senders. *Games and Economic Behavior*, 119:56–78, 2020.

[9] Moshe Babaioff, Robert Kleinberg, and Renato Paes Leme. Optimal mechanisms for selling information. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 92–109, 2012.

[10] Kapil Bawa and Robert Shoemaker. The effects of free sample promotions on incremental brand sales. *Marketing Science*, 23(3):345–363, 2004.

[11] Dirk Bergemann and Alessandro Bonatti. Selling cookies. *American Economic Journal: Microeconomics*, 7(3):259–294, 2015.

[12] Dirk Bergemann and Alessandro Bonatti. Markets for information: An introduction. *Annual Review of Economics*, 11(1):85–107, 2019.

[13] Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

[14] Dirk Bergemann and Martin Pesendorfer. Information structures in optimal auctions. *Journal of economic theory*, 137(1):580–609, 2007.

[15] Dirk Bergemann and Juuso Välimäki. Dynamic pricing of new experience goods. *Journal of Political Economy*, 114(4):713–743, 2006.

[16] Dirk Bergemann, Benjamin Brooks, and Stephen Morris. First-price auctions with general information structures: Implications for bidding and revenue. *Econometrica*, 85(1):107–143, 2017.

[17] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The design and price of information. *American economic review*, 108(1):1–48, 2018.

[18] Dirk Bergemann, Alessandro Bonatti, and Tan Gan. The economics of social data. *The RAND Journal of Economics*, 53(2):263–296, 2022.

[19] Omar Besbes and Marco Scarsini. On information distortions in online ratings. *Operations Research*, 66(3):597–610, 2018.

[20] Kostas Bimpikis, Davide Crapis, and Alireza Tahbaz-Salehi. Information sale and competition. *Management Science*, 65(6):2646–2664, 2019.

[21] Simon Board and Jay Lu. Competitive information disclosure in search markets. *Journal of Political Economy*, 126(5):1965–2010, 2018.

[22] Raphael Boleslavsky, Christopher S Cotton, and Haresh Gurnani. Demonstrations and price competition in new product release. *Management Science*, 63(6):2016–2026, 2017.

[23] Alessandro Bonatti, Munther Dahleh, Thibaut Horel, and Amir Nouripour. Selling information in competitive environments. *Journal of Economic Theory*, 216:105779, 2024.

[24] Etienne Boursier, Vianney Perchet, and Marco Scarsini. Social learning in non-stationary environments. In *International Conference on Algorithmic Learning Theory*, pages 128–129. PMLR, 2022.

[25] Yang Cai, Constantinos Daskalakis, and Christos Papadimitriou. Optimum statistical estimation with strategic data sources. In *Conference on Learning Theory*, pages 280–296. PMLR, 2015.

[26] Christophe Chamley. *Rational Herds: Economic Models of Social Learning*. Cambridge University Press, 2004.

[27] Junjie Chen, Minming Li, and Haifeng Xu. Selling data to a machine learner: Pricing via costly signaling. In *International Conference on Machine Learning*, pages 3336–3359. PMLR, 2022.

[28] Yiling Chen, Nicole Immorlica, Brendan Lucier, Vasilis Syrgkanis, and Juba Ziani. Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 27–44, 2018.

[29] Hsing Kenneth Cheng and Qian Candy Tang. Free trial or no free trial: Optimal software product design with network effects. *European Journal of Operational Research*, 205(2): 437–447, 2010.

[30] Vincent Conitzer, Curtis R Taylor, and Liad Wagman. Hide and seek: Costly consumer privacy in a market with repeat purchases. *Marketing Science*, 31(2):277–292, 2012.

[31] Davide Crapis, Bar Ifrach, Costis Maglaras, and Marco Scarsini. Monopoly pricing in the presence of social learning. *Management Science*, 63(11):3586–3608, 2017.

[32] Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani. Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 Conference on Innovations in Iheoretical Computer Science*, pages 317–324, 2015.

[33] Rachel Cummings, Hadi Elzayn, Emmanouil Pountourakis, Vasilis Gkatzelis, and Juba Ziani. Optimal data acquisition with privacy-aware agents. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 210–224. IEEE, 2023.

[34] Alexandre De Corniere and Greg Taylor. Data and competition: A simple framework. *The RAND Journal of Economics*, 56(4):494–510, 2025.

[35] Kimon Drakopoulos and Ali Makhdoumi. Providing data samples for free. *Management Science*, 69(6):3536–3560, 2023.

[36] Péter Eső and Balazs Szentes. Optimal information disclosure in auctions and the handicap auction. *The Review of Economic Studies*, 74(3):705–731, 2007.

[37] Alireza Fallah, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Bridging central and local differential privacy in data acquisition mechanisms. *Advances in Neural Information Processing Systems*, 35:21628–21639, 2022.

[38] Alireza Fallah, Michael I Jordan, Ali Makhdoumi, and Azarakhsh Malekian. On three-layer data markets. *arXiv preprint arXiv:2402.09697*, 2024.

[39] Alireza Fallah, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar. Optimal and differentially private data acquisition: Central and local mechanisms. *Operations Research*, 72(3):1105–1123, 2024.

[40] Matthew Gentzkow and Emir Kamenica. Competition in persuasion. *The Review of Economic Studies*, 84(1):300–322, 2016.

[41] Yiquan Gu, Leonardo Madio, and Carlo Reggiani. Data brokers co-opetition. *Oxford Economic Papers*, 74(3):820–839, 2022.

[42] Saram Han and Chris K Anderson. Customer motivation and response bias in online reviews. *Cornell Hospitality Quarterly*, 61(2):142–153, 2020.

[43] Amir Heiman, Bruce McWilliams, Zhihua Shen, and David Zilberman. Learning and forgetting: Modeling optimal product sampling over time. *Management Science*, 47(4):532–546, 2001.

[44] Shota Ichihashi. Online privacy and information disclosure by consumers. *American Economic Review*, 110(2):569–595, 2020.

[45] Shota Ichihashi. Competing data intermediaries. *The RAND Journal of Economics*, 52(3): 515–537, 2021.

[46] Shota Ichihashi. The economics of data externalities. *Journal of Economic Theory*, 196: 105316, 2021.

[47] Bar Ifrach, Costis Maglaras, Marco Scarsini, and Anna Zseleva. Bayesian social learning from consumer reviews. *Operations Research*, 67(5):1209–1221, 2019.

[48] Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272, 2019.

[49] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[50] Yifan Li, Xiaohui Yu, and Nick Koudas. Data acquisition for improving machine learning models. *arXiv preprint arXiv:2105.14107*, 2021.

[51] Guocheng Liao, Yu Su, Juba Ziani, Adam Wierman, and Jianwei Huang. The privacy paradox and optimal bias–variance trade-offs in data acquisition. *Mathematics of Operations Research*, 49(4):2749–2767, 2024.

[52] Rodrigo Montes, Wilfried Sand-Zantman, and Tommaso Valletti. The value of personal information in online markets with endogenous privacy. *Management Science*, 65(3):1342–1362, 2019.

[53] Roger B Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

[54] Phillip Nelson. Information and consumer behavior. *Journal of Political Economy*, 78(2): 311–329, 1970.

[55] Precisely. Data end user license agreement (april 2021). https://www.precisely.com/legal/licensing/data-end-user-license-agreement-april-2021/, 2021. Accessed 2026-02-09.

[56] Kaj Rosling. Inventory cost rate functions with nonlinear shortage costs. *Operations Research*, 50(6):1007–1017, 2002.

[57] Eden Saig, Inbal Talgam-Cohen, and Nir Rosenfeld. Delegated classification. *Advances in Neural Information Processing Systems*, 36:13200–13236, 2023.

[58] Carl Shapiro. Premiums for high quality products as returns to reputations. *The Quarterly Journal of Economics*, 98(4):659–679, 1983.

[59] Curtis R Taylor. Consumer privacy and the market for customer information. *RAND Journal of Economics*, pages 631–650, 2004.

Appendix A. Related Work

Data markets. Our work lies in the literature on data markets [see, e.g., 2, 12], and more specifically in the design of mechanisms for data acquisition to estimate an underlying population statistic [28, 32, 33, 37, 39, 51], which is closest to our setting, as well as mechanisms for learning machine learning models [5, 7, 25, 27, 50, 57]. In many of these works, the data buyer is assumed to know, observe, or control data quality through mechanism design or the design of sampling and evaluation strategies. In contrast, our work allows sellers to opt in to sharing samples and thereby influence the resulting equilibrium behavior.

Strategic information revelation in data markets. Strategic information revelation in data markets has been studied for various modes of information revelation. For example, sellers may disclose reviews or ratings [3, 19, 24, 26, 31, 42, 47]. Data owners may obfuscate the quality of their data to guard against price discrimination [4, 30, 44, 59]. The mode we study in our work is that of revealing information through providing free samples.

Information revelation through free samples. Previous work by [35] also studies free samples in data markets. They focus on the *hold up* problem, which involves preventing the buyer from free-riding on the seller's free samples. In contrast, our focus is on how free samples shape buyer's beliefs about the quality of data. There is a rich literature on free samples or trials in broader settings [10, 15, 21, 22, 29, 43, 54, 58].

Competition and externalities in data markets. Previous work has studied how data platforms shape competition among sellers [11, 18, 34, 38, 46]. There is also work studying platform level competition and coordination in data markets [1, 41, 45]. Finally, there is work studying competition among buyers in data markets [20, 23, 52]. Our work focuses on the effect of competition in data markets when there is an option to provide free samples.

Persuasion and information design. The role of free samples in shaping buyer's beliefs about data quality is closely related to persuasion and information design [13, 48, 49]. Particularly related is work on how competition increases information revelation [8, 40]. Our results can be viewed as an analogue where there are structural constraints on how signals can be generated; namely, through providing free samples, only the sample size can be controlled. Persuasion and information design has been studied in the context of auctions [6, 14, 16, 36], pricing [17], and mechanism design [9, 23].

Appendix B. Conclusion

This paper studies the strategic role of free trials in competitive data markets. We model free trials as a pre-auction disclosure decision: before a buyer arrives, each seller commits to releasing a limited number of free samples, after which the buyer runs a procurement mechanism to purchase data. To make the continuation game tractable, we analyze a simple mechanism that (approximately) optimizes the buyer's tradeoff between statistical accuracy and procurement cost. The resulting continuation policy induces a clear, belief-adjusted comparison of sellers based on both reported costs and posterior beliefs about data quality inferred from the free samples.

Our analysis delivers two main insights. First, free trials need not emerge in equilibrium: when the quality gap is sufficiently large, sellers may prefer to remain uninformative rather than risk revealing adverse information. Second, when competition is sufficiently intense, maximal disclosure becomes the unique prediction: with many sellers, the probability that at least one seller appears highly attractive under the buyer's selection rule increases, and free samples become a decisive signal for winning the procurement auction.

There are several natural directions for future work. One extension is to allow sellers' datasets to be heterogeneous not only in variance but also in mean, so that sellers' data may exhibit different biases or systematic shifts relative to the buyer's target. A second direction is to relax the assumption that free samples are i.i.d. draws from the same distribution as the purchased data. Finally, it would be valuable to extend the analysis beyond Gaussian data. In this paper the Gaussian assumption is used primarily to obtain a tractable likelihood for the sample variance and to derive likelihood-ratio bounds that drive the belief-updating and equilibrium arguments. Similar results should be attainable for broader families of distributions as long as one can derive suitable bounds on the relevant likelihood ratios (or comparable concentration bounds) for the statistic used to summarize free samples.

Appendix C. Proof of Theorem 1

We first find the optimal choice of weights $\boldsymbol{w}$ to produce the least variance estimator given a choice of $\boldsymbol{n}$ and $\boldsymbol{t}$. This will allow us to rewrite the mechanism design problem in terms of optimizing over $\boldsymbol{n}$ and $\boldsymbol{t}$.

Lemma 4 (Optimal weights to minimize variance) *Given the sequence of samples, purchases, payments, and means of posteriors $\boldsymbol{n}, \boldsymbol{t}, \bar{\boldsymbol{\sigma}}$, the vector of weights maximizing buyer utility U_b is $\boldsymbol{w}^*$ with*

$$w_i^* = \frac{n_i/\bar{\sigma}_i{}^2}{\sum_{i\in[K]} n_i/\bar{\sigma}_i{}^2}.$$

Proof [Proof of Theorem 4] Utility function U_b depends on $\boldsymbol{w}$ only through $\text{Var}(\boldsymbol{n}, \boldsymbol{w}; \boldsymbol{\pi}) = \sum_{i=1}^K w_i^2 \frac{\bar{\sigma}_i^2}{n_i}$. Hence the optimal choice of $\boldsymbol{w}$, given $\boldsymbol{n}, \boldsymbol{t}, \bar{\boldsymbol{\sigma}}$ minimizes $\sum_{i=1}^K w_i^2 \frac{\bar{\sigma}_i^2}{n_i}$ subject to the constraints that $\boldsymbol{w} \geq 0$ and $\sum_{i=1}^K w_i = 1$.

By the Cauchy-Schwarz inequality,

$$1 = \left(\sum_{i=1}^K w_i\right)^2 = \left(\sum_{i=1}^K \frac{\bar{\sigma}_i}{\sqrt{n_i}} w_i \cdot \frac{\sqrt{n_i}}{\bar{\sigma}_i}\right)^2 \leq \left(\sum_{i=1}^K \frac{\bar{\sigma}_i^2}{n_i} w_i^2\right)\left(\sum_{i=1}^K \frac{n_i}{\bar{\sigma}_i^2}\right)$$

$$\implies \sum_{i=1}^K \frac{\bar{\sigma}_i^2}{n_i} w_i^2 \geq \frac{1}{\sum_{i=1}^K \frac{n_i}{\bar{\sigma}_i^2}}.$$

The $\boldsymbol{w}^*$ with $w_i^* = \frac{n_i/\bar{\sigma}_i{}^2}{\sum_{i\in[K]} n_i/\bar{\sigma}_i{}^2}$ satisfies this lower bound with equality and hence is optimal. ■

Each seller's utility, Bayesian incentive compatibility (IC) and individual rationality (IR) have the same form as in a standard procurement auction. The utility of seller i with cost c_i when reported costs are $\tilde{\boldsymbol{c}}$ and purchasing sample size and payment rules are $\boldsymbol{n}$ and $\boldsymbol{t}$ is $U_i(\tilde{\boldsymbol{c}}; \boldsymbol{n}, \boldsymbol{t}; c_i) = t_i(\tilde{\boldsymbol{c}}) - c_i \boldsymbol{n}_i(\tilde{\boldsymbol{c}})$. The Bayesian IC and IR constraints are in terms of the expected utility $\bar{U}_i(c_i, c_i', \boldsymbol{n}, \boldsymbol{t}) =$

$\mathbb{E}_{\boldsymbol{c}_{-i}}[U_i(c_i, c_i^{'}, \boldsymbol{c}_{-i}; \boldsymbol{n}((c_i^{'}, \boldsymbol{c}_{-i})), \boldsymbol{t}((c_i^{'}, \boldsymbol{c}_{-i})))]$, which is the expected utility when seller i's realized cost is c_i, taking an expectation over other sellers' costs, when seller i reports cost $c_i^{'}$ and all other sellers truthfully report costs. We call this the *cost-interim utility* of the seller. We can write the cost-interim utility as $\bar{t}(c_i^{'}, c_i) - c_i\bar{n}(c_i^{'}, c_i)$ where $\bar{n}_i(c_i^{'}) = \mathbb{E}_{\boldsymbol{c}_{-i}}[n_i(c_i^{'}, \boldsymbol{c}_{-i})]$ and $\bar{t}_i(c_i^{'}) = \mathbb{E}_{\boldsymbol{c}_{-i}}[t_i(c_i^{'}, \boldsymbol{c}_{-i})]$ are the expected number of samples and payment received respectively by seller i when reporting cost $c_i^{'}$. We call $\bar{n}_i(c_i^{'})$ the *cost-interim sample-purchasing rule* and $\bar{t}_i(c_i^{'})$ the *cost-interim payment rule*.

By Myerson's lemma [53], a payment and sample purchasing rule is Bayesian Incentive Compatible (BIC) if and only if the following properties are satisfied:

1. Monotonicity: The cost-interim sample purchasing rule is non-increasing in ${c_i}'$. That is, more expected samples must not be purchased if a seller reports a higher cost.

2. Myerson payment: The cost-interim payment satisfies

$$\bar{t}_i(c_i^{'}) = c_i^{'}\bar{n}_i(c_i^{'}) + U_i(c_{\text{max}}) + \int_{c_i^{'}}^{c_{\text{max}}} \bar{n}_i(t)dt.$$

   The cost-interim IR constraint can be written as $U_i(c_{\text{max}}) \geq 0$. To design the BIC mechanism with the least payment made, we can set $U_i(c_{\text{max}}) = 0$.

As a consequence of the Myerson payment rule, we can write the expectation over each seller's cost c_i of the cost-interim payment in the following way. Note that this expectation is the interim payment—which is the expected payment the buyer provides to a seller, taking an expectation over the realization of all sellers' costs. We have:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{c}}[t_i(\boldsymbol{c})] &= \mathbb{E}_{c_i}[\bar{t}_i(c_i)] \\
&= \int_{c_{\text{min}}}^{c_{\text{max}}} c\bar{n}_i(c)f(c)dc + \int_{c_{\text{min}}}^{c_{\text{max}}} \left(\int_{c}^{c_{\text{max}}} \bar{n}_i(t)dt\right) f(c)dc \\
&= \int_{c_{\text{min}}}^{c_{\text{max}}} c\bar{n}_i(c)f(c)dc + \int_{c_{\text{min}}}^{c_{\text{max}}} \bar{n}_i(t) \left(\int_{c_{\text{min}}}^{c} f(c)dc\right) dt \\
&= \int_{c_{\text{min}}}^{c_{\text{max}}} \bar{n}_i(c) \left(c + \frac{F(c)}{f(c)}\right) f(c)dc \\
&= \mathbb{E}_{c_i}[\bar{n}_i(c_i)\psi(c_i)].
\end{aligned}$$

Hence, the expected payment made in the optimal BIC mechanism is the product of the expected cost-interim purchased sample size and the virtual cost.

With the Myerson payment, we can write the expected payment in the objective of the buyer's mechanism design problem $\mathbb{E}_{\boldsymbol{c}}[\lambda \sum_{i=1}^{K} t_i]$ as $\mathbb{E}_{\boldsymbol{c}}[\lambda \sum_{i=1}^{K} \bar{n}_i(c_i)\psi(c_i)]$, which is equal to $\mathbb{E}_{\boldsymbol{c}}[\lambda \sum_{i=1}^{K} n_i(\boldsymbol{c})\psi(c_i)]$. We can then write the buyer's optimal BIC mechanism problem as the following optimization problem over sample-purchasing rules subject to the monotonicity constraint (5):

$$\min_{\boldsymbol{n}} \quad \mathbb{E}_{\boldsymbol{c}}\left[\frac{1}{\sum_{i=1}^{K} n_i(\boldsymbol{c})/\bar{\sigma}_i^2} + \lambda \sum_{i=1}^{K} n_i(\boldsymbol{c})\psi(c_i)\right] \tag{4}$$

$$\text{subject to} \quad \mathbb{E}_{\boldsymbol{c}_{-i}}[\boldsymbol{n}_i(\boldsymbol{c}_{-i}, c_i)] \geq \mathbb{E}_{\boldsymbol{c}_{-i}}[\boldsymbol{n}_i(\boldsymbol{c}_{-i}, c_i')] \quad \text{for all } i \in [K],\ c_i \leq c_i' \tag{5}$$

Let us first minimize the objective (4), dropping the monotonicity constraint in (5). We will later show that the solution satisfies the monotonicity constraint under Assumption 1. We minimize the objective in (4) by minimizing it for every realized $\boldsymbol{c}$. Let $\sum_{i=1}^{K} n_i \psi(c_i) = B$. Then,

$$\sum_{i=1}^{K} n_i/\bar{\sigma}_i^2 = \sum_{i=1}^{K} \frac{1}{\bar{\sigma}_i^2 \psi(c_i)} n_i \psi(c_i) \leq \max_{i \in [K]} \frac{B}{\bar{\sigma}_i^2 \psi(c_i)}.$$

For any fixed value of payments, the smallest value of the variance is achieved by only purchasing samples from sellers with the least $\bar{\sigma}_i^2 \psi(c_i)$. Let i^* be such a seller. So the optimization problem becomes optimizing over the value of payments B:

$$\min_{B \geq 0} \frac{\bar{\sigma}_{i^*}^2 \psi(c_{i^*})}{B} + \lambda B.$$

The optimal value is $B = \bar{\sigma}_{i^*} \sqrt{\psi(c_{i^*})/\lambda}$. We consider the optimal mechanism that always chooses to purchase from one seller, breaking any ties by selecting uniformly at random. The number of samples purchased from the selected seller i^* is then n_{i^*}, chosen so that $n_{i^*} \psi(c_{i^*}) = \bar{\sigma}_{i^*} \sqrt{\psi(c_{i^*})/\lambda}$. In other words, $n_{i^*} = \bar{\sigma}_i / \sqrt{\lambda \psi(c_{i^*})}$.

Now let us show that the minimizer without constraint (5) satisfies the constraint nonetheless. We argue that the optimal sample purchasing rule is monotone in the reported cost as long as Assumption 1 holds (virtual costs are non-decreasing functions of costs). This implies the monotonicity constraint (5) is satisfied. Reporting a higher cost cannot increase chance of selection since the seller with the least $\bar{\sigma}_i^2 \psi(c_i)$ is selected. Additionally, the number of samples purchased from the selected seller is a non-increasing function of the virtual cost of the reported cost, and hence of the reported cost as well.

To go from the solution of the relaxed mechanism design problem to the actual mechanism design problem where the number of samples purchased is an integer, we round the allocation obtained in the relaxed problem down to the nearest integer.

The rounded-down sample purchase scheme remains monotone since it is obtained by applying the monotone floor function to the monotone relaxed allocation rule. Combined with the same Myerson payment rule, the resulting mechanism remains BIC. Due to the rounding, the resulting sample purchasing scheme might not be optimal, but we now show that it is approximately optimal.

Let $n^* = \bar{\sigma}_{i^*} / \sqrt{\lambda \psi(c_{i^*})}$ be the random variable indicating the number of samples purchased under the relaxation. Recall that the buyer utility is given by

$$-\frac{\bar{\sigma}_{i^*}^2}{n^*} - \lambda n^* \psi(c_{i^*}).$$

Note that rounding down the number of samples n^* increases the second term (as its absolute value decreases). Thus, the only source of suboptimality is that, due to the rounding, the buyer's variance increases. The rounding decreases the number of purchased samples by at most one. Therefore, the buyer's utility can decrease by at most

$$\bar{\sigma}_{i^*}^2 \left(\frac{1}{n^* - 1} - \frac{1}{n^*} \right) = \frac{\bar{\sigma}_{i^*}^2}{n^*(n^* - 1)}.$$

This is a multiplicative factor of $1/(n^* - 1)$ compared to the non-rounded utility.

Appendix D. Proof of Theorem 2

We claim that the strategy profile in which no seller shares any free samples in the first stage, and the buyer then runs the mechanism in Theorem 1, constitutes an approximate subgame perfect equilibrium. In the proof, we focus primarily on sellers' incentives in sharing free samples, and at the end choose parameters so that the mechanism in Theorem 1 achieves the desired approximation factor ζ.

For the sellers, we first derive a lower bound on the utility before deviation and then construct a distribution over seller variances that results in a strictly lower utility after unilateral deviation.

Lower bound on utility before deviation. This lower bound holds for more general symmetric data-sharing strategy profiles where every seller shares the same number of samples, regardless of the number of samples shared and shows a dependence of $\Omega(1/K^2)$ on the number of sellers K. In the special case of no samples shared, the lower bound scales with σ_0, where $\sigma_0 = \sqrt{\mu\sigma_L^2 + (1-\mu)\sigma_H^2}$ is the standard deviation according to the prior.

Lemma 5 (Utility in symmetric data-sharing strategy profiles) *Suppose every seller shares the same number of samples $m \in \{0, \ldots, M\}$. Under the assumption that the pdf of the cost generating distribution f satisfies $\ell \leq f(c) \leq L$ for every $c \in [c_{min}, c_{max}]$, the probability of the buyer buying samples from each seller i is $1/K$. The expected utility of each seller is at least $\bar{U}_{lb} := \frac{1}{K(K+1)L^2}$. Furthermore, when the number of samples shared $m = 0$, the expected utility of each seller is at least $\bar{U}_{lb}(0) := \sqrt{\frac{\mu\sigma_L^2+(1-\mu)\sigma_H^2}{\lambda\psi(c_{max})}}\bar{U}_{lb}$.*

Proof [Proof of Theorem 5] By symmetry, each seller has the same probability of winning. So the probability of the buyer buying samples from each seller i is $1/K$.

From the Myerson payment from Theorem 1, the expected utility of each seller is:

$$\mathbb{E}_{\boldsymbol{c}}[U_i(\boldsymbol{c})] = \mathbb{E}_{\boldsymbol{c}}[n_i(\boldsymbol{c})(\psi(c_i) - c_i)] = \mathbb{E}_{\boldsymbol{c}}[n_i(\boldsymbol{c})F(c_i)/f(c_i)] \geq \mathbb{E}_{\boldsymbol{c}}[n_i(\boldsymbol{c})(c_i - c_{\min})\ell/L]$$

Let N_0 be a constant that lower bounds the number of samples purchased from a seller if selected. Using the inequality $c_i \geq \min(c_1, \ldots, c_K)$,

$$\mathbb{E}_{\boldsymbol{c}}[U_i(\boldsymbol{c})] \geq \frac{\ell N_0}{L}\mathbb{E}_{\boldsymbol{c}}[\mathbb{1}\{\text{buyer buys from } i\}(\min(c_1, \ldots, c_K) - c_{\min})] \tag{6}$$

From Theorem 1, we know that N_0 is at least $\bar{\sigma}/\sqrt{\lambda\psi(c_{\max})}$, where $\bar{\sigma}$ is the mean standard deviation according to the posterior. We know in general that $\bar{\sigma} \geq \sigma_L$. For this lower bound on $\bar{\sigma}$, our assumption that $\sigma_L \geq \sqrt{\lambda\psi(c_{\max}))}$ implies that $N_0 \geq 1$. However, when no samples are shared, $\bar{\sigma}$ is always $\sqrt{\mu\sigma_L^2 + (1-\mu)\sigma_H^2} > \sigma_L$. So for $m \geq 1$, we use $N_0 = 1$ and for $m = 0$, we use $N_0 = \sqrt{\frac{\mu\sigma_L^2+(1-\mu)\sigma_H^2}{\lambda\psi(c_{\min})}}$.

Consider the lower-bound quantity from Equation (6). If we sum up the lower bound over all the sellers, we get the value $(\ell N_0/L)\mathbb{E}_{\boldsymbol{c}}[\min(c_1, \ldots, c_K)) - c_{\max}]$. This is using the fact that summing $\mathbb{E}_{\boldsymbol{c}}[\mathbb{1}\{\text{buyer buys from } i\}]$ over all $i \in [K]$ is 1. Additionally, this quantity is equal for every seller by symmetry. Therefore, we can write

$$\mathbb{E}_{\mathbf{c}}[\mathbb{1}\{\text{buyer buys from } i\}(\min(c_1, \ldots, c_K)) - c_{\min}] = \frac{1}{K}\mathbb{E}_{\boldsymbol{c}}[(\min(c_1, \ldots, c_K)) - c_{\min}]$$

Finally, we bound $\mathbb{E}_{\boldsymbol{c}}[(\min(c_1,\ldots,c_K)) - c_{\text{min}}]$ to obtain the lower bound on the expected utility. We again use the lower and upper bounds on cost densities.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{c}}[(\min(c_1,\ldots,c_K)) - c_{\text{min}}] &= \int_{c_{\text{min}}}^{c_{\text{max}}} \Pr_{\boldsymbol{c}}[\min(c_1,\ldots,c_K) \geq x]dx. \\
&= \int_{c_{\text{min}}}^{c_{\text{max}}} (1-F(x))^K dx. \\
&\geq \int_{c_{\text{min}}}^{c_{\text{max}}} \frac{1}{L}(1-F(x))^K f(x)dx. \\
&= \frac{1}{L(K+1)}.
\end{aligned}$$

Setting $N_0 = 1$ for $m \geq 1$ and $N_0 = \sqrt{\frac{\mu\sigma_L^2 + (1-\mu)\sigma_H^2}{\lambda\psi(c_{\text{max}})}}$ for $m = 0$, we get the lower bound on the expected utility in the lemma. ■

Upper bound on utility after deviation. We show that after deviating, the buyer's posterior belief that the seller's variance is σ_H^2 shifts: it decreases from $1-\mu$ to below $1-\mu-\Delta_L$ when σ_L^2 is realized, and increases to above $1-\mu+\Delta_H$ when σ_H^2 is realized. Here $\Delta_L, \Delta_H > 0$ are constants to be chosen later. We show that this shift occurs with high probability over the samples, and that this probability approaches one as $\sigma_H/\sigma_L \to \infty$.

This belief shift reduces the deviating seller's utility for both realized variance types, though for different reasons. When σ_L^2 is realized, the shift toward σ_L^2 reduces the utility upon selection (payment minus cost). When σ_H^2 is realized, the shift toward σ_H^2 reduces the probability of being selected.

Lemma 6 (Belief shift from sharing m samples) *For any $\Delta_L \in (0, 1-\mu)$, $\Delta_H \in (0, \mu)$, and $\delta \in (0,1)$, there exists $R_0 > 0$ depending on $\mu, \Delta_L, \Delta_H, \delta$, such that whenever $\sigma_H/\sigma_L > R_0$ and any $m \geq 2$ free samples are shared, the posterior $\pi_H := \Pr(\sigma^2 = \sigma_H^2 \mid S^2)$ satisfies:*

(i) $\Pr\big(\pi_H \leq 1-\mu-\Delta_L \;\big|\; \sigma^2 = \sigma_L^2\big) \geq 1-\delta$, *and*

(ii) $\Pr\big(\pi_H \geq 1-\mu+\Delta_H \;\big|\; \sigma^2 = \sigma_H^2\big) \geq 1-\delta$.

Proof [Proof of Theorem 6] Let S^2 be the sample variance of the free samples. The Bayesian belief update based on S^2 results in a posterior belief with probability of variance is σ_H^2 given by

$$\pi_H = \Pr(\sigma^2 = \sigma_H^2 | S^2) = \frac{1}{1 + \frac{\mu}{1-\mu}\Lambda_m(S^2)},$$

where $\Lambda_m(S^2) = f_{S^2|\sigma_L}(S^2)/f_{S^2|\sigma_H}(S^2)$ is the likelihood ratio based of the sample variance of m samples which can be written as

$$\Lambda_m = \left(\frac{\sigma_H^2}{\sigma_L^2}\right)^{(m-1)/2} \exp\left(-\frac{\sigma_H^2 - \sigma_L^2}{2\sigma_H^2\sigma_L^2}(m-1)S^2\right).$$

To obtain $\pi_H < 1-\mu-\Delta$, the likelihood ratio must satisfy

$$\Lambda_m > T_+ := \frac{(1-\mu)(\mu+\Delta)}{\mu(1-\mu-\Delta)}.$$

Under σ_L^2, let $Z = (m-1)S^2/\sigma_L^2 \sim \chi^2_{m-1}$. Then

$$\log \Lambda_m = (m-1)\log\frac{\sigma_H}{\sigma_L} - \frac{\sigma_H^2-\sigma_L^2}{2\sigma_H^2}\,Z.$$

The condition $\Lambda_m > T_+$ is equivalent to $Z < z_+$, where

$$z_+ := \frac{2\sigma_H^2}{\sigma_H^2-\sigma_L^2}\left((m-1)\log\frac{\sigma_H}{\sigma_L} - \log T_+\right).$$

Note that z_+ is well defined when $T_+ > 0$, which is true when $\Delta \in (0, 1-\mu)$. Additionally, z_+ is positive when σ_H/σ_L is chosen to be larger than T_+. In this case, as $\sigma_H/\sigma_L \to \infty$, $z_+ \to \infty$. As a result, $\Pr(Z \le z_+) \to 1$. Therefore, there is a choice of σ_H/σ_L large enough that $\Pr(Z \le z_+) \ge 1-\delta$.

To obtain $\pi_H > 1-\mu+\Delta$, the likelihood ratio must satisfy

$$\Lambda_m < T_- := \frac{(1-\mu)(\mu-\Delta)}{\mu(1-\mu+\Delta)}.$$

Under σ_H^2, let $Z' = (m-1)S^2/\sigma_H^2 \sim \chi^2_{m-1}$. Then

$$\log \Lambda_m = (m-1)\log\frac{\sigma_H}{\sigma_L} - \frac{\sigma_H^2-\sigma_L^2}{2\sigma_L^2}\,Z'.$$

The condition $\Lambda_m < T_-$ is equivalent to $Z' > z_-$, where

$$z_- := \frac{2\sigma_L^2}{\sigma_H^2-\sigma_L^2}\left((m-1)\log\frac{\sigma_H}{\sigma_L} - \log T_-\right).$$

Note that z_- is well defined when $T_- > 0$, which holds when $\Delta \in (0, \mu)$. As $\sigma_H/\sigma_L \to \infty$, so $z_- \to 0$. As a result, $\Pr(Z' \ge z_-) \to 1$. Therefore, there is a choice of σ_H/σ_L large enough that $\Pr(Z' \ge z_-) \ge 1-\delta$.

Setting R_0 large enough that both tail probabilities are at most δ completes the proof. ■

Upper bound on utility for low variance after deviation. Fix $\delta \in (0,1)$, $\Delta_L \in (0, 1-\mu)$, and $\Delta_H \in (0,\mu)$. We will state the specific values of these parameters later. Suppose σ_H/σ_L is large enough that, by Theorem 6, the posterior belief of σ_H^2 is at most $1-\mu-\Delta_L$ when σ_L^2 is realized and at least $1-\mu+\Delta_H$ when σ_H^2 is realized, each with probability at least $1-\delta$.

In the complementary event (probability at most δ), we use the trivial upper bounds: selection probability 1 and maximum payment $\sigma_H\sqrt{\psi(c_{\max})/\lambda}$.

When σ_L^2 is realized and the belief shift holds, we bound the selection probability by 1. The payment upon selection is at most $\sqrt{\mathbb{E}[\sigma^2 \mid \sigma_L^2]\,\psi(c_{\max})/\lambda}$, which is at most $\sqrt{((\mu+\Delta_L)\sigma_L^2 + (1-\mu-\Delta_L)\sigma_H^2)\,\psi(c_{\max})/\lambda}$.

When σ_H^2 is realized and the belief shift holds, we bound the utility upon selection by $\sigma_H \sqrt{\psi(c_{\max})/\lambda}$. It remains to bound the selection probability. The expected posterior variance is at least ${\sigma'}^2(\Delta_H) := (\mu - \Delta_H)\sigma_L^2 + (1-\mu+\Delta_H)\sigma_H^2$, which is strictly greater than $\sigma_0^2 = \mu\sigma_L^2 + (1-\mu)\sigma_H^2$. Whenever ${\sigma'}^2(\Delta_H)\,\psi(c_{\min}) > \sigma_0^2\,\psi(c_{\max})$, the deviating seller is never selected when σ_H^2 is realized.

This condition is equivalent to ${\sigma'}^2(\Delta_H)/\sigma_0^2 > \rho$, where $\rho := \psi(c_{\max})/\psi(c_{\min})$. Rearranging:

$$\frac{\sigma_0^2 + \Delta_H(\sigma_H^2 - \sigma_L^2)}{\sigma_0^2} > \rho \quad \iff \quad \frac{\sigma_H^2}{\sigma_L^2} > \frac{1+\mu(\rho-1)}{\Delta_H - (1-\mu)(\rho-1)}.$$

Setting $\Delta_H = \mu/2$ and choosing $\mu > \frac{2\rho}{1+2\rho}$ ensures the denominator is positive. Then for σ_H^2/σ_L^2 sufficiently large, the deviating seller is never selected when σ_H^2 is realized.

Putting the pieces together, the expected utility after deviation is at most

$$\mathbb{E}_{\boldsymbol{c}}[U_i(\boldsymbol{c})] \le \left(\mu\sqrt{(\mu+\Delta_L)\sigma_L^2 + (1-\mu-\Delta_L)\sigma_H^2} + \delta\,\sigma_H\right)\sqrt{\frac{\psi(c_{\max})}{\lambda}}.$$

We have fixed $\mu > \frac{2\rho}{1+2\rho}$ and $\Delta_H = \mu/2$. It remains to choose Δ_L and δ so that the upper bound above is strictly less than $\bar{U}_{\mathrm{lb}}(0)$. This requires

$$\mu\sqrt{(\mu+\Delta_L)\sigma_L^2 + (1-\mu-\Delta_L)\sigma_H^2} + \delta\,\sigma_H \le \sqrt{\frac{(\mu\sigma_L^2 + (1-\mu)\sigma_H^2)\,\ell\,(c_{\max}-c_{\min})}{K(K+1)L^2}}.$$

Setting $\delta = 1-\mu-\Delta_L$, this reduces to showing

$$\frac{(\mu+\Delta_L)\sigma_L^2 + (1-\mu-\Delta_L)\sigma_H^2}{\mu\sigma_L^2 + (1-\mu)\sigma_H^2} \le \xi,$$

where $\xi > 0$ is a constant depending on $c_{\min}, c_{\max}, \ell, L, K$. As $\Delta_L \to 1-\mu$ and $\sigma_H^2/\sigma_L^2 \to \infty$, the left-hand side tends to zero. Hence there exists $\epsilon > 0$ and $R_1 > 0$ such that setting $\Delta_L = 1-\mu-\epsilon$ and $\delta = \epsilon$ satisfies the inequality whenever $\sigma_H/\sigma_L > R_1$.

Finally, we choose σ_L large enough to satisfy (3) with α such that $1/(\alpha-1) < \zeta$, and also choose $\sigma_H > \sigma_L \cdot \max(R_0, R_1)$, where R_0 is from Theorem 6 to ensure belief shift holds with the chosen $\Delta_H, \Delta_L, \delta$. This ensures all conditions hold simultaneously, completing the proof.

Appendix E. Proof of Theorem 3

The key driver of the incentive to sharing the maximum number of samples is that it allows the mean variance under the posterior to be low enough which makes the seller more prone to being selected. There is a fundamental limit to how low the posterior mean can be if a seller shares fewer than M samples. This is given in the following lemma which establishes a lower bound on posterior mean as a function of number of free samples shared (m).

Lemma 7 (Lower bound on mean of posterior) *For any free sample set S_m^f of size $m \ge 2$, the mean of the posterior over variance (based on the sample variance S^2 of the free samples) is at least*

$$\sigma_{lb}^2(m) := \sigma_L^2 + \frac{\sigma_H^2 - \sigma_L^2}{1 + \frac{\mu}{1-\mu}\left(\frac{\sigma_H}{\sigma_L}\right)^{m-1}}.$$

Moreover for any $\sigma' > \sigma_{lb}$, there is a positive probability that the mean of the posterior is $\ge \sigma'$.

Proof [Proof of Theorem 7] The mean of the posterior distribution over variance can be written as

$$\begin{aligned}\mathbb{E}[\sigma^2|S^2] &= \sigma_L^2(1-\pi_H) + \sigma_H^2\pi_H \\ &= \sigma_L^2 + \pi_H(\sigma_H^2 - \sigma_L^2),\end{aligned}$$

where π_H is the probability of variance σ_H^2 under the posterior based on the sample variance S^2 of $S_m^f = \{X_1, \ldots, X_m\}$. By Bayes' theorem,

$$\pi_H = \Pr(\sigma^2 = \sigma_H^2|S^2) = \frac{1}{1 + \frac{\mu}{1-\mu}\Lambda_m(S^2)},$$

where $\Lambda_m(S^2) = f_{S^2|\sigma_L}(S^2)/f_{S^2|\sigma_H}(S^2)$ is the likelihood ratio based on the sample variance of m samples. Each likelihood function is the density of a scaled chi-squared distribution with $m-1$ degrees of freedom. Hence, the likelihood ratio can be written as:

$$\Lambda_m = \left(\frac{\sigma_H^2}{\sigma_L^2}\right)^{(m-1)/2} \exp\left(-\frac{\sigma_H^2 - \sigma_L^2}{2\sigma_H^2\sigma_L^2}(m-1)S^2\right).$$

Since $S^2 \geq 0$ and $\sigma_H^2 > \sigma_L^2$, the exponential term is non-positive, so

$$\Lambda_m \leq \left(\frac{\sigma_H^2}{\sigma_L^2}\right)^{(m-1)/2} = \left(\frac{\sigma_H}{\sigma_L}\right)^{m-1}.$$

This upper bound on Λ_m gives a lower bound on π_H:

$$\pi_H \geq \frac{1}{1 + \frac{\mu}{1-\mu}\left(\frac{\sigma_H}{\sigma_L}\right)^{m-1}},$$

which yields the lower bound $\sigma_{lb}^2(m)$ stated in the lemma.

For the moreover claim, fix any $\sigma_{lb} < \sigma' < \sigma_H$. There is a corresponding threshold $\Lambda' < (\sigma_H/\sigma_L)^{m-1}$ such that the posterior mean equals σ'^2 when $\Lambda_m = \Lambda'$. This corresponds to S^2 exceeding a strictly positive value, which has positive probability. ■

For any seller i consider the event

$$E_i^{\text{select}} = \{c_i \leq c_0, \bar{\sigma}_i \leq \sigma_0\},$$

where c_0, σ_0 are defined in the following way. Let $\sigma_0 = (\sigma_{lb}(M-1) + \sigma_{lb}(M))/2$, where σ_{lb} is the lower bound on posterior mean given by Theorem 7. And $c_0 \gneq c_{\min}$ satisfies the condition

$$\sigma_0\sqrt{\psi(c_0)} \leq \sigma_{lb}(M-1)\sqrt{\psi(c_{\min})}.$$

Such a c_0 exists since $\sigma_{lb}(M-1)$ is strictly bigger than σ_0 (given that $\sigma_{lb}(M-1) > \sigma_{lb}(M)$ as shown in Theorem 7) and since we assume the virtual cost function ψ to be continuous and increasing.

Lemma 8 *Conditioned on the event E^{select} for some seller i, no seller i' sharing $m < M$ free samples will be selected. For any seller i sharing M free samples, there is a positive probability of E_i^{select}.*

Proof [Proof of Theorem 8] For any seller i' sharing $m < M$ samples, the posterior mean satisfies $\bar{\sigma}_i \geq \sigma_{lb}(m)$ by Theorem 7. From that lemma, we also see that $\sigma_{lb}(m) \geq \sigma_{lb}(M-1) > \sigma_{lb}(M)$ since the lower bound established there is strictly decreasing as number of free samples increases. Hence $\bar{\sigma}_{i'}$ is strictly greater than σ_0 which is midpoint of $\sigma_{lb}(M-1)$ and $\sigma_{lb}(M)$. Additionally, $c_{i'} \geq c_{\text{min}}$. As a result, $\sigma_0\sqrt{\psi(c_0)} \leq \sigma_{lb}(M-1)\sqrt{\psi(c_{\text{min}})} < \bar{\sigma}_{i'}\sqrt{\psi(c_{i'})}$.

If there is a seller i with event E_i^{select} realized, then for this seller, $\bar{\sigma}_i\sqrt{\psi(c_i)} \leq \sigma_0\sqrt{\psi(c_0)}$. Comparing to seller i' sharing $m < M$ samples, $\bar{\sigma}_i\sqrt{\psi(c_i)} < \bar{\sigma}_{i'}\sqrt{\psi(c_{i'})}$ meaning that seller i' will not be selected.

Any seller i sharing M samples has a positive probability of having $\bar{\sigma}_i \leq \sigma_0$. This is due to σ_0 being strictly less than $\sigma_{lb}(M)$. Theorem 7 establishes a positive probability for this. Furthermore, there is a positive probability of $c_i \in [c_0/2, c_0]$ which is at least $\ell c_0/2(c_{\text{max}} - c_{\text{min}})$ from our assumption on the lower bound on densities of the cost distribution. Therefore, there is a positive probability of E_i^{select}. ■

A consequence of this lemma is that the probability of a seller sharing $m < M$ samples being selected decays exponentially in the number of sellers sharing M samples. This is because for the seller to be selected, the event E^{select} must not be realized for *any* of the M samples sharing sellers, and this event is independent for each of the sellers. This is stated in the following lemma.

Lemma 9 (Upper bound on expected utility when sharing $m < M$ samples) *Suppose J of the K sellers share M samples. Then the expected utility of a seller sharing $m < M$ samples is at most $(1-q)^J\sigma_H\sqrt{\psi(c_{\max})/\lambda}$, where for $q > 0$ is the probability of event E^{select} for a seller sharing M samples.*

Proof [Proof of Theorem 9] We show that the probability of a seller sharing $m < M$ samples being selected is at most $(1-q)^J$ and this results in the upper bound on utility assuming maximum utility upon selection. This upper bound on selection probability follows directly from Theorem 8. From Theorem 8, we know that if there is any seller i sharing M samples, there is a positive probability $\geq q$ of event E_i^{select} which results in the seller i' sharing $m < M$ samples not being selected. So for seller i' to be selected, each of the J sellers sharing M samples should not have event E^{select} realized. This event is independent across all sellers. So the probability of this event not occurring for any seller is $(1-q)^J$. ■

Next we show a lower bound on utility of a seller sharing M samples. This lemma shows that the expected utility of a seller sharing M samples is of the order $\Omega(1/J^2)$, where J is the number of sellers sharing M samples.

Lemma 10 (Lower bound on utility when sharing M samples) *Suppose J of the K sellers share M samples. Then the expected utility of a seller sharing M samples is at least*

$$\frac{(1-(1-q)^J)\ell}{L^2J^2}\frac{\frac{1}{J}-(1-\ell(c_0-c_{min}))^J}{(1-(1-\ell(c_0-c_{min}))^J)} \in \Omega(1/J^3),$$

where $q > 0$ is the probability of event E^{select} for a seller sharing M samples.

Proof [Proof of Theorem 10] Let $\mathcal{J}$ denote the index set of the J sellers sharing M samples. Consider the event $\bar{E}^{\text{select}} = \bigcup_{j\in\mathcal{J}} E_i^{\text{select}}$ which is the event that one of the J sellers sharing M

samples realizes event E^{select}. Whenever $\bar{E}^{\text{select}}$ occurs, one of the sellers in $\mathcal{J}$ gets selected by Theorem 8. The probability of $\bar{E}^{\text{select}}$ is at least $1-(1-q)^J$. We provide a lower bound on the expected utility by a lower bound on the expected utility conditioned on $\bar{E}^{\text{select}}$ times the probability of $\bar{E}^{\text{select}}$.

To bound the expected utility conditioned on $\bar{E}^{\text{select}}$, we perform an analysis similar to the analysis in Theorem 5 which expresses a lower bound on expected utility of a seller as in Equation (6). We extend this analysis to bound the conditional expected utility.

Note that conditioning on $\bar{E}^{\text{select}}$ retains the symmetry in distributions of sellers in $\mathcal{J}$. Note that the event $\bar{E}^{\text{select}}$ is symmetric for $\mathcal{J}$. That is, if a tuple $((c_j,\bar{\sigma}_j))_{j\in\mathcal{J}} \in \bar{E}^{\text{select}}$, then $((c_{\beta(j)},\bar{\sigma}_{\beta(j)})_{j\in\mathcal{J}} \in \bar{E}^{\text{select}}$ for every permutation β of $\mathcal{J}$. As a result the distribution of each $(c_j,\bar{\sigma}_j)$ conditioned on event $\bar{E}^{\text{select}}$ is the same for every $j \in \mathcal{J}$.

As we did in Theorem 5, using the payment derived in Theorem 1, we can write a lower bound on expected utility of a seller in J conditioned on the event $\bar{E}^{\text{select}}$ as

$$\mathbb{E}_{\boldsymbol{c}}[U_j(\boldsymbol{c})|\bar{E}^{\text{select}}] \geq \frac{\ell}{L}\mathbb{E}_{\boldsymbol{c}}[\mathbb{1}\{\text{buyer buys from } j\}(\min_{j\in\mathcal{J}} c_j - c_{\min})|\bar{E}^{\text{select}}]. \tag{7}$$

Consider the lower-bound quantity from Equation (7). If we sum up the lower bound over all the sellers, we get the value $\ell/L\mathbb{E}_{\boldsymbol{c}}[\min_{j\in\mathcal{J}} c_j - c_{\min}|\bar{E}^{\text{select}}]$. This uses the fact that summing $\mathbb{E}_{\boldsymbol{c}}[\mathbb{1}\{\text{buyer buys from } j|\bar{E}^{\text{select}}\}]$ over all $j \in \mathcal{J}$ is one. Additionally, this quantity is equal for every seller by symmetry since the conditional distribution of every seller's parameter, conditioned on $\bar{E}^{\text{select}}$ is the same. Therefore, we can write

$$\mathbb{E}_{\boldsymbol{c}}[U_j(\boldsymbol{c})|\bar{E}^{\text{select}}] \geq \frac{\ell}{LJ}\mathbb{E}_{\boldsymbol{c}}[(\min_{j\in\mathcal{J}} c_j - c_{\min}|\bar{E}^{\text{select}}]$$

Finally, we bound $\mathbb{E}_{\boldsymbol{c}}[(\min_{j\in\mathcal{J}} c_j - c_{\min}|\bar{E}^{\text{select}}]$ to obtain the lower bound on the expected utility. We again use the lower and upper bounds on cost densities:

$$\mathbb{E}_{\boldsymbol{c}}[\min_{j\in\mathcal{J}} c_j - c_{\min}|\bar{E}^{\text{select}}] = \mathbb{E}_{\boldsymbol{c}}[\min_{j\in\mathcal{J}} c_j - c_{\min}|\exists j \in \mathcal{J} \text{ s.t. } c_j < c_0, \bar{\sigma}_j < \sigma_0].$$

By independence of costs and sample distributions and the fact that the cost and sample distributions conditioned on $\bar{E}^{\text{select}}$ is the same for every seller in $\mathcal{J}$, the probability of a seller having cost lower than c_0 having posterior mean lower than σ_0^2 is $1/|J|$. Consequently,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{c}}[\min_{j\in\mathcal{J}} c_j - c_{\min}|\exists j \in \mathcal{J} \text{ s.t. } c_j < c_0, \bar{\sigma}_j < \sigma_0] &\geq \frac{1}{J}\,\mathbb{E}_{\boldsymbol{c}}[\min_{j\in\mathcal{J}} c_j - c_{\min}|\exists j \in \mathcal{J} \text{ s.t. } c_j < c_0] \\
&= \frac{1}{J}\int_{c_{\min}}^{c_{\max}} \Pr[\min_{j\in\mathcal{J}} c_j \geq x|\min_{j\in\mathcal{J}} c_j \leq c_0]dx \\
&= \frac{1}{J}\int_{c_{\min}}^{c_0} \frac{\Pr[\min_{j\in\mathcal{J}} c_j \geq x] - \Pr[\min_{j\in\mathcal{J}} c_j \geq c_0]}{\Pr[\min_{j\in\mathcal{J}} c_j \leq c_0]}dx \\
&= \frac{1}{J}\int_{c_{\min}}^{c_0} \frac{(1-F(x))^J - (1-F(c_0))^J}{1-(1-F(c_0))^J}dx
\end{aligned}$$

Using, $f(x) \le L$ for all $x \in [c_{\text{min}}, c_{\text{max}}]$,

$$\mathbb{E}_{\boldsymbol{c}}[\min_{j \in \mathcal{J}} c_j - c_{\text{min}} | \exists j \in \mathcal{J} \text{ s.t. } c_j < c_0, \bar{\sigma_j} < \sigma_0] \ge \frac{1}{J} \frac{\frac{1}{J} - (1 - \ell(c_0 - c_{\text{min}}))^J}{L\left(1 - (1 - L(c_0 - c_{\text{min}}))^J\right)}.$$

Finally, combining this lower bound on expected utility conditioned on $\bar{E}^{\text{select}}$ and the probability of $(1-(1-q)^J)$ of event $\bar{E}^{\text{select}}$ concludes the proof.

■

Using these lemmas, we now establish that the unique approximate equilibrium is one in which all sellers share the maximum number of samples.

Deviation from maximally informative strategy profile. Consider a unilateral deviation from the profile with all sellers sharing M samples to sharing m samples. From Theorem 5, we know that before deviation, the utility of each seller is at least $\Omega(1/K^2)$. After deviation, from Theorem 9, we know the expected utility is at most $O((1-q)^{K-1})$ for a $q > 0$ corresponding to the probability of event E^{select} for a non-deviating seller. Since deviation results in expected utility going from decaying quadratically in number of sellers K to decaying exponentially, for sufficiently large K, the utility after deviation is lower than the utility before deviation, making the maximally informative sample sharing strategy an equilibrium.

Deviation from any other sample sharing profile. Consider any other strategy profile where at least one seller shares strictly fewer than M samples. Consider the seller i' sharing $m < M$ samples with the least probability of being selected. We show if i' unilaterally deviates to sharing M samples, the expected utility of i' strictly increases.

Let us consider two cases for J, which is the number of sellers sharing M samples. The first case is that $J \ge K^{1/4}$. In this case, by Theorem 9, seller i's expected utility before deviation is $O((1-q)^{K^{1/4}})$. After deviation, since i' becomes one among the $J+1$ sellers sharing M samples, the expected utility becomes at least $\Omega(1/K^3)$ (since $J \le K$). So far large enough K, deviation strictly increases utility.

The other case is $J \le K^{1/4}$. Since i' is the one among the $K - J$ sellers with the least probability of being selected, i' has a probability of being selected, and hence expected utility, at most $O(1/(K - \sqrt{K}))$. After deviation, the expected utility becomes $\Omega(1/K^{3/4})$. Again for large enough K, deviation strictly increases utility.

Choosing a large enough K, larger than what is required for both cases, concludes the proof.