The Value of Data

Simone Galperti
UC, San Diego

Aleksandr Levkun
UC, San Diego

Jacopo Perego
Columbia University

August 11, 2021

Preliminary and Incomplete

Abstract

Personal data is an essential input of many modern industries. Yet, its value is hard to establish and formal markets for data are still lacking. Consider a platform that mediates trade between a seller and a population of buyers using individual *data records* of their personal characteristics. After formulating this as an information-design problem, we use linear-programming duality to characterize the unit value that the platform derives from each buyer's specific record. We find that this value differs from the payoff that the platform directly earns from the trade between the seller and a buyer, which would be a biased measure of her record's value. This bias reflects unaccounted externalities between records, which arise because the platform pools records to withhold information from the seller. We then characterize the platform's willingness to pay for more records—e.g., more buyers joining the platform—and for better records—e.g., more information about existing buyers. Our analysis establishes essential properties of the "demand side" of data markets. Our methods apply generally to a large class of principal-agents problems.

JEL Classification Numbers: C72, D82, D83

Keywords: Data, Information, Duality

We are thankful to S. Nageeb Ali, Alessandro Bonatti, Laura Doval, Elliot Lipnowski, Xiaosheng Mu, Joel Sobel, Denis Shishkin, Glen Weyl and seminar participants at Bonn University, Cambridge University, Columbia University, Penn State University, Pittsburgh University & CMU, UC Davis, and UC San Diego for useful comments. All remaining errors are ours.

> The first step toward valuing individual contributions to the data economy is measuring these (marginal) contributions. (Posner and Weyl, 2018, p. 244)

1 Introduction

Data is the "new oil" in modern economies and a topic of major debates about privacy. Yet, data is often not traded in formal markets, nor are individuals compensated when their personal data is used. Data is usually collected for free or, at best, bartered in exchange for online services. Either case may result in significant inefficiencies and misallocation. Many scholars and policymakers view establishing functioning data markets as essential for the digital economy to bring prosperity and stability to societies (Lanier, 2013; Posner and Weyl, 2018; Arrieta-Ibarra et al., 2018). A key challenge is to determine the value of an individual data record (Acquisti et al., 2016; Posner and Weyl, 2018). Is one consumer's data record more valuable than another's for an e-commerce platform? How much should each be paid?

This paper breaks new ground in two ways. First, we focus on the value of an *individual* data record, as opposed to the value of the entire database.[1] Second, we study this value for *intermediation problems*: An intermediary uses its data to strategically direct interactions between agents with conflicting interests by providing them with information. Such problems are ubiquitous due to the rise of digital platforms and "info-mediaries" (Acquisti et al., 2016): Besides e-commerce, examples include matching markets (like ride-sharing and navigation services) and auction-based markets (like ad auctions or eBay). However, the value of data when it is used by an intermediary needs to be established in fundamentally different ways than when it is used by a standard decision-maker to identify his best choices. This is because the intermediary may exploit its informational advantage to pool data records together thus creating externalities between them.[2] Our solution involves modeling this intermediation as an information-design problem and leveraging its structure as a linear program.

Consider an example. An online platform mediates the interactions between a population of buyers and a monopolist, which produces a good at zero marginal cost. For each buyer, who is assigned a unique identifier, the platform owns data consisting of a *record* of personal

[1] This is a key difference from recent work on data markets, such as Bergemann and Bonatti (2015) and Bergemann et al. (2018).

[2] Importantly, these externalities arise even when data records are statistically independent. As such, they differ from other important externalities highlighted in the literature that depend on correlation between records (e.g., see Acemoglu et al., 2021; Bergemann et al., 2021).

characteristics. There are two types of records, denoted by ω_1 and ω_2, depending on what the platform knows about the buyer. Concretely, suppose each one of these record types reveals whether the buyer's valuation for the seller's good is 1 or 2, respectively. We refer to the collection of buyers' records as the platform's database. Suppose it contains 3 million records of type ω_1 and 6 million of type ω_2. The seller knows only the composition of the database, denoted by $q = (3M, 6M)$. For each interaction, the platform sends a signal about ω to the seller so as to influence the price he charges (as in Bergemann et al., 2015). For instance, it could divide the buyers into market segments using their record—like their gender or age—and then tell the seller to which segment each buyer belongs.

We ask two main questions. First, how much value does the platform derive from each buyer's record and why? This individual contribution, denoted by $v^*(\omega)$, can offer a benchmark for compensating each buyer for her specific data. Second, how much is the platform willing to pay for more data? In this regard, the colloquial expression "having more data" can have two meanings: having *more records* in the database and thus more interactions between the corresponding buyers and the seller to mediate (e.g., because new buyers join the platform); having *better records* by observing more informative characteristics about existing buyers (e.g., because they are more active on the platform). Our framework allows us to analyze both and determine the platform's willingness to pay for more and for better records. This can guide intermediaries when acquiring data, taking into account its specific realizations. Answering these questions is standard if the platform itself is the seller and thus maximizes its profit by choosing a price for each buyer knowing ω. In this case, it effectively solves a decision problem and $v^*(\omega)$ is the profit for each interaction. Things change significantly for intermediation problems due to the conflict of interests.

Suppose, for example, that the platform's objective is to maximize the consumer surplus of each buyer, which goes against the seller's profits. To do so, it assigns each buyer with ω_1 to a subprime segment, s'; it assigns each buyer with ω_2 to s' or to a prime segment, s'', with equal probability. When told that the buyer is in s', the seller is indifferent between a price of 1 or 2, but picks 1 in favor of the platform; for s'', the optimal price is 2. Thus, the platform earns a payoff of 0.5 on average for each type ω_2 record, and a payoff of zero for each record of type ω_1. Do these payoffs capture the actual value the platform derives from each buyer's record? The answer is no. Perhaps counterintuitively, the most valuable records are those of buyers whose interaction with the seller yields the lowest payoff for the platform. Indeed, note that interactions with ω_2 yield a positive surplus only when pooled with interactions with ω_1 through s'. The records with ω_1 "help" to persuade the seller to charge a low price to some

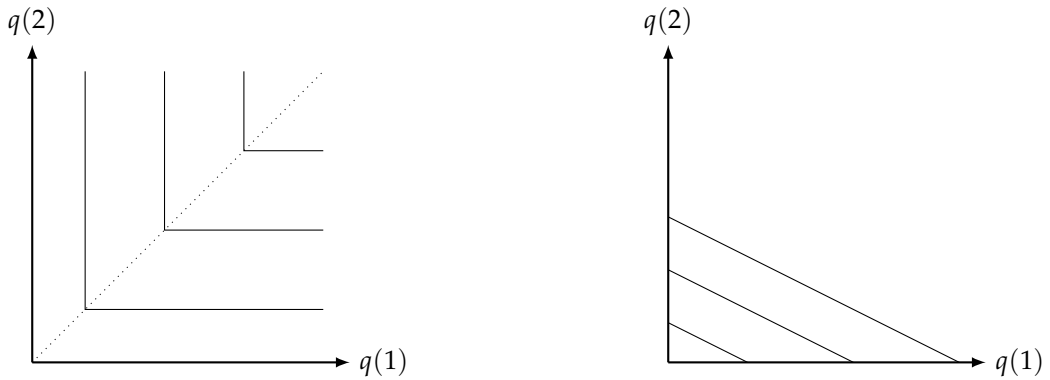
high-valuation buyer. Hence, they should not be worthless, even though each interaction with ω_1 by itself yields zero surplus. We will show that $v^*(\omega_1) = 1$, reflecting the fact that type-$\omega = 1$ records exert a positive information externality on interactions characterized by type-ω_2 records. By contrast, $v^*(\omega_2) = 0$ because interactions characterized by type-ω_2 records have to "repay" this externality to records of type ω_1.

This example illustrates three key takeaways. First, the payoff the platform *directly* obtains from a record provides a biased account of its value. This is in sharp contrast with decision problems, such as when the platform maximizes the seller's profits. The reason is that, due to conflicting interests, the platform pools buyers' records to produce partially informative signals, thereby using data of one interaction to influence the outcomes of other interactions. Second, the gap between the value and direct payoff of a buyer's record reflects information externalities with other records in the database. We characterize such externalities in terms of how a record helps (or not) the platform influence other interactions, and how the platform exploits the seller's incentives across interactions with its signals. Third, the optimal use of a buyer's record—hence, v^*—depends on the database composition q, as this determines the informational advantage and the platform. For instance, if in the example we swap the database composition and let $q = (6M, 3M)$, we get $v^*(\omega_1) = 0$ and $v^*(\omega_2) = 1$. We analyze how v^* varies with q and establish that, for example, the scarcer a type of record is, the higher its value.

We then use our framework to study the problem of acquiring more data. We can view acquiring data—whether to obtain more records or refine existing ones—as moving inside the space of possible databases and evaluate its effects using the platform's overall preference, which is pinned down by v^*.

For instance, imagine that our platform can expand its userbase and add *more* records to its database $q = (3M, 6M)$. How much should it be willing to pay for such new records? The answer has to rely on v^* (not direct payoffs), so it is zero for records with $\omega = 2$ and one for records with $\omega = 1$. If a new record's type is uncertain, we simply take the expectation of v^*. More generally, v^* characterizes intermediaries' marginal rate of substitution between types of records in the space of databases. Figure 1 illustrates our platform's indifference curves in this space. Interestingly, the same types of records are perfect complements when the platform maximizes the buyer's surplus (panel (a)), but perfect substitutes when it maximizes the seller's profit (panel (b)).[3] In fact, we find that indifference curves are convex if and only if an intermediation problem is non-trivial—i.e., its solution is *never* full disclosure, independently of q. This has immediate implications for which database an intermediary would choose given

[3]It is easy to see that for profit maximization $v^*(\omega_1) = 1$ and $v^*(\omega_2) = 2$ independently of q.



(a) Maximize Buyer's Surplus.

(b) Maximize Seller's Profit.

Figure 1: Iso-payoff curves in the platform's example.

a budget constraint and hence its demand functions for data.

Alternatively, the platform could obtain *better* records by refining the information it has about the corresponding buyers. For example, imagine some buyers have just joined the platform, which therefore knows their IDs but nothing about their valuation for the seller's good. In this case, how much is the platform willing to pay for refining their records? We establish that any i.i.d. refinement weakly increases the value of the records it refines. However, under some conditions, the platform's willingness to pay for such a refinement can be zero, even if it actively uses the new information it received. This is in sharp contrast with decision problems, where the gain from obtaining more information is positive if and only if it changes behavior of the decision maker. Similarly, we show that when the refinement is not i.i.d., the platform's willingness to pay for it could even turn negative.

Our analysis applies more generally to any setting where a principal mediates the interactions between multiple agents by providing them with information, which she produces using data she already owns.[4] We view each interaction's data as a physical input of this information production constrained by their quantity in the database. We use linear-programming duality to characterize the unit value v^* of these inputs, adapting classic work of Dorfman et al. (1987) and Gale (1989) to our class of problems. This approach allows us to handle the aforementioned complexities of assessing the value of data for intermediation problems, highlighting key differences from settings where data is used to solve decision problems.

Overall, this paper conducts a systematic analysis of the demand side of data markets. Getting the demand function for data right appears essential, for instance, when studying the wel-

[4] In a related project, we analyze the case where the principal has to first elicit the data from its sources.

fare effects of privacy policies, which have been hotly debated in recent years.[5] Our work also offers a benchmark to assess how privacy protection affects the value of data: Our principal's direct access to it is akin to no protection at all. A better understanding of the value of people's data may help improve the status quo of free data. Arrieta-Ibarra et al. (2018) list some of its pitfalls: jobs loss, wealth shift to the top, underproduction of the high-quality data that fuels AI productivity growth, and erosion of personal dignity. Lanier (2013) argues that compensating people for their data contributions to AI automation may be essential to save the middle class and ultimately democracy. Recently, "data unions" have emerged to intermediate individual sources of personal data and its user, which requires figuring out how much each should be paid.

1.1 Related Literature

Our work builds on the literature on information design (see, e.g., Bergemann and Morris, 2019, for a review). We formulate our mediator's "data-use" problem as an information-design problem and, as in Bergemann and Morris (2016), we use the revelation principle to express it as a linear program. Our analysis of the "data-value" problem—from which we derive the value of each data record—is shown to be equivalent to the linear-programming dual of the data-use problem.

Duality methods have been used to study information-design problems by Kolotilin (2018), Galperti and Perego (2018), Dworczak and Martini (2019), Dworczak and Kolotilin (2019), and Dizdar and Kováč (2020). Our work differs from these papers in two important ways. First, they exploit the dual as a tool to compute and study the solution to the original primal problem. We instead put the dual at the center of our analysis and use it to address an independent economic question—namely, what is the value of data?—which is of interest irrespective of the primal problem. Second, unlike these papers, we do not focus on single-receiver Bayesian persuasion problems or employ a "belief approach" (as in Kamenica and Gentzkow, 2011). Rather, we study general information-design problems with multiple agents interacting strategically through the notion of Bayes-correlated equilibrium. This connects our work to an earlier literature on the characterization of correlated equilibria in complete-information games, which includes Nau and McCardle (1990), Nau (1992), and Myerson (1997). Finally, duality methods have also been applied in the mechanism-design literature, at least since Myerson (1983) and Myerson (1984), and more recently to study and characterize informationally

[5] See, e.g., Scott-Morton et al. (2019); Crémer et al. (2019); Goldberg et al. (2021).

robust mechanisms (e.g., Du, 2018; Brooks and Du, 2020, 2021).

Our paper contributes to bridging the literature on information design and the growing literature on data markets (see Bergemann and Bonatti, 2019, for a review). Perhaps closest to our paper are Bergemann and Bonatti (2015) and Bergemann et al. (2018), who build on earlier contributions of Admati and Pfleiderer (1986, 1990) to study markets where a buyer purchases an information product and uses its signal realization to solve a decision problem. This involves two key differences from our approach. First, they focus on decision problems. By contrast, our potential buyer of data (i.e., the platform) uses it not to solve a decision problem, but to mediate strategic interactions. Second, they focus on information products (i.e. statistical experiments) and assign ex ante values to them, before any signal is realized. By contrast, we focus on data records, which on top of including information about a buyer also allow the platform to uniquely identify her and, thus, intermediate her interaction with the seller. Moreover, we assign an ex post value to such records. At an intersection of these questions, Frankel and Kamenica (2019) study the ex post value of information products in decision problems.

Finally, data markets also raise important questions about privacy (see Acquisti et al., 2016, for a review). Acemoglu et al. (2021) and Bergemann et al. (2021) examine the externalities and market distortions that occur between agents who supply to a common intermediary private data which is correlated between them. These externalities are conceptually distinct from the ones we highlight: In their context the intermediary uses the data of one agent to learn about another agent, in our context she uses the data of one agent to conceal the data of another agent. Calzolari and Pavan (2006) analyze information externalities between sequential interactions. We do not consider such interactions. Ali et al. (2020) examine when giving consumers control over their private data can help them benefit from personalized pricing. Our results provide a benchmark to understand how different privacy regulations could affect the value of consumers' data for its users. We explore this question in a related project.

2 Model

For ease of exposition, we present the model and analysis in the context where an e-commerce platform mediates the interactions of a group of sellers with a population of buyers, similarly to the introduction example.[6] Nonetheless, we keep our description fairly general, as the details are irrelevant and may distract from our main points. In fact, our approach and results apply

[6]Besides Bergemann et al. (2015), see also Elliott et al. (2020) who study a platform that tries to influence which prices competing sellers offer to their customers.

much more broadly. We discuss this and other aspects of the model in Section 5.

Let $i = 0$ denote the platform, which plays the role of the principal (it). Let $I = \{1, \dots, n\}$ be a set of sellers, who play the role of the agents (he). Let A_i be the finite set of seller i's actions. We can interpret a_i as seller i's choosing his product's price, features, or quality. The platform is used by a continuum of buyers (she), each interested in buying a product from the sellers. Each buyer's preference over the sellers' products is pinned down by a random variable θ, which is independently and identically distributed across buyers over a finite set Θ.

The platform has exclusive access to some data about each buyer. We think of this data as a concrete *record* of personal characteristics that is informative about θ—perhaps only partially. We assume throughout that a buyer's record provides information only about her θ, but not about any other buyer's θ.[7] There are different *types* of records—denoted by ω in some finite set Ω—depending on what the platform knows about the buyer. Let ω° denote the records whose data is fully uninformative about θ. Only the platform observes ω, which gives it an informational advantage over the sellers. Let $q \in \mathbb{R}^{\Omega}_{+}$ denote the platform's collection of buyers' records, where $q(\omega)$ are of type ω. We refer to q as the platform's *database*.

For each interaction between a buyer and the sellers, we leave her purchase decision given their actions implicit and embed it in the payoff functions of the sellers and the platform. For every ω and action profile $a = (a_1 \dots, a_n)$, let $u_i(a, \omega)$ be i's expected payoff conditional on the buyer's record. Let $\Gamma_\omega = \{I, (A_i, u_i(\cdot, \omega))_{i=0}^n\}$, which defines a complete-information game between the sellers. Thus, we may also refer to Γ_ω as a buyer-sellers interaction of type ω. The primitives $\Gamma = \{\Gamma_\omega\}_{\omega \in \Omega}$ and q are common knowledge.

Using its data, the platform mediates each interaction by privately conveying some information about its type to each seller so as to influence their actions. The sellers combine this information with Γ and q to form beliefs and choose actions. Our platform has full commitment power, similarly to the omniscient information designer in Bergemann and Morris (2019). Formally, the platform publicly commits to an information structure that, for each interaction, produces a private signal about ω for each seller i. By standard arguments (Myerson, 1983, 1984; Bergemann and Morris, 2016), we can focus on information structures in the form of recommendation mechanisms, where the platform privately recommends an action to each seller which he must find optimal to follow (obedience). A mechanism is then a function $x : \Omega \to \Delta(A)$, where $x(a|\omega)$ can be interpreted as the share of interactions of type ω that

[7] We make this assumption mostly for ease of exposition. Our model can accommodate correlation among records, as discussed in Section 4.2.2.

lead to recommendation profile a.[8] The formal problem is

$$\begin{aligned}\mathcal{U}_q: \qquad & \max_{x} \sum_{\omega\in\Omega, a\in A} u_0(a,\omega)x(a|\omega)q(\omega) \\ & \text{s.t. for all } i\in I \text{ and } a_i, a_i'\in A_i, \\ & \sum_{\omega\in\Omega, a_{-i}\in A_{-i}} \left(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\right) x(a_i,a_{-i}|\omega)q(\omega) \geq 0. \qquad (1)\end{aligned}$$

Each constraint (1) is equivalent to requiring that a_i maximizes seller i's expected utility conditional on the information conveyed by a_i given x and the database q. We denote any optimal mechanism by x_q^*. We define the *direct payoff* generated by each record of type ω as

$$u_q^*(\omega) \stackrel{\Delta}{=} \sum_{a\in A} u_0(a,\omega)x_q^*(a|\omega),$$

and the *total payoff* generated by the database as

$$U^*(q) \stackrel{\Delta}{=} \sum_{\omega\in\Omega} u_q^*(\omega)q(\omega).$$

We assume that $\mathcal{U}_q$ satisfies the following minor regularity property, which holds generically in the space of sellers' payoff functions.[9]

Assumption 1 (Non-degeneracy)**.** *Of the constraints (1) that define the feasible set of mechanisms x for $\mathcal{U}_q$, no more than $|A\times\Omega|$ are ever active at the same time.*

3 The Unit Value of Data

This section addresses our main question: How much value does the platform derive from each buyer's record and why? This individual contribution can offer a benchmark for compensating each buyer for her specific data. The next section will address two related questions that analyze two meanings of the colloquial expression "having more data." The platform could have more data in the sense of having *more* records in its database—thus having the ability to mediate more interactions between the corresponding buyers and the sellers. Or it could have more data in the sense of having *better* records—that is, observing more informative characteristics about existing buyers. Our framework allows us to analyze both meanings and determine the

[8] Note that we do not allow $x(\cdot|\omega)$ to differ between records of the same type ω, but this is without loss of generality.

[9] For more details, see Remark 1 in Appendix B.

platform's willingness to pay for more records and for better records. This can guide intermediaries when acquiring more data, taking into account its specific realizations.

It is useful to start by revisiting how we would answer these questions for standard decision problems. Suppose for a moment that the platform itself is a monopolistic seller (namely, there are no other sellers). For each interaction with a buyer it would choose a to maximize $u_0(a,\omega)$. What is the value of each record for the platform? To answer this, we would simply calculate the payoff from optimally using a record as an input in a decision, namely $u^*(\omega) = \max_{a\in A} u_0(a,\omega)$. How much is the platform willing to pay for one more record of type ω? The answer, in this case, would simply be $u^*(\omega)$. Finally, what is its willingness to pay for better records? We could assess the value of the information contained in a record by netting out the payoff of the optimal action if the platform had to use a record with fully uninformative data, namely, $u^*(\omega^\circ)$.[10] We could then aggregate $u^*(\omega) - u^*(\omega^\circ)$ across ω, as we normally do to evaluate information structures in standard decision problems. All these answers should look familiar, but they crucially rely on the key premise that $u^*(\omega)$ is indeed the right measure of the value of a record of type ω.

If instead the platform and the seller are distinct entities with different objectives, answering our questions requires to rethink that key premise, which is the focus of this section. For each buyer-sellers interaction, our platform continues to face a decision: which signals to send based on the buyer's record. However, these decisions are no longer independent across interactions and records. This is because the information a signal conveys about a buyer's record depends on which other records lead to the same signal.[11] The key implication is that the value of each record continues to be determined by how it is used to guide decisions, but this use is not confined to the interaction physically attached to that record. An interaction's outcome may not be determined by only its buyer's record, and this record may not determine only its interaction's outcome. This complicates the analysis because, as our introduction example illustrated, we can no longer use the direct payoff u^* to assess the value of a buyer's record. We instead have to systematically keep track of all the ways the platform uses each record in its database to mediate all interactions. This will allow us to simultaneously assess how much value it derives from each record and how it generates this value, thus answering our first question. At the same time, these values will determine the platform's willingness to pay for more data.

[10] The same point applies for other netting-out procedures, such as in Frankel and Kamenica (2019).

[11] Note that this dependence would arise even if the platform could not commit and we had to rely on some equilibrium notion.

Since we will refer to standard decision problems as a benchmark, it helps to note that they are effectively equivalent to a special case of our model. When all parties have aligned interests (i.e., u_i is an affine transformation of u_0 for all $i = 1, \ldots, n$), constraints (1) can be omitted, so it is as if the platform can directly choose all sellers' actions. Concretely, the previous case where the platform was also the seller is equivalent to our model where our platform maximizes the profits of a distinct monopolistic seller ($n = 1$). We will therefore use the terms *decision problem* and intermediation problem to refer to our model with aligned and conflicting interests, respectively.

3.1 The Data-Value Problem

Our approach builds on the observation that any information-design problem is a linear program. A standard economic interpretation is that linear programs describe the problem of optimally using some scarce inputs to produce some output (Dorfman et al., 1987, p. 39). We think of information design as a "data-use" problem, where the inputs are the records in the database and the output is the information conveyed by each mechanism in the form of recommendations. Following Dorfman et al. (1987, p. 39), we then exploit the dual of this data-use problem to evaluate each record.

We call this assignment task the *data-value* problem. Let $\lambda = (\lambda_1, \ldots, \lambda_n)$ where $\lambda_i : A_i \times A_i \to \mathbb{R}_+$ for all $i \in I$. For each i and (a, ω) define

$$t_i(a, \omega) \stackrel{\Delta}{=} \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \lambda_i(a_i' | a_i) \tag{2}$$

and $t(a, \omega) \stackrel{\Delta}{=} \sum_{i \in I} t_i(a, \omega)$. The data-value problem is

$$\begin{aligned} \mathcal{V}_q : \quad & \min_{v, \lambda} \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ & \text{s.t. for all } \omega \in \Omega, \\ & v(\omega) = \max_{a \in A} \Big\{ u_0(a, \omega) + t(a, \omega) \Big\}, \end{aligned} \tag{3}$$

We denote any optimal solution by (v_q^*, λ_q^*) and the induced functions t by t_q^*. By standard linear-programming arguments v_q^* is unique generically with respect to q (i.e., except on a set of qs with measure zero).[12] We refer to equation (3) as the *value formula*, which defines our main object of interest. The reason hinges on the next relation between the data-use and data-value problems and on the following interpretation.

[12] We will provide an independent economic interpretation of $\mathcal{V}_q$ in Section A.

Lemma 1. *For any database q, $\mathcal{V}_q$ is equivalent to the dual of $\mathcal{U}_q$. Thus, for every x_q^* and (v_q^*, λ_q^*)*

$$\sum_{\omega\in\Omega} v_q^*(\omega)q(\omega) = U^*(q) \overset{\Delta}{=} \sum_{\omega\in\Omega} u_q^*(\omega)q(\omega). \tag{4}$$

All proofs are in Appendix B.

This duality relation follows from basic linear-programming results. Yet, applied to our specific problem, it becomes the key to answering our economic questions. In $\mathcal{U}_q$, by choosing x the platform chooses a joint measure $\chi \in \mathbb{R}_+^{\Omega\times A}$, which must satisfy $\sum_{a\in A}\chi(a,\omega) = q(\omega)$; that is, the use of type-ω records to produce recommendations must exhaust their stock $q(\omega)$ in the database. Formally, $v(\omega)$ is the multiplier of this constraint, which is usually interpreted as the shadow price of the corresponding input through the thought experiment of adding a marginal unit of it. In fact, $v_q^*(\omega)$ is the derivative of $U^*(q)$ with respect to $q(\omega)$, as for any constrained optimization problem. However, it would be misleading to think that $v_q^*(\omega)$ captures only the value of a marginal record of type ω. In fact, Lemma 1 demonstrates that $v_q^*(\omega)$ captures the value of *each* record of type ω in the database, not just the marginal one. We will then refer to $v_q^*(\omega)$ as the *unit value* of a record of type ω (see also Gale (1989), p. 12). Note that $\mathcal{V}_q$ assigns such values to all records simultaneously and does not require to find x_q^*. Also, v_q^* can depend on q for intermediation problems but not for decision problems, as in this case $v_q^*(\omega) = \max_{a\in A} u_0(a,\omega)$ for all ω.

The rest of the paper characterizes the properties of v_q^*. Here, we can begin by establishing a lower bound for the value of a record. For $\omega \in \Omega$, let $CE(\Gamma_\omega)$ be the set of correlated equilibria of the game Γ_ω.

Lemma 2 (Lower Bound)**.** *For every q,*

$$v_q^*(\omega) \geq \bar{u}(\omega) \overset{\Delta}{=} \max_{y\in CE(\Gamma_\omega)} \sum_{a\in A} u_0(a,\omega)y(a), \qquad \omega \in \Omega.$$

As in Maskin and Tirole (1990), we say that x is *full-information incentive compatible* if $x(\cdot|\omega) \in CE(\Gamma_\omega)$ for all ω. If there is such an x that is also optimal, we say that $\mathcal{U}_q$ is a *full-disclosure problem*. By Lemma 1 and 2 this is the case if and only if $v_q^*(\omega) = \bar{u}(\omega)$ for all ω.

3.1.1 Interpretations and Use of v^*

We pause briefly to clarify how v_q^* can be interpreted and used in relation to our main questions.

Individual Contribution. By quantifying how much each record in the database contributes to the total payoff $U^*(q)$, v_q^* also offers a way to individually compensate each buyer as the

"owner" of her record. Paraphrasing Dorfman et al. (1987, p. 43), this interpretation is reminiscent of the operation of a competitive market where competition forces the platform to offer the "owner" of a record the full value to which her input gives rise, while competition among these "owners" drives down this value to the minimum consistent with this limitation. Gale (1989, Chapter 3.5) also shows how dual problems can deliver competitive prices of scarce inputs.

To see the importance of compensating data owners using v_q^* and not u_q^*, imagine that type-ω° records correspond to buyers who prevent the platform from tracking their characteristics (perhaps by browsing incognito). If they now allow tracking and reveal their valuation θ, how should the gain in the database value be allocated among them? As our introduction example illustrated, using u_q^* we may incorrectly allocate this gain only to the now-tracked buyers with high valuation (i.e., ω_2), while we should allocate it only to those with low valuation (i.e., ω_1).

More Records. Suppose the platform is offered a new buyer's record. Let $\rho \in \Delta(\Omega)$ be its belief about the record's type, which it will observe after acquiring the record. We can then define its willingness to pay for a new record as

$$v_q^*(\rho) \triangleq \sum_\omega \rho(\omega) v_q^*(\omega).$$

In fact, this is the expected marginal increase of the platform's total payoff from the new record.

Better Records. Imagine again that type-ω° records correspond to buyers who blocked tracking. How much would be platform be willing to pay to learn more about them? In other words, how can we quantify the value of the information contained in a record when the platform may use it to mediate multiple interactions? Similarly to standard decision problems, one way is to use the value of type-ω records net of the value of type-ω° records:

$$v_q^*(\omega) - v_q^*(\omega^\circ), \qquad \omega \in \Omega. \tag{5}$$

Despite the superficial similarity, this value of information for intermediation problems differs in important ways from decision problems, as we explain in Section 4.2.

3.2 Value Decomposition and Data Externalities

We now delve deeper into what determines the unit value of data records. Key to this will be comparing v_q^* with their direct payoffs u_q^*. We will show that the gap between them quantifies an externality between records. We will characterize this externality and argue that it is a defining feature of intermediation problems.

We start by showing that the value of each record can be decomposed into two parts: The direct payoff $u_q^*(\omega)$ and another component, $t_q^*(\omega)$, which captures the indirect effects a record

of type ω generates. This decomposition formalizes why and how u_q^* can bias our assessment of the actual value of each record for the platform.

Proposition 1. *For all $\omega \in \Omega$, $v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega)$ where*

$$t_q^*(\omega) \overset{\Delta}{=} \sum_{a \in A} t_q^*(a,\omega) x_q^*(a|\omega) \overset{a.e.}{=} \sum_{\omega' \in \Omega} \frac{\partial u_q^*(\omega')}{\partial q(\omega)} q(\omega') \tag{6}$$

This result highlights two aspects of the value of records. The first is that the indirect effects t_q^* are akin to an externality. By the last part of (6), $t_q^*(\omega)$ captures the marginal effect of a type-ω record on the direct payoff of *all* records. This externality is purely informational: By being in the database, each record affects the platform's informational advantage and hence the optimal way x_q^* in which the whole database is used. In fact, $\frac{\partial}{\partial q(\omega)} u_q^*(\omega') = \sum_a u_0(a,\omega') \frac{\partial}{\partial q(\omega)} x_q^*(a|\omega')$. Adjustments in x_q^* can arise because changing $q(\omega)$ can render x_q^* no longer feasible (i.e., obedient) or optimal.

Which records generate positive and which negative externalities?

Corollary 1. *$t_q^*(\omega) < 0$ for some ω if and only if $t_q^*(\omega') > 0$ for some ω'. Moreover, $t_q^*(\omega) < 0$ implies $u_q^*(\omega) > \bar{u}(\omega)$, while $u_q^*(\omega) < \bar{u}(\omega)$ implies $t_q^*(\omega) > 0$.*[13]

The first part shows that the externalities lead to cross-subsidization of value from records with $t_q^*(\omega) < 0$ to records with $t_q^*(\omega') > 0$. The latter's $v_q^*(\omega')$ exceeds $u_q^*(\omega')$, so they must extract this extra value from records with $t_q^*(\omega) < 0$. The second part of the corollary explains this cross-subsidization. Records with $t_q^*(\omega) < 0$ generate a direct payoff that exceeds the full-information payoff $\bar{u}(\omega)$, which requires that $u_0(a,\omega) > \bar{u}(\omega)$ and $x_q^*(a|\omega) > 0$ for some a. That is, the platform earns a payoff with type-ω records that would never be possible by fully disclosing such records, so it relies on pooling them with records of different types. This help from type-ω' records justifies why $t_q^*(\omega') > 0$ and their value exceeds $u_q^*(\omega')$. Conversely, if $t_q^*(\omega) < 0$, a record of type ω benefits from externalities caused by other records and hence has to "repay" them, which lowers its value. For the last part, we can interpret $u_q^*(\omega) < \bar{u}(\omega)$ as "sacrificing" type-ω records, as the platform could fully disclose them and ensure a payoff $\bar{u}(\omega)$. For this sacrifice to be worthwhile, such records must receive a compensation, explaining $t_q^*(\omega) > 0$. This last part offers a sufficient condition for $t_q^* \neq 0$ that is simple to check, but is not necessary. We provide another sufficient condition in Appendix C.

It is worth emphasizing that our externalities arise even though our records are statistically independent between buyers. As such, they differ from other data-driven externalities studied

[13] The corollary follows because Lemma 1 implies $\sum_{\omega \in \Omega} t_q^*(\omega) q(\omega) = 0$, and Lemma 2 and Proposition 1 imply $t_q^*(\omega) \geq \bar{u}(\omega) - u_q^*(\omega)$ for all ω.

in the literature. One distinction is that they never arise for decision problems, where $v_q^*(\omega) = u_q^*(\omega) = \bar{u}(\omega)$ and therefore $t_q^*(\omega) = 0$ for all $\omega \in \Omega$ and q.[14] By contrast, in Acemoglu et al. (2021) and Bergemann et al. (2021) externalities arise due to correlation between the data of different consumers. In our framework, we could have similar externalities *within* the record of a single interaction if there are multiple buyers and one buyer's observed characteristics are correlated with another buyer's unobserved characteristics. Thus, these are externalities between dimensions of a single record, which can also arise in decision problems.

A second aspect highlighted by Proposition 1 is that the externalities through u_q^* are tightly related to how the platform exploits the sellers' primitive incentives. By the first part of (6), we can view $t_q^*(\omega)$ as aggregating externalities that type-ω records generate by inducing specific actions a. These are inversely related to the payoff the platform gets, in the following sense.[15]

Corollary 2. *Suppose $x_q^*(a|\omega) > 0$ and $x_q^*(a'|\omega) > 0$. Then, $u_0(a,\omega) > u_0(a',\omega)$ if and only if $t_q^*(a,\omega) < t_q^*(a',\omega)$.*

Thus, inducing actions whose payoff exceeds $\bar{u}$ by more, for instance, requires paying larger externalities to other records. Since $t_q^*(a,\omega) \overset{\Delta}{=} \sum_{i\in I} t_{q,i}^*(a,\omega)$, we can view $t_{q,i}^*(a,\omega)$ as how much seller i contributes to the externality. Recall that $t_{q,i}^*(a,\omega)$ differs from zero only if $\lambda_{q,i}^*(a_i'|a_i) > 0$ for some a_i' (see (2)). By standard arguments (complementary slackness), $\lambda_{q,i}^*(a_i'|a_i) > 0$ only if

$$\sum_{\omega,a_{-i}} \Big(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\Big) x_q^*(a_i,a_{-i}|\omega) q(\omega) = 0; \tag{7}$$

the converse also holds generically in q. In words, $\lambda_{q,i}^*(a_i'|a_i) > 0$ if and only if seller i is indifferent between a_i and a_i' conditional on receiving recommendation a_i from x_q^*. This fact implies the following.

Corollary 3. *The sellers who contribute to the externality $t_q^*(\omega)$ are only those whom x_q^* renders indifferent with the actions it recommends using records of type ω (i.e., (7) holds).*

Note that this result differs from the immediate fact that optimal solutions of linear programs occur on the boundary of the feasible set, which here means that some obedience constraint must bind. As q varies, x_q^* and hence t_q^* may change. However, as long as λ_q^* does not change (see Proposition 3 below), how each seller contributes to $t_q^*(\omega)$ does not change. Section A

[14] The converse is not true: It is possible to construct examples where $t_q^*(\omega) = 0$ and $v_q^*(\omega) > \bar{u}(\omega)$ for all ω.

[15] This corollary follows from complementary slackness, which in our case says that $x_q^*(a|\omega) > 0$ implies $v_q^*(\omega) = u_0(a,\omega) + t_q^*(a,\omega)$.

explains further how the platform exploits the sellers to determine their contributions to these externalities.

3.2.1 Application – Data and Price Discrimination (Part I)

To better illustrate the importance of these data externalities, we consider a more general version of our introduction example, following Bergemann et al. (2015). There is only one seller ($n = 1$) who chooses the price a_1 for his product. For each buyer, θ is her valuation for the seller's product, where $\Theta = \{\theta_1, \ldots, \theta_K\}$, $K \geq 2$, and $\theta_k > 0$ is strictly increasing in the index k. Let ω_k be a record which fully reveals that $\theta = \theta_k$. Normalizing the seller's constant marginal cost to zero, his profit is a_1 if $\omega \geq a_1$ and zero otherwise: $u_1(a_1, \omega) = a_1 \mathbb{I}\{\omega \geq a_1\}$. Suppose the platform maximizes a weighted sum of profits and consumer surplus: $u_0(a_1, \omega) = \pi a_1 \mathbb{I}\{\omega \geq a_1\} + (1 - \pi) \max\{\omega - a_1, 0\}$, where $\pi \in [0, 1]$. Finally, let a_q be the price the seller would set conditional on knowing only the database composition q.

Proposition 2. *For $\pi \leq \frac{1}{2}$,*

$$v_q^*(\omega) = \begin{cases} (1-\pi)\omega & \text{if } \omega < a_q \\ \pi a_q + (1-\pi)(\omega - a_q) & \text{if } \omega \geq a_q; \end{cases}$$

moreover, $t_q^*(\omega) > 0$ for $\omega < a_q$ and $t_q^*(\omega) \leq 0$ for $\omega \geq a_q$. For $\pi \geq \frac{1}{2}$ we have $v_q^*(\omega) = u_q^*(\omega) = \pi\omega$ for all ω.

To understand this result, we note that the optimal x_q^* takes only two forms depending on π (see Appendix B). If $\pi \leq \frac{1}{2}$, the platform always maximizes the buyers' surplus subject to holding the seller's expected profits at a_q, as if $\pi = 0$. Thus, it is as if trade happens for every interaction, generating total surplus equal to ω, and only the buyers with valuation at least a_q contribute to guaranteeing this reservation profits for the seller. If $\pi \geq \frac{1}{2}$, the platform fully discloses all records. This allows the seller to perfectly price discriminate between buyers, so profits always equal the buyer's valuation and her surplus is zero.

Whenever the platform cares more about the buyers' surplus than the seller's profits, the direct payoff u_q^* provides a biased account of the value of each record. The result shows that this bias has a specific structure: t_q^* satisfies a single-crossing property in ω and this holds generally across q. That is, u_q^* is biased downward for low-valuation buyers (i.e., $\omega < a_q$) and upward for high-valuation buyers (i.e., $\omega \geq a_q$). This illustrates that ignoring the externalities we highlight may lead the platform to overcompensate high-valuation buyers for their data at the expense of low-valuation ones.

How does caring more about the buyers' surplus affect the value of their records? Simple algebra implies that, for $\pi \leq \frac{1}{2}$, lowering π decreases $v_q^*(\omega)$ if and only if the buyer has an intermediate valuation ($a_q \leq \omega < 2a_q$). Intuitively, this is because the buyers in this type of interaction contribute the most—as a share of their valuation—to funding the seller's guaranteed profits of a_q. By contrast, the records of buyers with very low valuations help to achieve positive surplus with other buyers, and the records of buyers with very high valuations just yield a large surplus. For $\pi \geq \frac{1}{2}$, $v_q^*(\omega)$ increases in π independently of ω. This is because the platform helps the seller extract the full surplus from each interaction, and the platform cares more about his profits.

4 Acquiring More Data

This section studies how the value of buyers' records v_q^* depends on the database composition q. This dependence allows us to assess the platform's willingness to pay for acquiring more data in the sense of more records (Section 4.1) and better records (Section 4.2).

4.1 More Records: Preferences over Databases

How do changes in the quantity of type-ω records affect their value? For example, do records become less valuable as they become more abundant in the database? To address these questions, we can think of the platform as a consumer of different types of goods called data records, where q is a bundle of such goods and U^* is its utility function. Then, $v_q^*(\omega)$ is akin to the marginal utility of type-ω records at q. We can also measure the marginal rate of substitution between records of type ω and ω' at q in the usual way, by letting $MRS_q(\omega, \omega') \overset{\text{a.e.}}{=} -\frac{v_q^*(\omega)}{v_q^*(\omega')}$. Thus, v_q^* fully characterizes the platforms' preferences.

A classic property in standard consumer theory is that marginal utilities are diminishing. Does the same hold for our platform? More generally, how does v_q^* vary with q? We first show that v_q^* is constant with respect to local, yet discrete, changes in q.

Proposition 3 (Stability)**.** *There exists a finite collection $\{Q_1, \ldots, Q_K\}$ of open, convex, and disjoint subsets of $\mathbb{R}_+^\Omega$ such that $\cup_k Q_k$ has full measure and, for every k, v_q^* is unique and constant for $q \in Q_k$.*

In fact, each Q_k is the interior of a cone in the space of databases $\mathbb{R}_+^\Omega$.[16] Importantly, v_q^* is

[16] It is easy to see that unit values are constant along the rays in the space of databases: If $q' = \alpha q$ for $\alpha > 0$,

constant even though the platform may adjust how she uses her data when q changes. Indeed, we can show that within each cone, while $v_q^*(\omega)$ is constant, the optimal $x_q^*(\omega)$ changes as a function of q (see Remark 1 in Appendix B). Intuitively, this is because x_q^* has to be fine-tuned to maximally exploit the sellers' incentives. By contrast, v_q^* depends only on which sellers' incentives are exploited, but not by how much (recall equation (2) and Corollary 3).

For global changes in q, we find that records of a given type become more valuable as they become scarcer. This establishes a "scarcity principle" for data and implies diminishing marginal utilities. For every q, define the share of records of each type by

$$\mu_q(\omega) \triangleq \frac{q(\omega)}{\sum_{\omega'} q(\omega')}, \qquad \omega \in \Omega.$$

Proposition 4 (Scarcity Principle)**.** *Consider databases q and q'. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$. Moreover, for every $\omega \in \Omega$ there exists $\bar{\mu}(\omega) < 1$ such that, if $\mu_q(\omega) > \bar{\mu}(\omega)$, then $v_q^*(\omega) = \bar{u}(\omega)$.*

Although each type of records can contribute to the platform's informational advantage in different ways, it always becomes more valuable as it becomes scarcer. This implies that $v_q^*(\omega)$ is weakly decreasing in $q(\omega)$—for any selection from the optimal solution correspondence of $\mathcal{V}_q$. Holding fixed the quantity of all other types of records, the platform's demand for type-ω records is downward sloping and converges to $\bar{u}(\omega)$ when $q(\omega)$ is sufficiently large. Equivalently, the individual contribution of type-ω records to the platform's payoff—hence, their owners' benchmark compensation—decreases as their quantity increases.

Another classic property in standard consumer theory is that marginal rates of substitution are diminishing, which means that the consumed goods are not perfect substitutes. Again, does the same hold for our platform? The answer is yes, unless it faces a trivial intermediation problem—that is, it is optimal to always fully disclose all records to the sellers, independently of q. The platform's preferences are always weakly convex, because $U^*(q)$ is always a weakly concave function of q.[17] However, in decision problems full disclosure is optimal regardless of q, which implies that $MRS_q(\omega, \omega') = -\frac{\bar{u}(\omega)}{\bar{u}(\omega')}$ for all ω and ω' and hence all types of records are perfect substitutes. Even in an intermediation problem full disclosure can be optimal for some particular q. The next result shows that, when this is the case, full disclosure is optimal for all q, so all types of records are again perfect substitutes.

---

then $v_q^* = v_{q'}^*$. This is because only the frequency of record types matters for the sellers' incentives.

[17] Concavity follows because, through (4), we can view U^* as the minimization of a family of linear functions in the "parameter" q (see, e.g., Theorem 5.5 in Rockafellar, 1970). It is related directly to the concavification results in Mathevet et al. (2020) and indirectly to the individual-sufficiency results in Bergemann and Morris (2016).

Proposition 5. *Fix* Γ. *Suppose* $\mathcal{U}_q$ *is a full-disclosure problem for some database* $q \in \mathbb{R}^{\Omega}_{++}$. *Then,* $v^*_{q'}(\omega) = \bar{u}(\omega)$ *for all* ω *and* $q' \in \mathbb{R}^{\Omega}_{+}$; *thus,* $\mathcal{U}_{q'}$ *is a full-disclosure problem for all* q'.

Proposition 5 has several implications. First, some types of records are imperfect substitutes if and only if it is never optimal to fully disclose all records. In this case, MRSs vary and $v^*_q \neq \bar{u}$ for all $q \in \mathbb{R}^{\Omega}_{++}$. Second, showing that it is not optimal to fully disclose some ω for some $q \in \mathbb{R}^{\Omega}_{++}$ suffices to show that full disclosure is never optimal for all $q \in \mathbb{R}^{\Omega}_{++}$. More generally, Γ is all one needs to know to establish the (sub)optimality of full disclosure.[18] Last but not least, we can identify whether the platform faces a non-trivial intermediation problem by detecting that it treats some types of records as imperfect substitutes.

Convexity of preferences implies that all intermediation problems lead to standard demand analysis. Choosing an optimal database subject to a budget constraint is a well-behaved problem. Given market price $p(\omega) > 0$ for every record type ω, the optimal q is characterized by

$$\max_{v \in v^*_q} \frac{v(\omega)}{v(\omega')} \geq \frac{p(\omega)}{p(\omega')} \geq \min_{v \in v^*_q} \frac{v(\omega)}{v(\omega')}, \qquad \omega, \omega' \in \Omega.^{19}$$

In this way, we can use v^*_q to characterize the platform's *demand functions* for data records, thus enabling a general study of the demand side of the "market for data." Which prices will prevail in this market is of course determined by the interplay of demand and supply. Under perfect competition, Dorfman et al. (1987) and Gale (1989) provide arguments for equality between v^*_q and equilibrium prices.

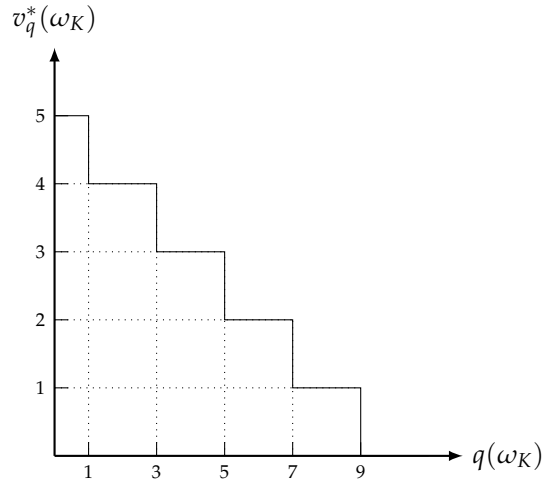
4.1.1 Application – Data and Price Discrimination (Part II)

We illustrate some of these concepts by specializing our analysis to the setting of Section 3.2.1. As before, we consider a single price-setting seller ($n = 1$) and assume that the platform maximizes a weighted sum of consumer surplus and this seller's profits. In this specification, recall that $\pi \in [0,1]$ is the weight the platform gives to the seller's profits and a_q is the price that the seller would charge conditional on knowing only the database composition q.

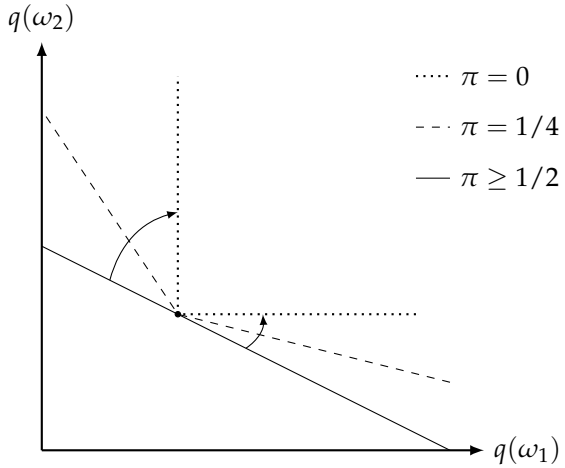
First, we show an example of a downward-sloping demand curve. The left panel of Figure 2 shows the value of records of type ω_K calculated using Proposition 2. This value is locally constant—as discussed in Proposition 3—and decreases as these records become more

[18] We provide sufficient conditions for this in terms of Γ in Appendix C.

[19] We slightly abuse notation by letting v^*_q stand for the *set* of optimal solutions at q. This condition is equivalent to $p \in \partial U^*(q)$, where $\partial U^*(q)$ is the superdifferential of U^* at q. Note that in the special case with a unitary budget and $p(\omega) = 1$ for all ω, choosing q is isomorphic to choosing an optimal prior in $\Delta(\Omega)$.



(a) Example of a demand curve. We set $\pi = 0$, $K = 10$, $\theta_k = k$ $(\forall k)$, $q(\omega_k) = 1$ $(\forall k < K)$.



(b) Example of indifference curves becoming more convex. We set $K = 2$, $\theta_k = k$ $(\forall k)$

Figure 2: Platform's demand and indifference curves

abundant—as discussed in Proposition 4. Moreover, the figure shows that, as $q(\omega_K)$ becomes sufficiently large, $v_q^*(\omega_K)$ reaches a lower bound, which in this specific case is 0.

Second, we explore how the substitutability between records changes as a function of π. When $\pi < \frac{1}{2}$, records become less substitutable as π decreases, that is, as the platform cares less about the seller' profit. When $\pi \geq \frac{1}{2}$, all types of records are perfect substitutes and $MRS_q(\omega, \omega') = -\frac{\omega}{\omega'}$ for all q, and thus it is constant in π.

Corollary 4. *Fix q and increase $\pi < \frac{1}{2}$. If $\omega, \omega' < a_q$, $MRS_q(\omega, \omega')$ is constant at $-\frac{\omega}{\omega'}$. If $\omega < a_q \leq \omega'$, $MRS_q(\omega, \omega')$ increases monotonically towards $-\frac{\omega}{\omega'}$ from below. If $\omega' > \omega \geq a_q$, $MRS_q(\omega, \omega')$ decreases monotonically towards $-\frac{\omega}{\omega'}$ from above.*

In words, as π increases towards $\frac{1}{2}$, for record types on the opposite side of a_q, the platform's indifference curve rotates counter-clockwise in the direction of perfect substitutability. For records on the same side of a_q, its indifference curve rotates clockwise in the direction of perfect substitutes. Thus, the indifference curves become "less convex" around the dimension $\omega = a_q$. In particular, at $\pi = 0$ records of type $\omega = a_q$ are perfect complements with every other type. These patterns are illustrated in the right panel of Figure 2, which shows the platform's indifference curves in the case with two types of records.

4.2 Better Records: The Value of Information for Intermediaries

This section considers the problem of a platform that can obtain more refined information about some of the records in its database. For example, it could learn new personal characteristics about a subset of the buyers. Intuitively, such a refinement transform the original records into records of different types, depending on what the platform learns about them. Therefore, refining records is equivalent to changing the original database into a new one, in which some of the original records are now of a differ type. Using the tools introduced so far, we can then assess how this refinement would change the value of the affected records and determine how much the platform would be willing to pay for it.

To do so, we first need to provide more structure to our notion of a database. We begin by explicitly modeling types of records using information partitions (e.g., as in Gentzkow and Kamenica, 2016). We describe the possible types of a record as cells of a finite partition Ω of $\Theta \times [0,1]$, such that each cell $\omega \in \Omega$ is a non-empty measurable subset of $\Theta \times [0,1]$. Let ζ be a random variable that is independent of θ and has uniform distribution λ on $[0,1]$. For each $\omega \in \Omega$, the probability of ω conditional on θ is then $\lambda(\{\zeta : (\theta,\zeta) \in \omega\})$. This allows us to consider types of records that are more refined than others. We do so by letting $\mathbf{\Omega} = \{\Omega^\circ, \Omega^1, \ldots, \Omega^K\}$ be a family of finite partitions of $\Theta \times [0,1]$ with the property that $\Omega^\circ = \{\omega^\circ\} = \Theta \times [0,1]$ and Ω^k is finer than Ω^{k-1} for all $k = 2, \ldots, K$. For example, Ω^k could correspond to the types of records that only contain the age of a buyer, while Ω^{k+1} could also contain her gender. Given $\mathbf{\Omega}$, we say that the space of databases $\mathbb{R}_+^\Omega$ is *rich* if $\Omega = \bigcup_{k=1}^K \Omega^k$. That is, if the platform can obtain refined information about a buyer whose record is of type $\omega \in \Omega$, the newly refined record would be of a type ω' that still belongs to Ω. That is, any refinement involves moving inside the space of databases itself.

When the platform refines a share $\alpha \in [0,1]$ of the type-ω records, it observes for each one of them an *independent* and identically distributed exogenous signal.[20] That is, each one of these records becomes of type ω' with probability

$$\sigma_\omega(\omega') = \frac{\lambda(\{\zeta : (\theta,\zeta) \in \omega \cap \omega'\})}{\lambda(\{\zeta : (\theta,\zeta) \in \omega\})}, \qquad \omega' \in \operatorname{supp} \sigma_\omega.$$

We denote such a refinement by (α, σ_ω). When the database q is refined according to (α, σ_ω), it transforms into a new database q_α. Indeed, by the Law of Large Numbers, this new database contains a quantity $q_\alpha(\omega) = (1-\alpha)q(\omega)$ of type-ω records and quantity $q_\alpha(\omega') = q(\omega') + \alpha\sigma_\omega(\omega')q(\omega)$ of type-ω' records, for each ω' such that $\sigma_\omega(\omega') > 0$. That is, there is no

[20] We discuss correlated refinements in Section 4.2.2.

uncertainty about the composition q_α of the new database, even though there is uncertainty about which records of type ω become of type ω'. The sellers know that the platform has refined its database according to (α, σ_ω) and, thus, they know the composition of the new database q_α.[21]

At a conceptual level, refining records generates a fundamental trade-off for the platform. A refinement by construction improves what the platform knows about the existing records. This means that the platform can better fine-tune what information it sends to the sellers, thus potentially achieving more with the refined records. However, a refinement also changes q and, thus, the sellers' beliefs about the buyers they interact with. This means that a recommendation mechanism x that was obedient before the refinement may no longer be obedient after it, thus potentially hurting the platform. Therefore, whether a refinement increases the value of refined records and ultimately benefits the principal is unclear. We begin by establishing the effects of a refinement on the value of the records it refines.

Proposition 6. *Fix q, $\alpha \in [0,1]$, and $\omega \in \Omega$. Let (α, σ_ω) be a refinement. The value of each refined record weakly increases in expectation:*

$$\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') - v^*_q(\omega) \geq 0.$$

This increase is smaller the larger α is. Moreover, it is equal to zero if there exists $a \in supp\, x^*_q(\cdot|\omega'')$ for $\omega'' = \omega$ and all $\omega'' \in supp\, \sigma_\omega$. The converse is true generically in q.

We find that the value of refined records weakly increases in expectation. That is, refined records contribute more to the total payoff than what their unrefined counterparts did. This is true even if, in the new database, the quantity of the types of records that result from the refinement has increased, and thus their value may decrease due to the scarcity principle (see Proposition 4). Proposition 6 also provides a sharp condition for the expected increase in the value of refined records to be exactly zero. There must be a *common* action profile that the platform induces with positive probability for records of type ω as well as for every type that records of type ω can turn into when refined. This condition shows that, due to the tradeoff previously illustrated, the value of refined records can be unaffected by the refinement even if the platform uses them in a different way, namely even if x^*_q changes.

While a refinement inevitably increases the value of the refined records, the previous discussion clarified that it can also decrease the value of unrefined ones. Therefore, it is natural

[21] Of course, in reality the platform may acquire information about existing buyers privately without the sellers' knowledge. Allowing for this requires enriching the model accordingly and introduces further complications, which we leave for future research.

to wonder whether the platform is always willing to pay a positive price for obtaining better records.

Corollary 5. *Fix q, $\alpha \in [0,1]$, and $\omega \in \Omega$. Let (α, σ_ω) be a refinement. The platform weakly benefits from the refinement, i.e., $U(q_\alpha) \geq U(q)$. The marginal benefit of the refinement is decreasing in α. Moreover, the benefit is equal to zero if there is $a \in supp\, x_q^*(\cdot|\omega'')$ for $\omega'' = \omega$ and all $\omega'' \in supp\, \sigma_\omega$. The converse is true generically in q.*

This result implies that the platform's willingness to pay for a refinement is always weakly positive. However, under the same condition established by Proposition 6, the platform may be unwilling to pay a strictly positive price for refining its records, despite acting on the information it receives. This is in sharp contrast with decision problems, where the gain from obtaining more information is positive if and only if it changes behavior of the decision maker.

The result also shows that at the margin, as the platform refines a larger α-share of type-ω records, she benefits less. This is because the expected value gain of the marginal record becomes smaller. This is akin to a decreasing marginal value of information, but with some qualifications. Our exercise is not to fix one buyer-seller interaction and gradually give the platform more information about that interaction, which is the usual way of thinking about the marginal value of information in a standard decision problem and does not change the platform's informational advantage because one interaction in infinitesimal. Our exercise is to fix the amount of information we give the platform for each interaction of a certain type and vary how many interactions we refine in this way, which changes its informational advantage. In fact, all these positive effects can become negative when refinements are correlated across records, as they change the platform's informational advantage in qualitatively different ways.

4.2.1 Application – Data and Price Discrimination (Part III)

As in the setting of Section 3.2.1, we assume there is a single seller. Moreover, we let $\pi = 0$ and, thus, assume that the platform maximizes consumers' surplus. Finally, let $\Omega = \{1, 2, \omega^\circ, \}$. As before, ω_1 and ω_2 are the record types that correspond to buyers whose valuation θ is 1 and 2, respectively; ω°, instead, corresponds to buyers' whose valuation is believed to be $\theta = 2$ with probability $h > \frac{1}{2}$ and $\theta = 1$ otherwise. For any database q that satisfies $(2h-1)q(\omega^\circ) < q(\omega_1) < q(\omega_2) + (2h-1)q(\omega^\circ)$, we have that $v_q^*(\omega_1) = 1$ and $v_q^*(\omega_2) = 0$, as in our introduction example (see Appendix D). Moreover, $v_q^*(\omega^\circ) = 1 - h$.

Consider an α-share of the type-ω° records. Denote by q_α the resulting database. Note that, for all $\alpha \in [0,1]$, the resulting database still satisfies $(2h-1)q_\alpha(\omega^\circ) < q_\alpha(\omega_1) < q_\alpha(\omega_2) +$

$x_q^*(a\|\omega)$	$\omega = 1$	$\omega = 2$	$\omega = \omega^\circ$
$a = 1$	1	$\frac{q(1)-(2h-1)q(\omega^\circ)}{q(2)}$	1
$a = 2$	0	$1 - \frac{q(1)-(2h-1)q(\omega^\circ)}{q(2)}$	0

Table 1: Optimal x_q^*.

$(2h-1)q_\alpha(\omega^\circ)$. Therefore, any refinement leaves the value of each refined record unchanged, since $v^*(\omega_1)(1-h) + v^*(\omega_2)h = v^*(\omega^\circ)$. Moreover, the platform is not willing to pay a positive price for such a refinement, since $U^*(q) = U^*(q_\alpha)$. This is despite the fact that, as q changes, the platform changes how it uses the data. This is illustrated in Table 1.

4.2.2 Correlation and General Refinements

One may wonder whether the results in this section depend on how records are refined. We can describe refinements that are not necessarily i.i.d. as follows. For every q and $\omega \in \Omega$, let $Q(\omega, q)$ be the set of all databases that can be reached from q by refining records of type ω: $Q(\omega, q)$ contains all $q' \in \mathbb{R}_+^\Omega$ that satisfy

- $q'(\omega) \leq q(\omega)$,
- $q'(\hat{\omega}) \geq q(\hat{\omega})$ if $\hat{\omega} \subset \omega$,
- $q'(\hat{\omega}) = q(\hat{\omega})$ if $\hat{\omega} \not\subseteq \omega$,
- $\sum_{\hat{\omega}\in\Omega} q'(\hat{\omega}) = \sum_{\hat{\omega}\in\Omega} q(\hat{\omega})$.

Then, given q, acquiring better data about type-ω records involves transitioning—possibly with some randomness—from q to some $q' \in Q(\omega, q)$. That is, any such data-acquisition process can be described by some distribution $\rho(\omega, q) \in \Delta(Q(\omega, q))$. While i.i.d. refinements induce such a distribution, other refinements may involve correlation in how the platform learns about different records of type ω. It is easy to see that for full-disclosure problems *any* refinement $\rho(\omega, q)$ increases, in expectation, both the value of the refined records and the platform's total payoff.[22] This is analogous to the result that the value of information in decision problems is always non-negative. However, this can fail for other intermediation problems, as the next example illustrates. The basic reason is that our acquisition of information changes not only

[22]This follows from the fact that $v_q^*(\omega)$ is independent of q for all $\omega \in \Omega$ and therefore only the marginal of ρ for each refined record matters.

how much the platform knows about each interaction, but also the degree and nature of the asymmetric information between the platform and the sellers implied by the commonly-known database composition. This highlights another important distinctive feature of the value of data in intermediation problems.

To illustrate, consider the same setting of Section 4.2.1. In addition, assume that $q(\omega_1) + q(\omega^\circ)(1-h) > q(\omega_2)$. Consider the following general refinement $\rho(\omega^\circ, q)$, which is arguably extreme but serves to make our point as clearly as possible. Suppose the platform is told that all its type-ω° records involve buyers with the same valuation. Thus, if refined, with probability $1-h$ they *all* become records of type 1 and with probability h they *all* become records of type 2. That is, $\rho(\omega^\circ, q)$ assigns positive probability to only q' and q'', where $q'(\omega_1) = q(\omega_1) + q(\omega^\circ)$, $q'(\omega_2) = q(\omega_2)$, $q''(\omega_1) = q(\omega_1)$, $q''(\omega_2) = q(\omega_2) + q(\omega^\circ)$, and $q'(\omega^\circ) = q''(\omega^\circ) = 0$. In this case, we have $v^*_{q'}(\omega_1) = 0$ and $v^*_{q''}(\omega_2) = 0$ (see Appendix D for the computations). This implies that

$$(1-h)v^*_{q'}(\omega_1) + hv^*_{q''}(\omega_2) - v^*_q(\omega^\circ) = -(1-h) < 0.$$

It is also easy to calculate that

$$U^*(q) = q(\omega_1) + q(\omega^\circ)(1-h) > \max\{q(\omega_2), q(\omega_1)\} = \max\{U^*(q'), U^*(q'')\}.$$

Therefore, this refinement has a strictly negative effect not only on the value of the refined records of type ω°, but also on the platform's overall payoff. Note that if the platform maximizes the seller's profits, $v^*_{\hat{q}}(\omega_1) = 1$, $v^*_{\hat{q}}(\omega_2) = 2$, and $v^*_{\hat{q}}(\omega^\circ) = 2h$ for all $\hat{q}$. Therefore, the *same* refinement $\rho(\omega^\circ, q)$ has a strictly positive effect on both the value of records of type ω° and the platform's overall payoff, as standard in decision problems.

The key is that a profit-maximizing platform treats each buyer-seller interaction independently and hence does not care about correlation in how it learns about their records. By contrast, a surplus-maximizing platform cares about such correlation, because it can have profound consequences on its informational advantage through the composition of its database.

To recap, if given the option to acquire better data about records of type ω by independently drawing more precise observations about each record, the platform always has a non-negative willingness to pay for such information, which may be decreasing in the scope of learning (Corollary 5). However, she may strictly prefer to not acquire better data that involves high correlation across type-ω records. In practice, for large databases it may be more reasonable that acquiring better data takes the form of random draws from a large population. The conceptual point remains that, unlike for decision-makers, for intermediaries information can have negative value.

5 Discussion

Our framework and results apply more broadly to any setting where a principal mediates interactions between multiple agents using data. We briefly explain this applicability here.

Principal's Actions and Agents' Data. For ease of exposition, we simplified the model in several ways. Neither changes the analysis or its interpretations. First, we can allow the principal to also choose an action $a_0 \in A_0$ for each mediated interaction. In this case, a mechanism x also has to specify a_0 for each ω. Second, we can allow each agent i to also observe privately some own data about the interaction he is in. For example, in our leading e-commerce context each seller can observe the quality or the history of customer reviews of his product. We can again model the realizations of such data with some finite set Ω_i, where each ω_i is ultimately an exogenous signal about some underlying payoff-relevant θ. Let $\Omega = \Omega_0 \times \ldots \times \Omega_n$ with typical element $\omega = (\omega_0, \ldots, \omega_n)$. The key assumption is that the principal also observes the private data of each agent—i.e., the entire $\omega = (\omega_0, \ldots, \omega_n)$—as does the omniscient designer in Bergemann and Morris (2016). Thus, now the whole vector ω defines a type of data record in the principal's database and characterizes each interaction that she mediates. Our proofs in Appendix B already take this more general setting into account.

A Comment on Terminology. Each Ω_i is akin to what is often called the set of party i's types. We can then view an interaction as being characterized by the profile ω of all its participants' types. We instead use "type" to refer to ω itself because our analysis focuses on how the principal mediates interactions based on what she knows about their characterizing data *as a whole*. This use is also consistent with viewing ω_0 as the principal's type when she alone observes data, as this type is defined by what data characterizes the interaction she is mediating.

Simultaneous vs Sequential Mediation. We interpret the principal as mediating all interactions in her database simultaneously. For instance, in the introduction example the platform may mediate all buyer-seller interactions simultaneously, where the seller sets one price for each market segment created by the platform. An equally valid interpretation is that the principal commits to a mechanism for the whole database and then interactions are drawn independently and mediated one at a time. In our example, the platform may create the market segments once and for all, then draw a buyer one at a time and tell the seller to which segment he or she belongs. Depending on the application, one interpretation may fit better.

Another Example: Routing Games. Our leading example has been that of an e-commerce platform mediating the interactions between buyers and sellers. We conclude this discussion by sketching another applications of the model. A navigation app uses data about routes'

conditions to direct traffic by providing drivers with information—such as recommended routes and travel times. Das et al. (2017) propose a simple way to model this complex problem. Suppose the app (principal) seeks to minimize congestion. We can think of an interaction as consisting of a group of drivers (agents) in some city who simultaneously choose, say, one of two routes between its residential and business district. For each route, the travel time increases in how many drivers choose it but at different rates (e.g., because one is a highway and one is surface streets); travel times also depend on some uncertain event (e.g., construction work), which is observed only by the app. A database is then the collection of the realized events for all interactions across the cities served by the app. For simplicity, suppose each interaction happens in a different city so that they are independent in all respects, including their uncertain event. If the database is large, its composition q should reflect the primitive distribution of this event (i.e., the probability of construction on a given route).

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2021): "Too Much Data: Prices and Inefficiencies in Data Markets," *Americal Economic Journal: Microeconomics*, Forthcoming.

Acquisti, A., C. Taylor, and L. Wagman (2016): "The Economics of Privacy," *Journal of Economic Literature*, 54, 442–92.

Admati, A. and P. Pfleiderer (1990): "Direct and Indirect Sale of Information," *Econometrica*, 58, 901–28.

Admati, A. R. and P. Pfleiderer (1986): "A Monopolistic Market for Information," *Journal of Economic Theory*, 39, 400–438.

Ali, S. N., G. Lewis, and S. Vasserman (2020): "Voluntary Disclosure and Personalized Pricing," *arXiv preprint arXiv:1912.04774v2*.

Arrieta-Ibarra, I., L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl (2018): "Should We Treat Data as Labor? Moving beyond "Free"," *AEA Papers and Proceedings*, 108, 38–42.

Bergemann, D. and A. Bonatti (2015): "Selling Cookies," *American Economic Journal: Microeconomics*, 7, 259–94.

——— (2019): "Markets for Information: An Introduction," *Annual Review of Economics*, 11, 85–107.

Bergemann, D., A. Bonatti, and T. Gan (2021): "The Economics of Social Data," *arXiv preprint arXiv:2004.03107v2*.

Bergemann, D., A. Bonatti, and A. Smolin (2018): "The Design and Price of Information," *American Economic Review*, 108, 1–48.

Bergemann, D., B. Brooks, and S. Morris (2015): "The Limits of Price Discrimination," *American Economic Review*, 105 (3).

Bergemann, D. and S. Morris (2016): "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 11, 487–522.

——— (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57(1), pp. 44-95).

Bertsimas, D. and J. N. Tsitsiklis (1997): *Introduction to Linear Optimization*, Athena Scientific.

Brooks, B. and S. Du (2020): "A Strong Minimax Theorem for Informationally-Robust Auction Design," *Working Paper*.

——— (2021): "Optimal Auction Design with Common Values: An Informationally-Robust Approach," *Econometrica*, 89(3), 1313–1360.

Calzolari, G. and A. Pavan (2006): "On the Optimality of Privacy in Sequential Contracting," *Journal of Economic Theory*, 130, 168–204.

Crémer, J., Y.-A. de Montjoye, and H. Schweitzer (2019): "Competition policy for the digital era," *European Commission*.

Das, S., E. Kamenica, and R. Mirka (2017): "Reducing Congestion through Information Design," in *2017 55th annual allerton conference on communication, control, and computing (allerton)*, IEEE, 1279–1284.

Dizdar, D. and E. Kováč (2020): "A Simple Proof of Strong Duality in the Linear Persuasion Problem," *Games and Economic Behavior*, 122, 407–412.

Dorfman, R., P. A. Samuelson, and R. M. Solow (1987): *Linear Programming and Economic Analysis*, Courier Corporation.

Du, S. (2018): "Robust Mechanisms Under Common Valuation," *Econometrica*, 86(5), 1569–1588.

Dworczak, P. and A. Kolotilin (2019): "The Persuasion Duality," *Available at SSRN 3474376*.

Dworczak, P. and G. Martini (2019): "The Simple Economics of Optimal Persuasion," *Journal of Political Economy*, 127, 1993–2048.

Elliott, M., A. Galeotti, and A. Koh (2020): "Market Segmentation through Information," *Working Paper*.

Frankel, A. and E. Kamenica (2019): "Quantifying information and uncertainty," *American Economic Review*, 109, 3650–80.

Gale, D. (1989): *The Theory of Linear Economic Models*, University of Chicago press.

Galperti, S. and J. Perego (2018): "A Dual Perspective on Information Design," *Available at SSRN 3297406*.

Gentzkow, M. and E. Kamenica (2016): "Competition in Persuasion," *Review of Economic Studies*, 84, 300–322.

Goldberg, S., G. Johnson, and S. Shriver (2021): "Regulating privacy online: An economic evaluation of the GDPR," *Available at SSRN*.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590 – 2615.

Kolotilin, A. (2018): "Optimal Information Disclosure: A Linear Programming Approach," *Theoretical Economics*, 13, 607 – 635.

Lanier, J. (2013): *Who Owns the Future?*, Simon & Schuster.

Maskin, E. and J. Tirole (1990): "The Principal-Agent Relationship with an Informed Principal: The Case of Private Values," *Econometrica*, 58, 379–409.

Mathevet, L., J. Perego, and I. Taneva (2020): "On Information Design in Games," *Journal of Political Economy*, 128, 1370–1404.

Milgrom, P. and C. Shannon (1994): "Monotone Comparative Statics," *Econometrica*, 157–180.

Myerson, R. B. (1983): "Mechanism Design by an Informed Principal," *Econometrica*, 51, 1767–1797.

——— (1984): "Two-Person Bargaining Problems with Incomplete Information," *Econometrica*, 52, 461–488.

——— (1997): "Dual Reduction and Elementary Games," *Games and Economic Behavior*, 21, 183–202.

Nau, R. F. (1992): "Joint Coherence in Games of Incomplete Information," *Management Science*, 38, 374–387.

Nau, R. F. and K. F. McCardle (1990): "Coherent Behavior in Noncooperative Games," *Journal of Economic Theory*, 50, 424–444.

Posner, E. and E. G. Weyl (2018): *Radical Markets: Uprooting Capitalism and Democracy for a Just Society*, Princeton University Press.

Rockafellar, R. T. (1970): *Convex analysis*, vol. 36, Princeton university press.

Scott-Morton, F., P. Bouvier, A. Ezrachi, B. Julien, R. Katz, G. Kimmelman, A. Melamed, and J. Morgenstern (2019): "Report of the Committee for the study of digital platforms, market structure and antitrust subcommittee," *George Stigler Center for the study of the economy and the state, University of Chicago Booth School of Business*.

Appendix

A A Gambling Perspective on Data Value and Externalities

To better understand the value of data records and the externalities between them, it will help to provide a stand-alone interpretation of the data-value problem $\mathcal{V}_q$. With minor adjustments, this interpretation applies to problems where the sellers also observe some data and the platform takes some action. In this part, we will fix $q \in \mathbb{R}^{\Omega}_{++}$ and so drop it from notation.

For our interpretation, it is convenient to rewrite the data-value problem in the following equivalent way. Let $b = (b_1, \ldots, b_n)$ be a profile such that $b_i : A_i \to \mathbb{R}_{++}$ for all i and $\ell = (\ell_1, \ldots, \ell_n)$ be a profile such that $\ell_i : A_i \to \Delta(A_i)$ for all i. Given (b, ℓ), for each $i \in I$ and (a, ω) define

$$t_i(a, \omega) = b_i(a_i) \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \ell_i(a_i' | a_i)$$

and $t(a, \omega) = \sum_{i \in I} t_i(a, \omega)$. The data-value problem can be rewritten as

$$\begin{aligned} \mathcal{V}_q : \quad & \min_{v, b, \ell} \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ & \text{s.t. for all } \omega \in \Omega, \\ & v(\omega) = \max_{a \in A} \Big\{ u_0(a, \omega) + t(a, \omega) \Big\}, \end{aligned} \tag{8}$$

As in the main text, we will denote any optimal solution of $\mathcal{V}_q$ by (v_q^*, b_q^*, ℓ_q^*).

A.1 Gambles Against the Agents

Our interpretation hinges on unpacking how the platform determines the sellers' contributions to the externalities between buyers' records. The value formula (8) reveals that she does so through her choice of b and ℓ, which fully pin down each $t(a, \omega)$ and hence ultimately $v(\omega)$. Recall that the platform wants to *minimize* the values of her records, so she would like to lower $t(a, \omega) = \sum_{i \in I} t_i(a, \omega)$ as much as possible for all (a, ω). Each term of $t_i(a, \omega)$ takes the form

$$b_i(a_i) \ell_i(a_i' | a_i) \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right),$$

which contributes to lowering $t_i(a, \omega)$ if and only if $\ell_i(a_i' | a_i) > 0$ and $u_i(a_i, a_{-i}, \omega) < u_i(a_i', a_{-i}, \omega)$. In words, if seller i knew his interaction's type ω and his opponents' offers a_{-i},

he would strictly prefer a_i' to a_i. In this case, offering a_i amounts to making a mistake from an ex-post viewpoint. We will then say that seller i regrets offering a_i.

Thus, inducing sellers to play actions they will regret emerges as an intrinsic goal of the platform's problem—together with maximizing her payoff u_0 of course. In this view, (b_i, ℓ_i) and the corresponding t_i become an exploitation strategy on the part of the platform against seller i. In order to induce regrettable actions, she must withhold some information from seller i about ω or a_{-i}. This explains why the platform may prefer partial disclosure, from the perspective of her data-value problem. In the end, the value $v(\omega)$ results from a trade-off between the payoff $u_0(a, \omega)$ and the return from inducing sellers to choose regrettable actions.

This return depends on the structure of b and ℓ, which define a family of gambles against the sellers. To see this, fix any (a, ω) and seller i. Then, $\ell_i(\cdot|a_i) \in \Delta(A_i)$ defines a lottery over prizes, where for each a_i' the prize is $u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)$ for the platform; the scaling term $b_i(a_i)$ defines the stakes that she bets on this lottery. Given her objective, the platform "wins" when $u_i(a_i, a_{-i}, \omega) < u_i(a_i', a_{-i}, \omega)$ and "loses" otherwise. Thus, we can interpret $t(a, \omega)$ as the overall expected prize from (b, ℓ). We can then think of $\mathcal{V}$ as a fictitious environment where money is a medium of exchange and the platform can write monetary gambling contracts with each seller. Such contracts are enforced through contingent-claim markets that determine prizes based on the interaction's type ω and outcome a.[23]

We can then link the externalities between buyers' records with how the platform chooses these gambles in $\mathcal{V}$. Negative externalities $t^* < 0$ correspond to gambles favorable to the platform, in the sense that wins exceed losses in expectation. This requires the help of other records to withhold information and thus induce the sellers to play actions they will regret. Conversely, positive externalities $t^* > 0$ correspond to gambles unfavorable to the platform, in the sense that losses exceed wins in expectation. Corollary 1 implies that, at the optimum, the platform chooses gambles that favor her for some records, but not for others. In fact, this stems from deeper constraints and trade-offs in the use of such gambles against the sellers.

A.2 Feasible Gambles and Trade-offs

The gambles the platform can use to exploit the sellers in $\mathcal{V}$ have specific features that help us understand the nature of the data-value problem.

Some of these features reflect structural properties of $\mathcal{V}$. While the prizes of each gamble are contingent on both ω and the entire a, for each seller i both b_i and ℓ_i can depend only of his a_i.

[23] See Nau (1992) for a related interpretation.

This constrains the platform's ability to tailor her gambles with each seller across records. These properties reflect in $\mathcal{V}$ key interdependences in $\mathcal{U}$: The independence of (b_i, ℓ_i) from a_{-i} reflects the interdependence in $\mathcal{U}$ between sellers' incentives; the independence of (b_i, ℓ_i) from ω reflects the non-separability of $\mathcal{U}$ across data records. To see this, suppose $\ell_i(\hat{a}_i|a_i) > 0$. Then, (b_i, ℓ_i) links the right-hand side of the value formula (8) for (a_i, a_{-i}, ω) and (a_i, a'_{-i}, ω'). In particular, if $u_i(a_i, a_{-i}, \omega) < u_i(\hat{a}_i, a_{-i}, \omega)$ but $u_i(a_i, a'_{-i}, \omega') > u_i(\hat{a}_i, a'_{-i}, \omega')$, the platform faces a trade-off in determining v, as she may not be able to use (b_i, ℓ_i) to lower $v(\omega)$ without also raising $v(\omega')$. This is another way to see why and how externalities arise between records. When committing to (b, ℓ) the platform has to take into account these effects of each (b_i, ℓ_i) across records. How she solves these trade-offs depends on the relative frequency of records in the database (hence q). Importantly, this transformation of non-separabilities in $\mathcal{U}$ into independence properties of (b, ℓ) is what enables $\mathcal{V}$ to assign values individually to each record.

In fact, the platform faces other restrictions in her ability to *jointly* exploit the sellers. It is intuitive that she would want to design (b, ℓ) so that $t(a, \omega) \leq 0$ for all (a, ω) with some strict inequality. This would guarantee a sure arbitrage against the sellers. Such gambles, however, are infeasible in the following sense. Recall that by complementary slackness $x^*(a|\omega) > 0$ implies $v^*(\omega) = u_0(a, \omega) + t^*(a, \omega)$. Thus, since every ω must induce some action profile for every x, action profiles that cannot be in the support of any obedient $x(\cdot|\omega)$ are irrelevant for determining $v^*(\omega)$. Given this, define

$$\mathbf{X} = \{(a, \omega) \in A \times \Omega : x(a|\omega) > 0 \text{ for some obedient } x\}.$$

Let $G(\mathbf{X})$ be the set of gambles that can be contingent only on pairs $(a, \omega) \in \mathbf{X}$ (formally, we restrict the functions b and ℓ to the subdomain $\mathbf{X}$). Note that restricting the platform to choosing from $G(\mathbf{X})$ in $\mathcal{V}$ is immaterial for its optimal solution, in the same way that restricting x to the domain $\mathbf{X}$ is immaterial in $\mathcal{U}$.

Proposition 7. *For every gamble $(b, \ell) \in G(\mathbf{X})$, if $t(a, \omega) < 0$ for some (a, ω), there must exist (a', ω') such that $t(a', \omega') > 0$.*

This property is closely related to a similar result in Nau (1992). For completeness Appendix B provides a proof, which relies on a dual characterization of $\mathbf{X}$ using Farkas' lemma.

The economic takeaway is that in her attempt to minimize values v by exploiting the sellers with (b, ℓ), the platform faces a fundamental trade-off, which is a hallmark of problem $\mathcal{V}$. Successfully exploiting the sellers for records of type ω with some outcome a requires paying

the cost of losing against them for records of some other type ω' or outcome a'. Note that this result is stronger than Corollary 1, as it refers to the deep structure of data-value problems for intermediaries. It also sheds light on how and how much they can actually manipulate sellers by conveying information.

B Proofs

All proofs in this appendix are for the general case where the agents observe private data in the form of $\omega_i \in \Omega_i$ and hence $\omega = (\omega_0, \omega_1, \ldots, \omega_n)$ (see Section 5). The special case where only the principal observes data obtains by having $|\Omega_i| = 1$ for all $i \in I$.

Proof of Lemma 1. We will formulate the problem directly in terms of choosing a measure $\chi \in \mathbb{R}_+^{A\times\Omega}$. Formally, the problem is

$$\begin{aligned}
\mathcal{U}_q : \quad & \max_{\chi} \sum_{\omega\in\Omega, a\in A} u_0(a,\omega)\chi(a,\omega) \\
& \text{s.t. for all } i \in I, \omega_i \in \Omega_i, \text{ and } a_i, a_i' \in A_i, \\
& \sum_{\omega_{-i}\in\Omega_{-i}, a_{-i}\in A_{-i}} \left(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\right)\chi(a_i,a_{-i},\omega) \geq 0, \qquad (9) \\
& \text{and for all } \omega \in \Omega, \qquad (10) \\
& \sum_{a\in A} \chi(a,\omega) = q(\omega).
\end{aligned}$$

It is convenient to express this problem in matrix form. Fix an arbitrary total ordering of the set $A \times \Omega$. We denote by $\mathbf{u}_0 \in \mathbb{R}^{A\times\Omega}$ the vector whose entry corresponding to (a,ω) is $u_0(a,\omega)$. For every player i, let $\mathbf{U}_i \in \mathbb{R}^{(A_i\times A_i\times\Omega_i)\times(A\times\Omega)}$ be a matrix thus defined: For each row $(a_i', a_i'', \omega_i') \in A_i \times A_i \times \Omega_i$ and column $(a,\omega) \in A \times \Omega$, let the corresponding entry be

$$\mathbf{U}_i((a_i', a_i'', \omega_i'), (a,\omega)) = \begin{cases} u_i(a_i', a_{-i}, \omega) - u_i(a_i'', a_{-i}, \omega) & \text{if } a_i' = a_i, \omega_i' = \omega_i \\ 0 & \text{else.} \end{cases}$$

Thus, $\mathbf{U}_i(a_i', a_i'', \omega_i')$ denotes the row labeled by (a_i', a_i'', ω_i') (which defines the corresponding obedience constraint) and $\mathbf{U}_i(a,\omega)$ denotes the column labeled by (a,ω). Define the matrix $\mathbf{U}$ by stacking all the matrices $\{\mathbf{U}_i\}_{i\in I}$ on top each other. Finally, define the indicator matrix $I \in \{0,1\}^{\Omega\times(A\times\Omega)}$ such that, for each row ω' and column (a,ω'),

$$I(\omega', (a,\omega)) := \begin{cases} 1 & \text{if } \omega' = \omega \\ 0 & \text{else.} \end{cases}$$

With this notation and treating q as a vector, $\mathcal{U}_q$ can be written as follows:

$$\begin{array}{lrcl} \max\limits_{\chi} & \mathbf{u}_0^T \chi & & \\ \text{s.t.} & \mathbf{U}\chi & \geq & \mathbf{0}, \\ & I\chi & = & q, \\ & \chi & \geq & \mathbf{0}. \end{array} \tag{11}$$

Given this, by standard linear-programming arguments (Bertsimas and Tsitsiklis (1997)) the dual of $\mathcal{U}_q$ can be written as

$$\min_{\lambda, v} \mathbf{0}^T \lambda + q^T v$$

subject to for all $i = 1, \ldots, n$, $a_i, a_i' \in A_i$, and $\omega_i \in \Omega_i$,

$$\lambda_i(a_i'|a_i, \omega_i) \geq 0,$$

$v(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$ (i.e., it is unconstrained), and for all $(a, \omega) \in A \times \Omega$

$$u_0(a, \omega) \leq v(\omega) - \sum_{i \in I} \left\{ \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \lambda_i(a_i'|a_i, \omega_i) \right\}.$$

The objective simplifies to

$$\min_{\lambda, v} \sum_{\omega \in \Omega} v(\omega) q(\omega).$$

The second set of constraints can be written as

$$v(\omega) \geq u_0(a, \omega) + \sum_{i \in I} \left\{ \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \lambda_i(a_i'|a_i, \omega_i) \right\}.$$

Finally, we express this dual in a form that is equivalent to $\mathcal{V}_q$ by exploiting the structure of the specific problem at hand. To this end, for every i and ω_i we can set $\lambda_i(a_i|a_i, \omega_i) = 1$ (or any strictly positive value) for all $a_i \in A_i$. Given this, for every i and $(a_i, \omega_i) \in A_i \times \Omega_i$, define

$$b_i(a_i, \omega_i) = \sum_{a_i' \in A_i} \lambda_i(a_i'|a_i, \omega_i),$$

which is strictly positive by construction. Also, for every i and $(a_i', a_i, \omega_i) \in A_i \times A_i \times \Omega_i$ define

$$\ell_i(a_i'|a_i, \omega_i) = \frac{\lambda_i(a_i'|a_i, \omega_i)}{b_i(a_i, \omega_i)},$$

which implies that $\ell_i(\cdot|a_i, \omega_i) \in \Delta(A_i)$. Letting $b = (b_1, \ldots, b_n)$ and $\ell = (\ell_1, \ldots, \ell_n)$ so defined, we have the dual of $\mathcal{U}_q$ is equivalent to

$$\min_{v, b, \ell} \sum_{\omega \in \Omega} v(\omega) q(\omega)$$

subject to for all (a, ω)

$$v(\omega) \geq u_0(a, \omega) + \sum_{i \in I} t_i(a, \omega),$$

where

$$t_i(a, \omega) = b_i(a_i, \omega_i) \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \ell_i(a_i' | a_i, \omega_i).$$

Since for every $\omega \in \Omega$ this constraint has to hold for all $a \in A$ and the data-value problem is the minimization problem, we conclude that each $v(\omega)$ has to satisfy

$$v(\omega) = \max_{a \in A} \{ u_0(a, \omega) + t(a, \omega) \},$$

where $t(a, \omega) = \sum_{i \in I} t_i(a, \omega)$. Thus, we obtain problem $\mathcal{V}_q$. □

Remark 1. *We can transform $\mathcal{U}_q$ to the standard form $\mathcal{U}_q^S$ which can be written as follows:*

$$\begin{array}{lrcl} \max\limits_{\chi, s} & \mathbf{u}_0 \chi & & \\ s.t. & \mathbf{U}\chi - s & = & \mathbf{0}, \\ & I\chi & = & q, \\ & \chi, s & \geq & \mathbf{0}, \end{array} \tag{12}$$

where each $s_i(a_i' | a_i, \omega_i)$ is a nonnegative slack variable. The dual of $\mathcal{U}_q^S$ coincides with the data-value problem $\mathcal{V}_q$. Note that $\mathcal{U}_q$ always has an optimal solution χ_q^*, which is generically unique and hence corresponds to an extreme point of the polyhedron of feasible χ. Moreover, this χ_q^* is an optimal solution of $\mathcal{U}_q^S$ as well. The extreme point χ_q^* is nondegenerate by Assumption 1 and characterized by a square nonsingular active-constraint submatrix $\mathbf{B}$ consisting of linearly independent rows of the stacked matrix $\left[\begin{array}{c|c} \mathbf{U} & -\mathbf{1} \\ \hline I & \mathbf{0} \end{array}\right]$, where $\mathbf{1}$ is the identity matrix. As illustrated in Chapter 4 of Bertsimas and Tsitsiklis (1997), given $\mathbf{B}$, we have

$$\left[\begin{array}{c} \chi_q^* \\ \hline s_q^* \end{array}\right] = \mathbf{B}^{-1} \left[\begin{array}{c} \mathbf{0} \\ \hline q \end{array}\right], \tag{13}$$

where s_q^* is the vector of optimal slack variables in $\mathcal{U}_q^S$. A corresponding solution of $\mathcal{V}_q$ is given by

$$\left[\begin{array}{c} v_q^* \\ \hline \lambda_q^* \end{array}\right] = \mathbf{u}_0 \mathbf{B}^{-1}. \tag{14}$$

It follows that as long as the optimal solutions of $\mathcal{U}_q$ and $\mathcal{V}_q$ are defined by the same extreme point defined by $\mathbf{B}$, χ_q^* varies with q, but (v_q^*, λ_q^*) does not.

Proof of Lemma 2. Fix an optimal solution (v_q^*, b_q^*, ℓ_q^*) of $\mathcal{V}_q$. For every q, $\omega \in \Omega$, and $x(\cdot|\omega) \in CE(\Gamma_\omega)$, by (3) we have

$$\begin{aligned}
v_q^*(\omega) &\geq \sum_{a\in A} u_0(a,\omega)x(a|\omega) + \sum_{a\in A} t(a,\omega)x(a|\omega) \\
&= \sum_{a\in A} u_0(a,\omega)x(a|\omega) \\
&\quad + \sum_{a\in A} \left\{ \sum_{i\in I} b_i^*(a_i,\omega_i) \sum_{\hat{a}_i\in A_i} \left(u_i(a_i,a_{-i},\omega) - u_i(\hat{a}_i,a_{-i},\omega)\right) \ell_i^*(\hat{a}_i|a_i,\omega_i) \right\} x(a|\omega) \\
&= \sum_{a\in A} u_0(a,\omega)x(a|\omega) \\
&\quad + \sum_{i\in I} \sum_{a_i,\hat{a}_i\in A_i} b_i^*(a_i,\omega_i)\ell_i^*(\hat{a}_i|a_i,\omega_i) \left\{ \sum_{a_{-i}\in A_{-i}} \left(u_i(a_i,a_{-i},\omega) - u_i(\hat{a}_i,a_{-i},\omega)\right) x(a|\omega) \right\} \\
&\geq \sum_{a\in A} u_0(a,\omega)x(a|\omega),
\end{aligned}$$

where the last inequality follows because any $x(\cdot|\omega) \in CE(\Gamma_\omega)$ is defined by the property that, for all $i \in I$ and $a_i, a_i' \in A_i$,

$$\sum_{a_{-i}\in A_{-i}} \left(u_i(a_i,a_{-i},\omega) - u_i(a_i',a_{-i},\omega)\right) x(a_i,a_{-i}|\omega) \geq 0.$$

Since $x(\cdot|\omega)$ is an arbitrary element of $CE(\Gamma_\omega)$, we conclude that $v_q^*(\omega) \geq \bar{u}(\omega)$. □

Proof of Proposition 1. By complementary slackness, $x_q^*(a,\omega) > 0$ implies $v_q^*(\omega) = u_0(a,\omega) + t_q^*(a,\omega)$. Hence,

$$v_q^*(\omega) = \sum_{a\in A} u_0(a,\omega)x_q^*(a|\omega) + \sum_{a\in A} t_q^*(a,\omega)x_q^*(a|\omega) = u_q^*(\omega) + t_q^*(\omega).$$

Suppose we start from database q, with $q(\omega) > 0$, and we increase the quantity of ω-datapoints from $q(\omega)$ to $\hat{q}(\omega)$, thus obtaining the database $\hat{q}$. We can write

$$U^*(\hat{q}) - U^*(q) = u_{\hat{q}}^*(\omega)[\hat{q}(\omega) - q(\omega)] + \sum_{\omega'\in\Omega} [u_{\hat{q}}^*(\omega') - u_q^*(\omega')]\hat{q}(\omega')$$

Dividing both sides by $\hat{q}(\omega) - q(\omega)$, taking limits as $\hat{q}(\omega) \to q(\omega)$, and using Lemma 1, we obtain that

$$\begin{aligned}
t_q^*(\omega) &= v_q^*(\omega) - u_q^*(\omega) = \frac{\partial U^*(q)}{\partial q(\omega)} - u_q^*(\omega) \\
&= \lim_{\hat{q}(\omega)\to q(\omega)} \frac{\sum_{\omega'\in\Omega}[u_{\hat{q}}^*(\omega') - u_q^*(\omega')]\hat{q}(\omega')}{\hat{q}(\omega) - q(\omega)} = \sum_{\omega'\in\Omega} \frac{\partial u_q^*(\omega')}{\partial q(\omega)} q(\omega')
\end{aligned}$$

$$= \sum_{\omega' \in \Omega, a \in A} u_0(a, \omega') \left(\lim_{\hat{q}(\omega) \to q(\omega)} \frac{[x^*_{\hat{q}}(a|\omega') - x^*_q(a|\omega')]}{\hat{q}(\omega) - q(\omega)} \right) \hat{q}(\omega') =$$

$$= \sum_{\omega' \in \Omega, a \in A} u_0(a, \omega') \frac{\partial x^*_q(a|\omega')}{\partial q(\omega)} q(\omega'),$$

where the existence of the derivative $\frac{\partial x^*_q(a|\omega')}{\partial q(\omega)}$ almost everywhere follows from (13). □

Proof Proposition 2 First, note that we can write $u_0(a_1, \omega) = \pi a_1 \mathbb{I}\{\omega \geq a_1\} + (1 - \pi) \max\{\omega - a_1, 0\}$ as

$$[a(2\pi - 1) + (1 - \pi)\omega]\mathbb{I}\{\omega \geq a\},$$

which is strictly increasing in a if and only if $\pi > \frac{1}{2}$. Let $\bar{x}^*$ be the profit-maximizing solution (i.e., for $\pi = 1$) and $\underline{x}^*$ be the surplus-maximizing solution (i.e., for $\pi = 0$).

Lemma 3. *$\bar{x}^*$ is optimal for all $\pi \geq \frac{1}{2}$ and $\underline{x}^*$ is optimal for all $\pi \leq \frac{1}{2}$.*

Proof. Fix any (non-trivial) q and $\pi \in (0, 1)$. Problem $\mathcal{U}_q$ involves maximizing

$$\sum_{\omega, a} u_\pi(a, \omega) x(a|\omega) q(\omega) = \sum_{\omega \geq a} [a(2\pi - 1) + (1 - \pi)\omega] x(a|\omega) q(\omega)$$

subject to constraints (1).

Suppose that $\pi > \frac{1}{2}$. Note that $\bar{x}^*$ is feasible and maximizes the objective function pointwise for every ω. Indeed, since $\bar{x}^*(\omega|\omega) = 1$, for every ω we have that $\bar{x}^*$ selects the highest $a \leq \omega$ for every ω, thereby maximizing $a(2\pi - 1)\mathbb{I}\{\omega \geq a\}$; it also maximizes $\sum_{a \leq \omega} \omega x(a|\omega)$ for every ω. We can invoke the Theorem of the Maximum to extend the optimality of $\bar{x}^*$ at $\pi = \frac{1}{2}$.

Suppose now that $\pi < \frac{1}{2}$. Now for each ω the objective is to pair ω with the smallest possible a and do so with the highest probability allowed by (1). This is what $\underline{x}^*$ essentially does. We can again invoke the Theorem of the Maximum to extend the optimality of $\underline{x}^*$ at $\pi = \frac{1}{2}$. □

We now derive the expression of $v^*_q(\omega)$ in the statement of the proposition. The case of $\pi \geq \frac{1}{2}$ follows immediately from the fact that $\bar{x}^*$ is full disclosure. Now suppose $\pi < \frac{1}{2}$. We will construct a candidate v^*_q and prove it solves $\mathcal{V}_q$ using strong duality. First, under $\underline{x}^*$ we have

$$\begin{aligned} U^*(q) &= \sum_{\omega, a} [\pi u_1(a, \omega) + (1 - \pi) u_0(a, \omega)] \underline{x}^*(a|\omega) q(\omega) \\ &= \pi \sum_{\omega, a} a \mathbb{I}\{\omega \geq a\} \underline{x}^*(a|\omega) q(\omega) \\ &\quad + (1 - \pi) \left[\sum_{\omega < a_q} \omega q(\omega) + \sum_{\omega \geq a_q} (\omega - a_q) q(\omega) \right]. \end{aligned}$$

Note that

$$\sum_{\omega,a} a\mathbb{I}\{\omega \geq a\}\underline{x}^*(a|\omega)q(\omega) = a_q \sum_{\omega \geq a_q} q(\omega),$$

because the left-hand side is the seller's expected profits under $\underline{x}^*$, which by construction equal to the expected profit from the fixed uninformed price a_q. Therefore, we can write

$$\begin{aligned} U^*(q) &= \pi a_q \sum_{\omega \geq a_q} q(\omega) + (1-\pi)\left[\sum_{\omega < a_q} \omega q(\omega) + \sum_{\omega \geq a_q} (\omega - a_q)q(\omega)\right] \\ &= (2\pi - 1)a_q \sum_{\omega \geq a_q} q(\omega) + (1-\pi)\sum_{\omega} \omega q(\omega). \end{aligned}$$

Now we construct (v_q^*, λ_q^*), we show that it satisfies all dual constraints and that it yields $\sum_\omega v_q^*(\omega)q(\omega) = U^*(q)$, which proves that (v_q^*, λ_q^*) is optimal by strong duality. Recall that, in general, for all (a, ω) the dual constraint reads as

$$v(\omega) \geq u_\pi(a, \omega) + \sum_{a'} [u_1(a, \omega) - u_1(a', \omega)]\lambda(a'|a).$$

Let $\lambda_q^*(a'|a) = 0$ for all $a' \neq a_q$. Let $\lambda_q^*(a_q|a) = 1 - 2\pi$ for all $a \in \operatorname{supp} \underline{x}(\cdot|\omega)$ for some ω and $\lambda_q^*(a_q|a) = 0$ otherwise. Given this, for $\omega < a_q$, the right-hand side of the dual constraint equals

$$\begin{cases} \pi a + (1-\pi)(\omega - a) + a\lambda_q^*(a_q|a) & \text{if } a \leq \omega \\ 0 & \text{if } a > \omega. \end{cases}$$

Given $\lambda_q^*(a_q|a)$, the first line always equals $(1-\pi)\omega > 0$. Therefore, for $\omega < a_q$ define

$$v_q^*(\omega) = (1-\pi)\omega.$$

For $\omega \geq a_q$, the right-hand side of the dual constraint equals

$$\begin{cases} \pi a + (1-\pi)(\omega - a) + (a - a_q)\lambda_q^*(a_q|a) & \text{if } a \leq \omega \\ -a_q\lambda_q^*(a_q|a) & \text{if } a > \omega. \end{cases}$$

Given $\lambda_q^*(a_q|a)$, the first line always equals

$$(2\pi - 1)a_q + (1-\pi)\omega = \pi a_q + (1-\pi)(\omega - a_q) > 0.$$

Therefore, for $\omega \geq a_q$ define

$$v_q^*(\omega) = (2\pi - 1)a_q + (1-\pi)\omega.$$

Note that by construction v_q^* satisfies all dual constraint and $\sum_\omega v_q^*(\omega)q(\omega) = U^*(q)$, as desired.

It follows immediately that for $\pi < \frac{1}{2}$ we have $t_q^*(\omega) > 0$ for $\omega < a_q$ and $t_q^*(\omega) \leq 0$ for $\omega \geq a_q$. □

Proof Proposition 3. By the formulation of $\mathcal{V}_q$ and Lemma 2, the polyhedron of feasible solutions of $\mathcal{V}_q$, denoted by $F(\mathcal{V}_q)$ does not contain a line because all dual variables are bounded from below. By Theorem 2.6 in Bertsimas and Tsitsiklis (1997), $F(\mathcal{V}_q)$ has at least one extreme point and at most finitely many of them by Corollary 2.1 in Bertsimas and Tsitsiklis (1997). By Theorem 4.4 in Bertsimas and Tsitsiklis (1997), $\mathcal{V}_q$ has at least one optimal solution. By Theorem 2.7 in Bertsimas and Tsitsiklis (1997), we can focus on solutions that are extreme points of $F(\mathcal{V}_q)$.

Fix q and suppose that the optimal solution (v_q^*, λ_q^*) of the dual of $\mathcal{U}_q$ is unique. As explained in Remark 1, there exists a submatrix $\mathbf{B}$ such that (v_q^*, λ_q^*) satisfies (14). Given Assumption 1, Theorem 3.1 and Exercise 3.6 in Bertsimas and Tsitsiklis (1997) imply that

$$\left[\begin{array}{c|c} \mathbf{U} & -\mathbf{1} \\ \hline I & \mathbf{0} \end{array}\right] \mathbf{B}^{-1} \begin{bmatrix} \mathbf{0} \\ q \end{bmatrix} \geq \begin{bmatrix} \mathbf{0} \\ q \end{bmatrix}.$$

The inequality is strict for each row of $\mathbf{U}$ that corresponds to $\lambda_{q,i}^*(a_i'|a_i,\omega_i) = 0$ (or, equivalently, $\ell_{q,i}^*(a_i'|a_i,\omega_i) = 0$):

$$\left[\mathbf{U}_i(a_i,a_i',\omega_i) \mid -\mathbf{1}_i(a_i,a_i',\omega_i)\right]\mathbf{B}^{-1}\begin{bmatrix}\mathbf{0}\\ q\end{bmatrix} > 0, \tag{15}$$

where $\mathbf{1}_i(a_i,a_i',\omega_i)$ is the row of the identity matrix $\mathbf{1}$ that corresponds to (i,a_i,a_i',ω_i). Note that for each row ω of the indicator matrix I (i.e., $I(\omega)$), which corresponds to variable $v_q^*(\omega)$, it automatically holds that $\left[I(\omega) \mid \mathbf{0}\right]\mathbf{B}^{-1}\begin{bmatrix}\mathbf{0}\\ q\end{bmatrix} = q(\omega)$. Similarly, for each row of $\mathbf{U}$ that corresponds to $\lambda_{q,i}^*(a_i'|a_i,\omega_i) > 0$ (or, equivalently, $\ell_{q,i}^*(a_i'|a_i,\omega_i) > 0$), it holds that $\left[\mathbf{U}_i(a_i,a_i',\omega_i) \mid -\mathbf{1}_i(a_i,a_i',\omega_i)\right]\mathbf{B}^{-1}\begin{bmatrix}\mathbf{0}\\ q\end{bmatrix} = 0$ as long as $\mathbf{B}$ identifies the optimal extreme point.

Now consider changes in q and note that it only enters the objective of $\mathcal{V}_q$. Each condition (15) defines an open set of q's in $\mathbb{R}_+^\Omega$ that satisfy it. Define $(v_{\mathbf{B}}^*, \lambda_{\mathbf{B}}^*)$ identified by $\mathbf{B}$ as in (14) and

$$Q(\mathbf{B}) = \{q : (15) \text{ holds for all } i \in I \text{ and } (a_i,a_i',\omega_i) \text{ s.t. } \lambda_{\mathbf{B},i}^*(a_i|a_i',\omega_i) = 0\}.$$

Note that $Q(\mathbf{B})$ is an open set because it is the intersection of finitely many open sets.

Now recall that there are only finitely many extreme points of the dual polyhedron of feasible solutions. Therefore, there are finitely many submatrices $\{\mathbf{B}_1,\ldots,\mathbf{B}_K\}$ such that each identi-

fies an optimal $(v^*_{\mathbf{B}_k}, \lambda^*_{\mathbf{B}_k})$ that is unique for all $q \in Q(\mathbf{B}_k)$. For all $k = 1, \ldots, K$, define $Q_k = Q(\mathbf{B}_k)$. By construction, each Q_k is open and $q, q' \in Q_k$ implies that $(v^*_q, \lambda^*_q) = (v^*_{q'}, \lambda^*_{q'})$. Since (v^*_q, λ^*_q) is generically unique with respect to q, it follows that $\mathbb{R}^\Omega_+ \setminus \cup_k Q_k$ has Lebesgue measure zero. □

Proof of Proposition 4. Fix $\mu_1, \mu_2 \in \Delta(\Omega)$. Let $\Omega^i = \{\omega \in \Omega : \mu_i(\omega) > \mu_j(\omega), j \neq i\}$, $i \in \{1, 2\}$, and $\Omega^3 = \Omega \setminus \Omega_1 \setminus \Omega_2$.

Let $X = \mathbb{R}^\Omega \times \mathbb{R}^{A_1 \times A_1}_+ \times \ldots \times \mathbb{R}^{A_n \times A_n}_+$. Associate the canonical component-wise order with X, with an exception that the order is reversed for $\omega \in \Omega^1$. X is a lattice, with a typical element (v, λ), where $v \in \mathbb{R}^\Omega$ and $\lambda \in \mathbb{R}^{A_1 \times A_1}_+ \times \ldots \times \mathbb{R}^{A_n \times A_n}_+$.

The data-value problem is equivalent to the problem $\max_{(v,\lambda) \in S} f(v, \lambda; \mu)$, where $f(v, \lambda; \mu) = -\sum_{\omega \in \Omega} v(\omega)\mu(\omega)$ and the feasible set $S \subset X$ is given by the inequalities

$$v(\omega) \geq u_0(a, \omega) + \sum_{i \in I} \sum_{a'_i \in A_i} (u_i(a_i, a_{-i}, \omega) - u_i(a'_i, a_{-i}, \omega))\lambda_i(a'_i | a_i).$$

We treat μ as a parameter. Note that S does not depend on μ. Furthermore, μ is an element of $(|\Omega| - 1)$-dimensional simplex, with which we associate the following partial order: $\mu' \geq \mu$ if $\mu'(\omega) \geq \mu(\omega)$ for $\omega \in \Omega^1$, $\mu'(\omega) \leq \mu(\omega)$ for $\omega \in \Omega^2$, and $\mu'(\omega) = \mu(\omega)$ for $\omega \in \Omega^3$. Note that $\mu_1 \geq \mu_2$ in accordance with this partial order.

We want to show that f is supermodular in (v, λ) and has increasing differences in $(v, \lambda; \mu)$. Observe that

$$\begin{aligned} f(v', \lambda'; \mu) + f(v'', \lambda''; \mu) &= -\sum_{\omega \in \Omega} v'(\omega)\mu(\omega) - \sum_{\omega \in \Omega} v''(\omega)\mu(\omega) \\ &= -\sum_{\omega \in \Omega} (v'(\omega) + v''(\omega))\mu(\omega) \\ &= -\sum_{\omega \in \Omega} (\max\{v'(\omega), v''(\omega)\} + \min\{v'(\omega), v''(\omega)\})\mu(\omega) \\ &= f((v', \lambda') \wedge (v'', \lambda''); \mu) + f((v', \lambda') \vee (v'', \lambda''); \mu). \end{aligned}$$

Then f is supermodular in (v, λ).

Fix $(v', \lambda') \geq (v, \lambda)$ and $\mu' \geq \mu$. Observe that

$$\begin{aligned} & (f(v', \lambda', \mu') - f(v, \lambda, \mu')) - (f(v', \lambda', \mu) - f(v, \lambda, \mu)) \\ = & \sum_{\omega \in \Omega} (v(\omega) - v'(\omega))(\mu'(\omega) - \mu(\omega)) \\ = & \sum_{\omega \in \Omega^1} (v(\omega) - v'(\omega))(\mu'(\omega) - \mu(\omega)) + \sum_{\omega \in \Omega^2} (v(\omega) - v'(\omega))(\mu'(\omega) - \mu(\omega)) \geq 0, \end{aligned}$$

where the inequality follows from the adapted partial orders. Then f has increasing differences in $(v, \lambda; \mu)$.

Finally, by Theorem 5 in Milgrom and Shannon (1994), $\arg\max_{(v,\lambda)\in S} f(v, \lambda; \mu)$ is monotone nondecreasing in μ. This monotone comparative statics coupled with generic uniqueness of (v_q^*, b_q^*, ℓ_q^*) with respect to q imply that if $\mu_q(\omega) > \mu_{q'}(\omega)$ for two databases q and q' then $v_q^*(\omega) \leq v_{q'}^*(\omega)$.

When only interactions of type ω are present in the database, that is, $\mu_q(\omega) = 1$, we have $v_q^*(\omega) = \bar{u}(\omega)$. Indeed, the definition of $\bar{u}(\omega)$ implies that it can be written as

$$\bar{u}(\omega) = \min_{b_\omega, \ell_\omega} \max_{a\in A} \{u_0(a, \omega) + t_{b_\omega,\ell_\omega}(a, \omega)\},$$

where $t_{b_\omega,\ell_\omega}(a, \omega) = \sum_{i\in I} b_{i,\omega}(a_i) \sum_{a_i'\in A_i} (u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega))\ell_{i,\omega}(a_i'|a_i)$, $b_\omega = (b_{1,\omega}, \ldots, b_{n,\omega})$, with $b_{i,\omega} : A_i \to \mathbb{R}_{++}$, and $\ell_\omega = (\ell_{1,\omega}, \ldots, \ell_{n,\omega})$, with $\ell_{i,\omega} : A_i \to \Delta(A_i)$.

For $\varepsilon > 0$, consider a set $M_\varepsilon(\omega)$ defined as $M_\varepsilon(\omega) = \{\mu \in \Delta(\Omega) : \mu(\omega') \in (0, \varepsilon) \text{ for } \omega \neq \omega', \mu(\omega) < 1\}$. By Proposition 3, there exists a finite collection $\{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$ of open, convex, and disjoint subsets of $\Delta(\Omega)$ such that $\cup_k \mathcal{P}_k$ has measure one and, for every k, (v_q^*, b_q^*, ℓ_q^*) is unique and constant for q, with $\mu_q \in \mathcal{P}_k$. Therefore, we can always find $\mathcal{P}_m \in \{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$, such that $\mathcal{P}_m \cap M_\varepsilon(\omega)$ is nonempty, open, and convex for all $0 < \varepsilon \leq \delta$, where $\delta > 0$. Then $v_q^*(\omega)$ is unique and constant for all $q \in \mathbb{R}_{++}^\Omega$, with $\mu_q \in \mathcal{P}_m \cap M_\delta(\omega)$. Let us refer to this constant as $\hat{u}(\omega)$. If $\hat{u}(\omega) = \bar{u}(\omega)$, then the result follows. Suppose, on the contrary, that $\hat{u}(\omega) \neq \bar{u}(\omega)$. We can always pick a sequence μ^n, $n \in \mathbb{N}$, from $\mathcal{P}_m \cap M_\delta(\omega)$ that converges to $\tilde{\mu}$, with $\tilde{\mu}(\omega) = 1$. Then for every $n \in \mathbb{N}$, $v_q^*(\omega) = \hat{u}(\omega)$ for every q, such that $\mu_q = \mu^n$. By the Berge's maximum theorem, (v_q^*, b_q^*, ℓ_q^*) is an upper-hemicontinuous correspondence and therefore has closed graph. Hence, $\hat{u}(\omega) \in v_q^*(\omega)$ for every q, with $\mu_q = \tilde{\mu}$. We obtain the desired contradiction, since $v_q^*(\omega) = \bar{u}(\omega)$ for such q. □

Proof of Proposition 5. Fix $q \in \mathbb{R}_{++}^\Omega$. Suppose that a FIC mechanism x_q^* is optimal. Then, we have

$$v_q^*(\omega) = u_q^*(\omega) + \sum_{a\in A} t_q^*(a, \omega) x_q^*(a|\omega) \geq u_q^*(\omega),$$

where the inequality follows from $x_q^*(\cdot|\omega) \in CE(\Gamma_\omega)$ for all ω. Since by Lemma 1 we must have $\sum_\omega v_q^*(\omega)q(\omega) = \sum_\omega u_q^*(\omega)q(\omega)$, it follows that $v_q^*(\omega) = u_q^*(\omega)$ for all ω. Finally, since x_q^* is optimal, it must be that $u_q^*(\omega) = \bar{u}(\omega)$ for all ω. Now, note that v_q^* defines a supporting hyperplane of the iso-payoff line of level $U^*(q)$ at q. The intercept of such an hyperplane on each ω-axis is $\hat{q}_\omega(\omega) = \frac{U^*(q)}{\bar{u}(\omega)}$ and $\hat{q}_\omega(\omega') = 0$ for $\omega' \neq \omega$. By definition, each $\hat{q}_\omega$ also belongs to the iso-payoff line of level $U^*(q)$ and therefore $U^*(q) = U^*(\hat{q}_\omega)$ for all ω. In other words, the intercepts of the hyperplane and the iso-payoff line coincide for all ω.

Now consider any $q' \in \mathbb{R}_{++}$, $q' \neq q$, that belongs to the supporting hyperplane of level $U^*(q)$ at q. By definition, we can obtain q' as a convex combination of intercepts $\hat{q}_\omega$ on each axis. Specifically, there exists $\beta \in \Delta(\Omega)$ such that $q'(\omega) = \beta(\omega)\hat{q}_\omega(\omega)$ for all ω. By concavity of $U^*(q)$ (Footnote 17), we must have that

$$U^*(q') = \sum_{\omega\in\Omega} v^*_{q'}(\omega)q'(\omega) \leq U^*(q) = \sum_{\omega\in\Omega} \beta(\omega)U^*(\hat{q}_\omega) = \sum_{\omega\in\Omega} \bar{u}(\omega)q'(\omega).$$

But since $v^*_{q'}(\omega) \geq \bar{u}(\omega)$ for all ω by Lemma 2, we must have $v^*_{q'}(\omega) = \bar{u}(\omega)$ for all ω. Then $v^*_{q''}(\omega) = \bar{u}(\omega)$ for all q'' that belong to the supporting hyperplane of level $U^*(q)$ at q. Finally, since v^*_q is invariant to scaling of q, it follows that $v^*_q(\omega) = \bar{u}(\omega)$ for all ω and all $q \in \mathbb{R}^\Omega_+$. □

Proof of Corollary 4. We have $MRS_q(\omega,\omega') = -\frac{\omega}{\omega'}$ for $\pi > \frac{1}{2}$ and all ω,ω'. Consider now $\pi < \frac{1}{2}$:

$$MRS_q(\omega,\omega') = \begin{cases} -\frac{\omega}{\omega'} & \text{if } \omega,\omega' < a_q \\ -\frac{(1-\pi)\omega}{\pi a_q + (1-\pi)(\omega' - a_q)} & \text{if } \omega < a_q \leq \omega' \\ -\frac{\pi a_q + (1-\pi)(\omega - a_q)}{\pi a_q + (1-\pi)(\omega' - a_q)} & \text{if } \omega,\omega' \geq a_q. \end{cases}$$

Thus, we have

$$\frac{\partial MRS_q(\omega,\omega')}{\partial \pi} = \begin{cases} 0 & \text{if } \omega,\omega' < a_q \\ \frac{\omega a_q}{[\pi a_q + (1-\pi)(\omega' - a_q)]^2} & \text{if } \omega < a_q \leq \omega' \\ -\frac{a_q(\omega' - \omega)}{[\pi a_q + (1-\pi)(\omega' - a_q)]^2} & \text{if } \omega,\omega' \geq a_q. \end{cases}$$

Finally, it is easy to see that $MRS_q(\omega,\omega') < -\frac{\omega}{\omega'}$ for $\omega < a_q \leq \omega'$ and that $MRS_q(\omega,\omega') > -\frac{\omega}{\omega'}$ for $\omega' > \omega \geq a_q$. □

Proof of Proposition 6. To build intuition, imagine the principal refines one ω-datapoint according to σ_ω, which does not change q by being infinitesimal. Since $u_i(a,\omega) = \mathbb{E}_\sigma[u_i(a,\omega')|\omega]$ for all $i = 0,1,\ldots,n$, using expression (3), by complementary slackness we get

$$v^*_q(\omega) = \sum_{\omega'\in\Omega} [u_0(a^*_q(\omega),\omega') + t^*_q(a^*_q(\omega),\omega')]\sigma_\omega(\omega'),$$

where $a^*_q(\omega)$ is any action profile in the support of $x^*_q(\cdot|\omega)$. By (3) again, this implies that such a refinement increases the expected value of the refined ω-datapoint:

$$\sum_{\omega'\in\Omega} v^*_q(\omega')\sigma_\omega(\omega') - v^*_q(\omega) \geq 0. \tag{16}$$

Note that if refining $\alpha q(\omega)$ of the current ω-datapoints according to σ_ω does not change the value of datapoints, then (16) implies the desired inequality.

Now suppose that acquiring better data changes the value of datapoints. That is, there exists a share $\alpha > 0$ such that refining $\alpha q(\omega)$ of the current ω-datapoints according to σ_ω leads to a new database q_α such that $v^*_{q_\alpha}(\omega') \neq v^*_q(\omega')$ for some $\omega' \in \operatorname{supp} \sigma_\omega$ or $\omega' = \omega$. Since the total quantity of datapoints does not change, we have that $\mu_{q_\alpha}(\omega) < \mu_q(\omega)$ and $\mu_{q_\alpha}(\omega') > \mu_q(\omega')$ for all $\omega' \in \operatorname{supp} \sigma_\omega$. By Proposition 4, it follows that $v^*_{q_\alpha}(\omega) \geq v^*_q(\omega)$ and $v^*_{q_\alpha}(\omega') \leq v^*_q(\omega')$ for all $\omega' \in \operatorname{supp} \sigma_\omega$. Now, note that for all α,

$$\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') \geq v^*_{q_\alpha}(\omega) \geq v^*_q(\omega), \tag{17}$$

where the first inequality follows from (16). This implies that the value of acquiring better data is always non-negative.

Now, suppose that there exists a common $\tilde{a} \in \operatorname{supp} x^*_q(\cdot|\omega)$ that satisfies $x^*_q(\tilde{a}|\omega'') > 0$ for all $\omega'' \in \operatorname{supp} \sigma_\omega$. By complementary slackness, it follows that for all $\omega'' \in \operatorname{supp} \sigma_\omega$, we have $v^*_q(\omega'') = u_0(\tilde{a}, \omega'') + t^*_q(\tilde{a}, \omega'')$. Therefore, by the scarcity principle,

$$\sum_{\omega'' \in \Omega} v^*_{q_\alpha}(\omega'')\sigma_\omega(\omega'') \leq \sum_{\omega'' \in \Omega} v^*_q(\omega'')\sigma_\omega(\omega'') = v^*_q(\omega) \leq v^*_{q_\alpha}(\omega),$$

which, combined with (17), implies the desired equality.

Conversely, suppose that for every $\hat{a} \in \operatorname{supp} x^*_q(\cdot|\omega)$ there exists $\omega' \in \operatorname{supp} \sigma_\omega$ that satisfies $x^*_q(\hat{a}|\omega') = 0$. If the solution to the data-value problem is unique for database q, then $x^*_q(\hat{a}|\omega') = 0$ implies $v^*_q(\omega') > u_0(\hat{a}, \omega') + t^*_q(\hat{a}, \omega')$ by strict complementary slackness. The desired strict inequality is then obtained.

Proof of Corollary 5. The directional derivative of U^* at q along the linear path from q to q_α is equal to

$$q(\omega)\left[\sum_{\omega' \in \Omega} v^*_q(\omega')\sigma_\omega(\omega') - v^*_q(\omega)\right].$$

The linear path from q to q_α can be parametrized as follows: for $t \in [0,1]$, define $q_t(\omega) = q(\omega) - t\alpha q(\omega)$, $q_t(\omega') = q(\omega') + t\alpha\sigma_\omega(\omega')q(\omega)$ for $\omega' \in \operatorname{supp} \sigma_\omega$, and $q_t(\omega'') = q(\omega'')$ for remaining ω''.

Note that $\sum_{\omega' \in \Omega} v^*_{q_t}(\omega')\sigma_\omega(\omega') - v^*_{q_t}(\omega)$ is non-negative by (16) and decreasing in t by the scarcity principle.

Finally, by the gradient theorem,

$$U^*(q_\alpha) - U^*(q) = \int_0^1 v^*_{q_t} \cdot \nabla q_t \, dt = \alpha q(\omega) \int_0^1 \left[\sum_{\omega' \in \Omega} v^*_{q_t}(\omega')\sigma_\omega(\omega') - v^*_{q_t}(\omega)\right] dt \geq 0,$$

where ∇q_t is the gradient of q_t with respect to t. □

Proof of Proposition 7 . We provide a proof for the general case where the principal can choose $a_0 \in A_0$ and each agent i can privately observe some own data $\omega_i \in \Omega_i$ about the interaction he is in. Fix $(a^*, \omega^*) \in \mathbf{X}$ and introduce $\mathbf{1}_{a^*,\omega^*}$ as a vector of size $|\mathbf{X}|$ with $\varepsilon > 0$ in the position indexed by (a^*, ω^*) and 0 in all other positions. Constitute a matrix $\mathbf{W}$ such that its rows are indexed by $(a, \omega) \in \mathbf{X}$, its columns are indexed by (i, a_i', a_i, ω_i), $i \in I$, and its entries are as follows:

$$\mathbf{W}\left((\tilde{a}, \tilde{\omega}), (i, a_i', a_i, \omega_i)\right) = 1\{a_i = \tilde{a}_i, \omega_i = \tilde{\omega}_i\}\left(u_i(a_i, \tilde{a}_{-i}, \omega_i, \tilde{\omega}_{-i}) - u_i(a_i', \tilde{a}_{-i}, \omega_i, \tilde{\omega}_{-i})\right).$$

By a variant of the Farkas' lemma, either there exists $\lambda \geq 0$, such that $\mathbf{W}\lambda \leq -\mathbf{1}_{a^*,\omega^*}$, or else there exists $\chi \geq 0$, such that $\mathbf{W}^T\chi \geq 0$, with $\chi^T\mathbf{1}_{a^*,\omega^*} > 0$. Now we show that the latter is true. Indeed, we can pick $\chi(a, \omega) = q(\omega)x(a|\omega)$, where x is obedient and satisfies $x(a^*|\omega^*) > 0$. We can find such x, since $(a^*, \omega^*) \in \mathbf{X}$. Then $\chi \geq 0$ and $\chi^T\mathbf{1}_{a^*,\omega^*} > 0$ are satisfied automatically. Finally, $\mathbf{W}^T\chi \geq 0$ corresponds exactly to the set of obedience constraints in $\mathcal{U}_q$ restricted to the subdomain $\mathbf{X}$.

Since any λ can be decomposed as $\lambda_i(a_i'|a_i, \omega_i) = b_i(a_i, \omega_i)\ell_i(a_i'|a_i, \omega_i)$, we conclude that there is no $(b, \ell) \in G(\mathbf{X})$ that satisfies $t(a, \omega) \leq 0$ for every $(a, \omega) \in \mathbf{X}$ and $t(a^*, \omega^*) < -\varepsilon$. The result then follows, since the choice of $(a^*, \omega^*) \in \mathbf{X}$ and $\varepsilon > 0$ was arbitrary. □

Proof of Proposition 8. We will argue by contradiction. Suppose $q \in \mathbb{R}^{\Omega}_{++}$ and $\mathcal{U}_q$ admits an FIC solution x_q^{**} and hence $x_q^{**}(\cdot|\tilde{\omega}) \in CE(\Gamma_{\tilde{\omega}})$ and $u_q^{**}(\tilde{\omega}) = \bar{u}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$. Then $v_q^{**}(\tilde{\omega}) = u_q^{**}(\tilde{\omega}) = \bar{u}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$ by Proposition 5.

Now suppose that (a, ω) satisfies both conditions in the statement of the proposition. For $(v_q^{**}, b_q^{**}, \ell_q^{**})$ to be feasible for $\mathcal{V}_q$, we must have for all $\tilde{\omega} \in \Omega$,

$$v_q^{**}(\tilde{\omega}) \geq u_0(a, \tilde{\omega}) + t_q^{**}(a, \tilde{\omega}).$$

Since $u_0(a, \omega) > \bar{u}(\omega) = v^{**}(\omega)$, we must have $t_q^{**}(a, \omega) < 0$. Therefore, there exists a pair $(i, \hat{a}_i)$ that satisfies $u_i(a_i, a_{-i}, \omega) < u_i(\hat{a}_i, a_{-i}, \omega)$ and $\ell_{q,i}^{**}(\hat{a}_i|a_i, \omega_i) > 0$. For such a pair $(i, \hat{a}_i)$, there exists $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$ with the properties listed in the proposition. Then, since $b_q^{**} > 0$ and $\ell_{q,i}^{**}(\hat{a}_i|a_i, \omega_i) > 0$,

$$\begin{aligned}
&\sum_{\tilde{a} \in A} u_0(\tilde{a}, \omega')x(\tilde{a}|\omega') + \sum_{\tilde{a} \in A} t_q^{**}(\tilde{a}, \omega')x(\tilde{a}|\omega') \\
\geq\ &\sum_{\tilde{a} \in A} u_0(\tilde{a}, \omega')x(\tilde{a}|\omega') \\
&+ b_{q,i}^{**}(a_i, \omega_i)\ell_{q,i}^{**}(\hat{a}_i|a_i, \omega_i)\left\{\sum_{\tilde{a}_{-i} \in A_{-i}} \left(u_i(a_i, \tilde{a}_{-i}, \omega') - u_i(\hat{a}_i, \tilde{a}_{-i}, \omega')\right)x(a_i, \tilde{a}_{-i}|\omega')\right\}
\end{aligned}$$

$$> \sum_{\tilde{a}\in A} u_0(\tilde{a},\omega')x(\tilde{a}|\omega') = v_q^{**}(\omega'),$$

where the first inequality follows because $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$. The strict inequality is incompatible with constraint (3) and delivers the desired contradiction. □

C A Sufficient Condition for Suboptimality of Full Disclosure

We provide a sufficient condition on Γ for suboptimality of full disclosure for the general case where the principal can choose $a_0 \in A_0$ and each agent i can privately observe some own data $\omega_i \in \Omega_i$ about the interaction he is in. Recall that if the principal fully reveals all ω, then she must be implementing a correlated equilibrium of the complete-information game Γ_ω for all ω, i.e., $x_q^*(\cdot|\omega) \in CE(\Gamma_\omega)$. The definition of CE in terms of inequalities can be adjusted to incorporate the principal's a_0.

Proposition 8. *Fix Γ. Suppose there exists (a,ω) that satisfies:*

(1) $u_0(a,\omega) > \bar{u}(\omega)$,

(2) for every agent i and action $\hat{a}_i$, such that $u_i(a_i,a_{-i},\omega) < u_i(\hat{a}_i,a_{-i},\omega)$, there exists an $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$ for some ω', with $\omega'_i = \omega_i$, that satisfies

$$\sum_{a\in A} u_0(a,\omega')x(a|\omega') = \bar{u}(\omega'),$$

$$\sum_{a_{-i}\in A_{-i}} \left(u_i(a_i,a_{-i},\omega') - u_i(\hat{a}_i,a_{-i},\omega')\right) x(a_i,a_{-i}|\omega') > 0.$$

Then $\mathcal{U}_q$ does not admit an FIC solution for any $q \in \mathbb{R}^{\Omega}_{++}$.

Condition (1) is clearly necessary: If for every datapoint ω every action profile a cannot deliver a payoff higher than the full-information payoff $\bar{u}(\omega)$, then it is clearly optimal for the principal to fully reveal every ω. Given an outcome (a,ω) with $u_0(a,\omega) > \bar{u}(\omega)$, there must be an agent who would have a profitable deviation from a_i to $\hat{a}_i$ if he knew (a_{-i},ω_{-i}). Otherwise, given a_0, the profile a_{-0} is a Nash Equilibrium of Γ_ω and hence $a_{-0} \in CE(\Gamma_\omega)$, which would imply $u_0(a,\omega) \leq \bar{u}(\omega)$. Then condition (2) requires that agent i's data ω_i is consistent with another datapoint ω'—so that he cannot tell ω and ω' apart based on his own data only—which admits a principal-preferred correlated equilibrium that also recommends i to play a_i and renders the deviation to $\hat{a}_i$ strictly suboptimal.

D Analysis of the Leading Example

This section presents the calculations that back up our statements regarding the leading example. We can ignore the buyers and build their decisions into the utility functions of the surplus-maximizing platform ($i = 0$) and the seller ($i = 1$). There are three types of datapoints, labeled by $\omega \in \{\omega_L, \omega_H, \omega^\circ\}$, where $\omega_H > \omega_L > 0$, and corresponding to whether the buyer's revealed valuation is ω_L, ω_H, or unknown to the platform. Suppose ω° turns into ω_H with probability h and ω_L with probability $1-h$. The prices the seller can charge are $a \in \{\omega_L, \omega_H\}$. The payoffs are $u_0(a,\omega) = \max\{\omega - a, 0\}$, $u_1(a,\omega) = a$ if $a \leq \omega$, and $u_1(a,\omega) = 0$ if $a > \omega$. Given this, we have $u_i(a,\omega^\circ) = hu_i(a,\omega_H) + (1-h)u_i(a,\omega_L)$ for $i = 0,1$. For completeness, we solve both the information-design problem and the data-value problem separately. Since our goal here is to only find the optimizers in both problems, we can work in the space of databases that satisfy $q(\omega_L) + q(\omega_H) + q(\omega^\circ) = 1$, so that $q(\omega) = \mu_q(\omega)$.

D.1 Information-Design Problem

The objective of the platform is

$$(\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H) + h(\omega_H - \omega_L)x(\omega_L|\omega^\circ)\mu_q(\omega^\circ).$$

The obedience constraints are

$$\begin{aligned}
-\omega_L x(\omega_H|\omega_L)\mu_q(\omega_L) + (\omega_H - \omega_L)x(\omega_H|\omega_H)\mu_q(\omega_H) + (h\omega_H - \omega_L)x(\omega_H|\omega^\circ)\mu_q(\omega^\circ) \geq 0,\\
\omega_L x(\omega_L|\omega_L)\mu_q(\omega_L) - (\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H) - (h\omega_H - \omega_L)x(\omega_L|\omega^\circ)\mu_q(\omega^\circ) \geq 0.
\end{aligned}$$

We consider two cases depending on whether $h\omega_H - \omega_L > 0$, or $h\omega_H - \omega_L \leq 0$. If $h\omega_H - \omega_L > 0$, or $h > \frac{\omega_L}{\omega_H}$, then from the second obedience constraint $x_q^*(\omega_L|\omega_L) = 1$. The first obedience constraint is then automatically satisfied. Since $h \in (0,1)$, it is always true that $\frac{h\omega_H - \omega_L}{\omega_H - \omega_L} < h$. Then the solution satisfies $x_q^*(\omega_L|\omega_H) = 0$ and $x_q^*(\omega_L|\omega^\circ) = \frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)}$, as long as $\frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)} \leq 1$. We conclude that the solution is as follows:

1. If $\mu_q(\omega_L) \leq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = 0, \text{ and } x_q^*(\omega_L|\omega^\circ) = \frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)};$$

2. If $\frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h) \geq \mu_q(\omega_L) \geq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = \frac{\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ)}{(\omega_H - \omega_L)\mu_q(\omega_H)}, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1;$$

3. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = 1, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1.$$

Now suppose that $h\omega_H - \omega_L \leq 0$, or $h \leq \frac{\omega_L}{\omega_H}$. Combining obedience constraints in the standard manner for communication problems with binary action, we get

$$\omega_L x(\omega_L|\omega_L)\mu_q(\omega_L) - (\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H) - (h\omega_H - \omega_L)x(\omega_L|\omega^\circ)\mu_q(\omega^\circ) \geq$$

$$\max\left\{\omega_H\mu_q(\omega_L) + (1-h)\omega_H\mu_q(\omega^\circ) - (\omega_H - \omega_L), 0\right\}.$$

It is immediate that $x_q^*(\omega_L|\omega^\circ) = x_q^*(\omega_L|\omega_L) = 1$, since this choice relaxes the platform's problem as much as possible. The obedience constraint then becomes

$$\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ) - \max\left\{\omega_H\mu_q(\omega_L) + (1-h)\omega_H\mu_q(\omega^\circ) - (\omega_H - \omega_L), 0\right\} \geq$$

$$(\omega_H - \omega_L)x(\omega_L|\omega_H)\mu_q(\omega_H)$$

We then conclude that the solution is as follows:

1. If $\mu_q(\omega_L) \leq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = \frac{\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ)}{(\omega_H - \omega_L)\mu_q(\omega_H)}, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1;$$

2. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then

$$x_q^*(\omega_L|\omega_L) = 1, x_q^*(\omega_L|\omega_H) = 1, \text{ and } x_q^*(\omega_L|\omega^\circ) = 1.$$

D.2 Data-Value Problem

Let $\lambda(a_1'|a_1) = b(a_1)l(a_1'|a_1)$. The data-value problem is then

$$\min_{v,\lambda} \mu_q(\omega_L)v(\omega_L) + \mu_q(\omega_H)v(\omega_H) + \mu_q(\omega^\circ)v(\omega^\circ),$$

subject to $\lambda(\omega_H|\omega_L), \lambda(\omega_L|\omega_H) \geq 0$,

$$\begin{aligned} v(\omega_L) &= \max\{\omega_L\lambda(\omega_H|\omega_L), -\omega_L\lambda(\omega_L|\omega_H)\} = \omega_L\lambda(\omega_H|\omega_L), \\ v(\omega_H) &= \max\{\omega_H - \omega_L - (\omega_H - \omega_L)\lambda(\omega_H|\omega_L), (\omega_H - \omega_L)\lambda(\omega_L|\omega_H)\} = \\ &\quad (\omega_H - \omega_L)\max\{1 - \lambda(\omega_H|\omega_L), \lambda(\omega_L|\omega_H)\}, \end{aligned}$$

$$v(\omega^\circ) = \max\{h(\omega_H - \omega_L) + (\omega_L - h\omega_H)\lambda(\omega_H|\omega_L), (h\omega_H - \omega_L)\lambda(\omega_L|\omega_H)\} =$$
$$h(\omega_H - \omega_L)\max\left\{1 - \frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)}\lambda(\omega_H|\omega_L), \frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)}\lambda(\omega_L|\omega_H)\right\}.$$

As we noted before, $\frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)} < 1$. Suppose that $h > \frac{\omega_L}{\omega_H}$. Then it is optimal to set $\lambda_q^*(\omega_L|\omega_H) = 0$ to relax the problem as much as possible. We then have

$$\begin{aligned}
v(\omega_L) &= \omega_L\lambda(\omega_H|\omega_L), \\
v(\omega_H) &= (\omega_H - \omega_L)\max\{1 - \lambda(\omega_H|\omega_L), 0\}, \\
v(\omega^\circ) &= h(\omega_H - \omega_L)\max\left\{1 - \frac{h\omega_H - \omega_L}{h(\omega_H - \omega_L)}\lambda(\omega_H|\omega_L), 0\right\}.
\end{aligned}$$

There are three candidates for optimal $\lambda(\omega_H|\omega_L)$, specifically, 0 and two kinks of the maxima in the expressions above, 1 and $\frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L} > 1$.

When $\lambda(\omega_H|\omega_L) = 0$, the objective is $S_0 := (1 - \mu_q(\omega_L) - \mu_q(\omega^\circ)(1-h))(\omega_H - \omega_L)$.

When $\lambda(\omega_H|\omega_L) = 1$, the objective is $S_1 := \mu_q(\omega_L)\omega_L + \mu_q(\omega^\circ)(1-h)\omega_L$.

When $\lambda(\omega_H|\omega_L) = \frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L}$, the objective is $S_f := \mu_q(\omega_L)\frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L}\omega_L$.

The following claims are true:

- $S_0 \leq S_1$ if and only if $\mu_q(\omega_L) + \mu_q(\omega^\circ)(1-h) \geq \frac{\omega_H - \omega_L}{\omega_H}$;
- $S_0 \leq S_f$ if and only if $\mu_q(\omega_L)\left(1 + \frac{h\omega_L}{h\omega_H - \omega_L}\right) + \mu_q(\omega^\circ)(1-h) \geq 1$;
- $S_1 \leq S_f$ if and only if $\mu_q(\omega_L)\frac{\omega_L}{h\omega_H - \omega_L} \geq \mu_q(\omega^\circ)$.

Figure 3 captures the resulting regions of $\mu_q(\omega_L)$ and $\mu_q(\omega^\circ)$ that correspond to the value of the problem being equal to one of S_0, S_1, and S_f.

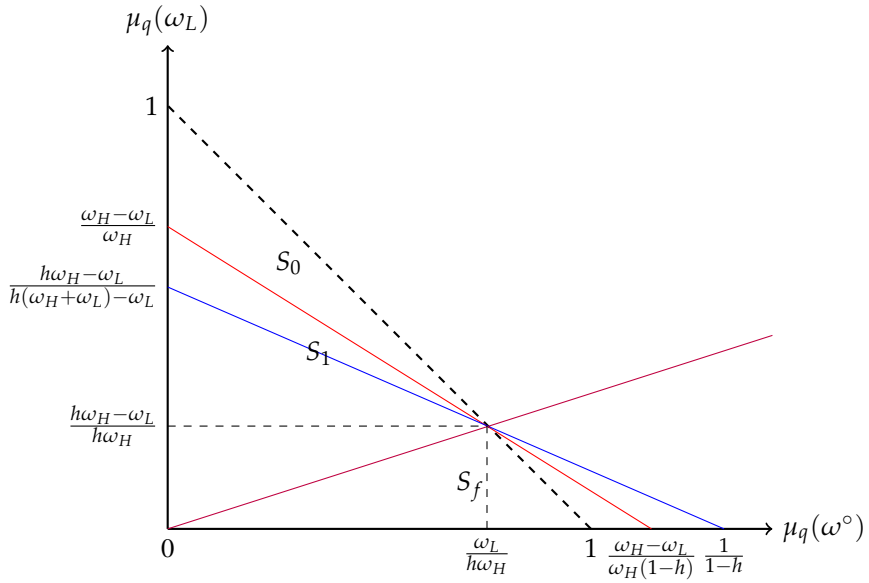


Figure 3: This figure pins down the minimum value of the data-value problem depending on μ_q when $h > \frac{\omega_L}{\omega_H}$. The red line corresponds to $S_0 = S_1$, the blue line corresponds to $S_0 = S_f$, and the purple line corresponds to $S_1 = S_f$.

We then conclude that the solution is as follows:

1. If $\mu_q(\omega_L) \geq \frac{\omega_H-\omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_H|\omega_L) = 0$ is optimal. The resulting values are
$$v_q^*(\omega_L) = 0, v_q^*(\omega_H) = \omega_H - \omega_L, \text{ and } v_q^*(\omega^\circ) = h(\omega_H - \omega_L);$$
2. If $\frac{h\omega_H-\omega_L}{\omega_L}\mu_q(\omega^\circ) \leq \mu_q(\omega_L) \leq \frac{\omega_H-\omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_H|\omega_L) = 1$ is optimal. The resulting values are
$$v_q^*(\omega_L) = \omega_L, v_q^*(\omega_H) = 0, \text{ and } v_q^*(\omega^\circ) = (1-h)\omega_L;$$
3. If $\mu_q(\omega_L) \leq \frac{h\omega_H-\omega_L}{\omega_L}\mu_q(\omega^\circ)$, then $\lambda_q^*(\omega_H|\omega_L) = \frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L}$ is optimal. The resulting values are
$$v_q^*(\omega_L) = \frac{h(\omega_H-\omega_L)}{h\omega_H-\omega_L}\omega_L, v_q^*(\omega_H) = 0, \text{ and } v_q^*(\omega^\circ) = 0.$$

Suppose now that $h \leq \frac{\omega_L}{\omega_H}$. Then immediately
$$v(\omega^\circ) = h(\omega_H - \omega_L) - (h\omega_H - \omega_L)\lambda(\omega_H|\omega_L).$$

$\lambda_q^*(\omega_L|\omega_H) = 0$ is optimal again. There are only two candidates for optimal $\lambda(\omega_H|\omega_L)$, specifically, 0 and 1. The solution then is as follows:

1. If $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_H|\omega_L) = 0$ is optimal. The resulting values are

$$v_q^*(\omega_L) = 0, v_q^*(\omega_H) = \omega_H - \omega_L, \text{ and } v_q^*(\omega^\circ) = h(\omega_H - \omega_L);$$

2. If $\mu_q(\omega_L) \leq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$, then $\lambda_q^*(\omega_H|\omega_L) = 1$ is optimal. The resulting values are

$$v_q^*(\omega_L) = \omega_L, v_q^*(\omega_H) = 0, \text{ and } v_q^*(\omega^\circ) = (1-h)\omega_L.$$

D.3 Summary

All the cases considered can be grouped into three scenarios based on $\mu_q(\omega_L)$ and $\mu_q(\omega^\circ)$.

Scenario 1. Suppose that $\mu_q(\omega_L) \leq \frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ)$. Note that this scenario appears only if $h > \frac{\omega_L}{\omega_H}$. The solution to the information-design problem is presented in Table 2.

$x_q^*(a\|\omega)$		ω		
		ω_L	ω_H	ω°
a	ω_L	1	0	$\frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)}$
	ω_H	0	1	$1 - \frac{\omega_L}{h\omega_H - \omega_L}\frac{\mu_q(\omega_L)}{\mu_q(\omega^\circ)}$

Table 2: Platform Example, x_q^* for Scenario 1.

The solution to the data-value problem is $\lambda_q^*(\omega_L|\omega_H) = 0$, $\lambda_q^*(\omega_H|\omega_L) = \frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L}$ and the unit values of datapoints are $v_q^*(\omega_L) = \frac{h(\omega_H - \omega_L)}{h\omega_H - \omega_L}\omega_L$, $v_q^*(\omega_H) = 0$, and $v_q^*(\omega^\circ) = 0$.

Scenario 2. Suppose that $\frac{h\omega_H - \omega_L}{\omega_L}\mu_q(\omega^\circ) \leq \mu_q(\omega_L) \leq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$. Note that the lower bound on $\mu_q(\omega_L)$ is meaningful only if $h > \frac{\omega_L}{\omega_H}$. The solution to the information-design problem is presented in Table 3.

$x_q^*(a\|\omega)$		ω		
		ω_L	ω_H	ω°
a	ω_L	1	$\frac{\omega_L\mu_q(\omega_L) - (h\omega_H - \omega_L)\mu_q(\omega^\circ)}{(\omega_H - \omega_L)\mu_q(\omega_H)}$	1
	ω_H	0	$\frac{\omega_H - \omega_L - \mu_q(\omega_L)\omega_H - \mu_q(\omega^\circ)(1-h)\omega_H}{(\omega_H - \omega_L)\mu_q(\omega_H)}$	0

Table 3: Platform Example, x_q^* for Scenario 2.

The solution to the data-value problem is $\lambda_q^*(\omega_L|\omega_H) = 0$, $\lambda_q^*(\omega_H|\omega_L) = 1$, and the unit values of datapoints are $v_q^*(\omega_L) = \omega_L$, $v_q^*(\omega_H) = 0$, and $v_q^*(\omega^\circ) = (1-h)\omega_L$.

Scenario 3. Suppose that $\mu_q(\omega_L) \geq \frac{\omega_H - \omega_L}{\omega_H} - \mu_q(\omega^\circ)(1-h)$. The solution to the information-design problem is presented in Table 4.

$x_q^*(a\|\omega)$		ω		
		ω_L	ω_H	ω°
a	ω_L	1	1	1
	ω_H	0	0	0

Table 4: Platform Example, x_q^* for Scenario 3.

The solution to the data-value problem is $\lambda_q^*(\omega_L|\omega_H) = \lambda_q^*(\omega_H|\omega_L) = 0$ and the unit values of datapoints are $v_q^*(\omega_L) = 0$, $v_q^*(\omega_H) = \omega_H - \omega_L$, and $v_q^*(\omega^\circ) = h(\omega_H - \omega_L)$. $\triangle$