The Value of Data Records

Simone Galperti
UC, San Diego

Aleksandr Levkun
UC, San Diego

Jacopo Perego
Columbia University

July 2, 2022

Abstract

Many e-commerce platforms use buyers' personal data to intermediate their transactions with sellers. How much value do such intermediaries derive from the data record of each single individual? We characterize this value and find that one of its key components is a novel externality between records, which arises when the intermediary pools some records to withhold the information they contain. Ignoring this can significantly bias the evaluations of data records. Our analysis has several implications about compensating individuals for the use of their data, guiding companies' investments in data acquisition, and more broadly studying the demand side of data markets. Our method combines modern information design with classic duality theory and applies to a large class of principal-agent problems.

JEL Classification Numbers: C72, D82, D83

Keywords: Value, Data, Record, Pooling Externality, Information, Duality

We thank Andrea Galeotti and five anonymous referees for their guidance. We are thankful to Dirk Bergemann and Emir Kamenica for their insightful comments as discussants of this paper. We thank S. Nageeb Ali, Alessandro Bonatti, Wouter Dessein, Laura Doval, Matt Elliott, Navin Kartik, Elliot Lipnowski, Alessandro Lizzeri, Xiaosheng Mu, Andrea Prat, Joel Sobel, Denis Shishkin, Rakesh Vohra, and Glen Weyl, as well as seminar participants at various universities for useful feedback. This research is supported by grants from the National Science Foundation (SES-2149289 and SES-2149315).

1 Introduction

Personal data is the "new oil" of modern economies. Search engines and social media platforms use it to sell targeted advertisement; e-commerce platforms use it to intermediate trade between buyers and sellers; job-matching platforms use it to match workers and employers. In each case, a large quantity of personal data fuels a multi-billion-dollar industry. How much of this total value is created by the data of each single individual? This basic question is at the core of some of the recent debates regarding the future of data markets, including how to design them to compensate individuals for their data, how to conduct demand analysis for data-brokers, and to what extent data is a source of market power (see, e.g., Federal Trade Commission, 2014; Stigler Report, 2019; Seim et al., 2022).[1]

Yet, the value of personal data is not well understood and can be hard to assess. For example, consider an e-commerce platform that mediates the trade between one seller and many buyers.[2] For each buyer, the platform owns a *data record*, which consists of a list of personal characteristics (gender, age, etc.). Suppose there are two types of records, ω_1 and ω_2, which reveal the buyer's willingness to pay for the seller's product is 1 and 2, respectively. The platform's database has more ω_2 records than ω_1 records. The seller knows the database composition, but has no other information about the buyers. How much value does the platform derive from each record? Of course, the answer depends on how it uses them, which in turn depends on its objective.

To start, consider the simple case where the platform's payoff is proportional to the seller's profits. The platform will then fully disclose all it knows about each buyer to the seller—for instance, using buyers' records to divide them into two market segments, each containing only buyers with the same willingness to pay. The seller then sets prices that extract all the surplus from each buyer. As a result, the platform earns a higher payoff with ω_2 records, which intuitively also have a higher value than ω_1 records. By contrast, consider the case where the platform's payoff is proportional to the buyers' surplus. In this case, it can do strictly better than full disclosure. For instance, it can add some buyers whose record is of type ω_2 to the segment that previously contained none, as long as the seller continues to set a price of 1 for this segment. As before, the platform earns a higher payoff with ω_2 records than with ω_1 records, whose payoff is zero. Yet, the latter are not worthless: They are essential to generate positive surplus with some buyers. In fact, we will show ω_1 records have the highest value for

[1] Posner and Weyl (2018) were among the first to advocate for data compensations. They also argued that "the first step toward valuing individual contributions to the data economy is measuring these contributions."

[2] This example is inspired by the model in Bergemann et al. (2015) and Elliott et al. (2021).

the platform despite generating the lowest payoff.

Our main contribution is to characterize the value of each data record for an *intermediary*—like the platform above—that uses data to influence the behavior of strategic agents to its own advantage by disclosing information. Our analysis reveals the value of a record is the sum of two components. The first is the payoff the intermediary *directly* obtains from this record (e.g., when the corresponding buyer trades with the seller in the platform's example). The second is a novel externality between records that arises when the intermediary pools them to withhold information from the agents. Therefore, the payoff generated by a record can be a misleading measure of its actual value. In the example above, when the platform maximizes buyers' surplus, ω_1 records exert a positive externality by allowing the platform to earn a positive payoff with some of the ω_2 records. For this reason, they have a positive value despite generating zero payoff. Records of type ω_2 instead exert a negative externality, because when pooled with ω_1 records, they make persuading the seller to set a low price harder—which explains why they are less valuable than ω_1 records.

These *pooling externalities* are a hallmark of intermediation problems: Conflicts of interest with and between the agents may induce the intermediary to withhold information from them, which creates the interdependencies between data records discussed above.[3] For example, the seller's belief after being told a buyer belongs to some segment depends jointly on all records pooled in that segment. These pooling externalities arise even when data records are statistically independent. As such, they are distinct from and complementary to the "learning" externalities that can arise when a buyer's record may be informative about another buyer's preferences, as highlighted by a recent literature on data markets.[4] In this paper, we switch off this channel and emphasize externalities that arise endogenously from how data is used, rather than from exogenous correlation.

We propose an inherently classical approach to determine the value of data records. We think of an intermediary as using inputs (data records) to produce output (information). Applying standard techniques from information design (Myerson (1982); Bergemann and Morris (2016)), this production problem acquires a linear structure. We then use duality methods to

[3]Withholding information is a common practice for many digital platforms. For example, Google's "quality score" pools people's searches to increase competition among advertisers (see, e.g., Sayedi et al., 2014); Uber conceals riders' destinations from drivers to increase riders' welfare; and Airbnb withholds hosts' profile pictures to decrease discrimination.

[4]Choi et al. (2019), Bergemann et al. (2022), Acemoglu et al. (2021), and Ichihashi (2021) study how consumers' incentives to disclose their personal data depend on the statistical correlation among them, and the effects of those incentives on market outcomes, such as consumers' participation and their welfare.

characterize the value of the data records, adapting the classic work of Dorfman et al. (1987) and Gale (1989). This approach is general and direct, as one can find those values without having to solve the primal production problem.

We use our characterization of the value of data records as a stepping stone to address three questions about data markets. First, consider the debate about whether and how a data market could compensate individuals for their personal data.[5] These questions involve complex considerations—such as how to design markets to facilitate this goal—which are outside the scope of this paper. However, we contribute to this debate by offering a useful benchmark—the value of a data record—against which to compare the compensation that individuals would obtain in practice. We discuss the desirable features of this benchmark and compare it with alternative ones. In an application that generalizes the platform example above, our benchmark implies that a flat compensation or one based on the platform's direct payoffs can overcompensate buyers with a higher willingness to pay at the expense of those with a lower willingness to pay.

Second, we contribute to the analysis of an intermediary's demand for data. A key step is to realize that its preferences over databases are pinned down by the values of data records. We find that the aforementioned pooling externalities introduce convexities in those preferences. This renders the intermediary akin to a consumer in standard consumer theory, opening the door to applying well-known analytical tools. In particular, we derive the demand function for records of a fixed type and find it is downward-sloping. We study the substitutability between types of records and characterize when they are imperfect substitutes or even complements. These properties establish a "scarcity principle" for data: Any intermediary places more value on scarcer types of records, both in absolute and in relative terms. We also note that assigning a value to each data record can be viewed as an internal accounting exercise on the part of the intermediary, which can guide its investment decisions of which types of records to acquire. Moreover, we show that a platform strictly benefits from merging two databases—for instance, via a takeover—if and only if our pooling externalities are present, which is directly revealed by the records' values at the original databases. These insights may inform regulatory policies.

Finally, we study how learning more about existing records affects their values and whether it benefits an intermediary. In this exercise, we fix the size of the database and refine some of its records by making them more informative (e.g., by observing new characteristics of the corresponding buyers in the platform example). We find that refining a record increases its

[5]For example, Seim et al. (2022) argue that "exploring ways to compensate internet users for the collection and use of their data is of utmost importance from the standpoint of efficiency and fairness, in addition to concerns about competition and privacy."

value in expectation, but can decrease the value of other records. These indirect effects are once again caused by the pooling externalities between records. Despite these mixed effects, overall, refining records never hurts the intermediary. We analyze the effects of refinements both at the intensive margin (Blackwell informativeness) and at the extensive margin (quantity of refined records).

Related Literature. This paper contributes to the burgeoning literature on data markets, recently reviewed by Bergemann and Bonatti (2019) and Bergemann and Ottaviani (2021). One of its strands studies the optimal "use" of databases. This often involves a single party who owns a database and designs information products to either sell them to some agents or to influence their behavior (e.g., Admati and Pfleiderer (1986, 1990); Bergemann and Bonatti (2015); Bergemann et al. (2015); Bergemann et al. (2018); Elliott et al. (2021); Yang (2022)). Our intermediary also owns a database and uses it to design information. However, our focus is not on the optimal use of the database, but on studying the value of each data record and its properties. Because our intermediary already owns the database, we abstract from important considerations about data privacy, which are the focus of another strand of the literature on data markets. In addition to the papers discussed in Footnote 4, see Ali et al. (2022) and Ichihashi (2020) for recent contributions to this literature and Acquisti et al. (2016) for a review. Calzolari and Pavan (2006) analyze information externalities between sequential interactions and the role of privacy.

Our methods build on the information-design literature, reviewed by Bergemann and Morris (2019). We formulate the intermediation problem as a linear program and then consider its dual to determine the value of data records. Others have used duality in information design as a method to solve the primal design problem (Kolotilin (2018); Galperti and Perego (2018); Dworczak and Martini (2019); Dworczak and Kolotilin (2019); Dizdar and Kováč (2020)). We instead use duality to address a distinct economic question of independent interest. The mechanism-design literature has also used duality methods at least since Myerson (1983; 1984), and more recently to study informationally robust mechanisms (e.g., Du (2018); Brooks and Du (2020, 2021)).

We contribute to the literature on the value of information in two ways: by focusing on intermediation problems rather than decision problems and by studying the ex-post value of data records rather than the ex-ante value of information. Frankel and Kamenica (2019) also take an ex-post perspective, but focus on decision problems and the ex-post value of information. A complementary literature studies properties of the cost—not the value—of information (e.g., see Morris and Strack, 2019; Bloedel and Zhong, 2021; Pomatto et al., 2021). These properties are conceptually distinct from those of the value of data records characterized in this paper.

2 Model

For ease of exposition, we present the model and analysis in a context similar to our example in the Introduction: An e-commerce platform (pronoun *it*) mediates interactions between many buyers (*she*) and a single seller (*he*). Our approach and results apply more broadly as discussed at the end of this section.

The seller has a finite set of actions A with typical element a, which can be interpreted as the price, quality, or other features of his product. The platform is used by a population of buyers, each with unit demand. A buyer's preference is pinned down by θ in a finite set Θ. Together, a and θ determine the buyer's final purchase decision, which we leave implicit and embed into the platform's payoff function $\hat{u}(a,\theta)$ and the seller's profit function $\hat{\pi}(a,\theta)$.

For each buyer, the platform has private access to a *data record* of her personal characteristics, which is informative about her θ. We model this record as the realization of an exogenous signal, denoted by ω in some finite set Ω. We refer to ω as the *type* of the buyer's record. We assume ω induces a belief over Θ and denote by $u(a,\omega)$ the expected payoff of the platform and by $\pi(a,\omega)$ the expected profit of the seller when he takes action a and the buyer's record is of type ω.[6]

The collection of all buyers' data records forms the platform's *database*. We denote a database by $q = (q(\omega))_{\omega\in\Omega}$, where $q(\omega)$ is the quantity of ω records. We think of each buyer as being "small" within the buyers' population, and thus allow $q(\omega)$ to vary continuously in the positive real line. The primitives A, Ω, u, π, and q are common knowledge.

The platform mediates each buyer-seller interaction by conveying information about the buyer's record to the seller to influence his action a. To do so, it commits once and for all to an information structure τ that consists of a finite space of signals S and a function $\tau : \Omega \to \Delta(S)$.[7] After observing τ and a signal $s \in S$, the seller updates his belief about the buyer and takes an action in the set $A(q,\tau,s) = \arg\max_{a\in A} \mathbb{E}(\pi(a,\omega)|q,\tau,s)$. We assume the seller breaks indifferences in favor of the platform. The platform's problem is to choose τ to maximize

$$\sum_{\omega\in\Omega, s\in S} q(\omega)\tau(s|\omega) \max_{a\in A(q,\tau,s)} u(a,\omega). \tag{1}$$

We refer to (1) as the platform's *intermediation problem*.

Our main question is as follows: How much value does the platform derive from each buyer's

[6] That is, $u(a,\omega) = \sum_{\theta\in\Theta} \hat{u}(a,\theta)\beta(\theta|\omega)$ and $\pi(a,\omega) = \sum_{\theta\in\Theta} \hat{\pi}(a,\theta)\beta(\theta|\omega)$, where $\beta(\cdot|\omega) \in \Delta(\Theta)$ is the belief induced by ω.

[7] Note that restricting $\tau(\cdot|\omega)$ to be the same between records of the same type ω is without loss of generality.

record and what are the properties of this value? Three aspects render this question non-trivial. First, the platform designs information structures to influence the decisions of someone else, which introduces an intermediate step between how it uses its data and the final payoff-relevant outcomes. Second, what information the platform can convey varies with what data records it has in its database. Third, we are interested in the individual value the platform obtains from each data record given its type—not the total payoff obtained from the whole database.

Generalizations. Our analysis applies to more general settings where an intermediary uses its data to mediate the strategic interaction of multiple agents (e.g., competing sellers). We can allow the intermediary to take contractible actions as a function of ω (e.g., monetary transfers). We can allow the agents to observe parts of the intermediary's data records (e.g., a seller may observe the buyer's gender but not her age). We can allow for ex-ante participation constraints (e.g., a seller may choose not to sell his product through the platform). These extensions are possible because they preserve the linear-programming structure of our baseline model. Our proofs in the Appendix already account for this more general case.

3 What Is the Value of a Data Record?

We begin by formulating our approach to calculating the values of data records (Section 3.1). We then analyze their properties (Section 3.2).

3.1 The Data-Value Problem

As is well known (Myerson, 1982; Bergemann and Morris, 2016), we can equivalently express the intermediation problem (1) in terms of direct recommendation mechanisms. That is, instead of choosing an information structure τ, the platform can directly recommend an action to the seller, which he must be willing to follow (obedience). We can formulate such a mechanism as a function $x : A \times \Omega \to \mathbb{R}_+$, where we interpret $x(a,\omega)$ as the quantity of ω records for which the seller is recommended a. Formally, the problem is

$$\begin{aligned} \mathcal{U}_q : \quad \max_{x} \quad & \sum_{\omega\in\Omega, a\in A} u(a,\omega)x(a,\omega) \\ \text{s.t.} \quad & \sum_{\omega\in\Omega} \big(\pi(a,\omega) - \pi(a',\omega)\big)\, x(a,\omega) \geq 0 \quad \text{for all } a,a' \in A \qquad (2) \\ & \sum_{a\in A} x(a,\omega) = q(\omega) \quad \text{for all } \omega \in \Omega. \qquad (3) \end{aligned}$$

The obedience constraint (2) is equivalent to requiring that the seller finds it optimal to follow the recommended action a conditional on the information it conveys, given x and q. We call x obedient if it satisfies (2). The resource constraint (3) requires that some recommendation be sent for every record in the database.[8] We denote any optimal mechanism by x_q^* and define two pieces of notation: the *direct payoff* generated by each record of type ω is

$$u_q^*(\omega) \triangleq \sum_{a \in A} u(a,\omega) \frac{x_q^*(a,\omega)}{q(\omega)},$$

and the *total payoff* generated by the database is

$$U^*(q) \triangleq \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega). \tag{4}$$

Hereafter, we assume $\mathcal{U}_q$ satisfies the following regularity property, which holds generically in the space of the seller's payoff functions π: No more than $|A \times \Omega|$ of the constraints in $\mathcal{U}_q$ are ever binding at the same time. Intuitively, this property rules out the possibility that the optimal x_q^* recommends an action that leaves the seller indifferent between too many alternative actions, which is not robust to perturbing their payoffs.

Our approach hinges on thinking about $\mathcal{U}_q$ as a production problem, where the inputs are the records in the database and the output is the information conveyed by the mechanism in the form of recommendations. Because $\mathcal{U}_q$ is a linear program, we can leverage this production interpretation to assign a value to each record (Dorfman et al. (1987), p. 39; Gale (1989), p. 12). These values can be computed through a problem related to $\mathcal{U}_q$, which is called its dual.

We refer to this dual as the *data-value* problem. It involves the choice of two objects. The first is $v : \Omega \to \mathbb{R}$, where $v(\omega)$ is the multiplier associated with the resource constraint (3) for ω records. The second is $\lambda : A \times A \to \mathbb{R}_+$, where $\lambda(a'|a)$ is the multiplier of the obedience constraint (2) when the seller is recommended a and considers deviating to a'. Applying basic linear-programming results (see Appendix A), the dual of $\mathcal{U}_q$ is

$$\begin{aligned} \mathcal{V}_q : \quad \min_{v,\lambda} \quad & \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ \text{s.t.} \quad & v(\omega) \geq u(a,\omega) + \sum_{a' \in A} \big(\pi(a,\omega) - \pi(a',\omega)\big) \lambda(a'|a), \quad \forall\, \omega, a. \end{aligned} \tag{5}$$

We denote an optimal solution by (v_q^*, λ_q^*). By standard arguments, v_q^* is unique generically with respect to q. Note that (v_q^*, λ_q^*)—and x_q^*, for that matter—ultimately depends only on the

[8] Note that choosing $x : A \times \Omega \to \mathbb{R}_+$ that satisfies (3) is equivalent to a more standard formulation where the platform chooses $\chi : \Omega \to \Delta(A)$ and we replace $x(a,\omega)$ with $\chi(a|\omega) q(\omega)$ everywhere in $\mathcal{U}_q$.

shares of the records of every type implied by q. The reason is that only the frequency of record types matters for the seller's incentives.[9]

The economic interpretation of $v_q^*(\omega)$ is key for us. It captures the value that any existing record of type ω in the database q generates for the platform. The structure of the data-value problem $\mathcal{V}_q$ implies $v_q^*(\omega)$ must satisfy

$$v_q^*(\omega) = \max_{a \in A} \left\{ u(a,\omega) + \sum_{a' \in A} \left(\pi(a,\omega) - \pi(a',\omega) \right) \lambda_q^*(a'|a) \right\}. \tag{6}$$

Thus, $v_q^*(\omega)$ is a measure of value because it is related to the payoffs of the platform and the seller (which could both be expressed in dollars, for instance). This value reflects how records are used. Indeed, complementary slackness requires that if the platform recommends action a for some ω records (i.e., $x_q^*(a,\omega) > 0$), then constraint (5) must hold with equality; hence, a maximizes the right-hand side of (6). In particular, this means that even if records of the same type ω lead to different recommendations, at the optimum they all have the same value $v_q^*(\omega)$. Hereafter, we refer to equation (6) as the *value formula*, whose analysis and explanation are the focus of the next subsection.

For these reasons, $v_q^*(\omega)$ captures the "infra-marginal" value of a record, namely the value of each existing record of type ω in the database q. In fact, Gale (1989) refers to the value of an input such as $v_q^*(\omega)$ as its *unit* value. Moreover, by strong duality,[10] the total payoff of the database $U^*(q)$ equals the sum the values of its records:

$$\sum_{\omega \in \Omega} v_q^*(\omega) q(\omega) = U^*(q) \stackrel{\Delta}{=} \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega). \tag{7}$$

The dual problem $\mathcal{V}_q$ can then be viewed as an internal accounting exercise: The platform allocates a share of the total payoff $U^*(q)$ to each record in the database according to the value it actually generated. In Section 3.3, this interpretation will allow us to establish a benchmark for how to compensate buyers for their specific data.

Our interest in v_q^* goes beyond this internal accounting interpretation. As for any constrained optimization (not just linear ones), $v_q^*(\omega)$ also captures the shadow value of adding a new record of type ω to the database. In fact, $v_q^*(\omega)$ is also equal to the derivative of $U^*(q)$ with respect to $q(\omega)$. Therefore, $v_q^*(\omega)$ captures not only the infra-marginal value of the existing records in the database, but also the marginal value of adding new ones. In Section 4.1, this interpretation will allow us to characterize the platform's demand for more data.

[9] Note that $\mathcal{V}_q$ finds all records' values simultaneously and does not require calculating x_q^* (see Section 3.3.1 for an illustration).

[10] See Theorem 4.4 in Bertsimas and Tsitsiklis (1997).

The rest of the paper characterizes the properties of v_q^* and their economic implications. As a preliminary step, we establish that the value of each record is bounded below by the payoff of fully revealing its type to the seller, and when this lower bound is achieved. Let the set of the seller's optimal actions if he knows ω be

$$A(\omega) \stackrel{\Delta}{=} \arg\max_a \pi(a,\omega).$$

Given this, define the *full-disclosure payoff* of a record of type ω as

$$\underline{v}(\omega) \stackrel{\Delta}{=} \max_{a \in A(\omega)} u(a,\omega), \qquad \omega \in \Omega.$$

Lemma 1. *For every database q, the value of records of type ω is bounded below by their full-disclosure payoff: $v_q^*(\omega) \geq \underline{v}(\omega)$. Moreover, $v_q^*(\omega) = \underline{v}(\omega)$ for all ω if and only if there is an optimal mechanism x_q^* that induces only actions which are optimal for the seller under full disclosure (i.e., $x_q^*(a,\omega) > 0$ only if $a \in A(\omega)$ for all ω).*

3.2 Value Decomposition and Pooling Externalities

What determines the value of a record? We show next that it can be decomposed into two parts: the direct payoff the record generates and an additional component, which captures the record's effects on the information the platform discloses about other records and thus on their direct payoffs. To formalize this, denote the second part of the value formula in (6) by

$$t_q^*(a,\omega) \stackrel{\Delta}{=} \sum_{a' \in A} \left(\pi(a,\omega) - \pi(a',\omega)\right) \lambda_q^*(a'|a) \qquad \forall a,\omega. \tag{8}$$

Proposition 1. *The value $v_q^*(\omega)$ of records of type ω satisfies*

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega),$$

where

$$t_q^*(\omega) \stackrel{\Delta}{=} \sum_{a \in A} t_q^*(a,\omega) \frac{x_q^*(a,\omega)}{q(\omega)} \stackrel{a.e.}{=} \sum_{\omega' \in \Omega} \frac{\partial u_q^*(\omega')}{\partial q(\omega)} q(\omega'). \tag{9}$$

This result explains why the direct payoff of a record $u_q^*(\omega)$ can be a biased measure of its value $v_q^*(\omega)$, as we illustrated in the Introduction. The direct payoff fails to take into account $t_q^*(\omega)$. This component is akin to an externality: It captures the effect that a record can exert on what the platform achieves with other records. To gain intuition, consider a buyer

called Ann whose record is of type ω. Simply by being in the database, Ann's record can affect what recommendations the platform sends for other records, whose types may be different from Ann's. Formally, this is shown by the second equality in (9): The externality $t_q^*(\omega)$ is equal to the marginal effect that Ann's record has on the direct payoff of all records (for almost all databases q). This effect comes about because the presence of Ann's record affects the seller's obedience constraint, and thus changes which mechanisms x are obedient or optimal.[11]

The first equality in (9) further illuminates the nature of this externality. It shows that $t_q^*(\omega)$ aggregates $t_q^*(a,\omega)$ across recommendations a, where $t_q^*(a,\omega)$ captures how Ann's record contributes to keeping the seller obedient whenever he is recommended a. To gain intuition, suppose $\pi(a,\omega) > \pi(a',\omega)$ in (8). If the seller knew he was trading with Ann, he would not want to deviate to a'. The platform can then pool Ann's record with other records of type ω' for which $\pi(a,\omega') < \pi(a',\omega')$ and still persuade the seller to take action a. In this case, Ann's record exerts a positive externality by helping the platform implement a with those records.[12] This pooling—which amounts to withholding information to exploit the seller's incentives—is thus at the heart of our externality. In fact, consider the special case where the platform reveals all information, which would be optimal, for instance, if the platform's and the seller's objectives are aligned.

Corollary 1. *If the platform's optimal information structure fully reveals the type of every record, no externality arises (i.e., $t_q^* = 0$) and the value of each record coincides with its direct payoffs (i.e., $v_q^* = u_q^*$).*[13]

These results lead to the following general point: The externality highlighted in this paper is a hallmark of intermediation problems. A distinguishing feature of these problems is that the intermediary and the agent(s) often have conflicting interests, which is captured by the obedience constraints in (2). To optimally manage these conflicts, the intermediary withholds information from the agent(s) by pooling data records, which creates the externalities. For these reasons, we refer to them as *pooling externalities*.

These externalities arise despite our assumption that each buyer's record is uninformative about the other buyers' preferences. As such, they are distinct from and complementary to the

[11] Indeed, note that $\frac{\partial}{\partial q(\omega)} u_q^*(\omega') = \sum_a u(a,\omega') \frac{\partial}{\partial q(\omega)} \left(\frac{x_q^*(a,\omega')}{q(\omega')} \right)$.

[12] Appendix D elaborates on this interpretation and how the platform exploits the seller to determine the externalities between records.

[13] If the optimal information structure is fully revealing, the corresponding mechanism satisfies $x_q^*(a,\omega) > 0$ only if $a \in A(\omega)$. This implies $\underline{v}(\omega) = u_q^*(\omega)$ and, moreover, that $\underline{v}(\omega) = v_q^*(\omega)$ by Lemma 1. Note that the converse of Corollary 1 is not true: In some examples, $t_q^*(\omega) = 0$ but $v_q^*(\omega) > \underline{v}(\omega)$ for all ω.

"learning" externalities discussed in Section 1, which arise because a buyer's record conveys information about other buyers. We switched off this channel to focus on externalities that arise endogenously from how data is used, rather than from exogenous correlation.

The rest of this subsection further characterizes these externalities. Which records generate positive and which negative externalities?

Corollary 2. *Fix a record type ω. If $t_q^*(\omega) < 0$, the direct payoff exceeds the full-disclosure payoff: $u_q^*(\omega) > \underline{v}(\omega)$. Conversely, if $u_q^*(\omega) < \underline{v}(\omega)$, then $t_q^*(\omega) > 0$. Moreover, $t_q^*(\omega) < 0$ for some ω if and only if $t_q^*(\omega') > 0$ for some ω'.*

To build intuition, consider records whose direct payoff exceeds their value (i.e., $t_q^*(\omega) < 0$). Why should this be the case? Corollary 2 shows the platform must earn a direct payoff $u_q^*(\omega)$ that would not be possible if it fully disclosed all ω records. Such a payoff must be achieved by pooling some ω records with records of different types. Accordingly, the value $v_q^*(\omega)$ reflects the fact that these records are being "helped" by others and do not generate $u_q^*(\omega)$ entirely on their own. There must be some other type of records—call it ω'—that generates a positive externality in favor of ω records (i.e., $t^*(\omega') > 0$). Corollary 2 shows this is the case when $u_q^*(\omega') < \underline{v}(\omega')$.[14] The platform effectively "sacrifices" ω' records, as their payoff is lower than what it could have obtained by simply revealing their type to the seller. Accordingly, their value $v_q^*(\omega')$ exceeds $u_q^*(\omega')$, as it accounts for the greater role they ultimately play.

Proposition 1 highlights that the pooling externalities are tightly related to how the platform exploits the seller's incentives by using its records. The next result complements this point by showing that if the platform recommends different actions using ω records, the records generating the highest contribution to the externality $t_q^*(\omega)$ must generate the lowest contribution to the direct payoff $u_q^*(\omega)$.

Corollary 3. *Suppose an optimal mechanism satisfies $x_q^*(a,\omega) > 0$ and $x_q^*(a',\omega) > 0$. Then, $t_q^*(a,\omega) > t_q^*(a',\omega)$ if and only if $u(a,\omega) < u(a',\omega)$.*[15]

3.3 A Benchmark for Data Compensation

An important open question in policy is how to compensate each individual for the collection and use of their personal data. The basic idea is that some part of the platform's total pay-

[14] Note that Corollary 2 indicates a simple sufficient condition for $t_q^* \neq 0$. Proposition C.1 in the Online Appendix provides another condition based on primitives.

[15] This follows from complementary slackness, namely $v_q^*(\omega) = u(a,\omega) + t_q^*(a,\omega)$ if $x_q^*(a,\omega) > 0$.

off (e.g., profits) would be distributed back to its users as a form of dividend for the monetization of their data. In this case, how much of this payoff will each receive? Answering this question involves complex considerations outside the scope of this paper—including which privacy-protection laws would grant consumers control over their data, or the extent to which these compensations are determined competitively. Nonetheless, our approach contributes to this debate by offering a benchmark against which to compare the individuals' actual compensations, however defined or calculated.

Building on the internal-accounting interpretation of the data-value problem $\mathcal{V}_q$ (Section 3.1), our benchmark consists in dividing the total payoff $U^*(q)$ in shares that reflect the value that each record actually generated, namely v_q^*. We highlight three features of this benchmark that may be desirable. First, it takes into account that different records may contain different information, which may reasonably lead to different compensations. Second, the benchmark takes into account that records with the same informational content (i.e., type) contribute equally to the platform's profits—even when used differently—which may reasonably lead to an equal compensation. Third, this benchmark is grounded in classic economic principles of marginal considerations and optimality. As such, it can be broadly applicable across different settings (e.g., in terms of the platform's industry or its users' data).

Our benchmark can complement others. For example, an alternative is to grant the same compensation to all records, regardless of their content and use. This approach has the advantage of being simple and operational, but ignores that some records may be more profitable than others. Another benchmark is to compensate records based on their direct payoff u_q^*. This approach recognizes that some records may deserve different compensations if they provide different information or are used differently. However, it ignores the indirect role that each record plays via the pooling externality highlighted in Section 3.2. As a result, higher compensations may be granted to less deserving records, as illustrated by the following application.

3.3.1 Application: Mediated Price Discrimination

We conclude this section with an application to the setting of Bergemann et al. (2015). Our goal is threefold: characterize the values of the records for a concrete setting, illustrate the implications of using them as a benchmark for compensation, and showcase how to use our duality approach.[16]

Let the seller's action a indicate the price he sets for his product. For each buyer, let θ be her

[16]Using similar arguments, one can characterize v_q^* in the setting with multiple sellers of Elliott et al. (2021).

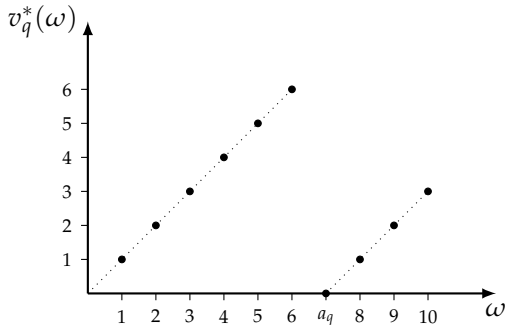


Figure 1: Values v_q^* when $\Omega = \{1, \ldots, 10\}$, $r = 0$, and q is such that $a_q = 7$.

strictly positive valuation for the product. Suppose the record type is fully informative about θ (i.e., $\theta = \omega$). Normalizing the seller's constant marginal cost to zero, his profit is a if $\omega \geq a$ and zero otherwise: $\pi(a, \omega) = a\mathbb{I}\{\omega \geq a\}$. The platform's payoff is a weighted sum of profits and consumer surplus: $u(a, \omega) = ra\mathbb{I}\{\omega \geq a\} + (1 - r)\max\{\omega - a, 0\}$, where $r \in [0, 1]$ captures the alignment between the platform and the seller's objectives. Finally, let a_q be the optimal price if the seller had to set the same price for all buyers in the database q.

Proposition 2. *For $r < \frac{1}{2}$, the value of a record is*

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

moreover, the externality $t_q^*(\omega)$ is strictly positive if and only if $\omega < a_q$. For $r \geq \frac{1}{2}$, the value of a record equals its direct payoff: $v_q^*(\omega) = u_q^*(\omega) = r\omega$ for all ω, and thus $t_q^*(\omega) = 0$.

The intuition is as follows. When $r \geq \frac{1}{2}$, the platform and the seller have sufficiently aligned objectives. As shown in the Appendix (Lemma A.2), the platform then withholds no information from the seller, who price discriminates all buyers by setting $a = \omega$. By Corollary 1, there are no externalities and the value of a record equals its direct payoff, which is a fraction r of the seller's profits ω. By contrast, when $r < \frac{1}{2}$, the platform and the seller have sufficiently misaligned objectives. To generate surplus for the buyers, the platform finds it optimal to withhold information from the seller, which requires pooling records. This creates externalities that induce a non-monotonicity in the record values: Higher records are not necessarily more valuable for the platform (Figure 1). In particular, under the optimal mechanism every buyer trades with the seller and generates a surplus of ω, but only those whose valuation exceeds a_q contribute to his profit (as in Bergemann et al. (2015)), which explains the shape of v_q^*.

With this result in mind, we return to the issue of compensating buyers for their data. Consider the case of $r < \frac{1}{2}$. Suppose a court rules that each record of type ω has to be compensated

based on the direct payoff it generated. Using our benchmark, we can ask to what extent these compensations reflect how much each record actually contributed to the platform's total payoff. This comparison reveals that low-valuation buyers (i.e., those with $\omega < a_q$) would be under-compensated (i.e., $t_q^*(\omega) > 0$), while high-valuation buyers would be over-compensated. Similarly, other biases would emerge if the court ruled that every record should get the same compensation. The point here is that knowing v_q^* is essential to identify these biases. Whether the court should correct them is ultimately a normative question, which is beyond the scope of this paper.

For the interested reader, we show next the key steps toward Proposition 2 as a way of illustrating the logic of our dual approach and its applicability. This highlights, for example, that the approach does not rely on computing the optimal mechanism x_q^*.

Proof of Proposition 2. Consider $\mathcal{V}_q$ and note that the right-hand side of constraint (5) is

$$ra\mathbb{I}\{\omega \geq a\} + (1-r)\max\{\omega - a, 0\} + \sum_{a'} \underbrace{[a\mathbb{I}\{\omega \geq a\} - a'\mathbb{I}\{\omega \geq a'\}]}_{\Delta\pi} \lambda(a'|a). \quad (10)$$

A few observations simplify the problem substantially. First, for every ω constraint (5) is strictly slack if $a > \omega$ and can therefore be ignored. If $a > \omega$, then (10) is weakly negative, whereas it is at least $r\omega > 0$ for $a = \omega$. Second, we can reduce the cases in which $\lambda(a'|a)$ can be positive using the fact that problem $\mathcal{V}_q$ seeks to minimize $v(\omega)$. If $a' < a \leq \omega$, then $\Delta\pi > 0$, and setting $\lambda(a'|a) > 0$ can only increase (10) and tighten constraint (5); so $\lambda_q^*(a'|a) = 0$ if $a' < a$.[17] Finally, because we can always set $\lambda \equiv 0$, it is useful to understand how the first part of (10) depends on a. For $a \leq \omega$ we can write $u(a,\omega) = a(2r-1) + (1-r)$, which is strictly increasing in a if and only if $r > \frac{1}{2}$.

The solution (v_q^*, λ_q^*) is immediate when $r > \frac{1}{2}$. In this case, $u(a,\omega)$ is maximal for $a = \omega$ and equals $r\omega$. This level is also a lower bound for (10), which can be achieved by setting $\lambda \equiv 0$. It follows that $v_q^*(\omega) = r\omega$ for all ω and $\lambda_q^* \equiv 0$, which implies no externalities (i.e., $t_q^* \equiv 0$). This solution extends to $r = \frac{1}{2}$ by continuity.

Finding (v_q^*, λ_q^*) requires a few more steps when $r < \frac{1}{2}$. Let ω_1 be the lowest type of records. Note $u(a,\omega)$ is maximal for $a = \omega_1$ and $u(\omega_1,\omega) > r\omega$ unless $\omega = \omega_1$. However, if $\omega > \omega_1$, there exists $a' > a = \omega_1$ such that $\Delta\pi < 0$, which allows us to reduce (10) by setting $\lambda(a'|\omega_1) > 0$ and to relax constraint (5). Thus, $\lambda \equiv 0$ is no longer optimal.

[17] This property of λ_q^* implies that, in problem $\mathcal{U}_q$, all constraints that involve deviating to a' when recommended $a > a'$ never bind at the optimum and can be ignored. A similar logic can substantially simplify the analysis of more general $\mathcal{U}_q$, which can be useful in applications with large sets A and Ω.

To find the optimal λ, we can use the observation that all a for which constraint (5) binds must give the same value of (10). That is, raising a above ω_1 must increase $\sum_{a'}[a - a'\mathbb{I}\{\omega \geq a'\}]\lambda(a'|a)$ at the same rate that it decreases $u(a,\omega)$, namely $2r-1$. It follows that

$$a\sum_{a'}\lambda(a'|a) - \sum_{a'}a'\mathbb{I}\{\omega \geq a'\}\lambda(a'|a) = a(1-2r) + const(\omega),$$

and hence $\sum_{a'}\lambda(a'|a) = 1-2r$. In addition, since $\sum_{a'}a'\mathbb{I}\{\omega \geq a'\}\lambda(a'|a)$ can depend at most on ω, we can write $\lambda(a'|a) = \hat{\lambda}(a')$ for every a, a'. This implies $v(\omega) = (1-r)\omega - \sum_{a'}a'\mathbb{I}\{\omega \geq a'\}\hat{\lambda}(a')$ for every ω. Substituting into the objective $\sum_{\omega}v(\omega)q(\omega)$ of $\mathcal{V}_q$, we get that $\hat{\lambda}$ has to maximize

$$\sum_{a'}\Big[\sum_{\omega}a'\mathbb{I}\{\omega \geq a'\}q(\omega)\Big]\hat{\lambda}(a')$$

subject to $\sum_{a'}\hat{\lambda}(a') = 1-2r$. Clearly, the optimal $\hat{\lambda}$ has to put all the "weight" $1-2r$ on the a' that maximizes $\sum_{\omega}a'\mathbb{I}\{\omega \geq a'\}q(\omega)$, which is a_q by definition. This implies $\lambda_q^*(a_q|a) = 1-2r$ for all $a \leq a_q$ and $\lambda_q^*(a'|a) = 0$ for all other a, a' (where we use the previous observation that $\lambda_q^*(a'|a) = 0$ if $a > a_q$).

We conclude that $v^*(\omega) = (1-r)\omega + (2r-1)a_q\mathbb{I}\{\omega \geq a_q\}$ and the externalities satisfy

$$t_q^*(\omega) = (1-2r) \times \begin{cases} \sum_a a\frac{x_q^*(a,\omega)}{q(\omega)} > 0 & \text{if } \omega < a_q \\ \sum_{a \leq a_q}[a-a_q]\frac{x_q^*(a,\omega)}{q(\omega)} \leq 0 & \text{if } \omega \geq a_q. \end{cases}$$

The first inequality follows because $a \geq \omega_1 > 0$; the second inequality follows because $a > a_q$ can never maximize (10) given λ_q^*, and hence, $x_q^*(a,\omega) = 0$ by complementary slackness. Note that we derived (v_q^*, t_q^*) and signed t_q^* without having to know x_q^*. □

4 The Demand for Data

What is the platform's willingness to pay for having "more data"? This colloquial expression can have two meanings. The first—analyzed in Section 4.1—is that the platform adds *more* records to the database, and hence mediates the interactions of more buyers with the seller. The second—analyzed in Section 4.2—is that the platform obtains *better* records; namely, it observes a more informative signal about θ for some buyers whose records already belong to the database.[18] In either case, having more data ultimately changes the database q. Hereafter, we

[18]The distinction between more and better records is consistent with that between *marketing lists* and *data appends*, the two main products traded in the data-brokerage industry (Federal Trade Commission, 2014). The

assume that how the platform changes q is publicly observed, and hence, q is always commonly known.[19] Building on Section 3, we can study the platform's willingness to pay for more data by analyzing how the records' values v_q^* depend on q. Alternatively, we can interpret the the following analysis as a comparative-statics exercise comparing the values of records between platforms that differ only in their databases.

4.1 More Records: Preferences over Databases

Analyzing the platform's willingness to pay for more records can shed light on properties of its demand for data records. For example, are demand curves downward sloping? Are data records complements or substitutes and, if so, why? We can view the platform as a "consumer" of records, whose utility function is U^*. Its preferences over databases are then fully characterized by v_q^*. Indeed, $v_q^*(\omega)$ is akin to the marginal utility of a record of type ω at q, which determines the platform's willingness to pay. We can also measure the substitutability between records of type ω and ω' at q by computing their marginal rate of substitution as usual, which satisfies $MRS_q(\omega, \omega') \overset{\text{a.e.}}{=} -\frac{v_q^*(\omega)}{v_q^*(\omega')}$.

As for a standard consumer, we find the platform's marginal utilities are diminishing: When records of a given type become more abundant, they become less valuable. This follows from the next result, which establishes a general "scarcity principle" for data. Given q, define the share of ω records by

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

Proposition 3 (Scarcity Principle)**.** *Consider databases q and q'. Fix ω. If $\mu_q(\omega) < \mu_{q'}(\omega)$, then $v_q^*(\omega) \geq v_{q'}^*(\omega)$. Moreover, $v_q^*(\omega)$ converges to the full-disclosure payoff $\underline{v}(\omega)$ as $\mu_q(\omega) \to 1$.*

This implies that $v_q^*(\omega)$ is weakly decreasing in $q(\omega)$. Hence, holding fixed the quantity of all other types of records, the platform's demand for ω records is downward sloping.

With regard to the marginal rate of substitution between records, we can easily see that it is always weakly diminishing. This is because $U^*(q)$ is a concave function of q.[20] The next

former allows companies to identify new customers with specific characteristics. The latter allows companies to learn new characteristics about existing customers.

[19] In reality, the platform may change its database privately without the sellers knowing exactly how. Allowing for this possibility introduces complications and requires enriching the model accordingly. We leave this for future research.

[20] To see why, imagine the platform mixes two databases q and q' to form a new one $\gamma q + (1-\gamma)q'$. Because

result characterizes when all records are perfect substitutes, namely $MRS_q(\omega, \omega')$ is constant in q for all ω, ω'. It implies that imperfect substitutabilities arise if and only if the platform withholds information and, hence, if and only if there are externalities between some records.

Proposition 4. *All records are perfect substitutes if and only if there is some database $q \in \mathbb{R}^{\Omega}_{++}$ at which it is optimal for the platform to fully disclose the type of every record. In this case, full disclosure is optimal for all databases.*

Combining these properties of the platform's preferences over databases leads to several implications. We highlight three. First, concavity of U^* leads to standard demand analysis. Given market prices for the different types of records, choosing an optimal database subject to a budget constraint is a well-behaved problem, whose solution q^* depends on usual relations between MRS_{q^*} and price ratios. Thus, we can use our values v^*_q to characterize the platform's demand function for records, enabling a general study of the demand side of data markets. This demand satisfies properties that can guide empirical analysis: For example, since $U^*(q)$ is homothetic, data records are normal goods and the optimal database q^* depends only on price ratios, not on the platform's budget. In particular, except in the case of perfect substitutability as in Proposition 4, there is always an open set of prices such that q^* contains multiple types of records. One may also use this demand to determine which prices will prevail in the market depending on supply conditions. Under perfect competition, Dorfman et al. (1987) and Gale (1989) provide arguments for equilibrium prices to equal v^*_q.

A second implication sheds light on a platform's incentives to merge databases. For instance, it may be interested in acquiring another platform to get access to its data, and a regulator may want to assess the potential consequences. Given databases q and q', merging them would lead to a new database $q + q'$.

Corollary 4. *For any databases q and q', the platform's total payoff satisfies $U^*(q + q') \geq U^*(q) + U^*(q')$, with equality if and only if $v^*_q = v^*_{q'}$ generically in q and q'.*

The inequality is intuitive. After the merger, the platform can treat the databases separately and always tell the seller whether a record belongs to q or q'. This would guarantee a total payoff of $U^*(q) + U^*(q')$. Together with Proposition 4, this corollary implies that merging databases can strictly benefit the platform only if the records are not perfect substitutes and, hence, it

it can always reveal to the seller from which original database each record comes, the total payoff from the new database is at least $\gamma U^*(q) + (1-\gamma)U^*(q')$. This argument is related directly to the concavification results in Mathevet et al. (2020) and indirectly to the individual-sufficiency results in Bergemann and Morris (2016).

withholds some information. Note that to predict whether a merger is strictly beneficial, one only needs to know the values of the records in the initial, separate databases (i.e., v_q^* and $v_{q'}^*$).

A third implication is to simplify assessing whether it is optimal for a platform to withhold information given a specific database q. By Proposition 4, the answer can depend on which types of records are in the database, but not on their specific quantities.[21] This is useful in applications. For instance, withholding information is optimal if, starting from full disclosure and *any* conveniently chosen $\hat{q} \in \mathbb{R}^{\Omega}_{++}$, one can show that it is strictly beneficial to conceal *any* type of records in *some* way. It is also possible to find conditions for the optimality of withholding information directly in terms of the primitives of the model (see, e.g., Proposition C.1 in the Online Appendix).

4.2 Better Records and Willingness to Pay for Information

We now turn to studying the consequences of refining existing records with more information. For instance, if a buyer's record originally contained only her age, refining it may entail observing her gender as well. How do refinements change the records' values? Do they always increase the platform's total payoff and command a positive willingness to pay?

We first formalize what a refinement is. We consider refining a single type of records, denoted by $\bar{\omega}$ hereafter. Intuitively, when the platform refines the record of a buyer, it privately observes a new exogenous signal about her underlying preference θ, which may change the belief $\beta(\cdot|\bar{\omega}) \in \Delta(\Theta)$ induced by $\bar{\omega}$ records. Formally, a refinement is then a distribution $\sigma \in \Delta(\Omega)$, where $\sigma(\omega)$ is the probability of observing a signal that updates $\beta(\cdot|\bar{\omega})$ to $\beta(\cdot|\omega)$ and, hence, transforms the refined record from type $\bar{\omega}$ to type ω. Therefore, σ has to satisfy Bayes' consistency: $\beta(\cdot|\bar{\omega}) = \sum_{\omega \in \operatorname{supp} \sigma} \sigma(\omega)\beta(\cdot|\omega)$. Fixing σ, let $\alpha \in [0,1]$ be the share of $\bar{\omega}$ records that are refined, each according to σ. Doing so transforms the original database q into a new one, denoted by q_α, as follows: The quantity of $\bar{\omega}$ records falls to $q_\alpha(\bar{\omega}) = (1-\alpha)q(\bar{\omega})$, while the quantity of ω records rises to $q_\alpha(\omega) = q(\omega) + \alpha q(\bar{\omega})\sigma(\omega)$ for $\omega \in \operatorname{supp} \sigma$. This is akin to assuming records are refined independently according to σ.[22] Note that the composition of q_α is certain, even though it is uncertain which records of type $\bar{\omega}$ become of type ω. Thus, it suffices that the seller knows how the platform refines its database (i.e., σ and α) to know its new composition.

[21] For this reason, the condition in Proposition 4 considers an interior q.

[22] If we interpret $\alpha q(\bar{\omega})$ as a finite but very large quantity of refined records, then with the usual abuse of the law of large numbers, the composition of the new database would be certain and reflect the underlying distribution σ as described. Section 5.1 discusses the case of correlated refinements.

The next result shows that, by changing the database composition, a refinement affects the value not only of the refined records, but also of those that are not being refined.

Corollary 5. *Fix any database q. Suppose a share α of $\bar{\omega}$ records is refined according to σ.*

- *The value of* refined *records increases in expectation: $\sum_{\omega\in\Omega} v^*_{q_\alpha}(\omega)\sigma(\omega) \geq v^*_q(\bar{\omega})$. This increase becomes smaller as α gets larger.*
- *The value of* unrefined *records of type $\bar{\omega}$ increases: $v^*_{q_\alpha}(\bar{\omega}) \geq v^*_q(\bar{\omega})$. The value of* unrefined *records of type $\omega \in \operatorname{supp}\sigma$ decreases: $v^*_{q_\alpha}(\omega) \leq v^*_q(\omega)$. Both effects become larger as α gets larger.*

Consider first refined records. On the one hand, the platform knows more about each of them, so it can better tailor its signals for the seller and achieve better outcomes. On the other hand, the value of the record types that result from the refinement falls, because they are more abundant (Proposition 3). The result shows that, in expectation, the first force always dominates. With regard to unrefined records, the change in values is a consequence of our pooling externalities and the scarcity principle (Proposition 3). They imply, for instance, that the platform's willingness to pay for Ann's record may change when it learns something new about another buyer, even if this news is uninformative about Ann.

Given these mixed effects of refinements on the values of records, one may wonder whether they always benefit the platform overall.

Proposition 5. *Fix any database q. Suppose a share α of $\bar{\omega}$ records is refined according to σ. The platform weakly benefits from the refinement: $U^*(q_\alpha) \geq U^*(q)$. This benefit is zero for all α if (and only if generically in q) there exists an action a such that $x^*_q(a,\omega) > 0$ for $\omega = \bar{\omega}$ and all $\omega \in \operatorname{supp}\sigma$. Finally, the refinement's marginal benefit decreases in α.*

The platform weakly benefits from the refinement for the following reason. A mechanism x that is obedient given q may no longer be obedient given q_α. However, there is always another mechanism $\hat{x}$ that is obedient given q_α and yields the same expected payoff for the platform as does x (see Lemma A.4 in the Appendix). The intuition is twofold. First, with $\hat{x}$, the platform commits to ignoring the information obtained from the refinement, by treating the refined records as x did before the refinement. Second, when considering the composition of the pool of refined records implied by σ, the seller forms the same belief about every buyer's θ as before the refinement (by the Bayes' consistency condition).

The second part of Proposition 5 also provides a sharp condition for when the platform's willingness to pay for a refinement is zero. This condition uniquely depends on how the platform uses records before the refinement. Specifically, at the original q, the platform must sometimes withhold information from the seller by simultaneously pooling $\bar{\omega}$ records with ω records for all $\omega \in \operatorname{supp} \sigma$. This implies it already uses the unrefined records to achieve the same outcomes as it would achieve after any realization of the refinement. In other words, the refinement provides extra information that the platform would strategically withhold anyway, so its willingness to pay for it is zero.[23] Moreover, under this condition, all records have the same value before and after the refinement.

Finally, Proposition 5 sheds light on the overall effects for the platform of obtaining more information, at both the intensive and extensive margin. With regard to the former, fixing the original q and α, suppose refinement σ' is more informative than σ (in Blackwell's sense). Then, the platform benefits weakly more from σ' than from σ; that is, $U^*(q'_\alpha) \geq U^*(q_\alpha)$, where q_α and q'_α are the new databases resulting from σ and σ'. This result is a consequence of the aforementioned Lemma A.4 and Proposition 5 and is reminiscent of classic results for decision problems (see, e.g., Moscarini and Smith, 2002; Varian, 2019). The extensive margin instead consists in fixing the refinement σ and increasing the share α of records to be refined. The last part of Proposition 5 shows that doing so has decreasing returns, which has no counterpart for decision problems.

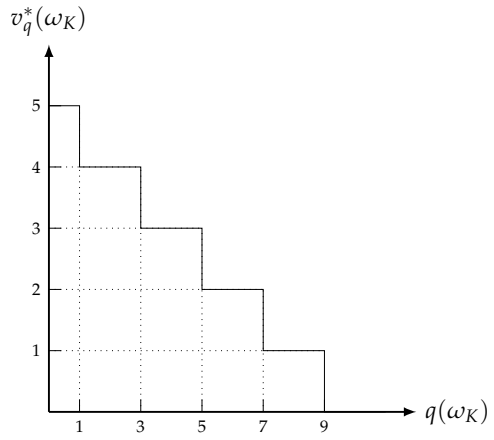
4.3 Application: More and Better Records

We illustrate some of the main ideas in this section by returning to the application of Section 3.3.1. Recall that the platform maximizes a weighted sum of the seller's profits (with weight r) and the buyers' surplus (with weight $1 - r$).

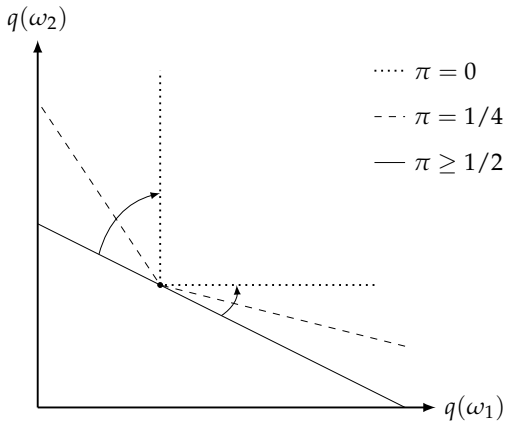
Demand Curve. Fix $r = 0$. Suppose $\Omega = \{\omega_1, \ldots, \omega_K\}$ and each record of type ω_k fully reveals that the buyer's $\theta = k$ for $k = 1, \ldots, K$. Figure 2(a) shows an example of the downward-sloping demand curve implied by Proposition 3, for ω_K records. As $q(\omega_K)$ increases, $v^*_q(\omega_K)$ decreases and eventually reaches its lower bound—the full-disclosure payoff $\underline{v}(\omega_K)$.[24] For $r = 0$, this lower bound is zero because under full disclosure the seller charges the price

[23] Note that this willingness to pay can be zero even if the platform is not indifferent between how it is using records of type $\bar{\omega}$ and $\omega \in \operatorname{supp} \sigma$ in the sense that $u^*_q(\bar{\omega}) \neq u^*_q(\omega)$ (which implies $x^*_q(\cdot, \bar{\omega}) \neq x^*_q(\cdot, \omega)$).

[24] The step-wise behavior of v^*_q is a consequence of the finiteness of $\Omega \times A$ and is a property we prove in Lemma A.3 in the Appendix.



(a) Demand curve when $r = 0$, $\theta_k = k$ ($\forall k$), $q(\omega_k) = 1$ ($\forall k < K$), and $K = 10$.



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

Figure 2: Platform's demand and indifference curves

$a = \omega_K$ and extracts all the buyer's surplus.

Indifference Curves. Suppose $\Omega = \{\omega_1, \omega_2\}$ and again each record of type ω_k fully reveals that $\theta = k$ for $k = 1, 2$. Figure 2(b) illustrates how the records' substitutability depends on the alignment between the platform's and the seller's objectives, namely r. When $r \geq \frac{1}{2}$, the platform fully discloses information and records are perfect substitutes (Proposition 4): Indeed, $MRS_q(\omega_1, \omega_2) = -\frac{\omega_1}{\omega_2}$ for all q. When $r < \frac{1}{2}$, instead, the platform withholds information and records are imperfect substitutes. In particular, as the platform's objective becomes less aligned with the seller's—that is, as r decreases toward 0—the indifference curves become more convex. At $r = 0$, records are perfect complements. This monotonic dependence of the records' substitutability on r extends to $K > 2$, as shown by Corollary A.1 in the Appendix.

Refinements. Fix $r = 0$ and suppose $\Omega = \{\omega_1, \omega_2, \bar{\omega}\}$. As before, let ω_1 and ω_2 reveal that θ is 1 or 2, respectively. Instead, assume $\bar{\omega}$ is partially informative and induces the belief that $\theta = 2$ with probability $h > \frac{1}{2}$ and $\theta = 1$ otherwise. Consider a database $\hat{q}$ that satisfies $\hat{q}(\bar{\omega}) < \hat{q}(\omega_1) < \hat{q}(\omega_2)$. Online Appendix E shows that, in this case, the values are $v_{\hat{q}}^*(\omega_1) = 1$, $v_{\hat{q}}^*(\omega_2) = 0$, and $v_{\hat{q}}^*(\bar{\omega}) = 1 - h$. Moreover, Table 1 reports an optimal mechanism $x_{\hat{q}}^*$. Consider an arbitrary database q satisfying the inequalities above. Suppose the platform refines a share α of $\bar{\omega}$ records by fully learning their underlying θ. That is, $\sigma(\omega_2) = h$ and $\sigma(\omega_1) = 1 - h$. Note that the resulting database q_α satisfies $q_\alpha(\bar{\omega}) < q_\alpha(\omega_1) < q_\alpha(\omega_2)$. As a consquence of the refinement, the platform changes how it uses some of the refined records—compare $x_q^*(\cdot, \bar{\omega})$ and $x_{q_\alpha}^*(\cdot, \omega_2)$—as well as the unrefined records of type ω_2—compare $x_q^*(\cdot, \omega_2)$ and $x_{q_\alpha}^*(\cdot, \omega_2)$. Nonetheless, the "if" condition in Proposition 5 holds and

$x^*_{\hat{q}}(a,\omega)$	ω_1	ω_2	$\bar{\omega}$
$a=1$	$\hat{q}(\omega_1)$	$\hat{q}(\omega_1)-(2h-1)\hat{q}(\bar{\omega})$	$\hat{q}(\bar{\omega})$
$a=2$	0	$\hat{q}(\omega_2)-\hat{q}(\omega_1)-(2h-1)\hat{q}(\bar{\omega})$	0

Table 1: Optimal $x^*_{\hat{q}}$ when $\hat{q}(\bar{\omega}) < \hat{q}(\omega_1) < \hat{q}(\omega_2)$.

the platform does not benefit from the refinement: $U^*(q) = U^*(q_\alpha)$. Moreover, the value of the refined record does not change in expectation: $v^*_{q_\alpha}(\omega_1)\sigma(\omega_1) + v^*_{q_\alpha}(\omega_2)\sigma(\omega_2) = v^*_q(\bar{\omega})$.

5 Discussion

5.1 Correlation between Records

In our analysis, we assumed each buyer's record is uninformative about other buyers' preferences. We did so to clarify that the pooling externality we highlighted arises not from exogenous correlation between records, but endogenously from how records are used. This assumption can be relaxed. For a fixed database, each buyer's record should already contain all the available information about her θ, which may include variables that refer to other individuals. For example, if Ann's income is predictive of Briana's θ, it should be listed in Briana's record. Suppose this assignment is done for each buyer. Then, by construction, Ann's record adds no information about Briana's preferences that is not already contained in Briana's record. Thus, our analysis for a fixed database q is unchanged.

The possibility that one buyer's data is informative about other buyers can have deeper implications when changing q. For example, consider the case of refinements. Suppose observing Ann's income requires updating Briana's record. Such a correlated refinement can induce a distribution over resulting databases (i.e., q_α) that have significantly different compositions, which in addition only the platform may observe. By contrast, the refinements in Section 4.2 led to a deterministic q_α that, from the seller's viewpoint, essentially involved the same composition of possible buyers' preferences as the original database q. One may then wonder whether the results in that section extend to the case of correlated refinements. The answer depends on how much the seller learns about the resulting database q_α, which in practice may depend on how publicly the platform refines records. To illustrate this point, Online Appendix F presents an example where the seller learns q_α and a refinement can strictly decrease both the value of the refined records and the platform's total payoff. In a nutshell, this is because cor-

related refinements can significantly change what information the platform has in its database and, hence, can use to influence the seller's behavior.

5.2 The Relation with Decision Problems

We briefly explain how our analysis would change if, instead of intermediation problems, we focused on decision problems in the tradition of Blackwell (1951, 1953). A decision-maker (DM) observes a signal realization ω with probability $q(\omega)$, which induces a belief $\beta(\cdot|\omega)$ about a state θ. She then chooses an action a to maximize her expected payoff $u(a,\omega)$. Let $x(a,\omega) \geq 0$ describe the joint probability that DM chooses action a when the signal is ω. Then, we can view DM as solving

$$\begin{aligned} \mathcal{D}_q : \qquad & \max_{x} \quad \sum_{\omega\in\Omega, a\in A} u(a,\omega)x(a,\omega) \\ & \text{s.t.} \quad \sum_{a\in A} x(a,\omega) = q(\omega) \quad \text{for all } \omega \in \Omega. \end{aligned}$$

Note that problem $\mathcal{D}_q$ is identical to $\mathcal{U}_q$ in Section 3.1, net of the obedience constraints. In this case, many of our questions have standard answers. In $\mathcal{D}_q$, the action DM optimally chooses given ω never depends on what she plans to choose given any other ω'. This implies x^* does not depend on q and $u^*(\omega) = \max_a u(a,\omega)$ for all ω. By formulating the dual of $\mathcal{D}_q$, we could derive the value of each signal realization ω. It is immediate to see that, in this case, $v^*(\omega) = u^*(\omega)$ for all ω and independently of q.[25] Thus, decision problems such as $\mathcal{D}_q$ never give rise to the kind of externalities highlighted in Section 3.2. In other words, the direct payoff from a signal realization gives a correct assessment of the value DM derives from it. Consequently, refinements in the sense of Section 4.2 would affect only the value of the refined signal realizations, while leaving all the other values unchanged.

This comparison with decision problems such as $\mathcal{D}_q$ further clarifies the role of obedience constraints in our analysis. These constraints are a defining feature of intermediation problems and create interdependencies in how data records are used, which then cause the pooling externality highlighted in this paper. Other types of constraints may introduce interdependencies in the use of data records, even in the context of decision problems. For example, one can add constraints to $\mathcal{D}_q$ that require DM to take specific actions with at least some probability. When these constraints are linear, our dual approach remains valid. We conjecture that, in this case,

[25] One can then use this value to calculate the net ex-post value of a piece of information as suggested by Frankel and Kamenica (2019).

v^* may differ systematically from u^* and may depend on q. We leave exploring such problems for future research.

6 Conclusion

This paper addressed the question of how much the data of a single individual contributes to fueling the business of firms in the digital economy. The paper proposes a unifying and classical approach and identifies a novel externality between data records. Our analysis has several implications: for compensating individuals for the collection and use of their data, for guiding companies' investments in data acquisition, and more broadly for studying the demand side of data markets, which may ultimately inform regulation.

Our approach can be used to study the value of data records in a broad range of intermediation problems, beyond our leading case of e-commerce platforms. Online advertisement slots are sold using data about keywords submitted by users of search engines or about the behavior of users on social-media platforms. Universities use data on past academic performance of their students for recommending them to potential employers so as to improve employment outcomes. Ride-sharing companies use riders' locations and requested destinations to match them with drivers. Navigation-service apps may use drivers' data to recommend routes and minimize congestion.

Nonetheless, our framework makes assumptions that may be restrictive in some other applications. For example, we considered settings with only one platform, thus ignoring the issue of cross-platform competition. We allowed our platform to send any information to the seller, whereas in practice privacy regulations may restrict how it can use its database. Finally, we assumed the platform has commitment power, which may be unrealistic in some settings. These extensions could break the linear-programming formulation of the problem and complicate the approach presented in this paper. These areas remain open for future research.

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 (2022): "Voluntary Disclosure and Personalized Pricing," *forthcoming, Review of Economic Studies*.

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 (2022): "The economics of social data," *The RAND Journal of Economics*, 53, 263–296.

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).

Bergemann, D. and M. Ottaviani (2021): "Information Markets and Nonmarkets," *Handbook of Industrial Organization, forthcoming*, 4.

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

Blackwell, D. (1951): "Comparison of Experiments," *Proc. Second Berkeley Sympos. on Mathematical Statistics and Probability*.

——— (1953): "Equivalent comparisons of experiments," *Annals of mathematical statistics*, 265–272.

Bloedel, A. W. and W. Zhong (2021): "The Cost of Optimally Acquired Information," *Working Paper*.

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.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2019): "Privacy and personal data collection with information externalities," *Journal of Public Economics*, 173, 113–124.

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, A. Koh, and W. Li (2021): "Market Segmentation through Information," *Working Paper*.

Federal Trade Commission (2014): *Data Brokers: A Call for Transparency and Accountability*, A Report by the Federal Trade Commission, May.

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*.

Ichihashi, S. (2020): "Online Privacy and Information Disclosure by Consumers," *American Economic Review*, 110, 569–95.

——— (2021): "The Economics of Data Externalities," *Journal of Economic Theory*.

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

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.

Morris, S. and P. Strack (2019): "The Wald Problem and the Relation of Sequential Sampling and Ex-Ante Information Costs," *Working Paper*.

Moscarini, G. and L. Smith (2002): "The law of large demand for information," *Econometrica*, 70, 2351–2366.

Myerson, R. B. (1982): "Optimal coordination mechanisms in generalized principal–agent problems," *Journal of Mathematical Economics*, 10, 67–81.

——— (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.

Pomatto, L., P. Strack, and O. Tamuz (2021): "The Cost of Information," *Working Paper*.

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

Sayedi, A., K. Jerath, and K. Srinivasan (2014): "Competitive Poaching in Sponsored Search Advertising and Its Strategic Impact on Traditional Advertising," *Marketing Science*, September, 33 (4), 586–608.

Seim, K., D. Bergemann, J. Cremer, D. Dinielli, C. C. Groh, P. Heidhues, D. Schaefer, M. Schnitzer, F. M. Scott Morton, and M. Sullivan (2022): "Market Design for Personal Data," *Policy Discussion Paper*, No. 6, Tobin Center for Economic Policy, Yale University, April.

Stigler Report (2019): "Stigler Committee on Digital Platforms," *Final Report*, available at https://research.chicagobooth.edu/stigler/media/news/committee-on-digitalplatforms-final-report, September.

Varian, H. (2019): *16. Artificial Intelligence, Economics, and Industrial Organization*, University of Chicago Press.

Yang, K. H. (2022): "Selling consumer data for profit: Optimal market-segmentation design and its consequences," *American Economic Review*, 112, 1364–93.

A Appendix

All proofs in this appendix are for the following model, which generalizes the one of Section 2. There are $n > 1$ sellers (agents), where their set is $I = \{1, \ldots, n\}$. Let $i = 0$ denote the platform (principal). Let A_i be the finite set of actions controlled by $i = \{0, 1, \ldots, n\}$ and let $A = A_0 \times \cdots \times A_n$. For each buyer, each seller may privately observe some data, which we model as an exogenous signal ω_i in some finite set Ω_i about the 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 platform also observes the private data of each seller—i.e., the entire $\omega = (\omega_0, \ldots, \omega_n)$—as does the omniscient designer in Bergemann and Morris (2016).[26] Thus, now the whole vector ω defines a type of data record in the platform's database q. For every $a = (a_0, \ldots, a_n)$, let $u(a, \omega)$ be the expected payoff of the platform and $\pi_i(a, \omega)$ be the expected profit of seller i calculated using the belief over θ defined by ω.

In this general case, we can formulate the platform's intermediation problem in terms of mechanisms $x : A \times \Omega \to \mathbb{R}_+$ as follows:

$$\begin{aligned}
\mathcal{U}_q : \qquad & \max_{x} \sum_{\omega \in \Omega, a \in A} u(a, \omega) x(a, \omega) \\
& \text{s.t. for all } i \in I, \omega_i \in \Omega_i, \text{ and } a_i, a_i' \in A_i, \\
& \sum_{\omega_{-i} \in \Omega_{-i}, a_{-i} \in A_{-i}} \left(\pi_i(a_i, a_{-i}, \omega) - \pi_i(a_i', a_{-i}, \omega) \right) x(a_i, a_{-i}, \omega) \geq 0, \quad \text{(A.1)} \\
& \text{and for all } \omega \in \Omega, \\
& \sum_{a \in A} x(a, \omega) = q(\omega). \quad \text{(A.2)}
\end{aligned}$$

General Derivation of the Data-Value Problem $\mathcal{V}_q$. We now derive the general version of problem $\mathcal{V}_q$ in Section 3.1. To this end, it is convenient to express $\mathcal{U}_q$ in matrix form. Fix an arbitrary total ordering of the set $A \times \Omega$. We denote by $\mathbf{u} \in \mathbb{R}^{A \times \Omega}$ the vector whose entry corresponding to (a, ω) is $u(a, \omega)$. For every player i, let $\mathbf{\Pi}_i \in \mathbb{R}^{(A_i \times A_i \times \Omega_i) \times (A \times \Omega)}$ 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

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

Thus, $\mathbf{\Pi}_i(a_i', a_i'', \omega_i')$ denotes the row labeled by (a_i', a_i'', ω_i') (which defines the corresponding obedience constraint) and $\mathbf{\Pi}_i(a, \omega)$ denotes the column labeled by (a, ω). Define the matrix $\mathbf{\Pi}$

[26] Online Appendix B illustrates how to incorporate participation constraints.

by stacking all the matrices $\{\mathbf{\Pi}_i\}_{i\in I}$ on top each other. Each row of this matrix corresponds to one obedience constraint of $\mathcal{U}_q$. Finally, define the indicator matrix $\mathbf{I} \in \{0,1\}^{\Omega\times(A\times\Omega)}$ such that, for each row ω' and column (a,ω'),

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

Treating q as a vector in $\mathbb{R}^{\Omega}_{+}$ and x as a vector in $\mathbb{R}^{A\times\Omega}$, we can then write $\mathcal{U}_q$ as

$$\begin{array}{lrcl} \max\limits_{x} & \mathbf{u}^T x & & \\ \text{s.t.} & \mathbf{\Pi} x & \geq & \mathbf{0}, \\ & \mathbf{I} x & = & q, \\ & x & \geq & \mathbf{0}. \end{array} \tag{A.3}$$

We can now invoke standard linear-programming arguments to derive the dual $\mathcal{V}_q$ of $\mathcal{U}_q$ (see, e.g., Bertsimas and Tsitsiklis (1997), p. 142). Let $v(\omega)$ be the dual variable corresponding to constraint (A.2) for each $\omega \in \Omega$ and $\lambda_i(a_i'|a_i,\omega_i)$ be the dual variable corresponding to the obedience constraint (A.1) for agent i when he has data ω_i, is recommended a_i, and considers deviation a_i'. Note that v can be viewed as a vector in $\mathbb{R}^{\Omega}$ and λ as a vector with as many entries as the rows of $\mathbf{\Pi}$. We can write $\mathcal{V}_q$ as

$$\begin{array}{lrcl} \min\limits_{\lambda,v} & \lambda^T\mathbf{0} & + & v^T q \\ \text{s.t.} & \mathbf{I}^T v & \geq & \mathbf{u} + \mathbf{\Pi}^T\lambda \\ & \lambda & \geq & \mathbf{0}. \end{array}$$

Thus, the objective simplifies to

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

The constraints on λ require that $\lambda_i(a_i'|a_i,\omega_i) \geq 0$ for all $i \in I$, $a_i, a_i' \in A_i$, and $\omega_i \in \Omega_i$. The first set of constraints requires that, for all $(a,\omega) \in A\times\Omega$,

$$v(\omega) \geq u(a,\omega) + \sum_{i\in I}\left\{\sum_{a_i'\in A_i}\left(\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega)\right)\lambda_i(a_i'|a_i,\omega_i)\right\}. \tag{A.4}$$

This $\mathcal{V}_q$ with $|I| = 1$ boils down to the formulation in Section 3.1.

A Remark on Non-degeneracy and the Structure of Solutions. In Section 2, we assumed that no more than $|A \times \Omega|$ of the constraints (2) are ever active at the same time. We now formalize that assumption following Bertsimas and Tsitsiklis (1997). Consider the polyhedron defined by the constraints in (A.3) and recall that $x \in \mathbb{R}_{+}^{A\times\Omega}$, which has dimensions $|A \times \Omega|$. A basic feasible solution of $\mathcal{U}_q$ is an x such that (*i*) all equality constraints are active, (*ii*) $|A \times \Omega|$ of the constraints active at x are linearly independent, and (*iii*) all constraints are satisfied. Formally, we assume the following.

Assumption 1 (Non-degeneracy)**.** *At every basic feasible solution x of problem $\mathcal{U}_q$ there are only $|A \times \Omega|$ active constraints.*

The next remark describes the structure of optimal solutions of $\mathcal{U}_q$ and $\mathcal{V}_q$.

Remark 1. *We can transform $\mathcal{U}_q$ into the standard form $\mathcal{U}_q^S$, which can be written as follows (Bertsimas and Tsitsiklis (1997), p. 53):*

$$\begin{aligned} \max_{x,s} \quad & \mathbf{u}^T x \\ s.t. \quad & \mathbf{\Pi} x - s = \mathbf{0}, \\ & \mathbf{I} x = q, \\ & x, s \geq \mathbf{0}, \end{aligned} \tag{A.5}$$

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 x_q^*, which is generically unique and hence corresponds to an extreme point of the polyhedron of feasible x. Moreover, this x_q^* is an optimal solution of $\mathcal{U}_q^S$ as well. The extreme point x_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 $\left[\begin{array}{c|c} \mathbf{\Pi} & -\mathbf{1} \\ \hline I & \mathbf{0} \end{array}\right]$, where $\mathbf{1}$ is the identity matrix. As illustrated in (Bertsimas and Tsitsiklis, 1997, Chapter 4), given $\mathbf{B}$, we have

$$\begin{bmatrix} x_q^* \\ \hline s_q^* \end{bmatrix} = \mathbf{B}^{-1} \begin{bmatrix} \mathbf{0} \\ \hline q \end{bmatrix}, \tag{A.6}$$

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

$$\begin{bmatrix} \lambda_q^* \\ \hline v_q^* \end{bmatrix} = \mathbf{u}\mathbf{B}^{-1}. \tag{A.7}$$

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

Generalization and Proof of Lemma 1. For the model with multiple agents, Lemma 1 takes the following form. Let $\Gamma_\omega = \{I, (A_i)_{i=0}^n, (\pi_i(\cdot,\omega))_{i=1}^n\}$ be the complete-information game between the agents defined by the primitive of the model given ω, where the principal is a dummy player. For $\omega \in \Omega$, let $CE(\Gamma_\omega)$ be the set of correlated equilibria of the game Γ_ω; that is, given $y \in \Delta(A)$, $y \in CE(\Gamma_\omega)$ if and only if, for all $i \in I$ and $a_i, a_i' \in A_i$,

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

Lemma A.1. *For every database q,*

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

Moreover, $v_q^*(\omega) = \underline{v}(\omega)$ for all ω if and only if there is an optimal mechanism x_q^* such that $\frac{x_q^*(\cdot,\omega)}{q(\omega)} \in CE(\Gamma_\omega)$ for all ω.

Proof. Fix an optimal solution (v_q^*, λ_q^*) of $\mathcal{V}_q$. For every q, $\omega \in \Omega$, and x such that $\frac{x(\cdot,\omega)}{q(\omega)} \in CE(\Gamma_\omega)$, by (A.4) we have

$$\begin{aligned}
v_q^*(\omega) &\geq \sum_{a \in A} u(a,\omega) \frac{x(a,\omega)}{q(\omega)} \\
&\quad + \sum_{a \in A} \left\{ \sum_{i \in I} \sum_{\hat{a}_i \in A_i} \left(\pi_i(a_i, a_{-i}, \omega) - \pi_i(\hat{a}_i, a_{-i}, \omega)\right) \lambda_i^*(\hat{a}_i | a_i, \omega_i) \right\} \frac{x(a,\omega)}{q(\omega)} \\
&= \sum_{a \in A} u(a,\omega) \frac{x(a,\omega)}{q(\omega)} \\
&\quad + \sum_{i \in I} \sum_{a_i, \hat{a}_i \in A_i} \lambda_i^*(\hat{a}_i | a_i, \omega_i) \left\{ \sum_{a_{-i} \in A_{-i}} \left(\pi_i(a_i, a_{-i}, \omega) - \pi_i(\hat{a}_i, a_{-i}, \omega)\right) \frac{x(a,\omega)}{q(\omega)} \right\} \\
&\geq \sum_{a \in A} u(a,\omega) \frac{x(a,\omega)}{q(\omega)},
\end{aligned}$$

where the last inequality follows from the definition of any element of $CE(\Gamma_\omega)$. Since $\frac{x(a,\omega)}{q(\omega)}$ is an arbitrary element of $CE(\Gamma_\omega)$, we conclude that $v_q^*(\omega) \geq \underline{v}(\omega)$.

To prove the second part of the lemma, first suppose $v_q^*(\omega) = \underline{v}(\omega)$ for all ω. Consider any mechanism x that satisfies (A.2), $\frac{x(a,\omega)}{q(\omega)} \in CE(\Gamma_\omega)$—which implies that x also satisfies (A.1)—and

$$\underline{v}(\omega) = \sum_{a \in A} u(a,\omega) \frac{x(a,\omega)}{q(\omega)}, \quad \omega \in \Omega.$$

It follows that $\sum_\omega v_q^*(\omega) q(\omega) = \sum_{a,\omega} u(a,\omega) x(a,\omega)$. Therefore, x is an optimal solution of $\mathcal{U}_q$ by strong duality. Conversely, suppose there is an optimal x_q^* such that $\frac{x_q^*(\cdot,\omega)}{q(\omega)} \in CE(\Gamma_\omega)$

for all ω. By strong duality and $v_q^*(\omega) \geq \underline{v}(\omega) = u_q^*(\omega)$ for all ω, we must have $v_q^*(\omega) = \underline{v}(\omega)$ for all ω. □

Proof of Proposition 1. For each $i \in I$ and $(a, \omega) \in A \times \Omega$, define

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

and $t(a,\omega) \stackrel{\Delta}{=} \sum_{i\in I} t_i(a,\omega)$. Denote the $t(a,\omega)$ resulting from λ_q^* by $t_q^*(a,\omega)$. By complementary slackness, $x_q^*(a,\omega) > 0$ implies $v_q^*(\omega) = u(a,\omega) + t_q^*(a,\omega)$. Hence,

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

Fix ω. Suppose we start from database q with $q(\omega) > 0$ and we increase $q(\omega)$ to $\hat{q}(\omega)$, thus obtaining 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 strong duality (i.e., (7)), we obtain that

$$\begin{aligned}
t_q^*(\omega) &= v_q^*(\omega) - u_q^*(\omega) = \frac{\partial U^*(q)}{\partial q(\omega)} - u_q^*(\omega) \\
&= \lim_{\hat{q}(\omega)\to q(\omega)} \frac{\sum_{\omega'\in\Omega}[u_{\hat{q}}^*(\omega') - u_q^*(\omega')]\hat{q}(\omega')}{\hat{q}(\omega) - q(\omega)} = \sum_{\omega'\in\Omega} \frac{\partial u_q^*(\omega')}{\partial q(\omega)} q(\omega') \\
&= \sum_{\omega'\in\Omega, a\in A} u(a,\omega') \left(\lim_{\hat{q}(\omega)\to q(\omega)} \frac{[x_{\hat{q}}^*(a,\omega')/\hat{q}(\omega') - x_q^*(a,\omega')/q(\omega')]}{\hat{q}(\omega) - q(\omega)} \right) q(\omega') = \\
&\stackrel{\text{a.e.}}{=} \sum_{\omega'\in\Omega, a\in A} u(a,\omega') \frac{\partial}{\partial q(\omega')} \left(\frac{x_q^*(a,\omega')}{q(\omega)} \right) q(\omega'),
\end{aligned}$$

where the existence of the derivative of $\frac{x_q^*(a,\omega')}{q(\omega)}$ almost everywhere follows from (A.6). □

Primal Solutions for the Application in Section 3.3.1. As mentioned after Proposition 2, for the setting of Section 3.3.1 the solution of problem $\mathcal{U}_q$ can take only two forms depending on r. Let $\bar{x}^*$ be the profit-maximizing solution (i.e., for $r = 1$) and $\underline{x}^*$ be the surplus-maximizing solution (i.e., for $r = 0$) (see Bergemann et al. (2015) for details).

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

Proof. First, note that we can write $u(a,\omega) = ra\mathbb{I}\{\omega \geq a\} + (1-r)\max\{\omega - a, 0\}$ as

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

which is strictly increasing in a if and only if $r > \frac{1}{2}$. Fix any (non-trivial) q and $r \in (0,1)$. Problem $\mathcal{U}_q$ involves maximizing

$$\sum_{\omega,a} u_r(a,\omega)x(a,\omega) = \sum_{\omega \geq a} [a(2r-1) + (1-r)\omega]x(a,\omega)$$

subject to constraints (2).

Suppose that $r > \frac{1}{2}$. Note that $\bar{x}^*$ is feasible and maximizes the objective function pointwise for every ω. Indeed, since $\bar{x}^*(\omega,\omega) = q(\omega)$, for every ω we have that $\bar{x}^*$ selects the highest $a \leq \omega$ for every ω, thereby maximizing $a(2r-1)\mathbb{I}\{\omega \geq a\}$; 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 $r = \frac{1}{2}$. Suppose now that $r < \frac{1}{2}$. Now for each ω the objective is to pair ω with the smallest possible a and do so with the highest probability allowed by (2). This is what $\underline{x}^*$ essentially does. We can again invoke the Theorem of the Maximum to extend the optimality of $\underline{x}^*$ at $r = \frac{1}{2}$. □

Uniqueness and Stability of the Values of Data Records. We establish the following uniqueness and stability properties of v_q^*, which we will use to prove other results below.

Lemma A.3. *There exists a finite collection $\{Q_1, \ldots, Q_M\}$ of open, convex, and disjoint cones in $\mathbb{R}_+^\Omega$ such that $\cup_m Q_m$ has full measure and, for every m, v_q^* is unique and constant for $q \in Q_m$.*

Proof. The argument leverages the structure of the polyhedron of feasible solutions of $\mathcal{V}_q$, denoted by $F(\mathcal{V}_q)$. By the formulation of $\mathcal{V}_q$ and Lemma 1, $F(\mathcal{V}_q)$ does not contain a line because all dual variables are bounded from below. By Theorem 2.6 and Corollary 2.1 in Bertsimas and Tsitsiklis (1997), $F(\mathcal{V}_q)$ has at least one and at most finitely many extreme points. 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 $F(\mathcal{V}_q)$.

Next, we characterize some properties of the dual solutions. Fix q and suppose that the optimal solution (v_q^*, λ_q^*) of $\mathcal{V}_q$ is unique. As explained in Remark 1, there exists a submatrix $\mathbf{B}$ such that (v_q^*, λ_q^*) satisfies (A.7). Given Assumption 1, Theorem 3.1 and Exercise 3.6 in Bertsimas and Tsitsiklis (1997) imply that

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

The inequality is strict for each row of $\mathbf{\Pi}$ that corresponds to $\lambda^*_{q,i}(a'_i|a_i,\omega_i) = 0$ (i.e., the corresponding obedience constraint is not binding):

$$\left[\mathbf{\Pi}_i(a_i,a'_i,\omega_i) \mid -\mathbf{1}_i(a_i,a'_i,\omega_i)\right]\mathbf{B}^{-1}\begin{bmatrix}\mathbf{0}\\ q\end{bmatrix} > 0, \tag{A.8}$$

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 $\mathbf{\Pi}$ that corresponds to $\lambda^*_{q,i}(a'_i|a_i,\omega_i) > 0$ (i.e., the corresponding obedience constraint is binding), it holds that $\left[\mathbf{\Pi}_i(a_i,a'_i,\omega_i) \mid -\mathbf{1}_i(a_i,a'_i,\omega_i)\right]\mathbf{B}^{-1}\begin{bmatrix}\mathbf{0}\\ q\end{bmatrix} = 0$ as long as $\mathbf{B}$ identifies the optimal extreme point.

Now consider how changes in q affect these optimality properties. Note that q only enters the objective of $\mathcal{V}_q$. Each condition (A.8) 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 (A.7) and

$$Q(\mathbf{B}) = \{q : (A.8) \text{ holds for all } i \in I \text{ and } (a_i,a'_i,\omega_i) \text{ s.t. } \lambda^*_{\mathbf{B},i}(a_i|a'_i,\omega_i) = 0\}.$$

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

Finally, recall that there are only finitely many extreme points of the dual polyhedron of feasible solutions. Therefore, there are finitely many submatrices $\{\mathbf{B}_1,\ldots,\mathbf{B}_M\}$ such that each identifies an optimal $(v^*_{\mathbf{B}_m},\lambda^*_{\mathbf{B}_m})$, where $v^*_{\mathbf{B}_m}$ is unique for all $q \in Q(\mathbf{B}_m)$. For all $m = 1,\ldots,M$, define $Q_m = Q(\mathbf{B}_m)$. By construction, each Q_m is open and $q,q' \in Q_m$ implies that $v^*_q = v^*_{q'}$. Since v^*_q is generically unique with respect to q, it follows that $\mathbb{R}^\Omega_+ \setminus \cup_m Q_m$ has Lebesgue measure zero. □

Proof of Proposition 3. The proof uses a monotone-comparative-static argument that relies on establishing convenient orders for the vectors μ, which describe the shares of record types in a database, and for the dual variables v and λ. Fix ω and suppose that $\mu_q(\omega) > \mu_{q'}(\omega)$ for databases q and q'. Let $\Omega^q = \{\hat{\omega} \in \Omega : \mu_q(\hat{\omega}) > \mu_{q'}(\hat{\omega})\}$, $\Omega^{q'} = \{\hat{\omega} \in \Omega : \mu_{q'}(\hat{\omega}) > \mu_q(\hat{\omega})\}$, and $\overline{\Omega} = \Omega \setminus \{\Omega^q \cup \Omega^{q'}\}$. Note that (v,λ) belongs to the set $Y = \mathbb{R}^\Omega \times \mathbb{R}_+^{A_1\times A_1\times\Omega_1} \times \ldots \times \mathbb{R}_+^{A_n\times A_n\times\Omega_n}$. Associate the canonical component-wise order with Y, except that this order is reversed for $\hat{\omega} \in \Omega^q$. Then, Y is a lattice. Treating μ_q and $\mu_{q'}$ as parameters of the problem, we order the set $\{\mu_q,\mu_{q'}\}$ as $\mu_q > \mu_{q'}$.

Now note that we can equivalently state the data-value problem as $\max_{(v,\lambda)\in S} f(v,\lambda;\mu)$, where $f(v,\lambda;\mu) = -\sum_{\omega\in\Omega} v(\omega)\mu(\omega)$ and the feasible set $S \subset Y$ is given by the inequalities

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

Next, we show that f is supermodular in (v,λ) and has increasing differences in $(v,\lambda;\mu)$ for $\mu \in \{\mu_q, \mu_{q'}\}$ given our order. Observe that for any (v',λ') and (v'',λ'')

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

Then f is supermodular in (v,λ). Fix $(v',\lambda') \geq (v,\lambda)$. Given $\mu_q > \mu_{q'}$, observe that

$$\begin{aligned} & (f(v',\lambda',\mu_q) - f(v,\lambda,\mu_q)) - (f(v',\lambda',\mu_{q'}) - f(v,\lambda,\mu_{q'})) \\ = & \sum_{\hat{\omega}\in\Omega} (v(\hat{\omega}) - v'(\hat{\omega}))(\mu_q(\hat{\omega}) - \mu_{q'}(\hat{\omega})) \\ = & \sum_{\hat{\omega}\in\Omega^q} (v(\hat{\omega}) - v'(\hat{\omega}))(\mu_q(\hat{\omega}) - \mu_{q'}(\hat{\omega})) + \sum_{\hat{\omega}\in\Omega^{q'}} (v(\hat{\omega}) - v'(\hat{\omega}))(\mu_q(\hat{\omega}) - \mu_{q'}(\hat{\omega})), \end{aligned}$$

which is non-negative given the partial orders we adopted. Then, f has increasing differences in $(v,\lambda;\mu)$.

It then follows that $\arg\max_{(v,\lambda)\in S} f(v,\lambda;\mu_q) \geq \arg\max_{(v,\lambda)\in S} f(v,\lambda;\mu_{q'})$ by Theorem 5 in Milgrom and Shannon (1994). This result and the generic uniqueness of $v^*_{\hat{q}}$ with respect to $\hat{q}$ imply that $v^*_q(\omega) \leq v^*_{q'}(\omega)$. That is, this monotonicity of $v^*_{\hat{q}}(\omega)$ holds for any selection $v^*_{\hat{q}}$ from the optimal solution correspondence of $\mathcal{V}_{\hat{q}}$.

We now prove the second part of the proposition. When only records of type ω are present in the database (i.e., $\mu_q(\omega) = 1$), we have $v^*_q(\omega) = \underline{v}(\omega)$. This follows immediately from Lemma A.1. We will show that $v^*_q(\omega) = \underline{v}(\omega)$ continues to hold in a neighborhood of $\mu_q(\omega) = 1$. For $\varepsilon > 0$, consider a set $M_\varepsilon(\omega)$ defined as $M_\varepsilon(\omega) = \{\mu \in \Delta(\Omega) : \mu(\omega') \in (0,\varepsilon)$ for $\omega \neq \omega', \mu(\omega) < 1\}$. By Lemma A.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 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 $\hat{v}(\omega)$. If $\hat{v}(\omega) = \underline{v}(\omega)$, then the result follows. Suppose, on the contrary, that $\hat{v}(\omega) \neq \underline{v}(\omega)$. 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 $n \in \mathbb{N}$, $v^*_q(\omega) = \hat{v}(\omega)$ for every q such that $\mu_q = \mu^n$. By the Berge's maximum theorem, (v^*_q, λ^*_q) is an upper-hemicontinuous correspon-

dence and therefore has a closed graph. Hence, $\hat{v}(\omega) \in v_q^*(\omega)$ for every q, with $\mu_q = \tilde{\mu}$. We obtain the desired contradiction, since $v_q^*(\omega) = \underline{v}(\omega)$ for such q.

Proof of Proposition 4. If all types of records are perfect substitutes, $MRS_q(\omega, \omega') = -\frac{v_q^*(\omega)}{v_q^*(\omega')}$ must be constant for all (ω, ω') and q. By Lemma A.1 and Proposition 3, it follows that $v_q^*(\omega) = \underline{v}(\omega)$ for all ω and q. Therefore, it is optimal to always fully disclose every record. □

Fix $q \in \mathbb{R}^{\Omega}_{++}$. Suppose that an optimal mechanism x_q^* involves full disclosure. Then, we have

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

where the inequality follows because $\frac{x_q^*(\cdot, \omega)}{q(\omega)} \in CE(\Gamma_\omega)$ for all ω. Since by strong duality we must have $\sum_\omega v_q^*(\omega) q(\omega) = \sum_\omega u_q^*(\omega) q(\omega)$ (see (7)), it follows that $v_q^*(\omega) = u_q^*(\omega)$ for all ω. Finally, since x_q^* is optimal, it must be that $u_q^*(\omega) = \underline{v}(\omega)$ 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 $\hat{q}_\omega(\omega) = \frac{U^*(q)}{\underline{v}(\omega)}$ 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 $\kappa \in \Delta(\Omega)$ such that $q'(\omega) = \kappa(\omega) \hat{q}_\omega(\omega)$ for all ω. By concavity of $U^*(q)$ (Footnote 20), we must have that

$$U^*(q') = \sum_{\omega \in \Omega} v_{q'}^*(\omega) q'(\omega) \leq U^*(q) = \sum_{\omega \in \Omega} \kappa(\omega) U^*(\hat{q}_\omega) = \sum_{\omega \in \Omega} \underline{v}(\omega) q'(\omega).$$

But since $v_{q'}^*(\omega) \geq \underline{v}(\omega)$ for all ω by Lemma A.1, we must have $v_{q'}^*(\omega) = \underline{v}(\omega)$ for all ω. Then $v_{q''}^*(\omega) = \underline{v}(\omega)$ 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,[27] it follows that $v_q^*(\omega) = \underline{v}(\omega)$ for all ω and all $q \in \mathbb{R}^{\Omega}_{+}$. □

Proof of Corollary 4. Since U^* is concave and homogeneous of degree 1, we have $U^*(q + q') = 2U^*\left(\frac{1}{2}q + \frac{1}{2}q'\right) \geq 2\left(\frac{1}{2}U^*(q) + \frac{1}{2}U^*(q')\right)$. If $v_q^* = v_{q'}^*$, then $v_q^* = v^*_{\frac{1}{2}q + \frac{1}{2}q'}$ by Proposition 3. Strong duality (i.e., (7)) implies that $U^*\left(\frac{1}{2}q + \frac{1}{2}q'\right) = \frac{1}{2}U^*(q) + \frac{1}{2}U^*(q')$. Conversely,

[27] It is easy to see that v_q^* is constant along the rays in the space of databases: If $q' = \delta q$ for $\delta > 0$, then $v_q^* = v_{q'}^*$.

if $U^*\left(\frac{1}{2}q+\frac{1}{2}q'\right)=\frac{1}{2}U^*(q)+\frac{1}{2}U^*(q')$, then since U^* is concave, it must be linear over the line defined by all convex combinations of q and q'. By homogeneity of degree 1, U^* must also be linear over the open cone that contains this line, which means that $v^*_{\hat{q}}$ is constant over this cone (see Footnote 27). Hence, v^*_q can differ from $v^*_{q'}$ only if one of them is on the boundary of this cone, which is a nongeneric condition. □

Application of Section 4.3: Indifference Curves for $K>2$. The next result characterizes how the records' substitutability depends on r when there are more than two types of records.

Corollary A.1. *Fix q and increase $r<\frac{1}{2}$. 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 toward $-\frac{\omega}{\omega'}$ from below. If $\omega'>\omega\geq a_q$, $MRS_q(\omega,\omega')$ decreases monotonically toward $-\frac{\omega}{\omega'}$ from above.*

In words, as r increases toward $\frac{1}{2}$, for record types on the opposite side of a_q the platform's indifference curves rotate counterclockwise in the direction of perfect substitutability. For records on the same side of a_q, its indifference curves rotate clockwise in the direction of perfect substitutes. Thus, the indifference curves become "less convex" around the dimension $\omega=a_q$. In particular, at $r=0$ records of type $\omega=a_q$ are perfect complements with every other type.

Proof. For $r>\frac{1}{2}$, we have $MRS_q(\omega,\omega')=-\frac{\omega}{\omega'}$ and all ω,ω'. Consider now $r<\frac{1}{2}$:

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

Thus, we have

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

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 Corollary 5. For this result, we assume that the agents observe no data (i.e., $|\Omega_i|=1$ for all $i=1,\ldots,n$). Fix q, the type of refined records $\bar{\omega}$, and the refinement σ. By the law

of iterated expectations and Bayes' consistency of σ, we have $u(a,\bar{\omega}) = \mathbb{E}_\sigma[u(a,\omega)|\bar{\omega}]$ and $\pi_i(a,\bar{\omega}) = \mathbb{E}_\sigma[\pi_i(a,\omega)|\bar{\omega}]$ for all i. Therefore , by (6)

$$\begin{aligned} v_q^*(\bar{\omega}) &= \max_{a\in A} \sum_{\omega\in\Omega} [u(a,\omega) + t_q^*(a,\omega)]\sigma(\omega) \\ &\leq \sum_{\omega\in\Omega} \max_{a\in A} [u(a,\omega) + t_q^*(a,\omega)]\sigma(\omega) = \sum_{\omega\in\Omega} v_q^*(\omega)\sigma(\omega). \end{aligned} \tag{A.9}$$

Thus, if refining $\alpha q(\bar{\omega})$ of the records of type $\bar{\omega}$ according to σ does not change the value of any record, then (A.9) implies the desired inequality. Now consider the other case: There exists a share $\alpha > 0$ such that refining $\alpha q(\bar{\omega})$ of the records of type $\bar{\omega}$ according to σ leads to a database q_α such that $v_{q_\alpha}^*(\omega) \neq v_q^*(\omega)$ for some $\omega \in \operatorname{supp}\sigma$ or $\omega = \bar{\omega}$. Since the total quantity of records does not change, we have that $\mu_{q_\alpha}(\bar{\omega}) < \mu_q(\bar{\omega})$ and $\mu_{q_\alpha}(\omega) > \mu_q(\omega)$ for all $\omega \in \operatorname{supp}\sigma$. By Proposition 3, it follows that $v_{q_\alpha}^*(\bar{\omega}) \geq v_q^*(\bar{\omega})$ and $v_{q_\alpha}^*(\omega) \leq v_q^*(\omega)$ for all $\omega \in \operatorname{supp}\sigma$ and that the effects on the unrefined records are increasing in α. Now, note that for all α,

$$\sum_{\omega\in\Omega} v_{q_\alpha}^*(\omega)\sigma(\omega) \geq v_{q_\alpha}^*(\bar{\omega}) \geq v_q^*(\bar{\omega}), \tag{A.10}$$

where the first inequality follows from (A.9). This implies that the effect of σ on the refined records is always non-negative and decreasing in α. □

Obedience and Refinements. For the next result, we again assume that the agents observe no data (i.e., $|\Omega_i| = 1$ for all $i = 1,\dots,n$).

Lemma A.4. *Consider databases q and q_α, where q_α is obtained from q by refining α records of type $\bar{\omega}$ according to σ. For every obedient x at q, there exists a obedient $\hat{x}$ at q_α such that*

$$\sum_{a\in A,\omega\in\Omega} u(a,\omega)x(a,\omega) = \sum_{a\in A,\omega\in\Omega} u(a,\omega)\hat{x}(a,\omega).$$

Proof. Fix q, x, and σ as stated. Recall that

$$q_\alpha(\omega) = \begin{cases} (1-\alpha)q(\omega) & \text{if } \omega = \bar{\omega} \\ q(\omega) + \alpha q(\bar{\omega})\sigma(\omega) & \text{if } \omega \in \operatorname{supp}\sigma \\ q(\omega) & \text{else.} \end{cases}$$

Define $\hat{x} : A \times \Omega \to \mathbb{R}_+$ as follows:

$$\hat{x}(\cdot,\omega) = \begin{cases} x(\cdot,\omega) + \alpha\sigma(\omega)x(\cdot,\bar{\omega}) & \text{if } \omega \in \operatorname{supp}\sigma \\ x(\cdot,\omega)(1-\alpha) & \text{if } \omega = \bar{\omega} \\ x(\cdot,\omega) & \text{else.} \end{cases}$$

To interpret $\hat{x}$, note that the previous definition is equivalent to

$$\frac{\hat{x}(\cdot,\omega)}{q_\alpha(\omega)} = \begin{cases} \frac{q(\omega)}{q_\alpha(\omega)}\frac{x(\cdot,\omega)}{q(\omega)} + \frac{q_\alpha(\omega)-q(\omega)}{q_\alpha(\omega)}\frac{x(\cdot,\bar{\omega})}{q(\bar{\omega})} & \forall \omega \in \operatorname{supp}\sigma \\ \frac{x(\cdot,\omega)}{q(\omega)} & \text{else.} \end{cases}$$

Thus, under $\hat{x}$, the probability that a record of type $\omega \in \operatorname{supp}\sigma$ leads to recommendations a (i.e., $\hat{x}(\cdot,\omega)/q_\alpha(\omega)$) is a convex combination of the probability that a record of that type led to a under x (i.e., $x(\cdot,\omega)/q(\omega)$) and the probability that a record of type $\bar{\omega}$ led to a under x (i.e., $x(\cdot,\bar{\omega})/q(\bar{\omega})$).

We begin by establishing that $\hat{x}$ is obedient at q_α. For any $i \in I$ and $a_i, a_i' \in A_i$,

$$\begin{aligned}
&\sum_{a_{-i},\omega} (\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega))\hat{x}(a_i,a_{-i},\omega) \\
=\ &\sum_{a_{-i}} (\pi_i(a_i,a_{-i},\bar{\omega}) - \pi_i(a_i',a_{-i},\bar{\omega}))x(a_i,a_{-i},\bar{\omega})(1-\alpha) \\
&+ \sum_{a_{-i},\omega\in\operatorname{supp}\sigma} (\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega))x(a_i,a_{-i},\omega) \\
&+ \sum_{a_{-i},\omega\in\operatorname{supp}\sigma} (\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega))x(a_i,a_{-i},\bar{\omega})\alpha\sigma(\omega) \\
&+ \sum_{a_{-i},\bar{\omega}\neq\omega\notin\operatorname{supp}\sigma} (\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega))x(a_i,a_{-i},\omega) \\
=\ &\sum_{a_{-i}} (\pi_i(a_i,a_{-i},\bar{\omega}) - \pi_i(a_i',a_{-i},\bar{\omega}))x(a_i,a_{-i},\bar{\omega})(1-\alpha) \\
&+ \sum_{a_{-i}} (\pi_i(a_i,a_{-i},\bar{\omega}) - \pi_i(a_i',a_{-i},\bar{\omega}))x(a_i,a_{-i},\bar{\omega})\alpha \\
&+ \sum_{a_{-i},\omega\neq\bar{\omega}} (\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega))x(a_i,a_{-i},\omega) \\
=\ &\sum_{a_{-i},\omega} (\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega))x(a_i,a_{-i},\omega) \geq 0
\end{aligned}$$

The first equality follows from the definition of $\hat{x}$. The second equality follows from the fact that $\sum_\omega \pi_i(a_i,a_{-i},\omega)\sigma(\omega) = \pi_i(a_i,a_{-i},\bar{\omega})$ by the Bayes' consistency conditions of the beliefs at $\bar{\omega}$ and $\omega \in \operatorname{supp}\sigma$. The last inequality follows from the assumption that x is obedient at q. We conclude that $\hat{x}$ is obedient at q_α. A similar progression allows us to argue that:

$$\begin{aligned}
\sum_{a,\omega} u(a,\omega)\hat{x}(a,\omega) &= \sum_{a} u(a,\bar{\omega})x(a,\bar{\omega})(1-\alpha) \\
&\quad + \sum_{a,\omega\in\operatorname{supp}\sigma} u(a,\omega)(x(a,\omega) + \alpha\sigma(\omega)x(a,\bar{\omega})) \\
&\quad + \sum_{a,\bar{\omega}\neq\omega\notin\operatorname{supp}\sigma} u(a,\omega)x(a,\omega) \\
&= \sum_{a,\omega} u(a,\omega)x(a,\omega),
\end{aligned}$$

which concludes the proof. □

Proof of Proposition 5. For this result, we again assume that the agents receive no data (i.e., $|\Omega_i| = 1$ for all $i = 1, \dots, n$). For the first part of the result, Lemma A.4 immediately implies that $U^*(q_\alpha) \geq U^*(q)$.

Now observe that the directional derivative of U^* at any $\hat{q}$ in the direction implied by applying refinement σ is equal to[28]

$$\sum_{\omega' \in \Omega} v^*_{\hat{q}}(\omega')\sigma(\omega') - v^*_{\hat{q}}(\bar{\omega}).$$

Note that if we replace $\hat{q}$ with q_α, this derivative is decreasing in α by Proposition 3. This proves that the refinement's marginal benefit decreases in α.

Finally, to prove the rest of the result, we first parametrize the linear path from q to q_α as follows: For $t \in [0,1]$, define $q_t(\bar{\omega}) = q(\bar{\omega}) - t\alpha q(\bar{\omega})$, $q_t(\omega) = q(\omega) + t\alpha\sigma(\omega)q(\bar{\omega})$ for $\omega \in \operatorname{supp}\sigma$, and $q_t(\omega') = q(\omega')$ for the remaining ω'. By the gradient theorem,

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

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

Now, suppose that there exists a common $\tilde{a}$ that satisfies $x^*_q(\tilde{a}, \bar{\omega}) > 0$ and $x^*_q(\tilde{a}, \omega) > 0$ for all $\omega \in \operatorname{supp}\sigma$. By complementary slackness, $v^*_q(\omega) = u(\tilde{a}, \omega) + t^*_q(\tilde{a}, \omega)$ for all $\omega \in \operatorname{supp}\sigma$ and for $\omega = \bar{\omega}$. Therefore, by Proposition 3,

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

which, combined with (A.10), implies that $\sum_{\omega \in \Omega} v^*_{q_\alpha}(\omega)\sigma(\omega) = v^*_{q_\alpha}(\bar{\omega})$ for all $\alpha \in [0,1]$. In turn, this implies that $U^*(q_\alpha) = U^*(q)$ for all $\alpha \in [0,1]$.

Conversely, suppose that for every $\hat{a}$ such that $x^*_q(\hat{a}, \bar{\omega}) > 0$, there exists $\omega \in \operatorname{supp}\sigma$ that satisfies $x^*_q(\hat{a}, \omega') = 0$. If the solution to the data-value problem is unique at q (which is the case generically), then $x^*_q(\hat{a}, \omega) = 0$ implies $v^*_q(\omega) > u(\hat{a}, \omega) + t^*_q(\hat{a}, \omega)$ by strict complementary slackness. This and Lemma A.3 imply that there exists $t' > 0$ such that $\sum_{\omega \in \Omega} v^*_{q_t}(\omega)\sigma(\omega) > v^*_{q_t}(\bar{\omega})$ for all $t \in [0, t']$. It follows that $U^*(q_\alpha) > U^*(q)$.

[28] Note that refining a record of type $\bar{\omega}$ implies losing that record for sure (which explains the coefficient -1 on $v^*_{\hat{q}}(\bar{\omega})$) and gaining a record of type ω with probability $\sigma(\omega)$.

Online Appendix (For Online Publication Only)

B Adding Participation Constraints.

This appendix enriches our baseline model by adding sellers' participation constraints and shows that they generate new externalities. Let $\underline{\pi}_i \geq 0$ be the (exogenously given) payoff of the outside option for seller i. This could be interpreted as the payoff of joining a competing platform or of opening a brick-and-mortar store. Varying $\underline{\pi}_i$ is thus a reduced-form way of capturing the intensity of the competition that the platform faces. Recall that a mechanism takes the form $x : A \times \Omega \to \mathbb{R}_+$. The platform solves

$$\begin{aligned}
&\max_x \sum_{\omega,a} u(a,\omega)x(a,\omega) \\
&\text{s.t. for all } i,\ a_i, a_i', \text{ and } \omega', \\
&\sum_{\omega,a_{-i}} \left(\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega)\right) x(a_i,a_{-i},\omega) \geq 0 \qquad \text{(B.1)} \\
&\sum_a x(a,\omega') = q(\omega') \qquad \text{(B.2)} \\
&\sum_{a,\omega} \pi_i(a,\omega)x(a,\omega) \geq \underline{\pi}_i \qquad \text{(B.3)}
\end{aligned}$$

We now consider the dual problem. As in the baseline model, let $\lambda_i(a_i'|a_i)$ and $v(\omega)$ be the multipliers of the obedience constraint (B.1) and feasibility constraint (B.2). Let ζ_i be the multiplier of seller i's participation constraint (B.3). The dual is then

$$\begin{aligned}
&\min_{v,\lambda,\zeta} \sum_{\omega\in\Omega} v(\omega)q(\omega) \\
&\text{s.t. for all } \omega \in \Omega \text{ and } a \in A, \\
&v(\omega) \geq u(a,\omega) + t(a,\omega) + l(a,\omega),
\end{aligned}$$

where $t(a,\omega)$ is defined as in the baseline model and $l(a,\omega) = \sum_i (\pi_i(a,\omega) - \frac{\underline{\pi}_i}{\sum_\omega q(\omega)})\zeta_i$. The structure of this dual is similar to that of the dual of the baseline model. From this formulation, we obtain a generalized version of the decomposition in Proposition 1: The value $v_q^*(\omega)$ of records of type ω satisfies

$$v_q^*(\omega) = u_q^*(\omega) + t_q^*(\omega) + l_q^*(\omega), \qquad \text{where} \qquad l_q^*(\omega) = \sum_{a\in A} l_q^*(a,\omega)\frac{x_q^*(a,\omega)}{q(\omega)}.$$

The first two components, $u_q^*(\omega)$ and $t_q^*(\omega)$, are identical to those in the baseline model and have the same interpretation. The last component captures the intuitive idea that the value of

type-ω records depends also on whether their use helps or hinders satisfying the participation constraints of the seller(s). An extension of the scarcity principle can be easily derived using the same logic of Proposition 3.

C A Sufficient Condition for Optimality of Withholding Information.

We provide a sufficient condition for optimality of withholding information in terms of the primitives $\Gamma_\omega = \{I, (A_i)_{i=0}^n, (\pi_i(\cdot,\omega))_{i=1}^n\}$ for $\omega \in \Omega$, excluding q. We consider the general case where the intermediary can choose $a_0 \in A_0$ and each agent i can privately observe some data $\omega_i \in \Omega_i$. Recall that if the intermediary always fully disclose all ω, then it must be implementing a correlated equilibrium of the complete-information game Γ_ω for all ω (i.e., $\frac{x_q^*(\cdot,\omega)}{q(\omega)} \in CE(\Gamma_\omega)$).

Proposition C.1. *Fix Γ. Suppose there exists (a,ω) that satisfies:*

(1) $u(a,\omega) > \underline{v}(\omega)$,

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

$$\sum_{a\in A} u(a,\omega')y(a) = \underline{v}(\omega'),$$

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

Then it is not optimal in $\mathcal{U}_q$ to always fully disclose all records for any $q \in \mathbb{R}^\Omega_{++}$.

Condition (1) is clearly necessary: If for every records of type ω every action profile a cannot deliver a payoff higher than the full-disclosure payoff $\underline{v}(\omega)$, then it is clearly optimal for the intermediary to fully disclose every ω. Given an outcome (a,ω) with $u(a,\omega) > \underline{v}(\omega)$, 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 $u(a,\omega) \leq \underline{v}(\omega)$. Then, condition (2) requires that agent i's data ω_i is consistent with another record ω' (so that he cannot tell ω and ω' apart based on his own data only) and that $\Gamma_{\omega'}$ admits a intermediary-preferred correlated equilibrium that also recommends i to play a_i and renders the deviation to $\hat{a}_i$ strictly suboptimal. Note that this condition is easy to check in applications starting from the best full-disclosure mechanism x.

Proof of Proposition C.1. . We will argue by contradiction. Suppose $q \in \mathbb{R}^{\Omega}_{++}$ and $\mathcal{U}_q$ admits a full-disclosure solution x^{**}_q and hence $x^{**}_q(\cdot|\tilde{\omega}) \in CE(\Gamma_{\tilde{\omega}})$ and $u^{**}_q(\tilde{\omega}) = \underline{v}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$. Then $v^{**}_q(\tilde{\omega}) = u^{**}_q(\tilde{\omega}) = \underline{v}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$ by Proposition 4.

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

$$v^{**}_q(\tilde{\omega}) \geq u(a, \tilde{\omega}) + t^{**}_q(a, \tilde{\omega}).$$

Since $u(a, \omega) > \underline{v}(\omega) = v^{**}_q(\omega)$, we must have $t^{**}_q(a, \omega) < 0$. Therefore, there exists a pair $(i, \hat{a}_i)$ that satisfies $\pi_i(a_i, a_{-i}, \omega) < \pi_i(\hat{a}_i, a_{-i}, \omega)$ and $\lambda^{**}_{q,i}(\hat{a}_i|a_i, \omega_i) > 0$. For such a pair $(i, \hat{a}_i)$, there exists a mechanism x such that $\frac{x(\cdot,\omega')}{q(\omega')} \in CE(\Gamma_{\omega'})$ with the properties listed in the proposition. Then, since $\lambda^{**}_{q,i}(\hat{a}_i|a_i, \omega_i) > 0$,

$$\begin{aligned}
& \sum_{\tilde{a} \in A} u(\tilde{a}, \omega') \frac{x(\tilde{a}, \omega')}{q(\omega')} + \sum_{\tilde{a} \in A} t^{**}_q(\tilde{a}, \omega') \frac{x(\tilde{a}, \omega')}{q(\omega')} \\
\geq & \sum_{\tilde{a} \in A} u(\tilde{a}, \omega') \frac{x(\tilde{a}, \omega')}{q(\omega')} \\
& + \lambda^{**}_{q,i}(\hat{a}_i|a_i, \omega_i) \left\{ \sum_{\tilde{a}_{-i} \in A_{-i}} \left(\pi_i(a_i, \tilde{a}_{-i}, \omega') - \pi_i(\hat{a}_i, \tilde{a}_{-i}, \omega') \right) \frac{x(a_i, \tilde{a}_{-i}|\omega')}{q(\omega')} \right\} \\
> & \sum_{\tilde{a} \in A} u(\tilde{a}, \omega') \frac{x(\tilde{a}, \omega')}{q(\omega')} = v^{**}_q(\omega'),
\end{aligned}$$

where the first inequality follows because $\frac{x(\cdot,\omega')}{q(\omega')} \in CE(\Gamma_{\omega'})$. The strict inequality is incompatible with the dual constraint (A.4) and delivers the desired contradiction. □

D Interpreting the Data-Value Problem

To further understand the values of data records and the pooling externalities, we provide a standalone interpretation of the data-value problem $\mathcal{V}_q$. This also sheds light on the forces, trade-offs, and constraints that determine the record values by shaping how the platform uses its data to influence the sellers. We focus on the case in which the agents have no data (i.e., $|\Omega_i| = 1$ for all i), but only minor adjustments are needed to also cover the more general case. We fix $q \in \mathbb{R}^{\Omega}_{++}$ and so drop it from notation. This interpretation links our work to an earlier literature on dual analysis of correlated equilibria (Nau and McCardle (1990); Nau (1992); Myerson (1997)).

We first rewrite $\mathcal{V}$ in the following equivalent way by exploiting the structure of the specific problem at hand. For every i, we can set $\lambda_i(a_i|a_i) = 1$ (or any strictly positive number) for all $a_i \in A_i$. Given this, for every i and $a_i \in A_i$, define

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

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

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

which implies that $\ell_i(\cdot|a_i) \in \Delta(A_i)$. After constructing $b = (b_1, \ldots, b_n)$ and $\ell = (\ell_1, \ldots, \ell_n)$ in this way, for each $i \in I$ and (a, ω) define

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

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

$$\begin{aligned} \mathcal{V}: \quad & \min_{v, b, \ell} \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ & \text{s.t. for all } \omega \in \Omega, \\ & v(\omega) = \max_{a \in A} \Big\{ u(a, \omega) + t(a, \omega) \Big\}, \end{aligned} \tag{D.1}$$

D.1 Gambles Against the Agents

Our interpretation hinges on unpacking how the platform determines the records' contributions $t(a, \omega)$ to the externality $t(\omega)$. By (D.1), it does so by choosing b and ℓ, which fully pin down $t(a, \omega)$ and hence $v(\omega)$. Recall that the platform wants to *minimize* the values of its records, so it 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

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

which contributes to lowering $t_i(a, \omega)$ if and only if $\ell_i(a_i'|a_i) > 0$ and $\pi_i(a_i, a_{-i}, \omega) < \pi_i(a_i', a_{-i}, \omega)$. That is, if seller i knew ω and his opponents' actions a_{-i}, he would strictly prefer deviating to a_i' than following recommendation a_i. In this case, playing a_i amounts to making a mistake from an ex-post viewpoint. We will say that seller i regrets offering a_i.

Thus, inducing sellers to make offers they will regret emerges as an intrinsic goal of the platform's problem—together with maximizing u of course. In this view, (b_i, ℓ_i) becomes an

exploitation strategy on the part of the platform against seller i. Inducing regrettable actions requires withholding information from seller i about ω or a_{-i}. This explains why the platform may prefer withholding information, but from the perspective of the data-value problem. In the end, $v(\omega)$ results from a trade-off between $u(a,\omega)$ and the return from inducing actions the sellers regret.

This return depends on the structure of b and ℓ, which can be interpreted as defining a family of gambles against the sellers. To see this, fix (a,ω) and seller i. Then, $\ell_i(\cdot|a_i) \in \Delta(A_i)$ defines a lottery whose prize for the platform is $\pi_i(a_i,a_{-i},\omega) - \pi_i(a_i',a_{-i},\omega)$ for each a_i'; the scaling term $b_i(a_i)$ captures the stake that the platform bets on this lottery. The platform "wins" when $\pi_i(a_i,a_{-i},\omega) < \pi_i(a_i',a_{-i},\omega)$ and "loses" otherwise. Thus, $t(a,\omega)$ is 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 record type ω and outcome a.[1]

We can then link how the platform chooses these gambles in $\mathcal{V}$ with the pooling externalities. Negative externalities $t^*(\omega) < 0$ correspond to favorable gambles, in the sense that the platform wins in expectation. This requires the help of other records to withhold information and induce the sellers to choose actions they will regret. Conversely, positive externalities $t^*(\omega) > 0$ correspond to unfavorable gambles. Corollary 2 implies that, at the optimum, the platform chooses gambles that favor it 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.

D.2 Feasible Gambles and Trade-offs

The feasible gambles in $\mathcal{V}$ have specific features that shed light on the data-value problem.

Some features reflect structural properties of $\mathcal{V}$. While the prizes of each gamble are contingent on ω and the entire a, for each seller i both b_i and ℓ_i can depend only on a_i. This limits the platform's ability to tailor its gambles across records and sellers. 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 value formula (D.1) for (a_i,a_{-i},ω) and (a_i,a'_{-i},ω'). In particular, if $\pi_i(a_i,a_{-i},\omega) < \pi_i(\hat{a}_i,a_{-i},\omega)$ but $\pi_i(a_i,a'_{-i},\omega') > \pi_i(\hat{a}_i,a'_{-i},\omega')$, the platform faces a trade-off because it

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

may not be possible 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 it solves the 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.

The platform faces other constraints in its ability to *jointly* exploit the sellers. Given $\mathcal{V}$, it is clear that it would want to choose (b, ℓ) so that $t(a, \omega) \leq 0$ for all (a, ω) with some strict inequality. Such gambles would guarantee a sure arbitrage against the sellers, but are infeasible in the following sense. By complementary slackness $x^*(a, \omega) > 0$ implies $v^*(\omega) = u(a, \omega) + t^*(a, \omega)$. Thus, since every ω must induce some a 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 $(a, \omega) \in \mathbf{X}$ (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, as restricting x to the domain $\mathbf{X}$ is immaterial in $\mathcal{U}$.

Proposition D.1. *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 we provide a proof below, which relies on a dual characterization of $\mathbf{X}$ using Farkas' lemma.

The economic takeaway is that in the attempt to minimize values v by exploiting the sellers with (b, ℓ), the platform faces a fundamental trade-off that is a hallmark of $\mathcal{V}$. 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'. This result sheds light on how and how much the platform can actually manipulate sellers by conveying information.

Proof of Proposition D.1. This proof is for the general case where each agent i can privately observe some own data $\omega_i \in \Omega_i$. 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. Construct 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) for $i \in I$, and its entries are as follows:

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

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 $x \geq 0$, such that $\mathbf{W}^T x \geq 0$, with $x^T \mathbf{1}_{a^*,\omega^*} > 0$. We show that the latter is true. Indeed, we can pick x to be a mechanism that is obedient and satisfies $x(a^*,\omega^*) > 0$. We can find such x, since $(a^*,\omega^*) \in \mathbf{X}$. Then $x \geq 0$ and $x^T \mathbf{1}_{a^*,\omega^*} > 0$ are satisfied automatically. Finally, $\mathbf{W}^T x \geq 0$ 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. □

E Data and Price Discrimination: Analysis

This section provides the calculations for Section 4.3 and the example in the Online Appendix F below. Recall that $a \in \{1,2\}$ and that $u(a,\omega) = \max\{\omega - a, 0\}$ and $\pi(a,\omega) = a\mathbb{I}\{\omega \geq a\}$ for $\omega \in \{\omega_1,\omega_2\}$. For $\bar{\omega}$, we have $u(a,\bar{\omega}) = hu(a,\omega_2) + (1-h)u(a,\omega_1)$ and $\pi(a,\bar{\omega}) = h\pi(a,\omega_2) + (1-h)\pi(a,\omega_1)$. For completeness, we solve both the intermediation (primal) problem $\mathcal{U}_q$ and the data-value (dual) problem $\mathcal{V}_q$.

Intermediation Problem. The objective function is

$$(\omega_2 - \omega_1)x(1,\omega_2) + h(\omega_2 - \omega_1)x(1,\bar{\omega}) = x(1,\omega_2) + hx(1,\bar{\omega}).$$

The obedience constraints are

$$\begin{aligned} -x(2,\omega_1) + x(2,\omega_2) + (2h-1)x(2,\bar{\omega}) &\geq 0, \\ x(1,\omega_1) - x(1,\omega_2) - (2h-1)x(1,\bar{\omega}) &\geq 0. \end{aligned}$$

Consider first the case of $h > \frac{1}{2}$. From the second constraint we get $x_q^*(1,\omega_1) = q(\omega_1)$. The first constraint is then automatically satisfied. Since $h \in (0,1)$, it is always true that $2h-1 < h$. The solution satisfies $x_q^*(1,\omega_2) = 0$ and $x_q^*(1,\bar{\omega}) = \frac{1}{2h-1}q(\omega_1)$, as long as $\frac{1}{2h-1}q(\omega_1) \leq q(\bar{\omega})$.

Now consider the case of $h \leq \frac{1}{2}$. Combining obedience constraints, we get

$$x(1,\omega_1) - x(1,\omega_2) - (2h-1)x(1,\bar{\omega}) \geq \max\left\{2q(\omega_1) + (1-h)2q(\bar{\omega}) - 1, 0\right\}.$$

It is immediate that $x_q^*(1,\bar{\omega}) = q(\bar{\omega})$ and $x_q^*(1,\omega_1) = q(\omega_1)$, since this relaxes the constraint as much as possible. The constraint then becomes

$$q(\omega_1) - (2h-1)q(\bar{\omega}) - \max\left\{2q(\omega_1) + (1-h)2q(\bar{\omega}) - 1, 0\right\} \geq x(1,\omega_2).$$

Data-Value Problem. The data-value problem is

$$\min_{v,\lambda} q(\omega_1)v(\omega_1) + q(\omega_2)v(\omega_2) + q(\bar{\omega})v(\bar{\omega}),$$

subject to $\lambda(2|1), \lambda(1|2) \geq 0$,

$$\begin{aligned} v(\omega_1) &= \max\{\lambda(2|1), -\lambda(1|2)\} = \lambda(2|1), \\ v(\omega_2) &= \max\{1 - \lambda(2|1), \lambda(1|2)\}, \\ v(\bar{\omega}) &= \max\{h + (1-2h)\lambda(2|1), (2h-1)\lambda(1|2)\} \\ &= h \max\left\{1 - \frac{2h-1}{h}\lambda(2|1), \frac{2h-1}{h}\lambda(1|2)\right\}. \end{aligned}$$

As we noted before, $\frac{2h-1}{h} < 1$. Suppose that $h > \frac{1}{2}$. Then, it is optimal to set $\lambda_q^*(1|2) = 0$ to relax the problem as much as possible. We then have

$$\begin{aligned} v(\omega_1) &= \lambda(2|1), \\ v(\omega_2) &= \max\{1 - \lambda(2|1), 0\}, \\ v(\bar{\omega}) &= h \max\left\{1 - \frac{2h-1}{h}\lambda(2|1), 0\right\}. \end{aligned}$$

There are three candidates for optimal $\lambda(2|1)$. When $\lambda(2|1) = 0$, the objective is $S_0 \stackrel{\Delta}{=} q(\omega_2) + hq(\bar{\omega})$. When $\lambda(2|1) = 1$, the objective is $S_1 \stackrel{\Delta}{=} q(\omega_1) + q(\bar{\omega})(1-h)$. When $\lambda(2|1) = \frac{h}{2h-1}$, the objective is $S_f \stackrel{\Delta}{=} q(\omega_1)\frac{h}{2h-1}$. The following claims are true. First, $S_0 \leq S_1$ if and only if $q(\omega_1) \geq q(\omega_2) + (2h-1)q(\bar{\omega})$. Second, $S_0 \leq S_f$ if and only if $q(\omega_1) \geq q(\omega_2)\frac{2h-1}{h} + (2h-1)q(\bar{\omega})$. Third, $S_1 \leq S_f$ if and only if $q(\omega_1) \geq (2h-1)q(\bar{\omega})$.

Suppose now that $h \leq \frac{\omega_1}{\omega_2}$. Then $v(\bar{\omega}) = h - (2h-1)\lambda(2|1)$ and $\lambda_q^*(1|2) = 0$ is again optimal. There are only two candidates for optimal $\lambda(2|1)$, specifically, 0 and 1.

Summary. All these cases lead to three scenarios in terms of q.

Scenario 1: $q(\omega_1) \leq (2h-1)q(\bar{\omega})$. Note that this requires $h > \frac{1}{2}$. Table 2 presents the optimal x_q^*.

$x_q^*(a,\omega)$		ω		
		ω_1	ω_2	$\bar{\omega}$
a	1	$q(\omega_1)$	0	$\frac{1}{2h-1}q(\omega_1)$
	2	0	$q(\omega_2)$	$q(\bar{\omega}) - \frac{1}{2h-1}q(\omega_1)$

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

The solution to the data-value problem is $\lambda_q^*(1|2) = 0$, $\lambda_q^*(2|1) = \frac{h}{2h-1}$ and the unit values are $v_q^*(\omega_1) = \frac{h}{2h-1}$, $v_q^*(\omega_2) = 0$, and $v_q^*(\bar{\omega}) = 0$.

Scenario 2: $(2h-1)q(\bar{\omega}) \leq q(\omega_1) \leq q(\omega_2) + (2h-1)q(\bar{\omega})$. Note that the lower bound on $q(\omega_1)$ is meaningful only if $h > \frac{1}{2}$. Table 3 presents the optimal x_q^*.

$x_q^*(a,\omega)$		ω		
		ω_1	ω_2	$\bar{\omega}$
a	1	$q(\omega_1)$	$q(\omega_1) - (2h-1)q(\bar{\omega})$	$q(\bar{\omega})$
	2	0	$q(\omega_2) - [q(\omega_1) - (2h-1)q(\bar{\omega})]$	0

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

The solution to the data-value problem is $\lambda_q^*(1|2) = 0$, $\lambda_q^*(2|1) = 1$, and the unit values are $v_q^*(\omega_1) = 1$, $v_q^*(\omega_2) = 0$, and $v_q^*(\bar{\omega}) = 1 - h$.

Scenario 3: $q(\omega_1) \geq q(\omega_2) + (2h-1)q(\bar{\omega})$. Table 4 presents the optimal x_q^*.

$x_q^*(a,\omega)$		ω		
		ω_1	ω_2	$\bar{\omega}$
a	1	$q(\omega_1)$	$q(\omega_2)$	$q(\bar{\omega})$
	2	0	0	0

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

The solution to the data-value problem is $\lambda_q^*(1|2) = \lambda_q^*(2|1) = 0$ and the unit values are $v_q^*(\omega_1) = 0$, $v_q^*(\omega_2) = 1$, and $v_q^*(\bar{\omega}) = h$.

F Correlated Refinements: An Example

We present an example of a correlated refinement that decreases the value of the refined records as well as the platform's total payoff.

The setting is as in the last part of Section 4.3, where $r = 0$ and $\Omega = \{\omega_1, \omega_2, \bar{\omega}\}$. The initial database q satisfies $q(\bar{\omega}) < q(\omega_1) < q(\omega_2)$ and $q(\omega_2) < q(\omega_1) + (1-h)q(\bar{\omega})$. Let $\alpha = 1$ and consider the following refinement, which is arguably extreme but serves to make our point as clearly as possible. Suppose the platform learns that all its records of type $\bar{\omega}$ involve buyers with the same θ. That is, with probability $1-h$ they *all* become records of type ω_1 and with probability h they *all* become records of type ω_2. Thus, with probability $1-h$ the

	ω_1	ω_2	$\bar{\omega}$	
v^*_q	1	0	$1-h$	$U^*(q) = q(\omega_1) + (1-h)q(\bar{\omega})$
$v^*_{q'}$	0	1	–	$U^*(q') = q'(\omega_2)$
$v^*_{q''}$	1	0	–	$U^*(q'') = q''(\omega_1)$

Table 5: Value of records and total payoffs for specific databases ($r = 0$)

new database is q' and satisfies $q'(\omega_1) > q'(\omega_2) > q'(\bar{\omega}) = 0$; with probability h, the new database is q'' and satisfies $q''(\omega_2) > q''(\omega_1) > q''(\bar{\omega}) = 0$. Table 5 reports the unit value of the records for these databases. The calculations are in Online Appendix E. The refinement strictly decreases both the unit value of the refined records and the platform's total payoff: $v^*_q(\bar{\omega}) > (1-h)v^*_{q'}(\omega_1) + hv^*_{q''}(\omega_2) = 0$ and $U^*(q) > U^*(q') > U^*(q'')$ because $q'(\omega_2) = q(\omega_2)$ and $q''(\omega_1) = q(\omega_1)$.

By contrast, the same refinement strictly increases both the unit value of the refined records and the platform's total payoff if the platform maximizes the seller's profits ($r = 1$). Indeed, we have $v^*_{\hat{q}}(\omega_1) = 1$, $v^*_{\hat{q}}(\omega_2) = 2$, and $v^*_{\hat{q}}(\bar{\omega}) = 2h$ for all $\hat{q} \in \{q, q', q''\}$. The key is that a profit-maximizing platform treats each buyer-seller interaction as independent of all other interactions, so it does not care about correlation in how it learns about records. Instead, a surplus-maximizing platform cares about such correlation, because it can have profound consequences on the information the platform has in its database and so can convey to (or withhold from) the seller. Specifically, for the original database q, the platform was pooling all records of type $\bar{\omega}$ with those of type ω_1. Now, if all records of type $\bar{\omega}$ become of type ω_2, their value drops to zero. This is because type-ω_2 records were already abundant in the original database and exhausted all possibilities of pooling them with records of type ω_1. If instead all records of type $\bar{\omega}$ become of type ω_1, their value again drops to zero. This is because now type-ω_1 records become too abundant in the new database and the ability to pool them with records of type ω_2 becomes worthless for the platform, since providing no information to the seller already induces him to charge the lowest price of 1 to all buyers.