The Value of Data Records

Simone Galperti
UC, San Diego

Aleksandr Levkun
UC, San Diego

Jacopo Perego
Columbia University

October 13, 2021

Abstract

Many online platforms intermediate trade between sellers and buyers using data records of the buyers' personal characteristics. How much value do such intermediaries derive from each record? Is this value higher for specific buyers? What are its properties? We find that an important component of the value of a data record is a novel externality that arises when a platform pools records to withhold information from the sellers. Ignoring this externality can significantly bias our understanding of the value of data records. We then characterize a platform's willingness to pay for more data, thereby establishing a series of basic properties of the demand side of data markets. Our analysis combines modern information design with classic duality methods and applies to a large class of principal-agent problems.

JEL Classification Numbers: C72, D82, D83

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

We are thankful to Dirk Bergemann for his insightful comments as a discussant of this paper and to S. Nageeb Ali, Alessandro Bonatti, Wouter Dessein, Laura Doval, Navin Kartik, Elliot Lipnowski, Alessandro Lizzeri, Xiaosheng Mu, Andrea Prat, Joel Sobel, Denis Shishkin, Rakesh Vohra, Glen Weyl, as well as seminar participants at various universities for useful feedback.

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

1 Introduction

Personal data is the "new oil" of modern economies. Markets for data have been rapidly developing and have fueled major policy debates (Federal Trade Commission, 2014). These markets also have spurred intense research to understand their unique properties (Bergemann and Ottaviani, 2021). However, many critical questions remain. Among them, what is the value of an individual's data for the firm using it? How does this value depend on the data's content and the firm's goals? Answering these questions can shed light on the demand side of data markets and on how people should be fairly compensated for their specific data (Lanier, 2013; Acquisti et al., 2016; Arrieta-Ibarra et al., 2018).

Understanding the value of data in modern economies raises new challenges. First, in many markets data is traded based on its specific informational content (Bergemann and Bonatti, 2019). Yet, standard theories almost exclusively evaluate information before it realizes. Second, data is often used by firms that act as *intermediaries*—like e-commerce marketplaces, search engines, and matching platforms—to strategically direct interactions between agents with conflicting interests. Yet, standard theories mostly evaluate information for single decision makers. To overcome these challenges, our approach combines modern information design with classic duality methods. We find that the value of data for intermediation problems differs fundamentally from standard decision problems. This is because, to manage conflicting interests, an intermediary may tailor the information it conveys to the agents by pooling data records, thus creating complex externalities between them.[1]

Consider an example. An online platform mediates the interactions between a population of buyers and a monopolist, who produces a good at zero marginal cost. For each buyer, the platform owns a data *record*, which consists of a list of the buyer's personal characteristics (gender, age, etc.). There are different types of records depending on how much the platform knows about the buyer. For simplicity, suppose type ω_k reveals that her valuation for the seller's good is k for $k \in \{1,2\}$. The collection of buyers' records forms the platform's database.

[1] This practice is widespread in 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 the riders' destinations from drivers to increase riders' welfare; and Airbnb withholds the host's profile picture to decrease discrimination.

Suppose its composition q consists of 3 million records of type ω_1 and 6 million of type ω_2. The seller knows only q. For each interaction, the platform sends a signal about ω to the seller so as to influence the price he charges.[2] Concretely, the platform may divide the buyers into market segments based on their records and tell the seller to which segment each buyer belongs. Our main goal is to determine how much value the platform derives from each buyer's record. This is immediate if the platform maximizes the seller's profits (e.g., because it keeps a share of it). This case is effectively the same as if the platform itself were the seller and directly set a price for each buyer knowing ω. Since this is akin to a decision problem, the value of a record is equal to the payoff the platform directly obtains conditional on ω.

The answer is no longer immediate when we consider other objectives of the platform. To illustrate, suppose it maximizes the buyers' surplus (e.g., because it cares about their loyalty). One way to do this is to assign each buyer whose record is of type ω_2 to either a subprime segment $\underline{s}$ or to a prime segment $\bar{s}$, with equal probability; instead, it assigns all buyers whose record is of type ω_1 to $\underline{s}$. The seller optimally sets a price of 1 for segment $\underline{s}$ and a price of 2 for $\bar{s}$. The expected payoff that the platform directly obtains from a record of type ω_1 is 0, while it is ½ for a record of type ω_2. Do these payoffs reflect the actual value the platform derives from each record? The answer is no. We will show that the actual values are $v^*(\omega_1) = 1$ and $v^*(\omega_2) = 0$. That is, the most valuable records for the platform are those that yield the lowest payoff. To see why, imagine two buyers, Ann and Bonnie, whose records are of type ω_1 and ω_2 respectively. Bonnie's record yields a positive payoff to the platform only when pooled with Ann's record through segment $\underline{s}$. In this case, Ann's record helps to persuade the seller to set a low price for Bonnie. Hence, Ann's record should not be worthless, even though Ann's interaction with the seller yields zero payoff to the platform. Indeed, $v^*(\omega_1) = 1$ reflects that Ann's record exerts a positive information externality on Bonnie's interaction. By contrast, $v^*(\omega_2) = 0$ reflects that we have to discount this externality from Bonnie's record.

Our main contribution is to characterize what determines the value of data records for intermediaries like the platform above. Our analysis delivers $v^*(\omega)$ as the *unit* value of every type-ω record in the database, leveraging the linear structure of intermediation problems. At the same time, $v^*(\omega)$ also equals the marginal effect on the platform's total payoff of adding type-ω records to the database. As the example showed, $v^*(\omega)$ can differ significantly from the payoff the platform *directly* obtains from a record because it pools records to tailor the information it conveys to the seller. We show that $v^*(\omega)$ is the sum of the platform's direct pay-

[2] This is in the spirit of Bergemann et al. (2015). Elliott et al. (2020) study a related problem with multiple horizontally differentiated sellers.

off from each type-ω record and the externalities caused by that record on other records and their interactions.[3] We relate these externalities to how the platform exploits the seller's incentives across interactions. We explain when these externalities are positive and negative. For instance, in price-discrimination settings—which generalize our example—they satisfy a single-crossing property as long as the platform cares more about the buyers' surplus than the seller's profit: The externality is positive for buyers whose valuation for the seller's product is low and negative for those whose valuation is high. This means that ignoring such externalities could lead to overcompensating the latter for their data at the expense of the former.

This characterization of the value of data records is a necessary step to study an intermediary's willingness to pay for more data. In our context, acquiring more data can have two meanings. In our example, (i) the platform can obtain *more* records for its database and, hence, gain the ability to mediate more interactions between the corresponding buyers and the seller (e.g., because new buyers join it); or (ii) our platform can obtain *better* records by observing more informative characteristics about existing buyers (e.g., because they become more active online).[4]

With regard to obtaining more records, a key insight is that the platform's preference over databases is pinned down by v^* as a function of their composition q. In particular, v^* determines the platform's willingness to pay for more records and the substitutability between types of records. We find that this willingness to pay is stepwise diminishing. Moreover, record types are imperfect substitutes (or even complements) if and only if the platform withholds some information from the seller. These properties establish a "scarcity principle" for data: In any intermediation problem, scarcer types of records are more valuable both in absolute and in relative terms. They also enable us to infer how the platform uses its data from observable features of its demand function, which can be derived using standard maximization subject to a budget constraint.

The platform's preference over databases is also useful to study its willingness to pay for better records. This is because obtaining more information about existing buyers changes their records' type and hence the database composition q. Imagine the platform refines the record of a buyer, called Cindy, by observing new characteristics about her. We show that such

[3] Importantly, these externalities arise even when data records are statistically independent. As such, they differ fundamentally from "learning" externalities highlighted by the literature (see below), which depend on the correlation between records.

[4] 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 former allows companies to identify new customers who have specific characteristics. The latter allows companies to learn new characteristics about existing customers.

refinements have a positive direct effect: Cindy's record becomes more valuable in expectation. Because they change q, refinements also have indirect effects: Unrefined records can become more or less valuable. These effects are due to the aforementioned externalities and exist even if Cindy's new characteristics provide no information about other buyers (i.e., refinements are independent). We find that, despite their negative effects on the value of some records, independent refinements always benefit the platform overall, which therefore has a positive willingness to pay for them. This benefit is decreasing in the extensive margin—namely, how many records of a given type are refined. The benefit becomes zero under a precise condition, even if the platform would act on the new information it gets and use the refined and unrefined records differently. This is in sharp contrast with decision problems, where getting information is strictly beneficial if and only if it changes optimal behavior. Another difference is that the platform's willingness to pay can be *negative* for refinements that are correlated between records.

Our analysis applies to any setting where an intermediary (principal) mediates interactions between multiple agents by providing them with information or by affecting incentives with its actions. We can also let the agents have some payoff-relevant data, as long as the intermediary has direct access to everybody's data.[5] We view the intermediary as using each interaction's data as an input to produce information or choose its actions. Once we see intermediation problems through this lens, it becomes natural to use linear-programming duality to characterize the value v^* of the data inputs, adapting the classic work of Dorfman et al. (1987) and Gale (1989).

This paper contributes to improving our understanding of the *demand side* of data markets. We believe that its insights are useful to informing empirical strategies for estimating the demand for data or to inferring from market observables how data is used. Progress in this area is essential for studying the welfare effects of critical policy interventions, such as new antitrust or privacy regulations.[6] Finally, a better understanding of the value of people's data may help improve on the status quo where they receive no compensation for it (Lanier, 2013; Arrieta-Ibarra et al., 2018; Jones and Tonetti, 2020).

Related Literature. This paper contributes to the burgeoning literature on data markets, comprehensively reviewed by Bergemann and Bonatti (2019) and Bergemann and Ottaviani (2021).

One of its strands studies the optimal "use" of a database. This often involves a single party—such as a platform or data broker—who owns a database and designs information prod-

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

[6] See, e.g., Stigler Report (2019); Crémer et al. (2019); Goldberg et al. (2021).

ucts for some agents—such as sellers, advertisers, or decision makers—to either charge a price or influence their behavior (or both). In Admati and Pfleiderer (1986, 1990), a platform sells signals (i.e., Blackwell experiments) about an asset to market traders. In Bergemann and Bonatti (2015), a platform sells segments of buyers to advertisers and charges a linear price based on the segment size. In Bergemann et al. (2018), a platform designs menus of signals to screen information buyers with heterogeneous priors. Yang (2020) studies a related problem in considerably richer settings. Our platform also owns a database and uses it to design information. However, our focus is not on the information products and their prices but on the value of their data inputs in the "upstream" market. These data records have two key features: Each record gives access to a buyer, on top of information about her, and each record can be valued ex post based on its specific content.

Another strand of the literature on data markets studies how to incentivize consumers to disclose their data. Choi et al. (2019), Acemoglu et al. (2021), and Ichihashi (2021) study the "learning" externalities that one consumer's disclosure has on others when their data is correlated. Bergemann et al. (2020) examine how this correlation affects consumers' incentives to participate in data markets and other market observables. These papers differ from ours in two ways. First, our platform is assumed to already have the database.[7] This offers a useful benchmark to study the effects of privacy regulations. Second, we isolate a new data externality, which stems from the platform's pooling records to withhold information from the sellers and arises even if consumers' records are statistically independent.

Our work is related to the literature on data privacy, reviewed by Acquisti et al. (2016). Calzolari and Pavan (2006) analyze information externalities between sequential interactions. Ali et al. (2020) examine when giving consumers control over their data can help them benefit from personalized pricing. Ichihashi (2020) finds that a multi-product platform can prefer not to use consumers' data for personalized pricing and maximizes profits via product recommendations.

Our methods build on the information-design literature, reviewed by Bergemann and Morris (2019). We formulate our "data-use" problem as a linear program, using standard arguments (Bergemann and Morris (2016)), and then consider its dual to obtain our "data-value" problem. Others have used duality to study information design (Kolotilin (2018); Galperti and Perego (2018); Dworczak and Martini (2019); Dworczak and Kolotilin (2019); Dizdar and Kováč (2020)). These papers exploit the dual to solve the primal design problem. We use the dual to address a distinct economic question of independent interest—what is the value of data?

[7] This may seem far-fetched but most data-brokers' transactions happen without the consumer's knowledge (Federal Trade Commission, 2014).

Unlike those papers, we also study problems with multiple agents through the notion of Bayes-correlated equilibrium. This links our work to an earlier literature on dual analysis of correlated equilibria (Nau and McCardle (1990); Nau (1992); Myerson (1997)). Finally, the mechanism-design literature has used duality methods at least since Myerson (1983; 1984), as well as more recently to study informationally robust mechanisms (e.g., Du (2018); Brooks and Du (2020, 2021)).

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 mediates interactions between buyers and sellers. Our approach and results apply much more broadly to settings where a principal influences the behavior of multiple strategic agents with information, its actions, or both. Section 5 discusses this and other aspects of the model.

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

The platform has access to some data about each buyer. We think of this data as a *record* of personal characteristics that is informative about her θ—perhaps only partially. We assume that each buyer's record is uninformative about the other buyers' θ.[8] There are different *types* of records—denoted by ω in some finite set Ω—depending on what the platform knows about the buyer. Thus, the content of each buyer's record is analogous to the realization of an exogenous signal about her underlying preference. Only the platform observes ω, which gives it an informational advantage over the sellers. Let $q \in \mathbb{R}_+^{\Omega}$ denote the collection of buyers' records, where $q(\omega)$ are of type ω. We refer to q as the platform's *database*.

For each interaction between a buyer and the sellers, we leave her purchase decision given their actions and θ implicit and embed it in the payoff functions of the sellers and the platform. For every ω and $a = (a_1 \dots, a_n)$, let $u_i(a, \omega)$ be i's expected payoff conditional on the buyer's

[8] We make this assumption to emphasize the novel aspects of our results. Our model can accommodate correlation among records (see Section 5.1).

record. Let $\Gamma_\omega = \{I, (A_i, u_i(\cdot, \omega))_{i=0}^n\}$, which defines a complete-information game between the sellers. We may also refer to Γ_ω as a buyer-sellers interaction of type ω. The primitives $\Gamma = \{\Gamma_\omega\}_{\omega \in \Omega}$ and q are common knowledge.

The platform mediates each interaction by privately conveying information about its type to each seller so as to influence their actions. The sellers combine this information with Γ and q to form beliefs and act. Our platform has full commitment power, similar to the omniscient information designer in Bergemann and Morris (2019). Formally, it publicly commits to an information structure that, for each interaction, produces a private signal about ω for each seller i. As is standard (Myerson, 1983, 1984; Bergemann and Morris, 2016), we can focus on information structures in the form of recommendation mechanisms, where the platform privately recommends an action to each seller that he must find optimal to follow (obedience). A mechanism is then a function $x : \Omega \to \Delta(A)$, where $x(a|\omega)$ can be interpreted as the share of interactions of type ω that lead to recommendation profile a.[9] Formally, the problem is

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

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

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

and the *total payoff* generated by the database as

$$U^*(q) \triangleq \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega). \tag{2}$$

We assume that $\mathcal{U}_q$ satisfies a minor regularity property, which holds generically in the space of sellers' payoff functions: No more than $|A \times \Omega|$ of the constraints (1) are ever active at the same time (see Remark 1 in Appendix A.2).

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

3 The Unit Value of Data

This section addresses our main question: How much value does the platform derive from each buyer's record and what are its properties? To get a sense of why the answer is nontrivial, it is useful to compare our problem $\mathcal{U}_q$ with standard decision problems. We can interpret $\mathcal{U}_q$ as a collection of decisions: For each buyer-sellers interaction, the platform uses its record to decide what to disclose about the buyer so as to influence the sellers' actions (i.e., $x(\cdot|\omega)$ for every ω).

To establish our benchmark, imagine that all parties have aligned interests: u_i is an affine transformation of u_0 for all $i = 1, \ldots, n$. Then, constraints (1) can be omitted and it is as if the platform *directly* controlled the sellers' actions. In this case, all the decisions in $\mathcal{U}_q$ are independent of one another. Indeed, $\mathcal{U}_q$ is separable across records: For each of them, the platform effectively faces a standard decision problem in which it chooses a to maximize $u_0(a, \omega)$ guided by the information in the record. For this reason, we will slightly abuse terminology and refer to our model with aligned interests as a *standard-decision problem*.

In this paper, however, we are mainly interested in instances of $\mathcal{U}_q$ where parties have conflicting interests, to which we will refer as *intermediation problems*. In this case, the platform can only influence the sellers' actions *indirectly*, subject to constraints (1). While it continues to face the collection of decisions in $\mathcal{U}_q$, these are no longer independent. That is, $\mathcal{U}_q$ is no longer separable across records because what information a signal conveys about one record depends on which other records lead to the same signal.[10] Consequently, while the value of each record continues to be determined by how it is used to guide decisions—like in standard-decision problems—this use is not confined to the interaction physically attached to that record—unlike in standard-decision problems. Thus, to answer our question, we need to systematically keep track of all the ways the platform uses each record to mediate all interactions and the resulting interdependencies.

This contrast between intermediation and standard-decision problems will be helpful to better understand our results and relate them to the classic work on the comparison of experiments under the decision-theoretic framework of Blackwell (1951; 1953) (see also Laffont (1989)). We will return to this point in Section 5.

[10] This dependence between signal decisions is orthogonal to our commitment assumption. It would arise even if our platform could not commit and we had to rely on some equilibrium notion.

3.1 The Data-Value Problem

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

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

$$t_i(a, \omega) \stackrel{\Delta}{=} \sum_{a_i' \in A_i} \left(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega) \right) \lambda_i(a_i' | a_i), \tag{3}$$

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

$$\begin{aligned} \mathcal{V}_q : \quad & \min_{v, \lambda} \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ & \text{s.t. for all } \omega \in \Omega, \\ & v(\omega) = \max_{a \in A} \Big\{ u_0(a, \omega) + t(a, \omega) \Big\}. \end{aligned} \tag{4}$$

We denote any optimal solution by (v_q^*, λ_q^*) and the induced functions t by t_q^*. By standard linear-programming arguments v_q^* is unique generically with respect to q. Note that v_q^* can depend on q for intermediation problems but not for standard-decision problems, as in this case $v_q^*(\omega) = \max_{a \in A} u_0(a, \omega)$ for all ω.

We refer to equation (4) as the *value formula*, which defines our main object of interest. The reason hinges on the next relation between the data-use and data-value problems and on the following interpretation. All proofs are in Appendix A.

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

$$\sum_{\omega \in \Omega} v_q^*(\omega) q(\omega) = U^*(q) \stackrel{\Delta}{=} \sum_{\omega \in \Omega} u_q^*(\omega) q(\omega). \tag{5}$$

This duality relation follows from basic linear-programming results. When applied to our specific problem, it becomes the key to answering our economic questions. In $\mathcal{U}_q$, every x defines a joint measure $\chi \in \mathbb{R}_+^{\Omega \times A}$, which must satisfy $\sum_{a \in A} \chi(a, \omega) = q(\omega)$; that is, the use of type-ω records to produce recommendations must exhaust their stock $q(\omega)$ in the database.

Formally, $v(\omega)$ is the multiplier of this constraint, which is usually interpreted as the shadow price of the corresponding input through the thought experiment of adding a marginal unit of it. In fact, $v_q^*(\omega)$ is equal to the derivative of $U^*(q)$ with respect to $q(\omega)$, as for any constrained optimization. However, it would be incorrect to think that $v_q^*(\omega)$ captures only the value of a marginal record of type ω. The linear structure of $\mathcal{V}_q$ and our value formula demonstrate that $v_q^*(\omega)$ is the value of *each* record of type ω in the database. Note that, by (4), $v_q^*(\omega)$ is measured in terms of the payoffs of the platform and the sellers. We will then call $v_q^*(\omega)$ the *unit value* of a record of type ω (see also Gale, 1989, p. 12). Note that $\mathcal{V}_q$ assigns such values simultaneously to all records and does not require finding x_q^*.

The rest of the paper characterizes the properties of v_q^* and their economic implications. Here, we begin with a useful lower bound. For $\omega \in \Omega$, let $CE(\Gamma_\omega)$ be the set of correlated equilibria of the game Γ_ω.

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

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

Lemma 1 and 2 imply that $v_q^*(\omega) = \bar{u}(\omega)$ for all ω if and only if there is an optimal x_q^* that satisfies $x_q^*(\cdot|\omega) \in CE(\Gamma_\omega)$ for all ω. In words, for such an x_q^* the platform fully discloses the buyer's record to the sellers for all interactions.

Unit Values and Individual Compensations

By quantifying how much a record contributes to the total payoff $U^*(q)$, v_q^* offers a *benchmark* for individually compensating each buyer as the "owner" of her record. In $\mathcal{V}_q$, we can view the platform as choosing v to minimize the total expenditure to compensate the buyers. However, the platform is constrained by equation (4), which imposes a lower bound for each buyer's compensation that takes into account how her record is used.[11] Paraphrasing Dorfman et al. (1987, p. 43), this interpretation is reminiscent of the operation of a competitive market where competition forces the platform to offer the "owner" of a record the full value to which her input gives rise, while competition among these "owners" drives down this value to the minimum consistent with this limitation. Gale (1989, Chapter 3.5) also shows how dual problems can deliver competitive prices of scarce inputs. In general, how much individuals will actually

[11] In fact, by complementary slackness $v_q^*(\omega) = u_0(a,\omega) + t_q^*(a,\omega)$ if $x_q^*(a|\omega) > 0$. We provide another independent economic interpretation of the data-value problem in Appendix B.

receive for their data can depend on the market structure, their bargaining power, and the need to incentivize them to disclose their data truthfully.

Nonetheless, to see the importance of guiding the compensation of data owners using v_q^* and not u_q^*, consider again our introduction example of a surplus-maximizing platform. Suppose it decides—perhaps forced by some regulation or court order—to compensate the buyers for their contribution to $U^*(q)$ by giving back some share δ to them. How δ is chosen and the compensations implemented is important in practice but irrelevant for the point we want to make here. The more fundamental question is how much each buyer should get. It seems that the answer should take into account each buyer's specific record. One could use u_q^*, which would result in incorrectly allocating $\delta U^*(q)$ only to the buyers with $\omega = \omega_2$ (each receiving $\delta u_q^*(\omega_2) = \delta\, 0.5$) because $u_q^*(\omega_1) = 0$. In fact, only the buyers with $\omega = \omega_1$ contribute to $U^*(q)$ because $v_q^*(\omega_2) = 0$. Thus, $\delta U^*(q)$ should be allocated entirely to these buyers (each receiving $\delta v_q^*(\omega_1) = \delta$).

3.2 Value Decomposition and Data Externalities

What determines the unit value of a record? Why and how are the direct payoffs u_q^* a biased measure of these values? We show next that the value of a record can be decomposed into two parts: its direct payoff and an additional component, which captures that record's effects on the information the platform discloses about other records and thus on their direct payoffs.

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

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

This result highlights that the effects captured by $t_q^*(\omega)$ are akin to an externality. Consider a buyer called Ann. Simply by belonging to the database, her record affects how the platform mediates the interactions that other buyers have with the sellers. Formally, $t_q^*(\omega)$ summarizes the marginal effect that Ann's record has on the direct payoff of other records (equation (6)). This externality is purely informational: Ann's record contributes to the information advantage that the platform has for all interactions it mediates and hence affects its decisions with other records through x_q^*. In fact, $\frac{\partial}{\partial q(\omega)} u_q^*(\omega') = \sum_a u_0(a, \omega') \frac{\partial}{\partial q(\omega)} x_q^*(a|\omega')$. Adjustments in x_q^* can arise because changing $q(\omega)$ can render x_q^* no longer feasible (i.e., obedient) or optimal.

This externality is a hallmark of intermediation problems. It arises when an intermediary tailors the information for the agents by pooling data records so as to manage conflicts of

interest. Indeed, the externality is absent in standard-decision problems, where it is optimal to fully disclose the type of each record.[12] It is worth emphasizing that this externality arises even if records are statistically independent. As such, it is distinct and complementary to the "learning" externalities discussed in Section 1, which arise because a buyer's record is informative about another buyer's preferences. This channel is intentionally switched off in our paper, which emphasizes externalities that arise endogenously from how data is used.

Which records generate positive and which records generate negative externalities?

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

The externalities lead to cross-subsidization of value from records with $t_q^*(\omega) < 0$ to records with $t_q^*(\omega') > 0$. Since the total payoff is fixed, records with $v_q^*(\omega') > u_q^*(\omega')$ must take their extra value from some other records. The second part of the corollary explains this cross-subsidization. Records with $t_q^*(\omega) < 0$ generate a direct payoff that exceeds the full-disclosure payoff $\bar{u}(\omega)$, which requires that $u_0(a,\omega) > \bar{u}(\omega)$ and $x_q^*(a|\omega) > 0$ for some a. That is, the platform earns a payoff with type-ω records that would never be possible by fully disclosing them, so it relies on pooling them with records of a different type. In this case, type-ω records do not "deserve" the full $u_q^*(\omega)$ and their value discounts the help received from other records. Conversely, this help from type-ω' records justifies why $t_q^*(\omega') > 0$ and their value exceeds $u_q^*(\omega')$. Finally, we can interpret $u_q^*(\omega) < \bar{u}(\omega)$ as "sacrificing" type-ω records, as the platform could fully disclose them and get $\bar{u}(\omega)$. For this sacrifice to be worthwhile, such records must receive compensation, explaining $t_q^*(\omega) > 0$. This last part offers a simple sufficient condition for $t_q^* \neq 0$. Appendix C provides another condition based on primitives.

Proposition 1 also highlights that the externalities through u_q^* are tightly related to how the platform exploits the sellers' incentives with its information. By the first part of (6), $t_q^*(\omega)$ aggregates externalities that type-ω records generate by inducing specific actions a. These are inversely related to the platform's resulting payoff, in the following sense.

Corollary 2. *Suppose $x_q^*(a|\omega) > 0$ and $x_q^*(a'|\omega) > 0$. Then, $u_0(a,\omega) > u_0(a',\omega)$ if and only if $t_q^*(a,\omega) < t_q^*(a',\omega)$.*[14]

[12] Whenever full disclosure is optimal, $u_q^*(\omega) = \bar{u}(\omega)$ and $v_q^*(\omega) = \bar{u}(\omega)$ as discussed after Lemma 2, so $t_q^*(\omega) = 0$. Note that the converse is not true: There are examples where $t_q^*(\omega) = 0$, but $v_q^*(\omega) > \bar{u}(\omega)$ for all ω.

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

[14] This follows from complementary slackness, namely $v_q^*(\omega) = u_0(a,\omega) + t_q^*(a,\omega)$ if $x_q^*(a|\omega) > 0$.

Thus, inducing actions whose payoff exceeds $\bar{u}$ by more, for instance, requires paying larger externalities to other records. Since $t_q^*(a,\omega) \stackrel{\Delta}{=} \sum_{i\in I} t_{q,i}^*(a,\omega)$, we can view $t_{q,i}^*(a,\omega)$ as how much seller i contributes to the externality. Recall that $t_{q,i}^*(a,\omega)$ differs from zero only if $\lambda_{q,i}^*(a_i'|a_i) > 0$ for some a_i' (see (3)). By standard arguments (complementary slackness), $\lambda_{q,i}^*(a_i'|a_i) > 0$ only if

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

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

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

Thus, we can also interpret $t_q^*(\omega)$ as aggregating the "cost of incentives" for the actions that the platform recommends using type-ω records. This cost is positive for seller i if recommending a with type-ω records hinders satisfying (7) because $u_i(a_i,a_{-i},\omega) < u_i(a_i',a_{-i},\omega)$, which then lowers $t_q^*(a,\omega)$ and hence $v_q^*(\omega)$. The opposite happens if recommending a with type-ω records helps satisfying (7) because $u_i(a_i,a_{-i},\omega) > u_i(a_i',a_{-i},\omega)$. Appendix B elaborates on this interpretation and how the platform exploits the sellers to determine their contribution to the externalities. Note that Corollary 3 differs from the immediate fact that optimal solutions of linear programs occur on the boundary of the feasible set, which here means that some obedience constraint must bind. Also, as q varies, x_q^* and hence t_q^* may change. However, as long as λ_q^* does not change, how each seller contributes to $t_q^*(\omega)$ does not change.

3.2.1 Application (Part I): Price Discrimination and the Externalities

To illustrate the importance of these data externalities, we consider a more general version of our example in the Introduction. There is only one seller ($n = 1$) who chooses the price a_1 for his product. For each buyer, θ is her valuation for the product. Let $\Omega = \{\omega_1,\ldots,\omega_K\} \subset \mathbb{R}_+$, $K \geq 2$, and ω_k be strictly increasing in k. Records of type ω_k fully reveal that $\theta = \omega_k$. Normalizing the seller's constant marginal cost to zero, his profit is a_1 if $\omega \geq a_1$ and zero otherwise: $u_1(a_1,\omega) = a_1 \mathbb{I}\{\omega \geq a_1\}$. The platform maximizes a weighted sum of profits and consumer surplus: $u_0(a_1,\omega) = \pi a_1 \mathbb{I}\{\omega \geq a_1\} + (1-\pi)\max\{\omega - a_1, 0\}$, where $\pi \in [0,1]$. Finally, let a_q be the optimal uniform monopoly price.

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

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

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

To understand this result, we note that x_q^* takes only two forms depending on π (see Appendix A.5). If $\pi \leq \frac{1}{2}$, the platform maximizes the buyers' surplus subject to holding the seller's expected profit at a_q, as when $\pi = 0$. Thus, it is as if trade happens for every interaction, generating total surplus equal to ω, and only the buyers with a product valuation of at least a_q contribute to guaranteeing this profit. If $\pi \geq \frac{1}{2}$, the platform fully discloses all records. This allows perfect price discrimination, so profits always equal the buyer's valuation and her surplus is zero.

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

How does caring more about the buyers' surplus affect the value of their records? By simple algebra, lowering $\pi \leq \frac{1}{2}$ decreases $v_q^*(\omega)$ if and only if the buyer has an intermediate valuation ($a_q \leq \omega < 2a_q$). Intuitively, for such records a larger share of the buyers' product valuation goes to fund the seller's guaranteed profits of a_q, which becomes more costly as their surplus becomes more important to the platform. By contrast, the records of buyers with low valuation help the platform achieve a positive surplus with other buyers, and the records of buyers with high valuation just yield a large surplus. For $\pi \geq \frac{1}{2}$, $v_q^*(\omega)$ increases in π independently of ω. This is because the platform helps the seller extract the full surplus from each interaction, and it cares more about doing so.

4 Willingness to Pay 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 obtains *more*

records in the database and, hence, it mediates more interactions between the buyers and sellers. The second—analyzed in Section 4.2—is that the platform obtains *better* records; namely, it observes more informative characteristics about existing buyers. In either case, having more data ultimately changes the database q, which is the basis for the sellers' beliefs. Hereafter, we assume that how the platform changes q is publicly observed and hence q is always commonly known.[15] Building on Section 3, we can then 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 comparative static exercises that show how the values of records vary between platforms which 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 the properties of the 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^*. The platform's preferences over databases are then fully characterized by v_q^*. Indeed, $v_q^*(\omega)$ is akin to the marginal utility of type-ω records 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 $MRS_q(\omega,\omega') \overset{\text{a.e.}}{=} -\frac{v_q^*(\omega)}{v_q^*(\omega')}$.

A classic property in standard consumer theory is that marginal utilities are diminishing. Does the same hold for the platform? More generally, how does v_q^* vary with q? We show that as records of a given type become more abundant, they become less valuable and do so stepwise. This follows from the next two results. The first establishes a general "scarcity principle" for data. Given q, define the share of type-ω records by

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

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, there exists $\bar{\mu}(\omega) < 1$ such that, if $\mu_q(\omega) > \bar{\mu}(\omega)$, then $v_q^*(\omega) = \bar{u}(\omega)$.*

This property holds generally, irrespective of the details of the intermediation problem. It

[15] Of course, in reality the platform may change its database privately without the sellers' knowing exactly how. Allowing for this introduces complications and requires enriching the model accordingly. We leave this for future research.

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 type-ω records is downward sloping and converges to $\bar{u}(\omega)$ when $q(\omega)$ is sufficiently large. Equivalently, the individual contribution of type-ω records to the platform's payoff—hence, their owners' benchmark compensation—decreases as their quantity increases.

This decline in value is stepwise because v_q^* is locally constant in q.

Proposition 4 (Stability)**.** *There exists a finite collection $\{Q_1, \ldots, Q_M\}$ of open, convex, and disjoint subsets of $\mathbb{R}_+^\Omega$ such that $\cup_m Q_m$ has full measure and, for every m, v_q^* is unique and constant for $q \in Q_m$.*

Each Q_m is the interior of a cone in the space of databases $\mathbb{R}_+^\Omega$.[16] Importantly, v_q^* is constant even though the platform may adjust how it uses its data when q changes. We can show that within each cone, while $v_q^*(\omega)$ is constant, the optimal $x_q^*(\omega)$ changes as a function of q (see Remark 1 in Appendix A.5). Intuitively, this is because x_q^* has to be fine-tuned to maximally exploit the sellers' incentives. By contrast, v_q^* depends only on which sellers' incentives are exploited, but not on how much (recall equation (3) and Corollary 3).

Returning to the platform's marginal rate of substitution between records, is it diminishing as in standard consumer theory? The answer is yes, at least weakly: The platform's preferences are always convex, because $U^*(q)$ is always a concave function of q.[17] However, in some cases records are perfect substitutes, namely $MRS_q(\omega, \omega')$ is constant. An example is when the platform faces a standard-decision problem, since then $v_q^*(\omega) = \bar{u}(\omega)$ for all ω. The next result characterizes which intermediation problems also lead to perfect substitutability between all records (i.e., $MRS_q(\omega, \omega')$ does not depend on q for all ω, ω').

Proposition 5. *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 every record. In this case, full disclosure is optimal for all $q \in \mathbb{R}_+^\Omega$.*

This result has several implications. First, suppose we can estimate the platform's demand functions by observing its transactions in the data market. Then, by detecting any imperfect substituatbility between record types, we can infer that the platform is withholding information

[16] It is easy to see that unit values are constant along the rays in the space of databases: If $q' = \alpha q$ for $\alpha > 0$, then $v_q^* = v_{q'}^*$. This is because only the frequency of record types matters for the sellers' incentives.

[17] Concavity follows because, by (5), we can view U^* as the result of minimizing a family of functions that are linear in q (Rockafellar (1970), Theorem 5.5). It is related directly to the concavification results in Mathevet et al. (2020) and indirectly to the individual-sufficiency results in Bergemann and Morris (2016).

from the sellers. We can do so even if we know nothing about the intermediation problem it faces (i.e., Γ). Indeed, by Proposition 5 some types of records are imperfect substitutes if and only if it is never optimal to fully disclose all records. More generally, the intermediary's transactions in a data market reveal properties of how it uses its database, which may be harder to observe.

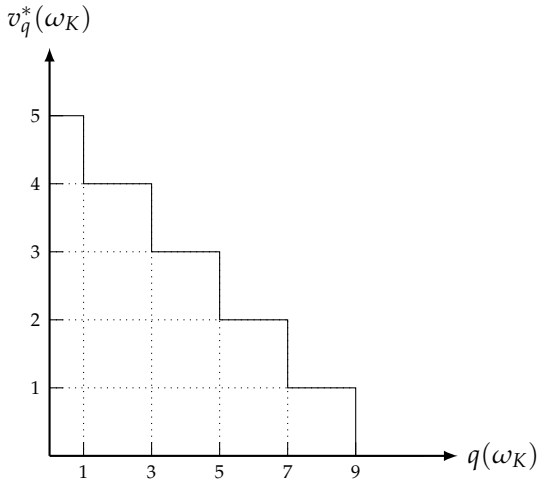
Another implication is that the optimality of withholding some information does not depend on the database composition. This simplifies assessing whether an intermediary will withhold information based on the primitives of a specific application (i.e., Γ). One way is to start from full disclosure and show that for some $q \in \mathbb{R}^{\Omega}_{++}$ we can do strictly better by sometimes concealing any type of records. Since the answer does not depend on q, we can pick it in any convenient way (e.g., uniform quantities). Alternatively, we can identify conditions for the optimality of withholding information directly in terms of Γ. Appendix C provides such a condition.

Last but not least, convexity of the platform's preferences leads to standard demand analysis. In particular, choosing an optimal database subject to a budget constraint is a well-behaved problem. Given market price $p(\omega) > 0$ for every ω, the optimal q is characterized by a generalized version of the usual tangency condition that deals with kinks in indifference curves:

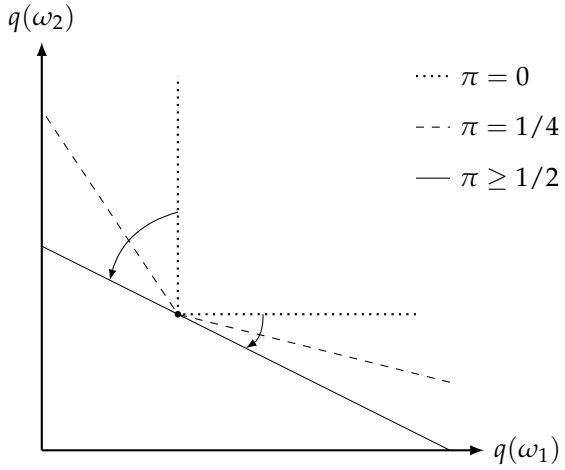
$$\max_{v \in v_q^*} \frac{v(\omega)}{v(\omega')} \geq \frac{p(\omega)}{p(\omega')} \geq \min_{v \in v_q^*} \frac{v(\omega)}{v(\omega')}, \qquad \omega, \omega' \in \Omega.^{18}$$

It is easy to see that if withholding information is optimal (i.e., $v_q^* \neq \bar{u}$ for some q), then there is an open set of prices for which the platform chooses a nontrivial database containing multiple types of records. More generally, we can use v_q^* to characterize the platform's demand functions for records, thus enabling a general study of the demand side of data markets. These functions satisfy some simple properties that may be useful for empirical analysis: Since $U^*(q)$ is homothetic, data records are normal goods and the optimal database composition depends only on price ratios, not on the platform's budget. Which prices will prevail in the market is of course determined by the interplay of demand and supply. Under perfect competition, Dorfman et al. (1987) and Gale (1989) provide arguments for equilibrium prices to equal v_q^*.

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



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



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

Figure 1: Platform's demand and indifference curves

4.1.1 Application (Part II): Demand Curve and Substitutability

Returning to the setting of Section 3.2.1, recall that the platform maximizes a weighted sum of the buyers' surplus and the profits of a single price-setting seller, where the latter receives weight $\pi \in [0, 1]$.

We first show an example of a downward-sloping demand curve. Figure 1(a) shows the value of records of type ω_K calculated using Proposition 2. This value is stepwise diminishing as these records become more abundant (Propositions 4 and 3). The figure also shows that as $q(\omega_K)$ becomes sufficiently large, $v_q^*(\omega_K)$ reaches a lower bound, which in this case is 0.

Next, we explore how the substitutability between records depends on π. When $\pi \geq \frac{1}{2}$, all types of records are perfect substitutes and $MRS_q(\omega, \omega') = -\frac{\omega}{\omega'}$ for all q, which is thus constant in π. When $\pi < \frac{1}{2}$ instead, records become more substitutable as the platform cares less about the buyers' surplus (i.e., π increases). Recall that a_q is the seller's optimal price if he knows only the database composition q.

Corollary 4. *Fix q and increase $\pi < \frac{1}{2}$. If $\omega, \omega' < a_q$, $MRS_q(\omega, \omega')$ is constant at $-\frac{\omega}{\omega'}$. If $\omega < a_q \leq \omega'$, $MRS_q(\omega, \omega')$ increases monotonically 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 π 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 $\pi = 0$ records of type $\omega = a_q$ are perfect complements with every other type. These patterns are illustrated in the right panel of Figure 1, which shows the platform's indifference curves in the case with two types of records.

4.2 Better Records and Willingness to Pay for Information

A platform can also change its database by refining some of its existing records with better information. For example, this could involve observing new personal characteristics about a subset of buyers. Intuitively, refining a record changes its type according to what the platform learns. How do such refinements change the value derived from each record? Do they always improve the platform's total payoff and consequently command a positive willingness to pay?

We first need to formalize what a refinement is. Recall that every buyer's record of type $\omega \in \Omega$ is informative about her θ, so it induces a belief $\beta_\omega \in \Delta(\Theta)$. A refinement of a record of type ω is a distribution $\sigma_\omega \in \Delta(\Omega)$ that satisfies the usual Bayes' consistency condition $\beta_\omega = \sum_{\omega' \in \Omega} \sigma_\omega(\omega')\beta_{\omega'}$. That is, any such refinement is equivalent to observing an exogenous signal that transforms a record of type ω into a record of type ω' with probability $\sigma_\omega(\omega')$. For instance, the original record may contain only the buyer's age, while the refined record may also contain her gender. When refining multiple records of type ω—in particular, a *share* $\alpha \in [0,1]$ of $q(\omega)$—each record is refined independently according to σ_ω.[19] Importantly, implicit in the definition there is the assumption that Ω is "rich" in the sense that it already contains all record types that can result from σ_ω. This allows us to use the platform's preferences over databases in $\mathbb{R}^\Omega_+$ characterized by v^*_q to assess the consequences of refinement σ_ω.

Consider refining a share α of type-ω records according to σ_ω. How does this change the unit value that the platform derives from its records? Such a refinement has both direct and indirect effects, as it affects both the records that are being refined and those that are not. The root of these interdependencies is the externality discussed in Section 3.2. Refining α of type-ω records changes the original database q into a new one, denoted by q_α, which contains fewer records of type ω and more records of the types ω' that result from the refinement (i.e., $\omega' \in \operatorname{supp} \sigma_\omega$). Thus, the unit value of the former records may increase and that of the latter decrease by the scarcity principle (Proposition 3). Formally, given $\alpha \in [0,1]$ and σ_ω, by the Law of Large Numbers $q_\alpha(\omega) = (1-\alpha)q(\omega)$ and $q_\alpha(\omega') = q(\omega') + \alpha\sigma_\omega(\omega')q(\omega)$ (where we can interpret $\alpha = 0$ as refining only one record since it is infinitesimal and $q_0 = q$). Note

[19] We discuss refinements that are correlated among records in Section 5.1.

that the composition q_α of the new database is certain, even though it is uncertain which records of type ω become of type ω'. Thus, it suffices that the sellers know that a database q has been refined according to σ_ω and α for them to know the resulting composition q_α.

Corollary 5. *Fix q. Suppose a share α of type-ω records is refined according to σ_ω.*

> Direct Effects: *The value of* refined *records increases in expectation. That is, we have $\sum_{\omega' \in \Omega} v^*_{q_\alpha}(\omega')\sigma_\omega(\omega') - v^*_q(\omega) \geq 0$. This increase shrinks as α gets larger.*
>
> Indirect Effects: *The value of* unrefined *records of type ω increases: $v^*_{q_\alpha}(\omega) \geq v^*_q(\omega)$. The value of* unrefined *records of type $\omega' \in \operatorname{supp}\sigma_\omega$ decreases: $v^*_{q_\alpha}(\omega') \leq v^*_q(\omega')$. Both these effects are larger as α gets larger.*

With regard to the refined records, the expected gain in their value can shrink but never turn into a loss, even if refining more records lowers the value of the record types that result from it. Intuitively, for each refined record the platform knows more about the corresponding interaction, so it can better tailor its signals for the sellers and achieve more with that record. However, the externalities that contribute to its value may now be smaller. The former positive aspect dominates the latter because we are considering independent refinements. This is no longer true if refinements are correlated between records (see Section 5.1). Finally, note that the direct effects of a refinement depend on q, so the net value of the information it adds to a record cannot be quantified in absolute terms (unlike for standard-decision problems).

Corollary 5 highlights a novel implication of people's decisions to disclose their data. Recall that v^*_q is a benchmark for compensating buyers for their specific record. We can interpret a refinement as a buyer's decision of whether to disclose more of her personal characteristics. Imagine a group of similar buyers—that is, whose records are of the same type—which includes Ann and Bonnie. Ann decides to disclose, expecting that her record will become more valuable and hence may result in a higher compensation (direct effect). Bonnie instead decides *not* to disclose, yet her record may also become more valuable but for different reasons (indirect effect). Moreover, a larger group of disclosing buyers decreases Ann's expected gain in value, but increases Bonnie's. The disclosing buyers can also cause the value of, say, Cindy's record to fall—hence, lower her compensation—if her record is of one of the types that can result from the refinement (indirect effect). Importantly, these effects happen even if the platform does not learn anything new about Bonnie and Cindy from what Ann and the other buyers disclosed—in contrast to the learning externalities discussed in the literature (see Section 3.2).

Given these mixed effects of refinements on the unit value of all records, it is unclear whether they benefit the platform overall. In fact, those effects reflect a fundamental trade-off that refin-

ing records can generate in intermediation problems (but not in standard-decision problems). On the one hand, knowing more about each refined record allows the platform to better tailor its signals for the sellers and possibly achieve more in those interactions. On the other hand, changing the database q can change the sellers' beliefs about each buyer, which in turn can weaken the platform's informational advantage and hence its ability to influence the sellers' actions. A mechanism x may be obedient before the refinement but not after it, which can hurt the platform. Nonetheless, we obtain the following.

Proposition 6. *Fix q. Suppose a share α of type-ω records is refined according to σ_ω. The platform weakly benefits from this refinement: $U^*(q_\alpha) - U^*(q) \geq 0$. Moreover, the benefit is zero for all $\alpha \in [0,1]$ if (and only if generically in q) there exists $a \in supp\, x_q^*(\cdot|\omega'')$ for $\omega'' = \omega$ and all $\omega'' \in supp\, \sigma_\omega$. Finally, the refinement's marginal benefit decreases in α.*

This implies that the platform's willingness to pay for a refinement is always weakly positive, so the positive effects on refined records always dominate the negative effects on other records. The platform's willingness to pay can be strictly negative for refinements that are correlated between records (see Section 5.1).

Proposition 6 provides a sharp condition for the willingness to pay for refinements to be zero, which depends only on the initial q. Given this q, there must be a common action profile that the platform induces with positive probability both for the original record to be refined and for every type that it can turn into when refined. Intuitively, this means that the platform is exploiting its information advantage to sometimes use the original record as if it was already refined, so refining it does not make it more valuable. In fact, under this condition all direct and indirect effects in Corollary 5 are zero. Importantly, note that the direct effect of refining a record can be zero even if the platform uses it differently after the refinement (i.e., even if $x_q^*(\cdot|\omega) \neq x_q^*(\cdot|\omega')$ and $u_q^*(\omega) \neq u_q^*(\omega')$ for some $\omega' \in \text{supp}\, \sigma_\omega$). Overall, the platform may be unwilling to pay a strictly positive price for refining its records, despite acting on the information it receives (i.e., changing x_q^*). This is different from standard-decision problems, for which a key insight is that more information is strictly beneficial if it changes the optimal choices.

Finally, Proposition 6 shows that the marginal benefit of a refinement is diminishing in the share of refined records. This may be reminiscent of classic results in standard decision problems where information has decreasing marginal returns (see, e.g., Moscarini and Smith, 2002; Varian, 2019). However, there is an important difference. Our exercise is not to gradually give the platform more information about one fixed interaction so that it can better learn the buyer's preferences. Focusing on this intensive margin is perhaps the most typical way of studying

$x_q^*(a\|\omega)$	ω_1	ω_2	$\omega = \omega^\circ$
$a = 1$	1	$\frac{q(\omega_1)-(2h-1)q(\omega^\circ)}{q(\omega_2)}$	1
$a = 2$	0	$1 - \frac{q(\omega_1)-(2h-1)q(\omega^\circ)}{q(\omega_2)}$	0

Table 1: Optimal x_q^*

returns from information, especially in standard decision problems (see Bergemann and Ottaviani, 2021, Section 2.5). We instead fix the amount of information we give the platform for each interaction (i.e., σ_ω) and vary how many interactions we *independently* refine in this way (i.e., α). As such, this extensive-margin exercise has constant returns for standard-decision problems but not for intermediation problems—again, due to the externalities documented in Section 3.2.[20]

4.2.1 Application (Part III): Refinements

We illustrate some of these points using the setting of Section 3.2.1 with a single price-setting seller. Suppose the platform maximizes the buyers' surplus ($\pi = 0$). As before, ω_1 and ω_2 are the record types that correspond to buyers whose valuation θ is 1 and 2; instead, ω° corresponds to buyers' whose valuation is believed to be $\theta = 2$ with probability $h > \frac{1}{2}$ and $\theta = 1$ otherwise. Fix any q that satisfies $q(\omega^\circ) < q(\omega_1) < q(\omega_2)$ in which case we have that $v_q^*(\omega_1) = 1$, $v_q^*(\omega_2) = 0$, and $v_q^*(\omega^\circ) = 1 - h$. Now, suppose we refine a share α of type-ω° records with a refinement σ_{ω° such that $\sigma_{\omega^\circ}(\omega_2) = h$ and $\sigma_{\omega^\circ}(\omega_1) = 1 - h$. As shown in Table 1, the platform changes how it uses the refined records—compare $x_q^*(\cdot|\omega^\circ)$ and $x_q^*(\cdot|\omega_2)$—as well as the unrefined records of type ω_2—note that $x_q^*(\cdot|\omega_2)$ depends on q. Nonetheless, the "if" condition in Proposition 6 holds. Therefore, for any $\alpha \in [0, 1]$ the platform's willingness to pay for the refinement as well as the expected increase in unit value of each refined record are zero: $U^*(q) = U^*(q_\alpha)$ and $v^*(\omega_1)\sigma_{\omega^\circ}(\omega_1) + v^*(\omega_2)\sigma_{\omega^\circ}(\omega_2) = v^*(\omega^\circ)$. Both are instead strictly positive if $q(\omega_1) < q(\omega^\circ)$ and $\alpha > 0$ is sufficiently small. See Appendix D for more details.

[20] In fact, given σ_ω the marginal effect of changing α on $U^*(q_\alpha)$ equals $\sum_{\omega' \in \Omega} v_{q_\alpha}^*(\omega')\sigma_\omega(\omega') - v_{q_\alpha}^*(\omega)$ (see the proof of Proposition 6 for details).

5 Discussion

5.1 Correlation Between Records and General Refinements

Throughout the paper we assumed that each buyer's record is uninformative about other buyers' preferences (i.e., records are independent). We did so to clarify that the interdependencies between the values of data records arise not from exogenous correlation, but endogenously from how the data is used. While this assumption may seem restrictive, it can be easily relaxed. For a fixed database, each buyer's record should already contain all the observations available to the platform that are relevant to that buyer, which may include variables that refer to other individuals. For example, if the average income in Ann's neighborhood is predictive of Ann's income, then it should be listed in Ann's record. Once this assignment is done for each buyer, conditional on her record any other record adds no information about her θ by construction. We can then replace our original independence assumption with this *conditional* independence assumption, and nothing changes in our analysis for a fixed database.

The possibility that one buyer's data is informative about other buyers has deeper implications with regard to refinements. For example, observing Bonnie's income may require updating her record as well as the record of her neighbor Ann. This introduces correlation in how records are updated, so it leads to more general refinements than those analyzed in Section 4.2. Nonetheless, we can continue to view such refinements as changing the platform's database and analyze their consequences using our tools (i.e., v_q^* over $\mathbb{R}_+^\Omega$). Ultimately, a refinement transforms the type of each affected record into a new one, so it changes the original q to another q'. This change can exhibit correlation between records; yet, for each of them the Bayes' consistency condition $\beta_\omega = \sum_{\omega' \in \Omega} \hat{\sigma}_\omega(\omega')\beta_{\omega'}$ must hold, where $\hat{\sigma}_\omega$ is the marginal distribution of the type changes for records of type ω.

Interestingly, correlated refinements can overturn some of the results from Section 4.2. For standard-decision problems, they always weakly increase both the records' unit values and the platform's total payoff. By contrast, for intermediation problems there are refinements that *decrease* the unit value of the refined records as well as the platform's total payoff. This is because they change more drastically not only what the platform knows about the buyers, but also the degree and nature of its information advantage over the sellers (recall that q is always commonly known).

We illustrate this possibility with an example. We use again the setting of Section 4.2.1 and assume that the platform maximizes buyers' surplus ($\pi = 0$). Let the initial q satisfy $q(\omega_\circ) <$

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

Table 2: Value of records and total payoffs for specific databases ($\pi = 0$)

$q(\omega_1) < q(\omega_2)$ and $q(\omega_2) < q(\omega_1) + (1-h)q(\omega^\circ)$. 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 type-ω° records involve buyers with the same valuation. Thus, if refined, 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, this refinement transforms the original database q into a new one: With probability $1-h$, the new database is q' and satisfies $q'(\omega_1) > q'(\omega_2) > q'(\omega^\circ) = 0$; with probability h, the new database is q'' and satisfies $q''(\omega_2) > q''(\omega_1) > q''(\omega^\circ) = 0$. Table 2 reports the value of each record for these databases (see Appendix D for details). We find that the refinement has a strictly negative effect on both the unit value of the refined records and the platform's total payoff. Indeed, note that $v^*_q(\omega^\circ) > (1-h)v^*_{q'}(\omega_1) + hv^*_{q''}(\omega_2) = 0$ and $U^*(q) > U^*(q') > U^*(q'')$ because $q'(\omega_2) = q(\omega_2)$ and $q''(\omega_1) = q(\omega_1)$. By contrast, if the platform maximizes the seller's profits ($\pi = 1$), we have $v^*_{\hat{q}}(\omega_1) = 1$, $v^*_{\hat{q}}(\omega_2) = 2$, and $v^*_{\hat{q}}(\omega^\circ) = 2h$ for all $\hat{q}$. Thus, the same refinement has a strictly positive effect on both the unit value of the refined records and the platform's total payoff. The key is that a profit-maximizing platform treats each buyer-seller interaction as an independent decision problem, 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 its information advantage through the composition of its database.

5.2 Standard-Decision Versus Intermediation Problems

It is instructive to briefly explain how our setting relates to the classic decision-theoretic framework of Blackwell (1951, 1953) and the ensuing literature (see, e.g., Laffont, 1989). In a standard decision problem there is an unknown state of nature $\theta \in \Theta$ that is drawn according to some distribution $\psi \in \Delta(\Theta)$. The decision maker observes an exogenous signal $\omega \in \Omega$ from a known experiment $e : \Theta \to \Omega$. Let $\tilde{u}_0(a, \theta)$ be the ex-post utility of the decision maker from the payoff-relevant action $a \in A$. Then, conditional on the signal realization ω, the expected payoff of the decision maker is $u_0(a, \omega) = \mathbb{E}_{\psi,e}(\tilde{u}_0(a, \theta)|\omega)$. Last but not least, the decision

maker directly chooses a.

Our setting shares this framework's basic elements. Our intermediary is the analogue of the decision-maker, but faces a fundamentally different decision. Instead of directly choosing a, the intermediary has to decide what to disclose about the exogenous signal ω so as to influence another agent's choice of a. The capacity to influence this choice depends on the intermediary's information advantage over the agent and is constrained by the agent's incentives.

Moreover, our intermediary mediates a *collection* of such problems whose respective θ and ω have already realized. While the frequency of problems whose signal is ω is common knowledge, the agent cannot identify which ones these are. Only the intermediary can: A data record of type ω allows the intermediary to identify a problem whose signal realization was ω.[21] While mathematically this collection of problems is the analogue of the usual prior distribution, this "frequentist" approach has two advantages. First, it is more descriptive: It allows us to think of data records as physical objects rather than mutually exclusive possibilities. This is important if we want to think about data records as being traded based on their specific content, as it is often the case in data markets (Bergemann and Bonatti, 2019). Second, this approach allows us to ask natural and practical questions—such as the effects of adding more records or refining existing records—which would be artificial with the standard approach. We can think of refining an existing record as observing the realization of an experiment, in line with the tradition following Blackwell. Hence, we can view Proposition 6 as studying the "value of information" in intermediation problems.

5.3 General Intermediation Problems

Our framework and results apply more broadly to any setting where a principal mediates interactions between multiple agents using data. For ease of exposition, we simplified the model in several ways. Neither changes the analysis or its interpretations. First, we can allow the principal to also choose an action $a_0 \in A_0$ for each mediated interaction. In this case, a mechanism x also has to specify a_0 for each ω. Second, we can allow each agent i to also privately observe some data about the interaction he is in. For example, in our leading e-commerce example the seller can observe the quality or the history of customer reviews of his product. We can again model the realizations of such data with some finite set Ω_i, where each ω_i is ultimately an ex-

[21] Rather than simultaneously mediating all problems in the collection, an equally valid interpretation is that the intermediary commits to a mechanism for the whole collection and then problems are drawn independently and mediated one at a time.

ogenous signal about some underlying payoff-relevant θ. Let $\Omega = \Omega_0 \times \ldots \times \Omega_n$ with typical element $\omega = (\omega_0, \ldots, \omega_n)$. The key assumption is that the principal also observes the private data of each agent—i.e., the entire $\omega = (\omega_0, \ldots, \omega_n)$—as does the omniscient designer in Bergemann and Morris (2016). Thus, now the whole vector ω defines a type of data record in the principal's database and characterizes each interaction that it mediates. Our proofs in Appendix A already take into account this more general setting.

5.4 Additional Examples

We conclude by sketching other possible applications of our model.

Navigation Services. A navigation app uses data about routes' conditions to direct traffic by providing drivers with information—such as recommended routes and travel times. Das et al. (2017) propose a simple way to model this complex problem. Suppose the app (principal) seeks to minimize congestion. We can think of an interaction as consisting of a group of drivers (agents) in some city who simultaneously choose, say, one of two routes between the residential and business district. For each route, the travel time increases in how many drivers choose it but at different rates (e.g., because one is a highway and one is surface streets); travel times also depend on some uncertain event (e.g., construction work), which is observed only by the app. For each city served by the app, the realization of this event defines its data record.

Ridesharing. A ridesharing platform mediates the interactions between n drivers and a population of potential riders who just landed at an airport. Each rider seeks to reach her final destination $\vartheta_d \in [0,1]$ and values a ride $\vartheta_v \geq 0$. Her preference is then pinned down by $\theta = (\vartheta_d, \vartheta_v)$. The platform knows ϑ_d and some personal characteristics of the rider, which are informative about ϑ_v. The drivers do not know anything about the riders and compete for them by posting a price a_i. Drivers have known preferences over final destinations—for instance, they are increasing in ϑ_d. Once an offer is accepted, the driver must honor it regardless of the final destination. The platform chooses what information about a rider's θ to disclose to the drivers so as to maximize a combination of the rider's and drivers' payoffs.

Online Advertisement. A population of individuals uses a search engine run by a platform. For each individual, it keeps a record that includes the searched keywords and some personal characteristics that are informative about her tastes in an horizontally differentiated product market, summarized by $\theta \in [0,1]$. There is a finite set $I \subseteq [0,1]$ of advertisers. The index $i \in I$ captures the advertiser's exogenous position in the product market—e.g., whether he advertises men's or women's apparel. The advertisers compete in a second-price auction to

display an ad to each individual. Advertiser's i expected profits from winning access to an individual decreases in the distance between θ and i. The platform chooses which information about each individual's θ to disclose to the advertisers so as to maximize its total payoff.

6 Conclusion

This paper explains what determines the individual value of specific data records and what its properties are. In doing so, it advances our understanding of the demand side of data markets, thus shedding light on how they work and how they may be affected by regulatory interventions. This can provide insights into a key part of the digital economy. To the best of our knowledge, our approach to assessing the value of data has not been used before, it is broadly applicable, and it lays the foundations on which more questions can be tackled by future research. One direction is to fully analyze specific applications, such as e-commerce or the settings sketched in Section 5.4. Another is to explicitly model some of the privacy regulations discussed in policy circles, or to consider richer and possibly private ways in which intermediaries can change their databases.

References

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

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

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

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

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

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

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

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

Bergemann, D., A. Bonatti, and T. Gan (2020): "The Economics of Social Data," *Cowles Foundation Discussion Papers*.

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.

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.

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

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

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

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

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

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

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

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

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

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*.

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

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*.

Jones, C. I. and C. Tonetti (2020): "Nonrivalry and the Economics of Data," *American Economic Review*, 110, 2819–58.

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

Laffont, J.-J. (1989): "The Economics of Uncertainty and Information," .

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

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.

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

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

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

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

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

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

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

Rockafellar, R. T. (1970): *Convex analysis*, 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.

Stigler Report (2019): "Stigler Committee on Digital Platforms," *Final Report*, September.

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

Yang, K. H. (2020): "Selling Consumer Data for Profit: Optimal Market-Segmentation Design and its Consequences," *Working Paper*.

Appendix

A Proofs

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

A.1 Proof of Lemma 1

We will formulate $\mathcal{U}_q$ in terms of choosing a measure $\chi \in \mathbb{R}_+^{A\times\Omega}$:

$$\begin{aligned}
\mathcal{U}_q : \quad & \max_{\chi} \sum_{\omega\in\Omega, a\in A} u_0(a,\omega)\chi(a,\omega) \\
& \text{s.t. for all } i \in I, \omega_i \in \Omega_i, \text{ and } a_i, a_i' \in A_i, \\
& \sum_{\omega_{-i}\in\Omega_{-i}, a_{-i}\in A_{-i}} \big(u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)\big)\, \chi(a_i, a_{-i}, \omega) \geq 0, \quad \text{(A.1)} \\
& \text{and for all } \omega \in \Omega, \\
& \sum_{a\in A} \chi(a,\omega) = q(\omega).
\end{aligned}$$

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

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

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

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

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

$$\begin{array}{lrcl} \max\limits_{\chi} & \mathbf{u}_0^T\chi & & \\ \text{s.t.} & \mathbf{U}\chi & \geq & \mathbf{0}, \\ & \mathbf{I}\chi & = & q, \\ & \chi & \geq & \mathbf{0}. \end{array} \tag{A.2}$$

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

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

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

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

$v(\omega) \in \mathbb{R}$ for all $\omega \in \Omega$, and for all $(a,\omega) \in A \times \Omega$

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

The objective simplifies to

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

The second set of constraints can be written as

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

Define the sum in this expression by $t(a\omega)$ for all (a,ω). Since for every $\omega \in \Omega$ this constraint has to hold for all $a \in A$ and we have a minimization problem, we conclude that each $v(\omega)$ has to satisfy

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

Thus, we obtain our data-value problem $\mathcal{V}_q$. □

A.2 Non-degeneracy and a Remark on the Structure of Solutions

In Section 2, we assumed that no more than $|A \times \Omega|$ of the constraints (1) 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.2) and recall that $\chi \in \mathbb{R}_{+}^{A\times\Omega}$, which has dimension $|A \times \Omega|$. A basic feasible solution of $\mathcal{U}_q$ is a χ such that (*i*) all equality constraints are active, (*ii*) $|A \times \Omega|$ of the constraints active at χ are linearly independent, and (*iii*) all constraints are satisfied. Formally, we assume the following.

Assumption 1 (Non-degeneracy)**.** *At every basic feasible solution χ 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 $\mathcal{U}_q$ to the standard form $\mathcal{U}_q^S$ which can be written as follows:*

$$\begin{array}{lrcl} \max\limits_{\chi,s} & \mathbf{u}_0\chi & & \\ s.t. & \mathbf{U}\chi - s & = & \mathbf{0}, \\ & I\chi & = & q, \\ & \chi, s & \geq & \mathbf{0}, \end{array} \tag{A.3}$$

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

$$\left[\frac{\chi_q^*}{s_q^*}\right] = \mathbf{B}^{-1}\left[\frac{\mathbf{0}}{q}\right], \tag{A.4}$$

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

$$\left[\frac{v_q^*}{\lambda_q^*}\right] = \mathbf{u}_0\mathbf{B}^{-1}. \tag{A.5}$$

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

A.3 Proof of Lemma 2

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

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

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

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

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

A.4 Proof of Proposition 1

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

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

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

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

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

$$t_q^*(\omega) \;=\; v_q^*(\omega) - u_q^*(\omega) = \frac{\partial U^*(q)}{\partial q(\omega)} - u_q^*(\omega)$$

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

where the existence of the derivative $\frac{\partial x^*_q(a|\omega')}{\partial q(\omega)}$ almost everywhere follows from (A.4).

A.5 Proof Proposition 2

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

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

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

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

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

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

subject to constraints (1).

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

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

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

$$
U^*(q) = \sum_{\omega,a} [\pi u_1(a,\omega) + (1-\pi)u_0(a,\omega)] \underline{x}^*(a|\omega) q(\omega)
$$

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

Note that

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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore, for $\omega \geq a_q$ define

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

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

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

A.6 Proof Proposition 4

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

Fix q and suppose that the optimal solution (v_q^*, λ_q^*) of the dual of $\mathcal{U}_q$ is unique. As explained in Remark 1, there exists a submatrix $\mathbf{B}$ such that (v_q^*, λ_q^*) satisfies (A.5). Given Assumption 1, Theorem 3.1 and Exercise 3.6 in Bertsimas and Tsitsiklis (1997) imply that

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

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

$$\left[\mathbf{U}_i(a_i, a_i', \omega_i) \mid -\mathbf{1}_i(a_i, a_i', \omega_i)\right] \mathbf{B}^{-1} \begin{bmatrix} \mathbf{0} \\ q \end{bmatrix} > 0, \tag{A.6}$$

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

Now consider changes in q and note that it only enters the objective of $\mathcal{V}_q$. Each condition (A.6) 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.5) and

$$Q(\mathbf{B}) = \{q : (A.6) \text{ holds for all } i \in I \text{ and } (a_i, a_i', \omega_i) \text{ s.t. } \lambda_{\mathbf{B},i}^*(a_i|a_i', \omega_i) = 0\}.$$

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

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

A.7 Proof of Proposition 3

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

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

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

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

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

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

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

Then f is supermodular in (v, λ).

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

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

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

Finally, by Theorem 5 in Milgrom and Shannon (1994), $\arg\max_{(v,\lambda)\in S} f(v, \lambda; \mu)$ is monotone nondecreasing in μ. This monotone comparative statics coupled with generic uniqueness of v_q^* with respect to q imply that if $\mu_q(\omega) > \mu_{q'}(\omega)$ for two databases q and q' then $v_q^*(\omega) \leq v_{q'}^*(\omega)$. That is, this monotonicity of $v_q^*(\omega)$ holds for any selection v_q^* from the optimal solution correspondence of $\mathcal{V}_q$.

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 $v_q^*(\omega) = \bar{u}(\omega)$. Indeed, the definition of $\bar{u}(\omega)$ implies that it can be written as

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

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

For $\varepsilon > 0$, consider a set $M_\varepsilon(\omega)$ defined as $M_\varepsilon(\omega) = \{\mu \in \Delta(\Omega) : \mu(\omega') \in (0, \varepsilon)$ for $\omega \neq \omega', \mu(\omega) < 1\}$. By Proposition 4, 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 $\hat{u}(\omega)$. If $\hat{u}(\omega) = \bar{u}(\omega)$, then the result follows. Suppose, on the contrary, that $\hat{u}(\omega) \neq \bar{u}(\omega)$. We can always pick a sequence μ^n, $n \in \mathbb{N}$, from $\mathcal{P}_m \cap M_\delta(\omega)$ that converges to $\tilde{\mu}$, with $\tilde{\mu}(\omega) = 1$. Then for every $n \in \mathbb{N}$, $v_q^*(\omega) = \hat{u}(\omega)$ for every q, such that $\mu_q = \mu^n$. By the Berge's maximum theorem, (v_q^*, λ_q^*) is an upper-hemicontinuous correspondence and therefore has a closed graph. Hence, $\hat{u}(\omega) \in v_q^*(\omega)$ for every q, with $\mu_q = \tilde{\mu}$. We obtain the desired contradiction, since $v_q^*(\omega) = \bar{u}(\omega)$ for such q.

A.8 Proof of Proposition 5

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 2 and Proposition 3, it follows that $v_q^*(\omega) = \bar{u}(\omega)$ for all ω and q. It follows that 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

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

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

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

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

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

A.9 Proof of Corollary 4

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

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

Thus, we have

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

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

A.10 Proof of Corollary 5

Fix ω, q, and a refinement σ_ω. Since $u_i(a,\omega) = \mathbb{E}_{\sigma_\omega}[u_i(a,\omega')|\omega]$ for all i, by (4) we have

$$\begin{aligned} v_q^*(\omega) &= \max_{a\in A} \sum_{\omega'\in\Omega} [u_0(a,\omega') + t_q^*(a,\omega')]\sigma_\omega(\omega') \\ &\leq \sum_{\omega'\in\Omega} \max_{a\in A} [u_0(a,\omega') + t_q^*(a,\omega')]\sigma_\omega(\omega') = \sum_{\omega'\in\Omega} v_q^*(\omega')\sigma_\omega(\omega'). \end{aligned} \quad \text{(A.7)}$$

Thus, if refining $\alpha q(\omega)$ of the original records of type ω according to σ_ω does not change the value of any record, then (A.7) implies the desired inequality. Now consider the other case: There exists a share $\alpha > 0$ such that refining $\alpha q(\omega)$ of the current records of type ω according to σ_ω leads to a new database q_α such that $v_{q_\alpha}^*(\omega') \neq v_q^*(\omega')$ for some $\omega' \in \operatorname{supp}\sigma_\omega$ or $\omega' = \omega$. Since the total quantity of records does not change, we have that $\mu_{q_\alpha}(\omega) < \mu_q(\omega)$ and $\mu_{q_\alpha}(\omega') > \mu_q(\omega')$ for all $\omega' \in \operatorname{supp}\sigma_\omega$. By Proposition 3, it follows that $v_{q_\alpha}^*(\omega) \geq v_q^*(\omega)$ and $v_{q_\alpha}^*(\omega') \leq v_q^*(\omega')$ for all $\omega' \in \operatorname{supp}\sigma_\omega$ and that the indirect effects are increasing in α. Now, note that for all α,

$$\sum_{\omega'\in\Omega} v_{q_\alpha}^*(\omega')\sigma_\omega(\omega') \geq v_{q_\alpha}^*(\omega) \geq v_q^*(\omega), \quad \text{(A.8)}$$

where the first inequality follows from (A.7). This implies that the direct effect of a refinement is always non-negative and decreasing in α.

A.11 Proof of Proposition 6

The directional derivative of U^* at any $\hat{q}$ in the direction σ_ω is equal to

$$\sum_{\omega'\in\Omega} v_{\hat{q}}^*(\omega')\sigma_\omega(\omega') - v_{\hat{q}}^*(\omega).$$

The linear path from q to q_α can be parametrized as follows: for $t \in [0,1]$, define $q_t(\omega) = q(\omega) - t\alpha q(\omega)$, $q_t(\omega') = q(\omega') + t\alpha\sigma_\omega(\omega')q(\omega)$ for $\omega' \in \operatorname{supp}\sigma_\omega$, and $q_t(\omega'') = q(\omega'')$ for remaining ω''. Note that $\sum_{\omega'\in\Omega} v^*_{q_t}(\omega')\sigma_\omega(\omega') - v^*_{q_t}(\omega)$ is non-negative by (A.7) and decreasing in t by the scarcity principle. Finally, by the gradient theorem,

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

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

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

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

which, combined with (A.8), implies that $\sum_{\omega''\in\Omega} v^*_{q_\alpha}(\omega'')\sigma_\omega(\omega'') = v^*_{q_\alpha}(\omega)$ 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 $\hat{a} \in \operatorname{supp} x^*_q(\cdot|\omega)$ there exists $\omega' \in \operatorname{supp}\sigma_\omega$ that satisfies $x^*_q(\hat{a}|\omega') = 0$. If the solution to the data-value problem is unique for database q—which is the case generically—then $x^*_q(\hat{a}|\omega') = 0$ implies $v^*_q(\omega') > u_0(\hat{a},\omega') + t^*_q(\hat{a},\omega')$ by strict complementary slackness. This and Proposition 4 imply that there exists $t' > 0$ such that $\sum_{\omega'\in\Omega} v^*_{q_t}(\omega')\sigma_\omega(\omega') > v^*_{q_t}(\omega)$ for all $t \in [0,1]$. It follows that $U^*(q_\alpha) > U^*(q)$.

B Interpreting the Data-Value Problem

To further understand the value of data records and the externalities between them, we provide a stand-alone interpretation of the data-value problem $\mathcal{V}_q$. With minor adjustments, this extends to the problems described in Section 5.3. We fix $q \in \mathbb{R}^\Omega_{++}$ and so drop it from notation.

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 $(a'_i, a_i) \in A_i \times A_i$ 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

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

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

$$\begin{aligned} \mathcal{V}: \quad & \min_{v,b,\ell} \sum_{\omega \in \Omega} v(\omega) q(\omega) \\ & \text{s.t. for all } \omega \in \Omega, \\ & v(\omega) = \max_{a \in A} \Big\{ u_0(a,\omega) + t(a,\omega) \Big\}, \end{aligned} \tag{B.1}$$

B.1 Gambles Against the Agents

Our interpretation hinges on unpacking how the platform determines the sellers' contributions to the externalities between records. By (B.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

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

which contributes to lowering $t_i(a,\omega)$ if and only if $\ell_i(a_i'|a_i) > 0$ and $u_i(a_i, a_{-i}, \omega) < u_i(a_i', a_{-i}, \omega)$. That is, if seller i knew ω and his opponents' offers a_{-i}, he would strictly prefer a_i' to a_i. In this case, offering a_i amounts to making a mistake from an ex-post viewpoint. We will 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_0 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 partial disclosure, but from the perspective of the data-value problem. In the end, $v(\omega)$ results from a trade-off between $u_0(a,\omega)$ and the return from inducing actions the sellers regret.

This return depends on the structure of b and ℓ, which define 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 $u_i(a_i, a_{-i}, \omega) - u_i(a_i', a_{-i}, \omega)$ for each a_i'; the scaling term $b_i(a_i)$ defines the stakes that it bets on this lottery. The platform "wins" when $u_i(a_i, a_{-i}, \omega) <$

$u_i(a_i', a_{-i}, \omega)$ 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 interaction's type ω and outcome a.[22]

We can then link how the platform chooses these gambles in $\mathcal{V}$ with the externalities between records. 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 make offers they will regret. Conversely, positive externalities $t^*(\omega) > 0$ correspond to unfavorable gambles. Corollary 1 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.

B.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 (B.1) for (a_i, a_{-i}, ω) and (a_i, a'_{-i}, ω'). In particular, if $u_i(a_i, a_{-i}, \omega) < u_i(\hat{a}_i, a_{-i}, \omega)$ but $u_i(a_i, a'_{-i}, \omega') > u_i(\hat{a}_i, a'_{-i}, \omega')$, the platform faces a trade-off because it 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

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

infeasible in the following sense. By complementary slackness $x^*(a|\omega) > 0$ implies $v^*(\omega) = u_0(a,\omega) + t^*(a,\omega)$. 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 B.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 B.1. This proof is for the general case where the principal can choose $a_0 \in A_0$ and each agent i can privately observe some own data $\omega_i \in \Omega_i$ about the interaction he is in. Fix $(a^*,\omega^*) \in \mathbf{X}$ and introduce $\mathbf{1}_{a^*,\omega^*}$ as a vector of size $|\mathbf{X}|$ with $\varepsilon > 0$ in the position indexed by (a^*,ω^*) and 0 in all other positions. Constitute a matrix $\mathbf{W}$ such that its rows are indexed by $(a,\omega) \in \mathbf{X}$, its columns are indexed by (i,a_i',a_i,ω_i), $i \in I$, and its entries are as follows:

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

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

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

C A Sufficient Condition for Optimality of Withholding Information

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

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

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

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

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

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

Then 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-information payoff $\bar{u}(\omega)$, then it is clearly optimal for the principal to fully reveal every ω. Given an outcome (a,ω) with $u_0(a,\omega) > \bar{u}(\omega)$, there must be an agent who would have a profitable deviation from a_i to $\hat{a}_i$ if he knew (a_{-i},ω_{-i}). Otherwise, given a_0, the profile a_{-0} is a Nash Equilibrium of Γ_ω and hence $a_{-0} \in CE(\Gamma_\omega)$, which would imply $u_0(a,\omega) \leq \bar{u}(\omega)$. Then condition (2) requires that agent i's data ω_i is consistent with another record ω'—so that he cannot tell ω and ω' apart based on his own data only—which admits a principal-preferred correlated equilibrium that also recommends i to play a_i and renders the deviation to $\hat{a}_i$ strictly suboptimal. 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 $u^{**}_q(\tilde{\omega}) = \bar{u}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$. Then $v^{**}_q(\tilde{\omega}) = u^{**}_q(\tilde{\omega}) = \bar{u}(\tilde{\omega})$ for all $\tilde{\omega} \in \Omega$ by Proposition 5.

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

$$v^{**}_q(\tilde{\omega}) \geq u_0(a, \tilde{\omega}) + t^{**}_q(a, \tilde{\omega}).$$

Since $u_0(a, \omega) > \bar{u}(\omega) = v^{**}_q(\omega)$, we must have $t^{**}_q(a, \omega) < 0$. Therefore, there exists a pair $(i, \hat{a}_i)$ that satisfies $u_i(a_i, a_{-i}, \omega) < u_i(\hat{a}_i, a_{-i}, \omega)$ and $\lambda^{**}_{q,i}(\hat{a}_i|a_i, \omega_i) > 0$. For such a pair $(i, \hat{a}_i)$, there exists $x(\cdot|\omega') \in CE(\Gamma_{\omega'})$ with the properties listed in the proposition. Then, since $\lambda^{**}_{q,i}(\hat{a}_i|a_i, \omega_i) > 0$,

$$\begin{aligned}
& \sum_{\tilde{a} \in A} u_0(\tilde{a}, \omega') x(\tilde{a}|\omega') + \sum_{\tilde{a} \in A} t^{**}_q(\tilde{a}, \omega') x(\tilde{a}|\omega') \\
\geq & \sum_{\tilde{a} \in A} u_0(\tilde{a}, \omega') x(\tilde{a}|\omega') \\
& + \lambda^{**}_{q,i}(\hat{a}_i|a_i, \omega_i) \left\{ \sum_{\tilde{a}_{-i} \in A_{-i}} \left(u_i(a_i, \tilde{a}_{-i}, \omega') - u_i(\hat{a}_i, \tilde{a}_{-i}, \omega') \right) x(a_i, \tilde{a}_{-i}|\omega') \right\} \\
> & \sum_{\tilde{a} \in A} u_0(\tilde{a}, \omega') x(\tilde{a}|\omega') = v^{**}_q(\omega'),
\end{aligned}$$

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

D Data and Price Discrimination: Analysis

This section provides the calculations for Section 4.2.1 and the example in Section 5.1. Recall that $a \in \{1, 2\}$ and that $u_0(a, \omega) = \max\{\omega - a, 0\}$ and $u_1(a, \omega) = a\mathbb{I}\{\omega \geq a\}$ for $\omega \in \{\omega_1, \omega_2\}$. For ω°, we have $u_i(a, \omega^\circ) = hu_i(a, \omega_2) + (1 - h)u_i(a, \omega_1)$ for $i = 0, 1$. For completeness, we solve both the information-design problem and the data-value problem.

Information Design. The objective function is

$$(\omega_2 - \omega_1)x(1|\omega_2)q(\omega_2) + h(\omega_2 - \omega_1)x(1|\omega^\circ)q(\omega^\circ) = x(1|\omega_2)q(\omega_2) + hx(1|\omega^\circ)q(\omega^\circ).$$

The obedience constraints are

$$-x(2|\omega_1)q(\omega_1) + x(2|\omega_2)q(\omega_2) + (2h - 1)x(2|\omega^\circ)q(\omega^\circ) \geq 0,$$

$$x(1|\omega_1)q(\omega_1) - x(1|\omega_2)q(\omega_2) - (2h-1)x(1|\omega^\circ)q(\omega^\circ) \geq 0.$$

Consider first the case of $h > \frac{1}{2}$. From the second constraint we get $x_q^*(1|\omega_1) = 1$. 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 $x_q^*(1|\omega^\circ) = \frac{1}{2h-1}\frac{q(\omega_1)}{q(\omega^\circ)}$, as long as $\frac{1}{2h-1}\frac{q(\omega_1)}{q(\omega^\circ)} \leq 1$.

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

$$x(1|\omega_1)q(\omega_1) - x(1|\omega_2)q(\omega_2) - (2h-1)x(1|\omega^\circ)q(\omega^\circ) \geq$$

$$\max\left\{2q(\omega_1) + (1-h)2q(\omega^\circ) - 1, 0\right\}.$$

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

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

Data Value. The data-value problem is

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

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

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

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

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

There are three candidates for optimal $\lambda(2|1)$. When $\lambda(2|1) = 0$, the objective is $S_0 \triangleq q(\omega_2) + hq(\omega^\circ)$. When $\lambda(2|1) = 1$, the objective is $S_1 \triangleq q(\omega_1) + q(\omega^\circ)(1-h)$. When $\lambda(2|1) = \frac{h}{2h-1}$, the objective is $S_f \triangleq q(\omega_1)\frac{h}{2h-1}$. The following claims are true. First, $S_0 \leq$

S_1 if and only if $q(\omega_1) \geq q(\omega_2) + (2h-1)q(\omega^\circ)$. Second, $S_0 \leq S_f$ if and only if $q(\omega_1) \geq q(\omega_2)\frac{2h-1}{h} + (2h-1)q(\omega^\circ)$. Third, $S_1 \leq S_f$ if and only if $q(\omega_1) \geq (2h-1)q(\omega^\circ)$.

Suppose now that $h \leq \frac{\omega_1}{\omega_2}$. Then $v(\omega^\circ) = h - (2h-1)\lambda(2|1)$ 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: $q(\omega_1) \leq (2h-1)q(\omega^\circ)$. Note that this requires $h > \frac{1}{2}$. Table 3 presents the optimal x_q^*.

$x_q^*(a\|\omega)$		ω		
		ω_1	ω_2	ω°
a	1	1	0	$\frac{1}{2h-1}\frac{q(\omega_1)}{q(\omega^\circ)}$
	2	0	1	$1 - \frac{1}{2h-1}\frac{q(\omega_1)}{q(\omega^\circ)}$

Table 3: 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 $v_q^*(\omega^\circ) = 0$.

Scenario 2: $(2h-1)q(\omega^\circ) \leq q(\omega_1) \leq q(\omega_2) + (2h-1)q(\omega^\circ)$. Note that the lower bound on $q(\omega_1)$ is meaningful only if $h > \frac{1}{2}$. Table 4 presents the optimal x_q^*.

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

Table 4: 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 $v_q^*(\omega^\circ) = 1 - h$.

Scenario 3: $q(\omega_1) \geq q(\omega_2) + (2h-1)q(\omega^\circ)$. Table 5 presents the optimal x_q^*.

$x_q^*(a\|\omega)$		ω		
		ω_1	ω_2	ω°
a	1	1	1	1
	2	0	0	0

Table 5: 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 $v_q^*(\omega^\circ) = h$.