Competitive Markets for Personal Data

Simone Galperti — UC, San Diego
Tianhao Liu — Columbia University
Jacopo Perego — Columbia University

March 29, 2024

Abstract

We study a competitive market where consumers own their data and can sell it to a platform. The platform pays each consumer a price for her data and—using this data as an input—intermediates her with a third-party merchant, from whom she can buy a product. We characterize the competitive equilibria of the data market and their ability to induce efficient data allocations. Our main results identify a novel inefficiency that leads this otherwise perfectly competitive market to fail. This inefficiency critically depends on the platform's role as an information intermediary. We provide three solutions to this market failure: a data union, data taxes, and more sophisticated data pricing policies.

JEL Classification Numbers: C72, D82, D83

Keywords: Consumer Data, Competitive Markets, Data Externalities, Data Union, Data Taxes.

We are thankful to Dirk Bergemann, Alessandro Bonatti, Laura Doval, Nima Haghpanah, Kevin He, Dana Foarta, Matthew Gentzkow, Alessandro Lizzeri, Alessandro Pavan, Paolo Siconolfi, Philipp Strack, Alexis Toda, Laura Veldkamp, Jidong Zhou, and seminar participants at numerous universities for helpful comments. This research is supported by grants from the NSF (Galperti: SES-2149289; Perego: SES-2149315). Galperti also gratefully acknowledges financial support from the UPenn's CTIC and the Warren Center for Network & Data Sciences.

1 Introduction

Consumer data has become a crucial productive input of the modern economy. It contributes to the success of many large industries, such as online advertisement and digital marketplaces. In these industries, firms use consumers' data to learn their tastes and offer them targeted advertisements or personalized products and services. While it is consumers themselves who are the primary suppliers of this data in the economy, they typically have limited control over how and by whom their data is used, and only rarely are they financially compensated in return (Federal Trade Commission, 2014). Such an arrangement could harbor inefficiencies and increase inequality (Bergemann et al., 2023). To combat these distortions, new legislation has been recently introduced across the world to give consumers more control over how their data is used.[1] This legislation creates the legal framework upon which *data markets* can emerge, where consumers have ownership over their personal data and firms compete to acquire and use it. What properties would such markets have? And which institutions should be designed to ensure they promote desirable outcomes?

This paper contributes to the ongoing discussion around these questions by studying a stylized model of a competitive data market. We present two main sets of results. First, we identify a novel inefficiency that can prevent the data market from inducing allocations that maximize welfare. This inefficiency stems from an externality that consumers exert on each other when selling their data, which is enabled by the platform's role as an information intermediary. Second, we discuss three potential solutions to this market failure. These involve, respectively, the establishment of a data union, the implementation of a data tax, or the creation of markets where the price of data can depend not only on its type but also on its intended use.

More specifically, our model features three sets of interacting agents: a heterogeneous population of consumers, an e-commerce platform, and a third-party merchant. Each consumer owns her data and can sell it to the platform. When this happens, the platform learns her type, pays her the market price for her data and, in addition, offers her a service. The service provided by the platform consists of intermediating this consumer with the merchant, from whom she can buy a product. As an intermediary, the platform can use the database it has acquired

[1] Most notably, the European Union's General Data Protection Regulation (GDPR) grants consumers the right to object to how firms use their data, to request it to be transferred to other firms, or to be deleted. In the United States, a growing list of States have passed bills with a similar scope.

from consumers to provide information to the merchant about these consumers' willingness to pay for the merchant's product, in the spirit of Bergemann et al. (2015). The main conceptual innovation of our model is that the platform's database is determined endogenously, as an equilibrium of the data market where the consumers and the platform interact. In particular, we assume that such a data market is perfectly competitive: That is, data prices are taken as given by the consumers and the platform and are pinned down by market clearing. This assumption is meant to shut down known distortions that may emerge from a platform's market power and focus, instead, on novel distortions that can persist even in a competitive economy.

Our main goal is to study the equilibrium properties of this competitive data market and, in particular, its ability to promote efficient data allocations. To this purpose, we identify necessary and sufficient conditions for efficiency and show how they depend on the platform's objective. In particular, we find that, when the platform's objective is sufficiently aligned with that of the merchant, the data market is efficient and consumers' welfare is maximized. In contrast—and perhaps counterintuitively—when the platform's objective is sufficiently aligned with that of the consumers, the equilibrium data allocation may be inefficient. In some cases, the data market can entirely unravel, resulting in no data being traded, which leads to the lowest possible welfare in the economy. We reinforce these negative findings by identifying sufficient conditions under which *all* equilibria of the economy are inefficient.

The cause of the inefficiency in this economy is that consumers exert an externality on each other when selling their data to the platform. This externality arises endogenously from the way the platform uses this data, which in turn depends on its objective. Specifically, when the platform cares relatively more about creating surplus for the consumers, it finds it optimal to withhold some information from merchants, to prevent excessive surplus extraction. To do so, the platform pools together consumers of different types, making it impossible for the merchant to fully ascertain the willingness to pay of each consumer in the pool. We show that this way of using the data also introduces a wedge between the individual benefit that a consumer enjoys when selling her data and joining such a pool, and the collective benefit that her presence in the pool creates for other consumers. Despite their competitive nature, the equilibrium data prices fail to fully account for this wedge and, thus, lead consumers to make decisions that are socially inefficient.[2]

[2]The idea of "pooling externality" was first introduced by Galperti, Levkun, and Perego (2023). Notice that it is distinct from and complementary to externalities that originate from exogenous correlation among consumers'

We then analyze three different institutions that can correct the aforementioned inefficiency. First, we introduce a new intermediary in the economy, called *data union*. A data union represents consumers by managing their data on their behalf. More specifically, it collects data from participating consumers, sells some of them to the platform, and distributes the proceeds of the sale back to the consumers, as compensation. We show that the data union helps consumers coordinate their decisions to sell the data, thus internalizing the externality described above. As a consequence, any equilibrium of the economy with the data union is efficient and consumers' welfare is maximized, regardless of the platform's objective. This result offers theoretical support to recent policy proposals that discuss the potential role that a data union could play in the data economy (e.g., see Posner and Weyl, 2018; Bergemann et al., 2023).

The second solution we consider consists of taxing the trade of data. Specifically, we introduce a *data tax*, which is levied on consumers who sell their data to the platform. This tax is "simple" in that it only depends on the type of data that is traded, and not on the identity of the consumer, or on how the data is used by the platform. When properly designed, such a data tax forces each consumer to internalize the effects that selling her data creates on the rest of the economy, above and beyond what can be done by equilibrium data prices. As a consequence, we show that any efficient allocation can be implemented by an equilibrium of the competitive economy with a budget-balanced data tax.

The third and final solution consists of letting the data price depend not only on the type of data that is traded, but also on its intended use by the platform. This solution is inspired by classic models of competitive economies with externalities such as Arrow (1969) and Laffont (1976). Additionally, it is broadly in the spirit of legislation like the aforementioned GDPR, which requires that the specific purpose for which consumer data is collected should be determined at the time of its collection (see, GDPR 2016/679 (39)). We show that this richer price system—or equivalently, the existence of different markets where to trade the same type of data depending on its intended use—guarantees the efficiency of the equilibria of the economy.

Related Literature. Our approach is rooted in a general-equilibrium tradition but leverages the recent progress of the information-design literature (for reviews of this literature, see, Bergemann and Morris, 2019; Kamenica, 2019). This allows us to offer a principled microfoundation of some of the key components of a data economy: How the data is *used* by the

data, which are analyzed by Choi et al. (2019), Acemoglu et al. (2022), and Bergemann et al. (2022a).

platform can be traced back to Bergemann et al. (2015); How the data is *valued* by the platform builds on Galperti et al. (2023); How the data is *priced* by the competitive market—a component that is novel to this paper—directly builds on this literature.

Our paper contributes to a recent literature that studies data markets and their properties. Particularly close to our work are a set of papers that identify data externalities and conditions under which they can lead to inefficient outcomes. As in Choi et al. (2019), Ichihashi (2021), Acemoglu et al. (2022), and Bergemann et al. (2022a), a consumer's decision of selling her data can create externalities on other consumers. Relative to these papers, we emphasize a novel market failure: It does not arise from exogenous correlation in consumers' data but, rather, from how the platform endogenously uses it. Indeed, to better emphasize the different nature of our inefficiency, we assume throughout that consumers' data is uncorrelated: That is, the platform learns nothing about a consumer when acquiring the data of another.

Our inefficiency builds instead on previous work by Galperti, Levkun, and Perego (2023). That paper characterizes how much the platform values the data of a single consumer in a larger database. It shows that the value of such a data record is the sum of two components: The direct payoff the platform earns from the underlying consumer and the indirect payoff her data record helps generate for the platform when using other data records. They refer to this second component as a "pooling" externalities, as it is non-zero only when the platform finds it optimal to pool data records together. Our paper pushes this agenda several steps forward. First, we model a competitive data market where consumers have ownership of their data. Second, we characterize the externalities that consumers create on each other when selling their data, rather than those that exist at the level of the platform's value for data. Third, we identify conditions under which the data market fails and we propose solutions that remedy such a failure.

More broadly, our paper contributes to a growing literature that studies the role of data in the modern economy (for reviews, see Acquisti et al., 2016; Bergemann and Bonatti, 2019; Bergemann and Ottaviani, 2021; Goldfarb and Tucker, 2023). Our model is stylized and purposefully abstracts from some other important aspects of a data economy, such as the rich dynamic interactions between the various constituencies of this economy (Chen, 2022), or the distortions that can emerge due to the non-rivalrous nature of consumer data (see, e.g., Varian, 2009; Jones and Tonetti, 2020; Farboodi et al., 2019), or those that emerge due to the repeated nature of online interactions (see, e.g., Taylor, 2004; Acquisti and Varian, 2005; Calzolari and Pavan, 2006; Doval and Skreta, 2023; Dworczak, 2020).

2 The Model

We present a stylized model of a data economy. It features a platform (*it*), a merchant (*he*), and a unit mass of consumers (*she*). The consumers can sell their personal data to the platform. The platform uses this data to provide information to the merchant about the consumers' preferences. Finally, the merchant charges a fee to each consumer in exchange for the product he produces. A discussion of the main modeling assumption appears in 2.1.

Formally, each consumer has a unit demand for the product sold by the merchant. We denote her willingness to pay by $\omega \in \Omega \subset \mathbb{R}_{++}$. Let $\bar{q} \in \Delta(\Omega)$ be the distribution of ω in the population and assume Ω is finite with $|\Omega| \geq 2$. Each consumer owns a *data record* that fully reveals her corresponding ω.[3]

The model has two periods. In the first period, the data markets are open. The platform and the consumers trade the data records at prices $p = (p(\omega))_{\omega\in\Omega} \in \mathbb{R}^{\Omega}$, which they take as given. On the demand side of these markets, the platform chooses how many records of each type to demand. Let $q = (q(\omega))_{\omega\in\Omega} \in \mathbb{R}^{\Omega}_{+}$ denote the composition of the *database* demanded by the platform, for which it pays a total of $\sum_{\omega\in\Omega} q(\omega)p(\omega)$. On the supply side, each consumer chooses whether to sell her record to the platform. If a type-ω consumer sells her record, the platform pays her the price $p(\omega)$ and later intermediates her with the merchant, as described below. Without loss of generality, we assume that consumers of the same type sell their records with the same probability, denoted by $z(\omega) \in [0,1]$. Conversely, if a type-ω consumer does not sell her record to the platform, she forgoes the opportunity to interact with the merchant and obtains a reservation utility of $r(\omega) \geq 0$.

In the second period, the product market is open. The platform uses the acquired database q—whose composition is publicly known—to mediate the interaction between the merchant and the subset of consumers who have sold their records. In particular, the platform acts as an information intermediary: It provides the merchant with information about the consumers in the

[3] As in Galperti et al. (2023), we interpret the data record as a list of identifiers (e.g., IP address, telephone number, etc.) and personal characteristics (e.g., gender, age, etc.). The former grants access to the consumer and, thus, the ability to intermediate her. The latter, instead, provides information about her type.

database. Formally, the platform solves a standard information-design problem where the relative frequency of consumers' types is given by q. The platform commits to an information structure that maps the record of each consumer in its database into random signals. Given the signal received, the merchant sets a fee $a \in A$ for the consumer, who then purchases the product if and only if the merchant's fee a is lower than her willingness to pay ω. Therefore, given ω and a, the consumer's and the merchant's second-period payoffs can be written as $u(a,\omega) = \max\{\omega - a, 0\}$ and $\pi(a,\omega) = a\mathbb{1}(\omega \geq a)$, respectively. Finally, the platform's payoff is a linear combination of the consumer's trading surplus and the merchant's profits, i.e., $v(a,\omega) = \gamma_u u(a,\omega) + \gamma_\pi \pi(a,\omega)$. We assume $\gamma_u, \gamma_\pi \geq 0$, with at least one strict inequality.

By standard results from the information-design literature (e.g., see Bergemann and Morris, 2016), the platform's problem in the second period can be formulated as choosing a recommendation mechanism $x : \Omega \to \Delta(A)$ that solves:

$$\begin{aligned} V(q) = \max_{x:\Omega\to\Delta(A)} \quad & \sum_{a,\omega} v(a,\omega)x(a|\omega)q(\omega) \\ \text{such that} \quad & \sum_{\omega} \big(\pi(a,\omega) - \pi(a',\omega)\big)x(a|\omega)q(\omega) \geq 0 \qquad \forall\, a, a' \in A. \end{aligned} \tag{$\mathcal{P}_q$}$$

Without loss of generality, we let $A = \Omega$.

To summarize, we have introduced four endogenous variables: prices p for the data records; the consumers' decisions z to supply their records; the platform's demanded database q; and the platform's mechanism x for problem $\mathcal{P}_q$. We define an equilibrium of this economy as follows.

Definition 1. *A profile (p^*, z^*, q^*, x^*) is an equilibrium of the competitive economy if*

(a). Given p^*, q^* solves the platform's problem in the first period, i.e.,

$$q^* \in \arg\max_{q\in\mathbb{R}^\Omega_+} V(q) - \sum p^*(\omega)q(\omega). \tag{1}$$

(b). Given q^*, x^* solves the platform's problem $\mathcal{P}_{q^*}$ in the second period.

(c). Given x^* and p^*, z^* solves the consumers' problem in the first period. That is, for all ω,

$$z^*(\omega) \in \arg\max_{\zeta\in[0,1]} \zeta\Big(p^*(\omega) + \sum_a x^*(a|\omega)u(a,\omega)\Big) + (1-\zeta)r(\omega).$$

(d). Data markets clear. That is, for all ω, $q^*(\omega) = z^*(\omega)\bar{q}(\omega)$.

Conditions (a) and (b) require that the platform acquires a database that maximizes its payoff taking prices as given, while anticipating it will use its data optimally in the second period. Condition (c) requires that each type-ω consumer choose $z(\omega)$ optimally, again taking prices as given and anticipating that the platform will acquire a database q^* and use it to implement mechanism x^*. Therefore, she sells her record at price $p(\omega)$ only if $p(\omega) + \sum_a u(a,\omega)x^*(a|\omega) \geq r(\omega)$, where $\sum_a u(a,\omega)x^*(a|\omega)$ captures her expected trading surplus. Finally, condition (d) requires that the demand of each type of record equals its supply. This last condition pins down data prices, in the spirit of a traditional competitive equilibrium. Proposition B.1 in the Online Appendix shows that an equilibrium exists.

Hereafter, we will refer to the two-period model just described as the "competitive economy." In the next sections, we will study the equilibria of this economy, their inefficiency, and possible remedies to it.

2.1 Discussion of Modeling Assumptions

Before proceeding, we briefly discuss our main modeling assumptions. Our economy features a single platform taking the prices of data records as given. This assumption has a substantive component and an expositional one. The substantive component is that the platform is a price taker and, thus, the data market is competitive. This is a distinguishing feature of this paper, and it allows us to shut down other potential sources of inefficiency—such has platform's market power—that have less to do with data as an input. The expositional component is that we focus on a single platform, rather than a finite number of identical ones. This substantially simplifies notation at little cost of generality. Galperti and Perego (2022) show how to model a competitive economy with multiple competing platforms.

The platform's objective is assumed to be linear in the consumers' trading surplus and the merchant's profits. This specification is especially tractable, while capturing key features of real-world two-sided markets (Xu and Yang (2023) offer a dynamic microfoundation of such an objective). The results of Section 4 do not depend on this assumption.

Three aspects of the consumer's problem have been simplified. First, the reservation utility $r(\omega)$ is exogenous. This assumption rules out settings where the consumer can bypass the platform and trade directly with the merchant. While not focusing on data markets, Bergemann and Bonatti (2023) study the interplay of on- and off-line interactions. Second, the consumer cannot participate in the platform's mechanism without revealing her type. That is, the data record bundles "access" to the consumer and information about her willingness to pay, which can be restrictive in some applications. With different goals than ours, Hidir and Vellodi (2021) and Ali et al. (2022) study models where these two aspects are unbundled. Third, selling the data record fully reveals the underlying consumer's type. Our analysis can be extended to records that are only partially informative of the consumer's type, a model of which is proposed by Galperti et al. (2023).

2.2 Efficiency Benchmark

How efficiently do the data markets allocate records between the consumers and the platform? To answer this, we need to introduce an efficiency benchmark. Let us refer to the pair (q, x) as an *allocation*. For each allocation, denote the sum of the payoff of the platform and the consumers as

$$\mathcal{W}(q, x) \triangleq \sum_{a,\omega} \Big(v(a, \omega) + u(a, \omega) \Big) x(a|\omega) q(\omega) + \sum_{\omega} \Big(\bar{q}(\omega) - q(\omega) \Big) r(\omega). \tag{2}$$

Note that, since data prices p only affect the distribution of payoffs between the platform and the consumers, this notion of welfare only depends on the allocation (q, x).

Definition 2. *An allocation (q°, x°) is* ***constrained efficient*** *if it solves*

$$\begin{aligned} W^\circ = \max_{q,x} \quad & \mathcal{W}(q, x) \\ \textit{such that} \quad & q \leq \bar{q}, \\ \textit{and} \quad & x \textit{ solves } \mathcal{P}_q. \end{aligned} \tag{$\mathcal{SB}$}$$

According to this efficiency benchmark, an allocation is constrained efficient if it maximizes the welfare function $\mathcal{W}(q, x)$ subject to two constraints. The first requires that the database q is feasible, i.e., it does not allocate to the platform more records than those that exist in the economy. The second constraint requires that the mechanism x is sequentially optimal for the platform given q.[4]

[4] A constrained efficient outcome always exists. This follows from the fact that $\mathcal{W}$ is continuous and, by Lemma B.1, the feasible set of outcomes in the planner's problem is nonempty and compact.

We briefly motivate the efficiency benchmark just introduced. Since the main goal of the paper is to demonstrate that equilibria of this economy can be inefficient, a less demanding efficiency benchmark is more desirable, as it makes such a negative result starker. Additionally, a less demanding benchmark allows us to ignore other potential sources of inefficiency that are rather standard and not special to data as an input. In particular, two aspects of Definition 2 make it less demanding. First, it focuses on "constrained" efficiency, i.e., it requires that the mechanism x be sequentially rational for the platform given q. Dropping this constraint, i.e., focusing on unconstrained efficiency, would lead us to detect an additional inefficiency that is merely driven by the fact that, in the first period, the platform cannot commit to a mechanism for the second period. Second, we exclude the merchant's payoff from the welfare function $\mathcal{W}$. Including the merchant's payoff would lead us to detect an additional inefficiency that is merely driven by the fact that the platform does not "sell" information to the merchant and, thus, cannot transfer the merchant's profit to the consumer. Thus, the consumer does not internalize the effects that selling her data create on the profit of the merchant.[5]

An additional useful feature of our welfare notion is that, in any equilibrium of the competitive economy, $\mathcal{W}(q^*, x^*)$ coincides with the consumers' welfare. This is because, in any equilibrium, the platform must earn a payoff of zero.[6] Therefore, any constrained efficient equilibrium also maximizes consumers' welfare.

3 The Inefficiency of the Data Market

In this section, we present a series of results that identify necessary and sufficient conditions for equilibrium efficiency and, by doing so, uncover what are the key drivers of inefficiency in our competitive economy.

We begin by introducing a notion that is instrumental for our analysis. What is the social cost—i.e., the decrease in the welfare function $\mathcal{W}$—that results from allocating an additional ω-record to the platform's database? Clearly, it is $r(\omega)$, i.e., the reservation utility that is lost

[5] Online Appendix D shows that our main qualitative results are unchanged when we consider "unconstrained" efficiency or a welfare function that includes the merchant profits.

[6] Indeed, notice that $V(q)$ is homogeneous of degree 1. If the platform earned a strictly positive payoff at q^*, it could profitably deviate by acquiring database $q' = \alpha q^*$, with $\alpha > 1$, which earns a payoff $V(q') - \sum_\omega p^*(\omega)q'(\omega) = \alpha\big(V(q^*) - \sum_\omega p^*(\omega)q^*(\omega)\big) > V(q^*) - \sum_\omega p^*(\omega)q^*(\omega)$.

by the corresponding consumer. What is the corresponding social benefit? To compute it, fix an arbitrary database q and consider the following maximization problem:

$$\begin{aligned} W(q) \triangleq \quad & \max_{x:\Omega\to\Delta(A)} \quad \sum_{a,\omega}(v(a,\omega)+u(a,\omega))x(a|\omega)q(\omega) \\ & \text{such that} \quad x \text{ solves } \mathcal{P}_q. \end{aligned}$$

We can think of the above as the problem of a planner who chooses a mechanism x to maximize the welfare of the platform and the consumers in database q. This planner is constrained to choose a mechanism that, given q, the platform would also be willing to implement. We denote by Ψ_q the set of supergradients of $W(q)$. In other words, just like a derivative, each $\psi_q(\omega)$ captures how $W(q)$ changes when we add an additional ω-record to database q. For this reason, $\psi_q(\omega)$ identifies the *social benefit* of allocating an additional ω-record into the platform's database.[7]

Our first result demonstrates how the social benefit and cost of data records can be used to characterize which allocations are constrained efficient.

Proposition 1. *An allocation (q,x) is constrained efficient if and only if x solves $\mathcal{P}_q$ and there exists a $\psi_q \in \Psi_q$ such that, for all ω,*

- $\psi_q(\omega) \geq r(\omega)$ *if* $q(\omega) > 0$,
- $\psi_q(\omega) \leq r(\omega)$ *if* $q(\omega) < \bar{q}(\omega)$.

It is clear that under any constrained efficient allocation, a record is allocated to the platform's database if and only if this record's social benefit exceeds its social cost. Perhaps less intuitively, these conditions are also sufficient. This will be key to characterize equilibrium efficiency in terms of the model primitives. To see why, fix an equilibrium (p^*,z^*,q^*,x^*) and denote by

$$U^*(\omega) \triangleq p^*(\omega) + \sum_a x^*(a|\omega)u(a,\omega) \tag{3}$$

the *private benefit* that a type-ω consumer obtains when selling her record to the platform. Notice that the equilibrium conditions require that $U^*(\omega) \geq r(\omega)$ if $q^*(\omega) > 0$, and that

[7] In Appendix A.1 we show $W(q)$ is concave in q and, therefore, Ψ_q is well-defined. Moreover, Lemma A.1 provides an analytical characterization of Ψ_q, which shows that Ψ_q is generically a singleton and it easy to compute. Since $W(q)$ can be written as the value of a linear program, its supergradients can be characterized using the duality approach developed by Galperti et al. (2023), as shown in the appendix.

$U^*(\omega) \leq r(\omega)$ if $q^*(\omega) < \bar{q}(\omega)$. Therefore, in light of Proposition 1, this equilibrium is constrained efficient if and only if the private and social benefits of data records are sufficiently "aligned." That is, if there is a $\psi_{q^*} \in \Psi_{q^*}$ such that $\psi_{q^*}(\omega) \geq r(\omega)$ if $U^*(\omega) \geq r(\omega)$ and, conversely, $\psi_{q^*}(\omega) \leq r(\omega)$ if $U^*(\omega) \leq r(\omega)$.

The key question is then, under what conditions on the model primitives, U^* and ψ_{q^*} are aligned. To address this question, it is useful to define $\sigma^*(\omega) \triangleq \psi_{q^*}(\omega) - p^*(\omega)$ and write

$$\psi_{q^*}(\omega) = p^*(\omega) + \sigma^*(\omega). \tag{4}$$

This decomposition of the social benefit—which is analogous to the definition of U^* in equation (3)—has a particularly useful economic interpretation: $p^*(\omega)$ and $\sigma^*(\omega)$ capture the marginal change in the platform's payoff and the consumers' trading surplus, respectively, that result from adding an ω-record to q^*. The following result formalizes this interpretation.

Lemma 1. *In any equilibrium (p^*, z^*, q^*, x^*), p^* is a supergradient of $V(q^*)$.*[8]

This result demonstrates that, due to the competitive nature of the data markets, the marginal change in the platform's payoff from acquiring an additional ω-records—formally, the supergradient of $V(q^*)$—must equal its cost $p^*(\omega)$. Intuitively, if that was not the case, the platform would strictly prefer to change her demand of ω-records. As a consequence of Lemma 1, the remaining component of the social benefit $\psi_{q^*}(\omega)$—namely $\sigma^*(\omega)$—must capture the marginal change in the trading surplus of all consumers, which results from adding an ω-record to the database q^*.

Comparing Equations (3) and (4) reveals that a necessary and sufficient condition for equilibrium efficiency is the alignment between the trading surplus that a type-ω consumer expects to receive when selling her record—namely, $\sum_a x^*(a, \omega) u(a, \omega)$—and the effect that this sale has on the trading surplus of *all* consumers—namely, $\sigma^*(\omega)$. Since the consumer internalizes the former but not the latter, her decision to sell her record may exert an externality on other consumers, thus introducing inefficiency in the economy.

The following result shows how the presence of these externalities, and thus the efficiency of the economy, depends on the model primitives and, in particular, hinges on the objective of the platform.

[8] We show $V(q)$ is concave in Appendix A.1 and provide analytical characterization of its supergradients in Lemma A.1.

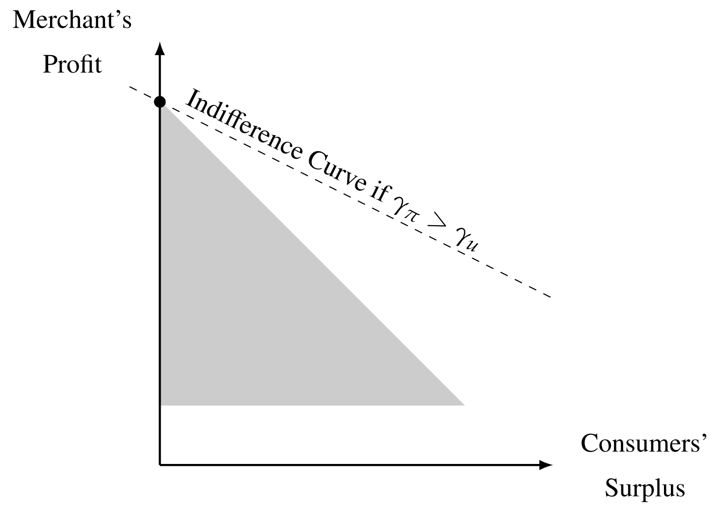


Figure 1: The trading surplus triangle. The dashed line depicts the platform's indifference curve when $\gamma_\pi > \gamma_u$.

Proposition 2. *If $\gamma_\pi > \gamma_u$, all equilibrium allocations of the competitive economy are constrained efficient and, therefore, maximize consumers' welfare. Conversely, if $\gamma_\pi \leq \gamma_u$, equilibrium allocations can be inefficient.*

Perhaps counterintuitively, if the platform cares relatively more about the merchant's profits, the social and private benefits of data records are aligned. Thus, equilibria are constrained efficient and consumers' welfare is maximal. Vice versa, if the platform cares relatively more about consumers' surplus, this alignment can break, leading to inefficiencies.

To gain intuition, suppose $\gamma_\pi > \gamma_u$ and consider an arbitrary database $q \neq 0$. In this case, the platform finds it optimal to reveal the ω of each consumer in the database to the merchant, allowing the merchant to extract all their surplus. To see this, notice that, by Bergemann et al. (2015, Theorem 1), the platform's problem $\mathcal{P}_q$ is equivalent to choosing a point in the triangle of Figure 1, which plots the set of pairs of merchant's profit and consumers' surplus that can be induced by any mechanism. Since the platform's payoff v is linear in $\pi(a, \omega)$ and $u(a, \omega)$, when $\gamma_\pi > \gamma_u$, the optimal mechanism maximizes the merchant's profits, leaving consumers with no surplus. As a consequence, $\sum_a x^*(a, \omega) u(a, \omega) = \sigma^*(\omega) = 0$. Therefore, consumers do not exert externalities on each other when selling their records, and all equilibria are constrained efficient.

Suppose instead $\gamma_\pi \leq \gamma_u$. In this case, for any q, the optimal mechanism x^* typically involves withholding some information from the merchant to prevent excessive surplus extraction (again, see Figure 1). To do so, the platform may pool a consumer's record with those of others. In this case, their payoffs become interdependent, as the presence of this record in the pool can affect the merchant's beliefs in a nontrivial way. Consequently, $\sigma^*(\omega)$ and $\sum_a x^*(a|\omega)u(a,\omega)$ can differ, leading to inefficiencies. The externality exerted by low-type consumers is typically positive. For example, consider a consumer of the lowest type, $\underline{\omega} \triangleq \min_\omega \Omega$. Her expected trading surplus is zero since the merchant would never charge a fee a lower than ω. Yet, $\sigma^*(\underline{\omega})$ can be strictly positive, since when this consumer is pooled with higher-type consumers, she helps them receive a lower fee and earn a positive surplus. Conversely, the externality exerted by high-type consumers is typically negative. For example, when the highest-type consumer sells her record, she may decrease the chances of other high-type consumers earning a positive surplus. Failure to internalize these externalities leads to inefficiencies.

This is the sense in which the market failure we emphasize is enabled by the role played by the platform as an information intermediary. It conceptually originates from the platform's incentive to withhold information to harness the conflicting objectives of the constituencies it intermediates. The latter is a distinguishing feature of multi-sided platforms, which explains why information withholding is so commonly observed in the marketplace.[9] Importantly, the market failure we emphasize does not conceptually originate from the absence of competition for the merchant and, instead, could arise even in a richer model with multiple competing merchants. For example, in a model where merchants compete in a second-price auction, Bergemann et al. (2022b) show that a platform has incentives to withhold information from them. This scenario would result in inefficiencies akin to those highlighted in this paper.

We conclude this discussion by finding sufficient conditions under which *all* equilibria are inefficient, thus sharpening the negative message of Proposition 2. To avoid trivial cases, let us focus on economies in which $W^\circ > R := \sum_\omega \bar{q}(\omega)r(\omega)$, that is, the constrained efficient allocation involves some trade.[10]

Corollary 1. *Let $\gamma_\pi \leq \gamma_u$ and suppose $W^\circ > R$. If $\gamma_u\underline{\omega} < r(\underline{\omega}) < (1+\gamma_u)\underline{\omega}$, then all*

[9] For instance, to avoid thin markets, Google Search withholds information from competing advertisers about the users' characteristics; Ridesharing platforms similarly withhold information from drivers about riders' destinations; etc.

[10] When $W^\circ = R$, all equilibria are constrained efficient. To see this notice that in any equilibrium (p^*, z^*, q^*, x^*), it must be that $R \leq \mathcal{W}(q^*, x^*) \leq W^\circ$. Therefore, if $W^\circ = R$, we have $\mathcal{W}(q^*, x^*) = W^\circ$.

equilibria are inefficient.

Corollary 1 gives a sufficient condition under which the positive externality discussed above causes all equilibria of the economy to be inefficient. First, we show that if $\gamma_u \underline{\omega} < r(\underline{\omega})$, the platform is unwilling to pay a price higher than $r(\underline{\omega})$ for $\underline{\omega}$-records. This implies that $U^*(\underline{\omega}) < r(\underline{\omega})$, since a consumer of type $\underline{\omega}$ necessarily earns a zero trading surplus when she sells her record. Thus, no such consumer sells her record. Second, we additionally show that, if $r(\underline{\omega}) < (1+\gamma_u)\underline{\omega}$, the social benefit of $\underline{\omega}$-records, $\psi_{q^*}(\underline{\omega})$, exceed its private cost, $r(\underline{\omega})$. As a result, a trade that would be socially beneficial does not occur in equilibrium, generating an inefficiency.

Under the sufficient condition of Corollary 1, the $\underline{\omega}$-type consumers exert a positive externality on other consumers, which they fail to internalize, thus leading to inefficiencies. Conversely, Corollary A.1 in Appendix A.1 provides alternative sufficient conditions under which the inefficiency in the economy is caused by a negative externality exerted by higher-type consumers. In general, both positive and negative externalities exist, as illustrated by the next example.

3.1 An Example

We now provide an example to illustrate why the data economy can be inefficient. Suppose there are two types of consumers, $\Omega = \{1, 2\}$, and $\bar{q}(2) > \bar{q}(1)$. All consumers have the same reservation utility: $r(\omega) = \bar{r} \in (0, 1)$ for all ω. The platform only cares about consumers' trading surplus: $\gamma_u > \gamma_\pi = 0$. To avoid uninteresting cases, we assume that $\bar{r} < \frac{1+\gamma_u}{2}$ so that some trade is required to achieve constrained efficiency.[11] We will show that all equilibria of this economy are inefficient.

We first characterize the constrained efficient allocations. Let (q°, x°) be such that $q^\circ(1) = q^\circ(2) = \bar{q}(1)$ and $x^\circ(1|\omega) = 1$ for all ω. In other words, the platform is given the records of all the low-type consumers and an equal amount of high-type ones. The platform then pools all these records in the same segment, inducing the merchant to charge the lowest fee (i.e., $a = 1$) to all consumers in the database. We argue that (q°, x°) is the unique constrained efficient allocation. To see why, consider any other database $q > 0$ and notice that an optimal mechanism x_q given q is to set $x_q(1|\omega) = \min\{q(1), q(2)\}/q(\omega)$ for all ω. That is, the

[11] When $\bar{r} \geq \frac{1+\gamma_u}{2}$, the absence of trade—i.e., $q_0 = (0, 0)$—is constrained efficient. That is, $W^\circ = R = \bar{r}$. In this case, any equilibrium allocation (q^*, x^*) is constrained efficient since $W^\circ \geq W(q^*, x^*) \geq \bar{r}$.

platform creates the largest possible segment with an equal quantity of low- and high-type consumers, who are then charged the lowest fee. Thus, the allocation (q, x_q) induces a welfare of $\mathcal{W}(q, x_q) = (1+\gamma_u)\min\{q(1), q(2)\} + (1 - q(1) - q(2))\bar{r}$. To maximize $\mathcal{W}(q, x_q)$, any constrained efficient allocation (q°, x°) must satisfy $q^\circ(1) = q^\circ(2)$. Since by assumption $\bar{q}(1) < \bar{q}(2)$ and $\bar{r} < \frac{1+\gamma_u}{2}$, setting $q^\circ(1) = q^\circ(2) = \bar{q}(1)$ uniquely maximizes $\mathcal{W}(q, x_q)$. For future reference, note that welfare under the constrained efficient allocation is $W^\circ = \bar{r} + \bar{q}(1)(1 + \gamma_u - 2\bar{r})$.

We now characterize all equilibria and show that they are inefficient. Let (p^*, z^*, q^*, x^*) be an equilibrium. Let a_{q^*} be the fee the merchant would charge if the platform did not provide him with any information besides the database composition. Using Lemma A.1 in Appendix A, we can compute the marginal change in the consumers' trading surplus that results from adding an ω-record to q^*, which is

$$\sigma^*(\omega) = \omega - a_{q^*}\mathbb{1}(\omega \geq a_{q^*}), \tag{5}$$

and the equilibrium price, which is

$$p^*(\omega) = \gamma_u \sigma^*(\omega) = \gamma_u\Big(\omega - a_{q^*}\mathbb{1}(\omega \geq a_{q^*})\Big). \tag{6}$$

Since $1 \leq a_{q^*} \leq 2$, we have $0 \leq \sigma^*(\omega) \leq 1$ and $0 \leq p^*(\omega) \leq \gamma_u$, for all ω.

Case 1, $\bar{r} > \gamma_u$: Inefficiently Low Trade. When $\bar{r} > \gamma_u$, the only equilibrium involves no trading. To show this, we first argue that any equilibrium involves no trade of 1-records: $q^*(1) = 0$. Indeed, type-1 consumers always earn a zero trading surplus when they sell their records. Moreover, by Equation (6), the price $p^*(1)$ can be no higher than γ_u. Therefore, their net payoff is no higher than γ_u, which is strictly smaller than $\bar{r}$ by assumption. Thus, type-1 consumers do not sell their data in any equilibrium: $q^*(1) = 0$. Next, we argue that this implies $q^*(2) = 0$. To see why, suppose $q^*(2) > 0$. Since $q^*(1) = 0$, we must have $a_{q^*} = 2$, and thus Equation (6) imply that $p^*(2) = 0$. Moreover, since $q^*(1) = 0$, type-2 consumers will be perfectly discriminated against and earn a zero trading surplus. Since type-2 consumers get a zero net payoff when they sell their data, they must be unwilling to do so, contradicting $q^*(2) > 0$. Therefore, $q^* = (0, 0)$ is the only database compatible with equilibrium. Under this complete market unraveling, any equilibrium allocation (q^*, x^*) must yield $\mathcal{W}(q^*, x^*) = \bar{r} < W^\circ$ and, thus, the inefficiency is as severe as it could be.

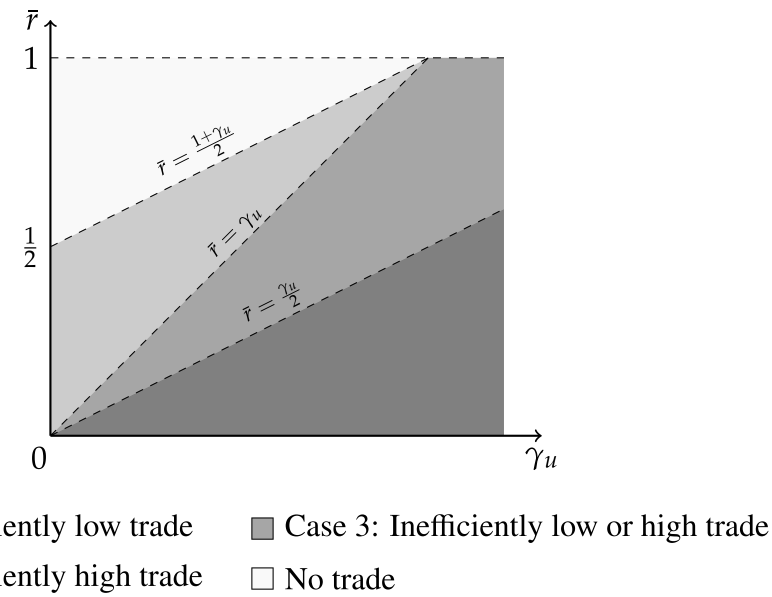
Why are these equilibria inefficient? By selling her record, a low-type consumer could create a positive externality: The platform would pool this consumer with a high-type one, thus creating a social benefit of $1 + \gamma_u$, which by assumption is larger than $2\bar{r}$, i.e., the sum of the reservation utilities of these two consumers. For this trade to happen, however, the price $p^*(1)$ has to exceed the reservation utility $\bar{r}$. Unfortunately, the platform is unwilling to pay such a price, since its value for a 1-record is, at most, $\gamma_u < \bar{r}$. $\triangle$

Case 2, $\bar{r} < \frac{\gamma_u}{2}$: Inefficiently High Trade. When $\bar{r} < \frac{\gamma_u}{2}$, the unique equilibrium involves $q^*(1) = \bar{q}(1)$ and $q^*(2) = \min\{\bar{q}(2), \frac{\bar{q}(1)}{\bar{r}}\}$. To see this, note first that by the no-profit condition the equilibrium prices must satisfy $p^*(1) + p^*(2) \geq \gamma_u$. If not, the platform could acquire a pair of low- and high-type records, pool them in the same segment, and generate a payoff of γ_u and hence a positive profit. Therefore, either $p^*(1)$ or $p^*(2)$ must exceed $\frac{\gamma_u}{2}$. We argue that $p^*(2) \leq \frac{\gamma_u}{2}$. Otherwise, all high-type consumers would strictly prefer to sell, leading to an uninformed merchant's price of $a_{q^*} = 2$, which by Equation (6) implies $p^*(2) = 0$, a contradiction. Since then $p^*(1) \geq \frac{\gamma_u}{2}$, the low-type consumers strictly prefer to sell and, thus, $q^*(1) = \bar{q}(1)$. As in the constrained efficient allocation, the optimal mechanism involves $x^*(1|2) = \min\{\bar{q}(1), q^*(2)\}/q^*(2)$: The platform creates a segment that includes all the low-type consumers and as many high-type ones as possible subject to inducing the fee $a = 1$. Since $\bar{r} < 1$, we must have $q^*(2) > q^*(1)$. Otherwise, the expected trading surplus of a high-type consumer selling her record would be 1 and, thus, all such consumers would sell their records, leading to a contradiction. Given such q^*, note that $a_{q^*} = 2$ and, by Equation (6), $p^*(2) = 0$. Thus, $q^*(2) = \min\{\bar{q}(2), \frac{\bar{q}(1)}{\bar{r}}\}$. We conclude that the unique equilibrium is inefficient because $\mathcal{W}(q^*, x^*) = (1 + \gamma_u)\bar{q}(1) + \max\{0, \bar{r} - \bar{q}(1)(1 + \bar{r})\} < W^\circ$.

In this equilibrium, inefficiently too many high-type consumers sell their records to the platform: $q^*(2) > q^\circ(2)$. They are attracted by the possibility of buying the merchant's product at the lowest fee. However, when an additional high-type consumer sells her record, she exerts a negative externality on other consumers, by undermining their chances of buying the product at the lowest fee. On the one hand, such a consumer gains individually from selling her record, since $\sum_a x^*(a|2)u(a, 2) = q^*(1)/q^*(2) \geq \bar{r}$. On the other hand, she does not help increase the aggregate trading surplus for all consumers. Indeed, $\sigma^*(2) = 0$ (by Equation (5)).[12] $\triangle$

[12] One may wonder whether a negative price for 2-records could correct this inefficiency, as it would disincentivize selling these records. Such a price is incompatible with equilibrium behavior, as the platform would then

Figure 2: Equilibrium Inefficiency in the Example of Section 3.1



We cover the case of $\frac{\gamma_u}{2} \leq \bar{r} \leq \gamma_u$ in Appendix C. It shares similar intuitions to the previous two cases and features multiple equilibria with either inefficiently low or inefficiently high trade. In summary, all equilibria of this simple economy are inefficient as represented in Figure 2.

4 Remedies to the Market Failure

This section discusses three market designs that provide a remedy to the inefficiency of the competitive economy. The first establishes a "data union" that manages consumers' data on their behalf. The second introduces data taxes in the competitive economy. The third renders data markets "more complete."

4.1 Data Unions

We first introduce a new intermediary in the economy, called a *data union*.[13] A data union manages consumers' data records on their behalf and with the objective of maximizing their

demand an infinite quantity of 2-records. More generally, Corollary A.1 in Appendix A provides sufficient conditions in an arbitrary economy for similar inefficiencies.

[13] Data unions have also been discussed in Posner and Weyl (2018) and Bergemann et al. (2023).

welfare: It collects records from participating consumers, sells some or all of them to the platform, and distributes the proceeds back to the consumers, as compensation. Thus, the union coordinates consumers' actions, by unilaterally deciding which records should be sold to the platform and how consumers should be compensated. We show that a data union implements allocations that are constrained efficient. This result offers theoretical support to recent policy proposals that discuss the potential role that a data union could play in the data economy (e.g., see Posner and Weyl, 2018; Bergemann et al., 2023).

More formally, our data union operates as follows. Consumers voluntarily decide whether to become members of the union and relinquish their records to it. Given the collected database $\hat{q}$, the union sells part of it, $q \leq \hat{q}$, to the platform at a price that extracts all the platform's expected payoff $V(q)$. Finally, the union distributes the proceeds $V(q)$ to its members through transfers $p \in \mathbb{R}^{\Omega}$ such that $\sum_{\omega} p(\omega)\bar{q}(\omega) \leq V(q)$. The union's objective is to maximize the welfare of its members. We assume that, if a consumer becomes a union member, she retains her reservation utility $r(\omega)$ unless her record is sold to the platform. This assumption is critical for our result. A by-product of this assumption is that, without loss of generality, all consumers join the union, i.e., $\hat{q} = \bar{q}$. Thus, the problem of the data union can be written as follows:

$$\begin{aligned}
\max_{(p,q,x)} \quad & \sum_{\omega} p(\omega)\bar{q}(\omega) + \sum_{a,\omega} u(a,\omega)x(a|\omega)q(\omega) + \sum_{\omega}(\bar{q}(\omega) - q(\omega))r(\omega) \\
\text{such that} \quad & q \leq \bar{q}, \\
\text{and} \quad & x \text{ solves } \mathcal{P}_q, \\
\text{and} \quad & \sum_{\omega} p(\omega)\bar{q}(\omega) \leq V(q), \\
\text{and} \quad & p(\omega) + \tfrac{q(\omega)}{\bar{q}(\omega)} \textstyle\sum_a u(a,\omega)x(a|\omega) + \left(1 - \tfrac{q(\omega)}{\bar{q}(\omega)}\right) r(\omega) \geq r(\omega).
\end{aligned}$$

The first two constraints are familiar from Definition 2. The third constraint requires that the data union does not run a deficit. The last constraint guarantees consumers' participation. By participating, a type-ω consumer receives compensation $p(\omega)$; with probability $\frac{q(\omega)}{\bar{q}(\omega)}$, her record is sold to the platform, in which case she also earns a trading surplus of $\sum_a u(a,\omega)x(a|\omega)$; with remaining probability, her record is not sold and she gets her reservation utility $r(\omega)$. Notice that, in this specification, all consumers are entitled to a payment $p(\omega)$, regardless of whether their data records are used. This assumption is immaterial for the result.

Proposition 3. *Let (p^*, q^*, x^*) be a solution to the data union's problem. The allocation*

(q^*, x^*) is constrained efficient. Conversely, if (q°, x°) is a constrained efficient allocation, there exists p° such that $(p^\circ, q^\circ, x^\circ)$ is a solution to the data union's problem.

There are two main differences from the competitive data economy in Section 3. First, while consumers can decide to leave the union, they have no say in whether their data records are sold to the platform. This allows the union to coordinate consumers in a way that competitive markets cannot. Second, the union has bargaining power vis-a-vis the platform and the consumers: The prices p are chosen by the union rather than being determined by market clearing. As a result, the union both internalizes the externalities discussed in the previous section and can properly compensate consumers for their participation.

It is worth noting that the data union's objective of maximizing consumers' welfare is not essential for its ability to induce constrained efficiency allocations. If the union maximized the platform's payoff—or, equivalently, if the platform was the union—then it can be shown that the induced allocations would still be constrained efficient. Clearly, such a union would have no incentive to distribute the proceeds of its activity back to the consumers. Indeed, while allocations would be constrained efficient, consumers' welfare would be minimized—i.e., equal to $R = \sum_\omega \bar{q}(\omega) r(\omega)$, with all consumers earning their reservation utility.

4.2 Data Taxes

While the data union helps achieve constrained-efficient allocations, it involves a concentration of bargaining power in a single institution. One may wonder if constrained-efficient allocations can be decentralized in our original competitive economy. This section provides a positive answer by introducing data taxes levied on consumers.[14]

These taxes work as follows. Whenever a type-ω consumer sells her record to the platform, she pays a "data tax" $\tau(\omega) \in \mathbb{R}$ to the government. We can interpret $\tau(\omega) \leq 0$ as a subsidy paid by the government. To make the problem interesting, suppose the government cannot run a deficit, i.e., $\sum_\omega q(\omega)\tau(\omega) \geq 0$. Besides taxes, all other components of the model are as in Section 2. The next result shows that any constrained efficient allocation can be supported as an equilibrium of the economy with taxation.

[14] As usual, the data tax could also be levied on the platform and, in equilibrium, passed on to the consumers.

Proposition 4. *Let (q°, x°) be a constrained-efficient allocation. There exists a profile of taxes τ^*, of prices p^*, and of consumer choices z^*, such that $(p^*, z^*, q^\circ, x^\circ)$ is an equilibrium of the economy with taxation τ^* and the government does not run a deficit.*

We can explicitly characterize the taxes and the equilibrium that supports any constrained efficient allocation (q°, x°). Let p^* be a supergradient of $V(q^\circ)$ and define

$$\tau^*(\omega) \triangleq p^*(\omega) + \sum_a x^\circ(a|\omega)u(a,\omega) - r(\omega). \tag{7}$$

Additionally, define $z^*(\omega) \triangleq q^\circ(\omega)/\bar{q}(\omega)$ for all ω. It is easy to check that $(p^*, z^*, q^\circ, x^\circ)$ is an equilibrium of the economy with taxation τ^*. First, since p^* is a supergradient of $V(q^\circ)$, q° must solve the platform's problem in the first period. Moreover, since (q°, x°) is constrained efficient, x° must solve $\mathcal{P}_{q^\circ}$, i.e., the platform's problem in the second period. Third, all consumers are indifferent between selling or not their data records to the platform since, if they sell, they earn $p^*(\omega) + \sum_a x^\circ(a|\omega)u(a,\omega) - \tau^*(\omega) = r(\omega)$. Finally, by the definition of z^*, data markets clear. In this equilibrium, consumer welfare equals $R = \sum_\omega \bar{q}(\omega)r(\omega)$ while the platform's payoff is zero. Since the allocation is constrained efficient, it must be that the government runs a budget surplus of $W^\circ - R$. If these proceeds are distributed to the consumers (in a way that does not affect their behavior, e.g., in a lump-sum manner), then consumer surplus is maximized in equilibrium. The data tax corrects the inefficiency of the competitive economy by making consumers indifferent between selling or not their data records. In this case, any constrained efficient allocation can be supported in equilibrium. It is instructive to see this in the context of a concrete example.

Example of a Data Tax. Consider Case 1 in Section 3.1. We argued that low-type consumers, whose records would be socially beneficial to sell, are not sufficiently remunerated by the competitive market, leading to inefficiency. Our taxation subsidizes these consumers just enough to make them indifferent between selling or not: $\tau^*(1) = \gamma_u - \bar{r} < 0$. This subsidy is financed by taxing the high-type consumers: $\tau^*(2) = 1 - \bar{r}$. The equilibrium prices are $p^*(1) = \gamma_u$ and $p^*(2) = 0$. Under these taxes and prices, the constrained efficient allocation (q°, x°) can be supported in equilibrium and the government gets a proceed of $\bar{q}(1)(1 + \gamma_u - 2\bar{r}) > 0$. △

4.3 Lindahl Pricing

Finally, we show how the inefficiency of the competitive economy can be corrected by enriching the price system for data records. We do so by allowing that the terms of trade between a consumer and the platform involve not only whether she sells her record, but also how the platform intends to use it.[15] That is, the price of a record can depend not only on its type ω, but also on the fee a that the platform will recommend to the merchant. The basic logic behind this approach is that the aforementioned inefficiency stems from the externalities generated by how data records are used, so market participants should have ways to take those externalities into account in their trades. As such, our approach is similar to classic ways of modeling competitive economies with externalities (e.g., Arrow (1969) and Laffont (1976)).[16] For this reason, we refer to this setting as the Lindahl economy.

More formally, this economy features one market for each pair (a, ω) with an associated price $p(a, \omega)$ at which ω-records can be traded for use a. A type-ω consumer decides in which market to sell her ω-record, if any. That is, for all a, she chooses the probability of selling her record to the platform for use a, denoted by $z(a, \omega) \in [0, 1]$. As in our baseline economy, the platform takes prices as given and chooses a database q and a mechanism x, with $x(a|\omega)q(\omega)$ representing its demand of ω-records in market (a, ω). It is implicit in the trade agreement between the platform and the consumers that, if the platform acquires a record for use a, it has to deliver on this promise. The platform's problem can then be written as

$$\begin{aligned} \max_{q,x} \quad & \sum_{a,\omega} \Big(v(a,\omega) - p(a,\omega)\Big) x(a|\omega) q(\omega) \qquad (8) \\ \text{such that} \quad & \sum_{\omega} \big(\pi(a,\omega) - \pi(a',\omega)\big) x(a|\omega) q(\omega) \geq 0 \qquad \forall\, a, a' \in A \end{aligned}$$

It is instructive to compare the platform's problem in the Lindahl economy with the one in the baseline economy (conditions (a) and (b) in Definition 1). They only differ insofar as the Lindahl economy has richer markets, with prices that can depend on a and not just on ω.[17]

[15] This is reminiscent of the European Union's general data protection regulation (GDPR), which requires that "the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data" (see, GDPR 2016/679 (39)).

[16] In particular, we follow Bonnisseau et al. (2023).

[17] In particular, the timing of the two economies is the same. To see this, note that we could have equivalently written conditions (a) and (b) in Definition 1 as Equation (8), with the additional restriction that $p(a, \omega) = p(\omega)$ for all (a, ω).

The equilibrium definition in the Lindahl economy follows from Definition 1.

Definition 3. *A profile (p^*, z^*, q^*, x^*) is an equilibrium of the Lindahl economy if*

(a) Given p^*, (q^*, x^*) solves the platform's problem in (8).

(b) Given p^*, z^* solves the consumers problem: for all ω,

$$z^*(\cdot,\omega) \in \arg \max_{z \in \mathbb{R}^A_+ \text{ s.t. } \sum_a z(a) \leq 1} \sum_a z(a)\big(p^*(a,\omega) + u(a,\omega)\big) + (1 - \sum_a z(a))r(\omega).$$

(c) Markets clear: $x^*(a|\omega)q^*(\omega) = z^*(a,\omega)\bar{q}(\omega)$ for all ω and a.

For the Lindhal economy, we will use the following notion of effciency.

Definition 4. *An allocation $(q^\dagger, x^\dagger)$ is **unconstrained efficient** if it solves*

$$\begin{aligned} W^\dagger = \max_{q,x} \quad & \mathcal{W}(q,x) \\ \textit{such that} \quad & q \leq \bar{q}, \\ \textit{and} \quad & \sum_\omega \big(\pi(a,\omega) - \pi(a',\omega)\big)x(a|\omega)q(\omega) \geq 0 \qquad \forall\, a, a' \in A \end{aligned} \tag{$\mathcal{FB}$}$$

Compared to the notion of constrained efficiency (Definition 2), an unconstrained efficient allocation (q, x) requires x to be optimal for the planner and not necessarily for the platform. Besides this, the planner's problem is the same. Indeed, notice that, in Definition 2, x was already required to be obedient since x solved $\mathcal{P}_q$. Therefore, the welfare of an unconstrained efficient allocation is weakly higher than that of a constrained efficient allocation: $W^\dagger \geq W^\circ$.

The next result shows that the equilibria of the Lindahl economy are unconstrained efficient.

Proposition 5. *Let (p^*, z^*, q^*, x^*) be an equilibrium of the Lindahl economy. The allocation (q^*, x^*) is unconstrained efficient. Moreover, consumer welfare equals $W^\dagger$. Conversely, any unconstrained efficient allocation $(q^\dagger, x^\dagger)$ can be supported as an equilibrium of the Lindahl economy.*[18]

The richness of the price system allows the equilibria of this economy not only to avoid the inefficiency highlighted in Section 3, but to achieve unconstrained efficiency. For the same

[18] Since an unconstrained efficient allocation always exists, this result implies the existence of an equilibrium of the Lindahl economy.

reasons as before, the platform still earns zero profit, so $W^{\dagger}$ is also the consumer welfare. The following example illustrates how the richer price system helps inducing efficient allocations.

Example of a Lindahl economy. Consider again the example of Section 3.1. Since $\gamma_{\pi} = 0$, a mechanism x is optimal for the platform if and only if it is also optimal for the planner. Therefore, the constrained- and unconstrained-efficient allocations coincide, leading to a welfare of $W^{\circ} = W^{\dagger} = \bar{r} + \bar{q}(1)(1 + \gamma_u - 2\bar{r})$. Moreover, as in Section 3.1, the unconstrained-efficient allocation $(q^{\dagger}, x^{\dagger})$ is unique and given by $q^{\dagger}(\omega) = \bar{q}(1)$ and $x^{\dagger}(1|\omega) = 1$ for all ω. Recall that in the baseline economy all equilibria are inefficient. By contrast, we now discuss an equilibrium of the Lindahl economy and show it is unconstrained efficient. Let (p^*, z^*, q^*, x^*) be defined as follows. First, let $(q^*, x^*) = (q^{\dagger}, x^{\dagger})$, i.e., the candidate equilibrium supports the unconstrained-efficient allocation. Second, for all ω, let $z^*(1, \omega) = \frac{\bar{q}(1)}{\bar{q}(\omega)}$ and $z^*(2, \omega) = 0$, so that z^* and (q^*, x^*) clear the data markets. Finally, let prices be $p(a = 2, \omega) = 0$, for all ω, $p^*(a = 1, \omega = 1) = \gamma_u + (1 - \bar{r})$, and $p^*(a = 1, \omega = 2) = -(1 - \bar{r})$. To see that this is an equilibrium of the Lindahl economy, note that given prices p^*, type-1 consumers strictly prefer to sell their record in market $(a = 1, \omega = 1)$. Type-2 consumers, instead, are indifferent between not selling and selling in market $(a = 1, \omega = 2)$. Finally, the platform maximizes $(\gamma_u + 1 - \bar{r})(x(1|2)q(2) - x(1|1)q(1))$ subject to $x(1|1)q(1) \geq x(1|2)q(2)$. Therefore, the platform cannot make a positive payoff, and (q^*, x^*) achieves the maximum of 0 given p^*.

It is worth noting the crucial role of the richer price system. The price $p^*(a = 1, \omega = 1)$ incorporates the positive externality that a low-type consumer generates when selling her record. This high price paid by the platform is financed by the high-type consumers, who have to pay to participate in the platform's mechanism. The platform uses their payments to acquire low-type records. That is, it is as if high-type consumers who participate in the platform's mechanism subsidize the participation of low-type consumers. Notice that the equilibrium exists even if $p^*(a = 1, \omega = 2) < 0$. Despite the negative price, the platform does not have an incentive to acquire an arbitrary quantity of such records. This is because, to fulfill the terms of trade for this market, the platform has to guarantee the merchant is willing to charge a low fee $a = 1$ to the type-2 records it acquires. $\triangle$

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022): "Too much data: Prices and inefficiencies in data markets," *American Economic Journal: Microeconomics*, 14, 218–256.

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

Acquisti, A. and H. R. Varian (2005): "Conditioning prices on purchase history," *Marketing Science*, 24, 367–381.

Ali, S. N., G. Lewis, and S. Vasserman (2022): "Voluntary Disclosure and Personalized Pricing," *forthcoming, Review of Economic Studies*.

Arrow, K. J. (1969): "The Organization of Economic Activity: Issues Pertinent to the Choice of Market versus Non-Market Allocation," *The Analysis and Evaluation of Public Expenditures: the PPB System*, Joint Economic Committee, Congress of the United States, Washington, D.C., 47–64.

Bergemann, D. and A. Bonatti (2019): "Markets for Information: An Introduction," *Annual Review of Economics*, 11, 85–107.

——— (2023): "Data, competition, and digital platforms," *arXiv preprint arXiv:2304.07653*.

Bergemann, D., A. Bonatti, and T. Gan (2022a): "The economics of social data," *The RAND Journal of Economics*, 53, 263–296.

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

Bergemann, D., J. Crémer, D. Dinielli, C.-C. Groh, P. Heidhues, M. Schafer, M. Schnitzer, F. M. S. Morton, K. Seim, and M. Sullivan (2023): "Market design for personal data," *Yale J. on Reg.*, 40, 1056.

Bergemann, D., T. Heumann, S. Morris, C. Sorokin, and E. Winter (2022b): "Optimal Information Disclosure in Classic Auctions," *American Economic Review: Insights*, 4, 371–88.

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," in *Handbook* *of industrial organization*, Elsevier, vol. 4(1), 593–672.

Bertsimas, D. and J. N. Tsitsiklis (1997): *Introduction to linear optimization*, vol. 6, Athena scientific Belmont, MA.

Böhm, V. (1975): "On the continuity of the optimal policy set for linear programs," *SIAM Journal on Applied Mathematics*, 28, 303–306.

Bonnisseau, J.-M., E. L. del Mercato, and P. Siconolfi (2023): "Existence of an equilibrium in arrowian markets for consumption externalities," *Journal of Economic Theory*, 209, 105638.

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

Chen, D. (2022): "The market for attention," *Available at SSRN 4024597*.

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.

Doval, L. and V. Skreta (2023): "Purchase history and product personalization," *RAND Journal of Economics*, Forthcoming.

Dworczak, P. (2020): "Mechanism design with aftermarkets: Cutoff mechanisms," *Econometrica*, 88, 2629–2661.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp (2019): "Big data and firm dynamics," in *AEA papers and proceedings*, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, vol. 109, 38–42.

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

Galperti, S., A. Levkun, and J. Perego (2023): "The Value of Data Records," *Review of Economic Studies*, rdad044.

Galperti, S. and J. Perego (2022): "Competitive Markets for Personal Data," *Available at SSRN 4309966*.

Goldfarb, A. and C. Tucker (2023): *The Economics of Privacy*, NBER Conference Volume.

Hidir, S. and N. Vellodi (2021): "Privacy, Personalization and Price Discrimination," *Journal of the European Economic Association*, 19, 1342–1363.

Ichihashi, S. (2021): "The economics of data externalities," *Journal of Economic Theory*, 196, 105316.

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

Kamenica, E. (2019): "Bayesian persuasion and information design," *Annual Review of Economics*, 11, 249–272.

Laffont, J. J. (1976): "Decentralization with Externalities," *European Economic Review*, 359–375.

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

Taylor, C. R. (2004): "Consumer privacy and the market for customer information," *Rand Journal of Economics*.

Varian, H. R. (2009): "Economic Aspects of Personal Privacy," in *Internet Policy and Economics*, ed. by W. H. Lehr and L. M. Pupillo, New York: Springer.

Xu, W. and K. H. Yang (2023): "Informational Intermediation, Market Feedback, and Welfare Losses," *Working Paper*.

Appendix

A Proofs

A.1 Proofs for Section 3

As a result of the linear specification of the platform's payoff, $v(a,\omega) = \gamma_u u(a,\omega) + \gamma_\pi \pi(a,\omega)$, we can explicitly compute $V(q)$ and $W(q)$ for all q, and characterize their supergradients. To do so, fix an arbitrary database $q \geq 0$ and let $w(q) \triangleq \sum_\omega \omega q(\omega)$ be the gains from trade and $\pi_m(q) \triangleq \max_a \sum_\omega \pi(a,\omega) q(\omega)$ be the payoff of the uninformed merchant who only knows the composition of q.

As discussed in the second paragraph after Proposition 2 and shown in Figure 1, when the platform's payoff v has the linear specification, the optimal mechanism x^* either maximizes merchant's profit (when $\gamma_\pi > \gamma_u$) or maximizes consumers' surplus (when $\gamma_u > \gamma_\pi$). Therefore, we can explicitly compute the platform's value V and the planner's value W, depending on the relative magnitude of γ_u and γ_π:

$$V(q) = \begin{cases} \gamma_\pi w(q), & \gamma_\pi > \gamma_u \\ \gamma_\pi \pi_m(q) + \gamma_u (w(q) - \pi_m(q)), & \gamma_\pi \leq \gamma_u \end{cases},$$

$$W(q) = \begin{cases} \gamma_\pi w(q), & \gamma_\pi > \gamma_u \\ \gamma_\pi \pi_m(q) + (1+\gamma_u)(w(q) - \pi_m(q)), & \gamma_\pi \leq \gamma_u \end{cases}. \tag{A.1}$$

Since $w(q)$ is linear and $\pi_m(q)$ is convex, V and W are concave functions and their supergradients are well-defined. Next, we characterize these supergradients. Let $A_q \triangleq \arg\max_a \sum_\omega \pi(a,\omega) q(\omega)$ be the set of optimal monopoly prices under q. Let $\phi_a \in \mathbb{R}^\Omega$ be defined as $\phi_a(\omega) \triangleq \gamma_u \omega + (\gamma_\pi - \gamma_u) a \mathbb{1}(\omega \geq a)$. Define $\tilde{\Phi}_q(\gamma_u, \gamma_\pi) \subset \mathbb{R}^\Omega$ as:

$$\tilde{\Phi}_q(\gamma_u,\gamma_\pi) \triangleq \begin{cases} (\gamma_\pi \omega)_{\omega \in \Omega}, & \gamma_\pi > \gamma_u \\ cov\{\phi_a : a \in A_q\}, & \gamma_\pi \leq \gamma_u \end{cases},$$

where cov refers to convex hull. Moreover, let

$$\Phi_q(\gamma_u,\gamma_\pi) \triangleq \{\phi \in \mathbb{R}^\Omega : \exists \tilde{\phi} \in \tilde{\Phi}_q(\gamma_u,\gamma_\pi) \text{ s.t. } \phi(\omega) = \tilde{\phi}(\omega), \text{ if } q(\omega) > 0 \\ \text{and } \phi(\omega) \geq \tilde{\phi}(\omega), \text{ if } q(\omega) = 0\}.$$

Lemma A.1.

1. *The set of supergradients of* $V(q)$ *at* q *is* $\Phi_q(\gamma_u, \gamma_\pi)$.

2. *The set of supergradients of* $W(q)$ *at* q *is* $\Psi_q = \Phi_q(\gamma_u, \gamma_\pi)$ *if* $\gamma_\pi > \gamma_u$ *and* $\Psi_q = \Phi_q(1 + \gamma_u, \gamma_\pi)$ *if* $\gamma_\pi \leq \gamma_u$.

Proof. When $\gamma_\pi > \gamma_u$, by definition of w we know that $\partial_q V(q) = (\gamma_\pi \omega)_{\omega \in \Omega}$ when $q \gg 0$. When $q(\omega) = 0$ for some ω, $\phi(\omega)$ can be any value greater than $\gamma_\pi \omega$ because the domain of q is $\mathbb{R}^\Omega_+$. This proves the $\gamma_\pi > \gamma_u$ case.

When $\gamma_\pi \leq \gamma_u$, we need to examine the subgradient of $\pi_m(q)$. Note that the optimal monopoly price can only be one of $a \in \Omega$. Therefore, π_m is the maximum of finitely many linear functions of q. For such a function, the subgradient at q is the convex hull of the subgradient of all functions achieving the maximum value at q. A function achieves the maximum value at q if and only if it is an element of A_q. Therefore, we have $\partial_q \pi_m(q) = cov\{\pi(a_q, \cdot) : a_q \in A_q\}$. This concludes $\partial_q V(q) = cov\{\phi_a : a \in A_q\}$ when $q \gg 0$. Similar to the previous case, when $q(\omega) = 0$ for some ω, $\phi(\omega)$ can be any value greater than the ones given above because the domain of q is $\mathbb{R}^\Omega_+$. This proves the first claim of Lemma XX.

The second claim of Lemma XX follows immediately, since, given Equation (A.1), if $\gamma_\pi > \gamma_u$, $W(q) = V(q)$; if $\gamma_\pi \leq \gamma_u$, Ψ_q can be derived as described above by replacing γ_u with $1 + \gamma_u$. □

Note that Φ_q and Ψ_q are singletons when A_q is a singleton and $q \gg 0$, so the supergradients are generically unique. Moreover, Theorem 5.2 in Bertsimas and Tsitsiklis (1997) shows that the set of supergradients of the value function of a linear program coincides with the set of solutions of its dual. Therefore, $\Phi_q(v)$ is also the set of solutions to the dual problem of $\mathcal{P}_q$:

$$\begin{aligned} \min_{\phi, \lambda} \quad & \sum_\omega \phi(\omega) q(\omega) \\ \text{such that} \quad & \phi(\omega) \geq v(a, \omega) + \sum_{\hat{a}} (\pi(a, \omega) - \pi(\hat{a}, \omega)) \lambda(\hat{a}|a) \qquad \forall a, \omega \qquad (\mathcal{P}'_q(v)) \\ \text{and} \quad & \lambda(\hat{a}|a) \geq 0 \qquad \forall \hat{a}, a \end{aligned}$$

We will use this fact in the following proofs.

It is with noting that Lemma A.1 extends Proposition 2 in Galperti et al. (2023). In their setting, Galperti et al. (2023) characterize the solution to problem $\mathcal{P}'_q(v)$ when it is unique, i.e., when A_q is a singleton and $q \gg 0$. We complement their results by providing a full characterization of the solutions even when it is not unique, and also by characterizing the supergradients of W, i.e., the marginal value of data to the planner under the incentive constraint of the platform.

With these results in mind, we proceed to prove Proposition 1 and Lemma 1.

Proof of Proposition 1. Suppose $\gamma_\pi > \gamma_u$. It is easy to see that in this case the planner's solution to problem $\mathcal{SB}$ (see Definition 2) is $q(\omega) = 0$ if $\gamma_\pi \omega < r(\omega)$ and $q(\omega) = \bar{q}(\omega)$ if $\gamma_\pi \omega > r(\omega)$. Taken together, q is constrained efficient if and only if $q(\omega) > 0$ implies $\gamma_\pi \omega \geq r(\omega)$ while $q(\omega) = 0$ implies $\gamma_\pi \omega \leq r(\omega)$. From Lemma A.1 we know that $\psi_q(\omega) = \gamma_\pi \omega$ if $q(\omega) > 0$ and $\psi_q(\omega) = [\gamma_\pi \omega, \infty)$ if $q(\omega) = 0$. This completes the proof.

Conversely, suppose $\gamma_\pi \leq \gamma_u$. In this case, the constraint that requires x to be sequentially optimal for the platform can be relaxed, and the planner's problem $\mathcal{SB}$ coincides with $\mathcal{FB}$ (see Definition XXX). Thus, a data allocation (q, x) is constrained efficient if and only if it solves problem $\mathcal{FB}$. The dual problem of $\mathcal{FB}$ can be formulated as:

$$
\begin{aligned}
(\mathcal{FB}') : \min_{\mu, \lambda} \quad & \sum_\omega \mu(\omega) \bar{q}(\omega) \\
\text{such that} \quad & \mu(\omega) \geq \gamma_\pi \pi(a, \omega) + (\gamma_u + 1) u(a, \omega) + \sum_{\hat{a}} (\pi(a, \omega) - \pi(\hat{a}, \omega)) \lambda(\hat{a}|a) \qquad \forall a, \omega \\
\text{and} \quad & \mu(\omega) \geq r(\omega) \qquad \forall \omega \\
\text{and} \quad & \lambda(\hat{a}|a) \geq 0 \qquad \forall \hat{a}, a
\end{aligned}
$$

We first show the "only if" direction of the proposition. Take any efficient allocation (q, x). Then the planner's value can also be written as:

$$
\begin{aligned}
\sum_\omega \Big(\bar{q}(\omega) - q(\omega) \Big) r(\omega) + \quad \max_{\chi \in \mathbb{R}_+^{A \times \Omega}} \quad & \sum_{a, \omega} \Big(v(a, \omega) + u(a, \omega) \Big) \chi(a, \omega) \\
\text{such that} \quad & \sum_a \chi(a, \omega) = q(\omega), \qquad \forall \omega \in \Omega \\
\text{and} \quad & \sum_\omega \big(\pi(a, \omega) - \pi(\hat{a}, \omega) \big) \chi(a, \omega) \geq 0 \qquad \forall\, a, \hat{a} \in A
\end{aligned}
$$

Let (μ, λ) be a solution of $\mathcal{FB}'$. Then (μ, λ) is also feasible for problem $\mathcal{P}_q'(v + u)$, which is the dual of the maximization problem above. By strong duality, the planner's value can be written as $\sum_\omega \mu(\omega) \bar{q}(\omega)$. Therefore, we have:

$$
\sum_\omega \mu(\omega) \bar{q}(\omega) \leq \sum_\omega \Big(\bar{q}(\omega) - q(\omega) \Big) r(\omega) + \sum_\omega \mu(\omega) q(\omega).
$$

Meanwhile, since $\mu(\omega) \geq r(\omega)$, we must also have the other direction of the inequality, so we conclude:

$$\sum_\omega \mu(\omega)\bar{q}(\omega) = \sum_\omega \Big(\bar{q}(\omega) - q(\omega)\Big) r(\omega) + \sum_\omega \mu(\omega) q(\omega).$$

This equality has two implications. First, (μ, λ) achieves the minimum of $\mathcal{P}'_q(v+u)$, so $\mu \in \Phi_q(v+u)$. Second, $\mu(\omega) = r(\omega)$ whenever $q(\omega) < \bar{q}(\omega)$. Taking $\psi_q = \mu$, we conclude the "only if" direction.

Next, we prove the "if" direction. Let (ψ, λ) be a solution to $\mathcal{P}'_q(v+u)$ that satisfies the condition. Take $\mu := \max\{\psi, r\}$. Then, (μ, λ) is feasible for $\mathcal{FB}'$ and, thus, $\sum_\omega \mu(\omega)\bar{q}(\omega)$ is an upper bound for the planner's value by weak duality. Next, we argue q achieves this value. Since (μ, λ) is a solution to $\mathcal{P}'_q(v+u)$, by strong duality, we know that under q, the planner's value is

$$\sum_\omega q(\omega)\mu(\omega) + \sum_\omega r(\omega)(\bar{q}(\omega) - q(\omega)).$$

From the proposition's condition, when $q(\omega) = \bar{q}(\omega)$, $\psi(\omega) \geq r(\omega)$ and thus $\psi(\omega)q(\omega) = \mu(\omega)\bar{q}(\omega)$; when $q(\omega) = 0$, $\psi(\omega) \leq r(\omega)$ and thus $r(\omega)(\bar{q}(\omega) - q(\omega)) = \mu(\omega)\bar{q}(\omega)$; when $0 < q(\omega) < \bar{q}(\omega)$, we have $\psi(\omega) = r(\omega) = \mu(\omega)$. These imply that:

$$\sum_\omega q(\omega)\psi(\omega) + \sum_\omega r(\omega)(\bar{q}(\omega) - q(\omega)) = \sum_\omega \mu(\omega)\bar{q}(\omega).$$

This means that allocation q achieves the upper bound of the planner's value and thus q is constrained efficient. □

Instead of proving Lemma 1, we prove a slightly stronger result below.

Lemma A.2. *Given an arbitrary v (not necessarily linear in u and π). $q \leq \bar{q}$ solves the platform's first stage problem* (1) *if and only if $p \in \Phi_q(v)$.*

Proof. We first observe that the platform's first-stage problem (1) is essentially to choose (q, x) given price p, or equivalently choosing $\chi(a, \omega) = x(a|\omega)q(\omega)$ without any feasibility constraint. This is because in the first-stage the platform can choose as many data records as it demands. Therefore, its dual problem is formulated as:

$$\begin{aligned}
\min_{\lambda} \quad & 0 \\
\text{such that} \quad & \sum_{\hat{a}} (\pi(\hat{a}, \omega) - \pi(a, \omega))\lambda(\hat{a}|a) \geq v(a, \omega) - p(\omega) \qquad \forall a, \omega \qquad \text{(A.2)} \\
\text{and} \quad & \lambda(\hat{a}|a) \geq 0 \qquad \forall \hat{a}, a
\end{aligned}$$

In other words, the platform's optimal payoff is zero if (A.2) is feasible, and it is infinite otherwise.

To show the "only if" direction, note that $q \leq \bar{q}$ solving (1) implies that $V(q) = \sum_\omega p(\omega)q(\omega)$. If not, the platform can scale up q proportionally to earn an infinite payoff. By strong duality, Problem (A.2) is feasible. Take any feasible solution λ, and consider (ϕ, λ) where $\phi = p$. Next we argue (ϕ, λ) is an optimal solution to $\mathcal{P}'_q$. Suppose not, then since (ϕ, λ) is feasible to $\mathcal{P}'_q$, we must have $V(q) < \sum_\omega \phi(\omega)q(\omega)$, but this contradicts $V(q) - \sum_\omega p(\omega)q(\omega) = 0$.

To show the "if" direction, suppose (p, λ) is an optimal solution to $\mathcal{P}'_q$. This means Problem (A.2) is feasible. Therefore, the platform's optimal payoff is 0 given p. By strong duality, we have $V(q) = \sum_\omega p(\omega)q(\omega)$. This means that q gives the platform a payoff of 0. Therefore, q is a solution to the platform's problem (1) given price p. □

Note that Lemma A.2 implies Lemma 1 since in any equilibrium (p^*, z^*, q^*, x^*), we must have $q^* \leq \bar{q}$ and q^* solves (1) given p^*.

Proof of Proposition 2. We only prove the case of $\gamma_\pi > \gamma_u$ here. The other case is discussed in detail below. Consider any equilibrium (p^*, z^*, q^*, x^*). We know that $q^*(\omega) > 0$ implies $p^*(\omega) \geq r(\omega)$ and $q^*(\omega) < \bar{q}(\omega)$ implies $p^*(\omega) \leq r(\omega)$. By Lemma 1, we know $p^* \in \Phi_{q^*}$. Since $\Psi_{q^*} = \Phi_{q^*}$ by Lemma A.1, taking $\psi_{q^*} = p^*$ we conclude that q^* is constrained efficient by Proposition 1. □

Proof of Corollary 1. Let (p^*, z^*, q^*, x^*) be an equilibrium. If $q^* \equiv 0$, then there is no trade, which is inefficient by assumption. Therefore, assume $q^* \neq 0$. Next, we argue $q^*(\underline{\omega}) = 0$. By Lemma 1 and Lemma A.1, $p^*(\underline{\omega}) \leq \gamma_u \underline{\omega} < r(\underline{\omega})$. Since a type-$\underline{\omega}$ consumer will earn a zero trading surplus by selling her record, she strictly prefer not to do so. However, this is inefficient since, given $q^* \neq 0$ and Lemma A.1, the marginal value of $\underline{\omega}$ to the planner is no lower than $(\gamma_u + 1)\underline{\omega} > r(\underline{\omega})$. By Proposition 1 we conclude. □

Next, we introduce another set of sufficient conditions under which all equilibria of the competitive economy are inefficient. Let ω_1 be the highest type in Ω and ω_2 be the second-highest one. In Corollary A.1, we assume that, if all type-ω_1 consumers sell their records, the merchant strictly prefers to set a fee $a = \omega_1$. This is guaranteed by the requirement that $\bar{q}(\omega_1)\omega_1 > \omega \sum_{\omega' \geq \omega} \bar{q}(\omega)$ for all $\omega < \omega_1$. Additionally, we assume that if the uninformed merchant fee is $a_q = \omega_1$, the platform wants to pool type-ω_2 and type-ω_1 consumers, to induce a lower fee from the merchant. This is guaranteed by the requirement that $r(\omega_2) < \gamma_u \omega_2$. Under these assumptions, we show that for a range of $r(\omega_1)$, type-ω_1 consumers will earn an expected surplus higher than their marginal contribution. That is, they exert a negative externality on other consumers when selling their records. This corresponds to Case 2 in the example of Section 3.1.

Corollary A.1. *Consider $\gamma_\pi \leq \gamma_u$. Assume $\bar{q}(\omega_1)\omega_1 > \omega \sum_{\omega' \geq \omega} \bar{q}(\omega)$ for all $\omega < \omega_1$ and $r(\omega_2) < \gamma_u \omega_2$. Then no equilibrium is efficient if $\gamma_\pi \omega_1 < r(\omega_1) < \gamma_\pi \omega_1 + \omega_1 - \omega_2$.*

Proof. Let (p^*, z^*, q^*, x^*) be a competitive equilibrium. We want to show it is not constrained efficient. Note that since $\gamma_\pi \leq \gamma_u$, solving $\mathcal{SB}$ is equivalent to solving $\mathcal{FB}$. In particular, (q^*, x^*) must maximize consumer surplus as well as social welfare in order to be constrained efficient. There are four cases to discuss.

- Case 1: Suppose $q^*(\omega_2) < \bar{q}(\omega_2), q^*(\omega_1) < \bar{q}(\omega_1)$. To see that the equilibrium is inefficient, note that we can marginally increase and pool ω_2 and ω_1 and create a market of proposition $(1 - \frac{\omega_2}{\omega_1}, \frac{\omega_2}{\omega_1})$, with a recommended price $a = \omega_2$. The cost of this market is $r(\omega_2)(1 - \frac{\omega_2}{\omega_1}) + r(\omega_1)\frac{\omega_2}{\omega_1} < \gamma_\pi \omega_2 + (\gamma_u + 1)\frac{\omega_2}{\omega_1}(\omega_1 - \omega_2)$, where the right-hand side is the social benefit of the market. Therefore, the social welfare strictly increases by introducing this market and, thus, the equilibrium is inefficient.

- Case 2: Suppose $q^*(\omega_2) < \bar{q}(\omega_2), q^*(\omega_1) = \bar{q}(\omega_1)$. In this case, by assumption, the unique uninformed merchant fee for market q^* is $a_{q^*} = \omega_1$. Therefore, by Lemma A.1, $\psi_{q^*}(\omega_2)$ is at least $(\gamma_u + 1)\omega_2 > r(\omega_2)$. By Proposition 1, we conclude that q^* is inefficient.

- Case 3: $q^*(\omega_2) = \bar{q}(\omega_2), q^*(\omega_1) < \bar{q}(\omega_1)$. Note that, in this case, we must have $x^*(\omega_1|\omega) = 0$ for all $\omega < \omega_1$. If not, some type-ω consumers will sell their records while left with nothing, which violates social optimality. Next, we argue that (q^*, x^*) is not constrained efficient. First, if $q^*(\omega_1) = 0$, then $\psi_{q^*}(\omega_1)$ is at least $\gamma_\pi \omega_2 + (\gamma_u + 1)(\omega_1 - \omega_2) = \gamma_\pi \omega_1 + (\gamma_u - \gamma_\pi + 1)(\omega_1 - \omega_2) > r(\omega_1)$. Therefore, by Proposition 1, q^* cannot be efficient. Second, suppose $q^*(\omega_1) > 0$. If $x^*(\omega_1|\omega_1) > 0$, then by the observation, ω_1 is perfectly revealed with positive probability since $x^*(\omega_1|\cdot) = 0$ for all other types. However, by excluding these type-ω_1 consumers from the platform's database, their welfare increases by $r(\omega_1) > \gamma_\pi \omega_1$ while all other markets (with fee smaller than ω_1) are not affected. This implies that the social welfare strictly increases. Therefore, (q^*, x^*) cannot be constrained efficient. Third, suppose $x^*(\omega_1|\omega_1) = 0$. Then, on the platform ω_1 consumers get at least $\omega_1 - \omega_2$, and they receive a payment of at least $\gamma_\pi \omega_1$, by Lemma 1. Together, since $\omega_1 - \omega_2 + \gamma_\pi \omega_1 > r(\omega_1)$, it must be that $q^*(\omega_1) = \bar{q}(\omega_1)$, a contradiction. Therefore, (q^*, x^*) cannot be constrained efficient.

- Case 4: Suppose $q^*(\omega_2) = \bar{q}(\omega_2)$ and $q^*(\omega_1) = \bar{q}(\omega_1)$. In this case, by our assumption, the unique uninformed merchant fees is $a_{q^*} = \omega_1$. Therefore, we have $\psi_{q^*}(\omega_1) = \gamma_\pi \omega_1 < r(\omega_1)$. By Proposition 1, (q^*, x^*) is inefficient.

□

A.2 Proofs for Section 4

Next, we provide proofs for all results in Section 4. Note that none of the results in that section require the assumption that the platform's payoff is linear in u and π.

Proof of Proposition 3. *"Only If" Direction.* Let (q^*, x^*) be constrained efficient. First, we argue that (q^*, x^*) is a solution of a relaxed version of the data union's problem, which we obtain by discarding the consumers' participation constraints. Additionally, note that the constraint $\sum_\omega \hat{q}(\omega) p(\omega) \leq V(q)$ must hold with equality and, thus, can be substituted into the data union's objective. By doing so, prices p do not appear in the relaxed problem, which can be written as

$$\begin{aligned}
\max_{(q,x)} \quad & \sum_{a,\omega} \Big(v(a,\omega) + u(a,\omega) \Big) x(a|\omega) q(\omega) + \sum_{\omega} (\bar{q}(\omega) - q(\omega)) r(\omega) \\
\text{such that} \quad & q \leq \bar{q}, \\
\text{and} \quad & x \text{ solves } \mathcal{P}_q.
\end{aligned}$$

Note that this relaxed problem coincides with the planner's problem $\mathcal{SB}$. Thus, any solution to the relaxed problem is constrained efficient and must yield a value that is weakly higher than the value of the data union's problem. To complete the proof, we find prices p^* such that, given (q^*, x^*), the consumers' participation constraints are satisfied and $\sum_\omega \hat{q}(\omega) p(\omega) = V(q)$.

To this end, let $p^*(\omega) = \tilde{p}(\omega) + t(\omega)$ with $\tilde{p}(\omega) = \frac{q^*(\omega)}{\bar{q}(\omega)} \Big(r(\omega) - \mathbb{E}_{x^*}(u(a,\omega)) \Big)$, where $t(\omega)$ will be pinned down later. If $t(\omega) = 0$, all type-ω consumers would be indifferent between joining the union or not, and in particular, $z^*(\omega) = 1$ is optimal. In this case, the union's budget is:

$$\begin{aligned} G(q^*, x^*) &= V(q^*) - \sum_{\omega} \bar{q}(\omega)\tilde{p}(\omega) \\ &= \sum_{a,\omega} \Big(v(a,\omega) + u(a,\omega)\Big) x^*(a|\omega) q^*(\omega) - \sum_{\omega} q^*(\omega) r(\omega). \end{aligned}$$

Since (q^*, x^*) is constrained efficient, $G(q^*, x^*) \geq 0$. To see this, we add $\sum_\omega \bar{q}(\omega) r(\omega)$ on both sides of this inequality. On the left hand side, we obtain the value of the planner's objective at (q^*, x^*), which is weakly higher than $R = \sum_\omega \bar{q}(\omega) r(\omega)$.

Finally, to guarantee $\sum_\omega \hat{q}^*(\omega) p^*(\omega) = V(q^*)$, we can uniformly distribute $G(q^*, x^*)$ to the consumers. Specifically, let $t(\omega) = G(q^*, x^*)$ (recall that $\sum_\omega \bar{q}(\omega) = 1$). Therefore, since $z^*(\omega) = 1$ was optimal under $\tilde{p}(\omega)$, it is still optimal under $p^*(\omega) \geq \tilde{p}(\omega)$.

Thus, we constructed a profile (p^*, q^*, x^*) that is feasible for the data union. Moreover, since (q^*, x^*) solves the relaxed problem, it must be that (p^*, q^*, x^*) solves the data union's problem.

("If" Direction). Let (p^*, q^*, x^*) be a solution to the data union's problem but suppose it is not constrained efficient. Let (q°, x°) be a constrained efficient allocation. We have that:

$$\begin{aligned} &\sum_{a,\omega} \Big(v(a,\omega) + u(a,\omega)\Big) x^*(a|\omega) q^*(\omega) - \sum_{\omega} q^*(\omega) r(\omega) \\ <&\sum_{a,\omega} \Big(v(a,\omega) + u(a,\omega)\Big) x^\circ(a|\omega) q^\circ(\omega) - \sum_{\omega} q^\circ(\omega) r(\omega) = W^\circ \end{aligned} \tag{A.3}$$

By the "only if" direction, we know there exist p° such that $(p^\circ, q^\circ, x^\circ)$ is feasible for the data union and achieves W°. This contradicts (p^*, q^*, x^*) being a solution to the data union's problem. □

Proof of Proposition 4. □

Proof of Proposition 5. *Step 1:* Let (p^*, q^*, x^*, z^*) be a Lindahl equilibrium. We first prove that (q^*, x^*) must solve $\mathcal{FB}$. Since (q^*, x^*) solves $\mathcal{P}'$, we have that

$$\begin{aligned} &\sum_{a,\omega} v(a,\omega) x^*(a|\omega) q^*(\omega) - \sum_{a,\omega} v(a,\omega) x(a|\omega) q(\omega) \\ \geq &\sum_{a,\omega} p^*(a,\omega) x^*(a|\omega) q^*(\omega) - \sum_{a,\omega} p^*(a,\omega) x(a|\omega) q(\omega) \end{aligned} \tag{A.4}$$

for all (q,x) that satisfies obedience. Similarly, by the maximization problem of type-ω consumers, we get

$$\sum_a u(a,\omega)z^*(a,\omega)+r(\omega)\big(1-\sum_a z^*(a,\omega)\big)-\sum_a u(a,\omega)z(a,\omega)-r(\omega)\big(1-\sum_{a,} z(a,\omega)\big)$$
$$\geq -\sum_a p^*(a,\omega)z^*(a,\omega)+\sum_a p^*(a,\omega)z(a,\omega)$$

for all $z(a,\omega)\in\mathbb{R}^A_+$ such that $\sum_a z(a,\omega)\leq 1$. Summing over consumers of the same type and across types, we get that for all (q,x) such that $q\leq\bar{q}$:

$$\sum_{a,\omega} u(a,\omega)x^*(a|\omega)q^*(\omega)-\sum_\omega r(\omega)q^*(\omega)-\sum_{a,\omega} u(a,\omega)x(a|\omega)q(\omega)+\sum_\omega r(\omega)q(\omega)$$
$$\geq -\sum_{a,\omega} p^*(a,\omega)x^*(a|\omega)q^*(\omega)+\sum_{a,\omega} p^*(a,\omega)x(a|\omega)q(\omega). \tag{A.5}$$

Equations (A.4) and (A.5) jointly imply that for all (q,x) satisfying feasibility and obedience:

$$\begin{aligned} &\sum_{a,\omega}\big(v(a,\omega)+u(a,\omega)\big)x^*(a|\omega)q^*(\omega)-\sum_\omega r(\omega)q^*(\omega)\\ \geq\quad &\sum_{a,\omega}\big(v(a,\omega)+u(a,\omega)\big)x(a|\omega)q(\omega)-\sum_\omega r(\omega)q(\omega).\end{aligned}$$

Therefore, (q^*,x^*) solves $\mathcal{FB}$.

Step 2: We now prove that for any allocation (q^*,x^*) that solves $\mathcal{F}B$, there is a (p^*,z^*) such that (p^*,q^*,x^*,z^*) is a Lindahl equilibrium. First of all, notice that $\mathcal{F}B$ admits an optimal solution. Second, we can define $p^*(a,\omega)=r(\omega)-u(a,\omega)$ for all a,ω, so that each ω consumer is indifferent across all possible $z(\cdot,\omega)$ and we can therefore assume to choose z^* such that $z^*(\cdot,\omega)\bar{q}(\omega)=x^*(\cdot|\omega)q^*(\omega)$.

We can equivalently rewrite $\mathcal{FB}$ in terms of χ:

$$\begin{aligned}(\mathcal{FB}'):\quad &\max_{\chi\in\mathbb{R}_+^{A\times\Omega}} && \sum_{a,\omega}\Big(v(a,\omega)+u(a,\omega)\Big)\chi(a,\omega)+\sum_\omega\Big(\bar{q}(\omega)-\sum_a\chi(a,\omega)\Big)r(\omega)\\ &\text{such that} && \sum_a\chi(a,\omega)\leq\bar{q}(\omega),\qquad\forall\omega\in\Omega\\ &\text{and} && \sum_\omega\big(\pi(a,\omega)-\pi(\hat{a},\omega)\big)\chi(a,\omega)\geq 0\qquad\forall\, a,\hat{a}\in A\end{aligned}$$

Since (q^*,x^*) is a first-best efficient allocation, we know $\chi^*(a,\omega):=x^*(a|\omega)q^*(\omega)$ solves $\mathcal{FB}'$. Define $z^*(a,\omega)=\chi^*(a,\omega)/\bar{q}(\omega)$. Since χ^* is an optimal solution to $\mathcal{FB}'$, by strong duality, we know its dual admits an optimal solution $(\mu^*(\omega),\lambda^*(\hat{a}|a))$. Define $p^*(a,\omega)=\mu^*(\omega)+r(\omega)-u(a,\omega)$.

We first argue that given p^*, $z^*(\omega)$ is optimal for type-ω consumers. When $\mu^*(\omega) = 0$, we have $p^*(a,\omega) = r(\omega) - u(a,\omega)$. Thus, type-$\omega$ consumers are indifferent between keeping the data and selling it in market (a,ω) for all $a \in A$. Therefore, $z^*(\cdot,\omega)$ is optimal. When $\mu^*(\omega) > 0$, by complementary slackness, we have that $\sum_a z^*(a,\omega) = 1$. Therefore, no type-ω consumer keeps the data. Since selling the record in market (a,ω) for any $a \in A$ gives the consumer a payoff of $\mu^*(\omega) + r(\omega)$. They are indifferent between different a and, thus, $z^*(\omega)$ is optimal.

Next, we argue that χ^* solves the platform's problem given p^*. We first show that the platform's payoff is non-positive under p^*. To show this, we only need to show the dual problem of the platform's problem is feasible. The dual feasible set is given by:

$$\sum_{\hat{a}}(\pi(\hat{a},\omega) - \pi(a,\omega))\lambda(\hat{a}|a) \geq v(a,\omega) - p^*(a,\omega)$$

$$= v(a,\omega) + u(a,\omega) - \mu^*(\omega) - r(\omega)$$

for all a,ω, with $\lambda \geq 0$. But we know this is feasible because λ^* satisfies these constraints. Given dual feasibility, weak duality implies:

$$\sum_{a,\omega}(v(a,\omega) - p^*(a,\omega))\chi(a,\omega) \leq 0$$

for all χ that is feasible to the platform.

Finally, by strong duality we have:

$$\sum_{a,\omega}\Big(v(a,\omega) + u(a,\omega)\Big)\chi^*(a,\omega) - \sum_{a,\omega}\chi^*(a,\omega)r(\omega) = \sum_{\omega}\mu^*(\omega)\bar{q}(\omega).$$

This implies:

$$\sum_{a,\omega}(v(a,\omega) - p^*(a,\omega) + \mu^*(\omega))\chi^*(a,\omega) = \sum_{\omega}\mu^*(\omega)\bar{q}(\omega).$$

By complementary slackness we know $\sum_{a,\omega}\mu^*(\omega)\chi^*(a,\omega) = \sum_{\omega}\mu^*(\omega)\bar{q}(\omega)$, which implies:

$$\sum_{a,\omega}(v(a,\omega) - p^*(a,\omega))\chi^*(a,\omega) = 0.$$

Therefore, we conclude χ^* solves the platform's problem given p^*. □

Online Appendix (For Online Publication Only)

B Equilibrium Existence

In this section, we prove the existence of an equilibrium of the competitive economy, allowing for arbitrary specification of v. We start by showing that the solution correspondence of $\mathcal{P}_q$ has nice properties.

Lemma B.1.

1. *The solution correspondence $x^*(q)$ of $\mathcal{P}_q$ is nonempty-valued, compact-valued, and upper-hemicontinuous.*
2. *$V(q)$ is continuous in q.*

Proof. Fix q. Note that $\mathcal{P}_q$ can be reformulated as:

$$\begin{aligned} \max_{\chi \geq 0} \quad & \sum_{a,\omega} v(a,\omega)\chi(a,\omega) \\ \text{such that} \quad & \sum_{\omega} \big(\pi(a,\omega) - \pi(a',\omega)\big)\chi(a,\omega) \geq 0 \qquad \forall\, a, a' \in A. \\ \text{and} \quad & \sum_{a} \chi(a,\omega) = q(\omega) \qquad \forall\, \omega \in \Omega \end{aligned} \tag{B.1}$$

In this problem, the objective is continuous in χ and the feasible set is nonempty (because $\chi(\omega,\omega) = q(\omega)$ is always feasible) and compact. Therefore, the solution correspondence is nonempty- and compact-valued. The continuity of $\chi^*(q)$ is a special feature of the linear programming and the fact that q only appears on the right-hand side of the constraints. This follows from Theorem 2 in Böhm (1975). Since χ^* is continuous in q, the optimal policy admits a continuous choice. This implies that $V(q)$ is continuous in q, since the objective is continuous in χ.

Note that x is a solution to $\mathcal{P}_q$ if and only if $\chi(a,\omega) := x(a|\omega)q(\omega)$ is a solution to (B.1), we claim that $x^*(q)$ is upper-hemicontinuous. It is clear that x^* is closed-valued, so we only need to show it has a closed graph. Take any $(q_n, x_n) \to (q, x)$ such that $x_n \in x^*(q_n)$, we want to show $x \in x^*(q)$. Note that $\chi_n(a,\omega) \to \chi(a,\omega) := x(a|\omega)q(\omega)$. By continuity of χ^* we know $\chi \in \chi^*(q)$ and thus $x \in x^*(q)$. $\square$

In light of Lemma B.1, we are ready to prove the existence of an equilibrium.

Proposition B.1. *An equilibrium of the competitive economy exists.*

Proof. We start by introducing a correspondence whose fixed points characterize the set of competitive equilibria. Let $P = [-M, M]^{|\Omega|}$ be the space of possible equilibrium prices, where M is chosen to be large so that any possible equilibrium prices are within that range. Let $Q \times X$ be the space of feasible data allocations. Taken together, $P \times Q \times X$ is a nonempty, compact, and convex set. Define a correspondence $F : P \times Q \times X \rightrightarrows P \times Q \times X$ such that $(p', q', x') \in F(p, q, x)$ if:

1. x' solves problem $\mathcal{P}_q$.

2. q' solves the consumers' problem given (p, x).[19]

3. p' is such that q solves the platform's first-stage problem (1).

Note that (p, q, x) is a competitive equilibrium if and only if it is a fixed point of F. Therefore, a competitive equilibrium exists if F admits a fixed point. Toward this, we first prove the following claim and then apply Kakutani's fixed point theorem.

Claim. *F is nonempty-valued, convex-valued, and has a closed graph.*

Proof of the Claim. We first show that F is nonempty-valued. Fix any (p, q, x). By Lemma B.1, $\mathcal{P}_q$ admits a solution x'; given (p, x), the consumers' problem always has a solution q'; given q, since $\mathcal{P}_q$ admits an optimal solution, by strong duality $\mathcal{P}'_q$ also admits an optimal solution. Lemma A.2 then implies that a price p' under which q solves the platform's problem exists. Therefore, $(p', q', x') \in F(p, q, x)$.

Next we show F is convex-valued. Note that by definition of F, given (p, q, x), the choice of p', q', and x' are independent with each other. Therefore, it is sufficient to check convexity for each dimension. If x' and x'' both solve $\mathcal{P}_q$, clearly any convex combination also solves it; If q' and q'' both solve the consumers' problem, then any convex combination also solves the consumers' problem. To see this, if under (p, x) consumer ω has a strict preference, then $q'(\omega) = q''(\omega)$. if under (p, x) consumer ω is indifferent, then any $q(\omega)$ is optimal. If under

[19]Formally, we should impose market clearing saying that $z' = q'/\bar{q}$ solves the consumers' problem. We skip this step to abbreviate notation.

both p' and p'', q solves the platform's first-stage problem, then by Lemma A.2 we know both p' and p'' are solutions to $\mathcal{P}'_q$. Therefore, any convex combination of them is still a solution to $\mathcal{P}'_q$. Again by Lemma A.2, q solves the platform's first-stage problem under that convex combination.

Finally, we argue F has a closed graph. Suppose $(p_n, q_n, x_n) \to (p, q, x)$, $(p'_n, q'_n, x'_n) \to (p', q', x')$, and $(p'_n, q'_n, x'_n) \in F(p_n, q_n, x_n)$. We want to show $(p', q', x') \in F(p, q, x)$. By Lemma B.1, we know the solution correspondence of $\mathcal{P}_q$ is upper-hemicontinuous, so x' is a solution to $\mathcal{P}_q$; To see q' solves the consumers' problem, note that for all ω and $z \in [0, \bar{q}(\omega)]$:

$$q'_n(\omega)(p_n(\omega) + \sum_a u(a,\omega)x_n(a|\omega)) + (\bar{q}(\omega) - q'_n(\omega))r(\omega)$$
$$\geq z(p_n(\omega) + \sum_a u(a,\omega)x_n(a|\omega)) + (\bar{q}(\omega) - z)r(\omega)$$

By continuity we get:

$$q'(\omega)(p(\omega) + \sum_a u(a,\omega)x(a|\omega)) + (\bar{q}(\omega) - q'(\omega))r(\omega)$$
$$\geq z(p(\omega) + \sum_a u(a,\omega)x(a|\omega)) + (\bar{q}(\omega) - z)r(\omega)$$

Therefore, q' is optimal for the consumers given (p, x); To see under p', q solves the platform's problem, note that for all $\tilde{q} \geq 0$:

$$V(q_n) - \sum_\omega p'_n(\omega)q_n(\omega) \geq V(\tilde{q}) - \sum_\omega p'_n(\omega)\tilde{q}(\omega)$$

Since V is continuous by Lemma B.1, taking limit we get:

$$V(q) - \sum_\omega p'(\omega)q(\omega) \geq V(\tilde{q}) - \sum_\omega p'(\omega)\tilde{q}(\omega).$$

This completes the proof that $(p', q', x') \in F(p, q, x)$. □

Using the Claim, we can apply Kakutani's fixed-point theorem to F and conclude that F admits a fixed point. Therefore, a competitive equilibrium exists. □

C Complete Equilibrium Characterization for Section 3.1

In this section, we characterize the entire set of equilibria for our example from Section 3.1. We first note that in order for the platform's problem to admit a solution, we must have $p^*(1) \geq$

$0, p^*(2) \geq 0, p^*(1) + p^*(2) \geq \gamma_u$. Moreover, in order for the platform to trade, we must have $p^*(1) + p^*(2) = \gamma_u$.

Case 1: $2\bar{r} - 1 < \gamma_u < \bar{r}$**.** The unique equilibrium allocation is no trade, i.e., $q^*(\omega) = 0$ for all ω, and it is supported by a price vector p^* satisfying:

$$p^*(1) \in [0, \bar{r}] \qquad \text{and} \qquad p^*(2) \in [\max\{0, \gamma_u - p^*(1)\}, \bar{r}]. \tag{C.1}$$

Next we explain why this is the solution. Note that type-1 consumers are willing to sell only if $p^*(1) \geq \bar{r} > \gamma_u$. However, in this case we cannot have $p^*(1) + p^*(2) = \gamma_u$. Therefore, the unique equilibrium allocation is $q^*(\omega) = 0$. It can be supported by $x^*(\omega|\omega) = 1$. It can be checked that with the prices in (C.1), it is optimal for the consumers not to sell and for the platform not to buy. Any other price will induce some type of consumers to strictly prefer selling.

Case 2: $\gamma_u > 2\bar{r}$**.** There is a unique equilibrium such that: $p^*(1) = \gamma_u$ and $p^*(2) = 0$; $z^*(1) = 1$ and $z^*(2) = \min\{1, \frac{\bar{q}(1)}{\bar{r}\bar{q}(2)}\}$; $q^*(1) = \bar{q}(1)$ and $q^*(2) = \min\{\bar{q}(2), \frac{\bar{q}(1)}{\bar{r}}\}$; $x^*(1|1) = 1$ and $x^*(1|2) = \frac{q^*(1)}{q^*(2)}$. It can be easily checked that this is an equilibrium. To show uniqueness, note that since $\gamma_u > 2\bar{r}$, at least one type has a strict incentive to sell because $p^*(1) + p^*(2) \geq \gamma_u$. If type-1 has a strict incentive, we have $q^*(2) \geq \min\{\bar{q}(2), \frac{\bar{q}(1)}{\bar{r}}\}$ since $p^*(2) \geq 0$, but this requires $p^*(1) = \gamma_u, p^*(2) = 0$; if type-2 has a strict incentive, in order for the platform to be willing to buy, we must have $p^*(1) = \gamma_u, p^*(2) = 0$.

Case 3: $\bar{r} < \gamma_u \leq 2\bar{r}$**.** It can be easily checked that both equilibria of Case 1 and Case 2 continue to be an equilibrium in this case. Next we argue those are all possible equilibria. On one hand, for the equilibria with no trade, the price has to satisfy (C.1), otherwise some type will have a strict incentive to sell. On the other hand, given any equilibrium with trade, we must have $p^*(1) + p^*(2) = \gamma_u$. Moreover, we must have $q^*(1) > 0$, otherwise the platform is not willing to buy $\omega = 2$ at a positive price while type-2 consumers are not willing to sell at 0 price. If $q^*(1) > 0$, we must have $q^*(2) \geq \min\{\bar{q}(2), \frac{q^*(1)}{\bar{r}}\}$ because $p^*(2) \geq 0$ and the platform will choose $x^*(1|1) = 1, x^*(1|2) = \frac{q^*(1)}{q^*(2)}$. Since $q^*(2) > q^*(1)$, in order for the platform to be willing to buy, it must be the case that $p^*(1) = \gamma_u, p^*(2) = 0$. The unique equilibrium with trade then follows.

Case 4: $\gamma_u = \bar{r}$**.** In this case, the equilibria with no trade is the same as Case 1. The equilibria with trade satisfy $p^*(1) = \gamma_u = \bar{r},\ p^*(2) = 0$ with

$$0 < q^*(1) \le \bar{q}(1),\ q^*(2) = \min\{\bar{q}(2), \frac{q^*(1)}{\bar{r}}\}.$$

It is easy to check these are equilibria. To see these capture all equilibria with trade, we can follow the same argument as Case 3 to derive the unique equilibrium price under trade: $p^*(1) = \gamma_u = \bar{r},\ p^*(2) = 0$. With these prices, since type-1 consumers are indifferent, any $0 \le q^*(1) \le \bar{q}(1)$ is optimal for them. $q^*(2)$ is then pinned down by the indifference condition of type-2 consumers.

To sum up, the constrained efficient allocation $q^\circ(1) = q^\circ(2) = \bar{q}(1)$ can never be an equilibrium, so all equilibria of this competitive economy are inefficient.

D Social Welfare

In the main text, we focused on a notion of welfare that excludes the merchant's profit (see Equation (2)). In this section, we show that an analogous result to Proposition 2 holds if we allow the welfare function to include the merchant's profit. We refer to this as the "social welfare," defined as

$$SW(q,x) = \sum_{a,\omega} \Big(v(a,\omega) + u(a,\omega) + \pi(a,\omega) \Big) x(a|\omega) q(\omega) + \sum_{\omega} \Big(\bar{q}(\omega) - q(\omega) \Big) r(\omega). \tag{D.1}$$

In light of this, the new efficiency benchmark is as follows.

Definition 5. *An allocation (q°, x°) is **constrained socially efficient** if it solves*

$$\begin{aligned} \max_{q,x} \quad & SW(q,x) \\ \textit{such that} \quad & q \le \bar{q}, \\ \textit{and} \quad & x \textit{ solves } \mathcal{P}_q. \end{aligned}$$

Remark D.1. *Define "unconstrained social efficiency" by replacing the second constraint in Definition 5 with obedience constraints, in the spirit of Definition 4. Note that the notion of "constrained social efficiency" and "unconstrained social efficiency" are equivalent. This is because when trading-off π and u, the planner uses weights $1 + \gamma_\pi$ and $1 + \gamma_u$ and the platform uses weights γ_π and γ_u. Therefore, as shown in Figure 1, their optimal choice of x coincides given any database q. In this sense, their incentives are aligned in choosing x.*

Indeed, whenever the platform's incentive in choosing x is aligned with the planner's preference, constrained efficiency is equivalent to unconstrained efficiency. For example, when the planner does not take into account the merchant's payoff as in the main text, the platform's incentive is aligned with the planner when either $\gamma_\pi \geq 1 + \gamma_u$ or $\gamma_u \geq \gamma_\pi$. In these cases, the positive statements of Proposition 2, 3, and 4 can be equivalently stated in terms of unconstrained efficiency (cf. Definition 4).

Under this new criterion of social efficiency, we have the following result, which extends Proposition 2 to this more demanding efficiency benchmark.

Proposition D.1. *Let (p^*, z^*, q^*, x^*) be an equilibrium of the competitive economy. If $\gamma_\pi > \gamma_u$ and, in addition, $r(\omega) \notin [\gamma_\pi \omega, (1+\gamma_\pi)\omega)$ for all ω, the equilibrium allocation (q^*, x^*) is constrained socially efficient. Otherwise, the equilibrium allocation can be socially inefficient.*

Proof. We prove sufficiency here. Let (p^*, z^*, q^*, x^*) be a competitive equilibrium. By Proposition 2, we know the equilibrium allocation (q^*, x^*) is constrained efficient. Moreover, following the argument in the proof of Proposition 2, we also know that $x^* = \hat{x}$, where $\hat{x}(\omega|\omega) = 1$ is the full-disclosure mechanism, is the unique optimal mechanism for the platform given any q. Therefore,

$$\begin{aligned} q^* \in \quad & \arg\max_{q \leq \bar{q}} \sum_{a,\omega} \Big(v(a,\omega) + u(a,\omega) \Big) \hat{x}(a|\omega) q(\omega) - \sum_{\omega} r(\omega) q(\omega) \\ = \quad & \arg\max_{q \leq \bar{q}} \sum_{\omega} \Big(\gamma_\pi \omega - r(\omega) \Big) q(\omega). \end{aligned}$$

The solution to this problem is $q^*(\omega) = \bar{q}(\omega)$ if $\gamma_\pi \omega > r(\omega)$, $q^*(\omega) = 0$ if $\gamma_\pi \omega < r(\omega)$, and $q^*(\omega) \in [0,1]$ if $\gamma_\pi \omega = r(\omega)$. The constrained socially efficient allocation (q°, x°) also features $x^\circ = \hat{x}$. Therefore, the solution of the planner's problem is $q^\circ(\omega) = \bar{q}(\omega)$ if $(1+\gamma_\pi)\omega > r(\omega)$, $q^\circ(\omega) = 0$ if $(1+\gamma_\pi)\omega < r(\omega)$, and $q^\circ(\omega) \in [0,1]$ if $(1+\gamma_\pi)\omega = r(\omega)$. When $r(\omega) \notin [\gamma_\pi \omega, (1+\gamma_\pi)\omega)$ for all ω, the equilibrium allocation (q^*, x^*) is also a solution to the planner's problem, and thus constrained socially efficient. □

Intuitively, if we take into account the merchant's profit, the inefficiency can arise from two sources. The first one is the pooling externality discussed in the main text. When $\gamma_\pi > \gamma_u$, the only optimal mechanism for the platform given any q is full disclosure, so the pooling externality disappears. The second one is a traditional externality. Since the platform does not take into account the merchant's payoff, it refuses to buy data when the price is high, even when trade is still socially optimal. When the sufficient condition of the proposition is not satisfied, the equilibrium can be inefficient.

Next, we illustrate the two sources of inefficiency using the example of Section 3.1. We will denote the constrained efficient allocation by (q°, x°). We also denote the equilibrium allocation in Case 1 (inefficiently low trade) by (q^*_L, x^*_L) and in Case 2 (inefficiently high trade) by (q^*_H, x^*_H). These are characterized in Section 3.1.

We first argue that in both cases, the social welfare of the equilibrium, $SW(q^*, x^*)$, is strictly lower than $SW(q^\circ, x^\circ)$. As before, this is originated from the pooling externality. Using the characterizations in Section 3.1, we can directly compute:

$$SW(q^\circ, x^\circ) = \bar{q}(1)(3+\gamma_u) + \bar{r}(\bar{q}(2) - \bar{q}(1)),$$

$$SW(q^*_L, x^*_L) = \bar{r} < SW(q^\circ, x^\circ),$$

$$SW(q^*_H, x^*_H) = \bar{q}(3+\gamma_u) + \bar{r}\max\{0, \bar{q}(2) - \frac{\bar{q}(1)}{\bar{r}}\} < SW(q^\circ, x^\circ).$$

The take is that, even if we measure efficiency using social welfare (Equation (D.1)), the equilibria are still suboptimal compared to the constrained efficient allocation. One may suspect that in Section 3.1, the inefficiency is an artifact that we did not take into account the merchant's profit, but as we highlight here, that is not the case.

In addition to the pooling externality, there is a new source of inefficiency: since in this case we have $(1+\gamma_\pi)\omega > \bar{r} > \gamma_\pi\omega = 0$, even (q°, x°) is not constrained socially efficient. The social welfare is maximized at $q^\bullet(\omega) = \bar{q}(\omega)$ and $x^\bullet(1|1) = 1, x^\bullet(1|2) = \frac{\bar{q}(1)}{\bar{q}(2)}$, which gives a social welfare of

$$SW(q^\bullet, x^\bullet) = \bar{q}(1)(\gamma_u + 1) + 2(\bar{q}(2) - \bar{q}(1)) > SW(q^\circ, x^\circ).$$

Therefore, the constrained efficient allocation is not constrained socially efficient. This additional gap is created by the fact that the profit of the merchant is not taken into account by the platform or the consumers. This is a traditional externality that can arise even without the informational friction discussed in our paper. For instance, consider the case where there is only one type of consumers $\omega = 1$ with $0 < r(1) < 1$. The platform's objective has $\gamma_u > \gamma_\pi = 0$. Then the constrained socially efficient allocation is $q^\bullet(1) = 1$, but the only equilibrium is no trade.