Competitive Markets for Personal Data

Simone Galperti
UC, San Diego

Tianhao Liu
Columbia University

Jacopo Perego
Columbia University

March 19, 2024

Abstract

We study the optimal design of competitive markets for personal data. In a data market, consumers own their data and sell it to a platform in exchange for a price and a service, which consists of being interacted with a third-party merchant, from whom they can buy a product. We study the equilibrium properties of this market and its ability to promote efficient data allocations. Our main result identifies a novel inefficiency that leads this otherwise perfectly competitive market to fail. We show how the underlying inefficiency critically depends on the platform's role as an information intermediary. We then discuss three solutions to this market failure: establishing a data union, implementing data taxes, and allowing the price of data to depend on more than just the type of data being traded.

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, Nima Haghpanah, Kevin He, Dana Foarta, Matthew Gentzkow, Alessandro Lizzeri, 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, for instance, 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 imperfect arrangement could harbor inefficiencies and increase inequality (Seim et al., 2022). 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 sufficient and necessary conditions under which this otherwise perfectly competitive data market can fail. This inefficiency stems from an externality that consumers exert on each other when selling their data, which relates to and builds upon earlier work by Galperti, Levkun, and Perego (2023). 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

[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.

she can buy a product. As an intermediary, the platform can use the database it has acquired from consumers to provide information to the merchant about these consumers' willingness to pay for the merchant's product, just as 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 generate 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 sufficient and necessary conditions for efficiency and show how they ultimately 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 can become inefficient. In some cases, the data market can entirely unravel, resulting in no data being traded and, thus, the lowest possible consumer welfare. 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 can 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. However, 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

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; Seim et al., 2022).

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 it 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 micro-

is distinct from and complementary to externalities that originate from exogenous correlation among consumers' data, which are analyzed by Choi et al. (2019), Acemoglu et al. (2022), and Bergemann et al. (2022).

foundation of some of the key components of a data economy: How the data is *used* by the 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. (2022), 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).

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.

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, she is paid price $p(\omega)$ by the platform and is later intermediated 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 $\zeta(\omega) \in [0,1]$. 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

[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.

the 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 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 $v(a,\omega) = \gamma_u u(a,\omega) + \gamma_\pi \pi(a,\omega)$, i.e., a linear combination of the consumer's trading surplus and the merchant's profits. We assume $\gamma_u, \gamma_\pi \geq 0$, with at least one strict inequality.

By standard arguments (e.g., 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 ζ 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 the competitive economy as follows.

Definition 1. *A profile (p^*, ζ^*, 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^*, ζ^* solves the consumers' problem in the first period. That is, for all ω,

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

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

Conditions (a) and (b) require that the platform acquires a database that maximizes its payoff given prices, while anticipating it will use its data optimally in the second period. Condition (c) requires that each type-ω consumer chooses $\zeta(\omega)$ optimally, 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.[4]

2.1 Efficiency Benchmark

How efficient is the competitive economy? In our baseline analysis, the welfare notion that we use to measure efficiency is the aggregate payoff of the platform and the consumers, namely

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

Note that, since the prices p of data records only affect the distribution of payoffs between the platform and the consumers, aggregate welfare only depends on the database q and the mechanism x. We refer to the pair (q,x) as an *allocation* of the economy.

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

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

In this benchmark, an allocation is constrained efficient if it maximizes welfare subject to two constraints. The first requires that the database is feasible, i.e., it does not allocate to the

[4] In the appendix, we show that an equilibrium of the competitive economy exists (Proposition B.1). While somewhat expected, this result is not immediate because—unlike in the typical Walrasian equilibrium—the consumers' and the platform's payoffs depend both on the data allocation q and on the platform's mechanism x.

platform more records than those that exist in the economy. The second constraint requires the mechanism x to be sequentially optimal for the platform given q.[5]

We briefly motivate the efficiency benchmark of Definition 2. As anticipated in the introduction, our main goal in the paper will be to show that, under certain conditions, the equilibrium of the competitive economy is inefficient. Towards this goal, a less demanding efficiency benchmark is more desirable, as it makes such a negative result starker and allows us to ignore sources of inefficiencies that are neither new nor surprising. In particular, there are two aspects of Definition 2 that make it less demanding. First, we focus on "constrained" efficiency, i.e., we require that the mechanism x is sequentially rational for the platform given q. Dropping this constraint would lead us to detect inefficiencies that are 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 calculation. To be constrained efficient, an outcome has to maximize only the sum of the platform's and the consumers' payoffs as opposed to the sum of all agents' payoffs. Were we to include the merchant's payoff, we would detect a rather standard source of inefficiency, which materializes in the downstream interaction between the platform and the merchant. Specifically, the platform does not sell information to the merchant and, thus, it does not fully internalize how the choice of x affects the merchant's payoff (except when $\gamma_\pi = 1$). We relegate to Appendix D the analysis of constrained efficiency when $\mathcal{W}(q,x)$ also includes the merchant's payoff.

An additional 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 since $\mathcal{W}(q^*,x^*) = W^\circ$.

[5] We note that a constrained efficient outcome always exists. This is because $\mathcal{W}$ is continuous and, by Lemma B.1, the feasible set of outcomes in the planner's problem is nonempty and compact.

[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 $V(q') - \sum_\omega p^*(\omega)q'(\omega) = \alpha\big(V(q^*) - \sum_\omega p^*(\omega)q^*(\omega)\big) > V(q^*) - \sum_\omega p^*(\omega)q^*(\omega)$.

2.2 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. The substantive aspect of this assumption is that the platform is a price taker and, thus, the economy is competitive—a distinguishing characteristic of our paper. The focus on a single platform, rather than a finite number of identical ones, is expositional, as it simplifies notation at little cost of generality. Galperti and Perego (2022) show how to model a competitive economy with finitely many competing platforms.

The platform's objective is assumed to be linear in the consumers' trading surplus and the merchant's profits. This specification guarantees tractability, while capturing key features of real-world two-sided markets (see Xu and Yang (2023) for a dynamic microfoundation of such an objective). All 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, for example, settings where the consumer can bypass the platform and trade directly with the merchant. While with a different goal, Bergemann and Bonatti (2023) studies the combination of on- and offline 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. In some settings, this assumption is restrictive. See Ali et al. (2022) for a model where these two aspects are unbundled. Third, selling the data record fully reveals the underlying consumer's type. By construction, this implies that consumers' data is uncorrelated: That is, i.e., if a consumer sells her data, the platform does not learn anything about other consumers. This is an important departure relative to, e.g., Acemoglu et al. (2022) and allows us to focus on a novel source of market failure.[7]

3 The Inefficiency of the Data Economy

In this section, we present a series of results that identify necessary and sufficient conditions for equilibrium efficiency and uncover what are the key drivers of the inefficiency of our com-

[7] 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) (Section 4).

petitive economy.

We begin by introducing a notion that is instrumental for our analysis: the social benefit of allocating a data record to the platform's database. To compute it, fix an arbitrary database q and consider the following maximization problem:

$$W(q) \triangleq \max_{x:\Omega\to\Delta(A)} \sum_{a,\omega}(v(a,\omega)+u(a,\omega))x(a|\omega)q(\omega) \quad \text{such that} \quad x \text{ solves } \mathcal{P}_q(v). \tag{3}$$

We can think of (3) 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 $W(q)$. Each $\psi_q(\omega)$ captures how $W(q)$ changes when we add an additional ω-record to database q. In other words, $\psi_q(\omega)$ identifies the *social benefit* of allocating an additional ω-record into the platform's database.[8]

Our first result demonstrates how the social benefit of a data record 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 consumer sells her record to the platform if and only if its social benefit exceeds its private 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^*,ζ^*,q^*,x^*) and denote by

$$U^*(\omega) \triangleq p^*(\omega) + \sum_a x^*(a,\omega)u(a,\omega) \tag{4}$$

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

[8] 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. In particular, it shows that Ψ_q is generically a singleton and easy to compute.

$U^*(\omega) \leq r(\omega)$ 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, 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 $\xi_{q^*} \triangleq \psi_{q^*} - p^*$ and write

$$\psi_{q^*}(\omega) = p^*(\omega) + \xi^*(\omega). \tag{5}$$

This decomposition of the social benefit—which is analogous to the definition of U^* in equation (4)—has an important economic interpretation: $p^*(\omega)$ and $\xi^*(\omega)$ 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^*, ζ^*, q^*, x^*), p^* is a supergradient of $V(q^*)$.*[9]

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 marginal cost $p^*(\omega)$. Thus, the remaining component of the social benefit—namely $\xi^*(\omega)$—captures the marginal change in the trading surplus of all consumers resulting from adding an additional ω-record to the database q^*.

By comparing Equations (4) and (5), it is clear that a sufficient and necessary condition for equilibrium efficiency is the alignment between $\sum_a x^*(a,\omega)u(a,\omega)$ and $\xi^*(\omega)$. The reason is that, when a type-ω consumer sells her record to the platform, she expects to earn a trading surplus of $\sum_a x^*(a,\omega)u(a,\omega)$ but does not internalize the effect she imposes on the trading surplus of all other consumers—namely $\xi^*(\omega)$. In other words, her decision to sell her record may exert an externality on other consumers, thus introducing inefficiency in the economy.

The following result shows that the presence of these externalities, and thus the efficiency of the economy, hinges on the platform's objective.

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.*

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

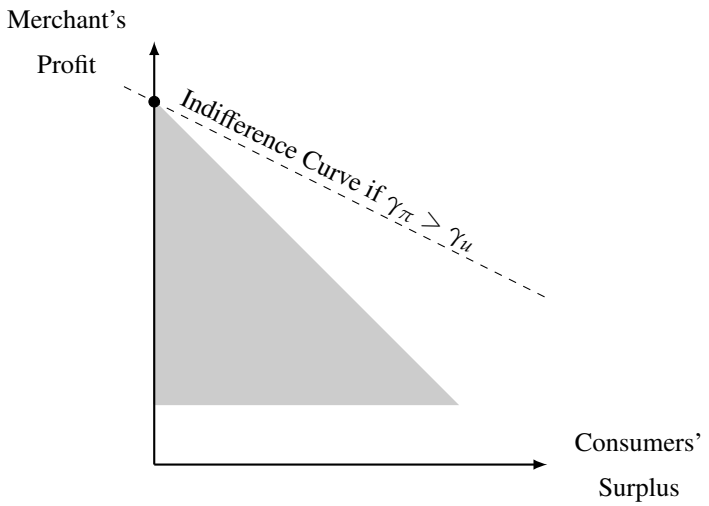

Figure 1: The trading surplus triangle. The dashed line depicts the platform's indifference curve when $\gamma_\pi > \gamma_u$.

Perhaps counterintuitively, if the platform cares relatively more about the merchant's profits, the social and private benefits of data records are aligned. Thus, any equilibrium in this case is constrained efficient and consumers' welfare is maximized. 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, both $\sum_a x^*(a,\omega)u(a,\omega)$ and, a fortiori, $\xi^*(\omega)$ equal zero. 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 pooling, that is, the platform withholds some information from the merchant to prevent excessive surplus extraction (again, see Figure 1). Consequently, $\xi^*(\omega)$ and $\sum_a x^*(a|\omega)u(a,\omega)$ can differ. In other words, when an ω-consumer sells her record she may fail to internalize the externality she creates on other consumers. The externality exerted by low-type consumers is

typically positive. For example, consider a consumer of type $\underline{\omega} \triangleq \min_{\omega} \Omega$. The expected trading surplus of this consumer—namely, $\sum_a x^*(a|\underline{\omega})u(a,\underline{\omega})$—is zero since the merchant will never want to charge a fee a lower than ω. Yet, $\xi^*(\underline{\omega})$ 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.

We conclude this discussion by finding sufficient conditions under which *all* equilibria are inefficient, thus sharpening the negative part 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 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 $r(\underline{\omega})$. 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.

[10] When $W^\circ = R$, all equilibria are constrained efficient. Indeed, given any equilibrium (p^*,ζ^*,q^*,x^*), since the platform always earns a zero payoff, the consumers' surplus equals $\mathcal{W}(q^*,x^*)$, which is bound to be between R and W°. When $W^\circ = R$, all these values are equal so constrained efficiency is always attained.

3.1 An Example

We conclude this section by illustrating with a simple example why the data economy can be inefficient. We consider an economy with only two types of consumers, $\Omega = \{1,2\}$, and $\bar{q}(2) > \bar{q}(1)$. All consumers have the same reservation utility, i.e., $r(\omega) = \bar{r} \in (0,1)$ for all ω. The platform only cares about consumers' trading surplus, i.e., $\gamma_u > \gamma_\pi = 0$. 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.

To do so, we first characterize constrained efficient allocations for this economy. 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 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 of the economy and show that they are inefficient. To this purpose, let (p^*, ζ^*, q^*, x^*) be an equilibrium and denote by a_{q^*} the fee the merchant would charge if the platform would not provide him with any additional information besides the database composition. Using Lemma A.1 in Appendix A, we can compute two variables that will be useful in our equilibrium characterization. The marginal change in the consumers' trading surplus that result from adding an ω-record to q^* is given by

$$\xi^*(\omega) = \omega - a_{q^*}\mathbb{1}(\omega \geq a_{q^*}). \tag{6}$$

[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 $W^\circ \geq W(q^*, x^*) \geq \bar{r}$.

The equilibrium price, instead, is given by

$$p^*(\omega) = \gamma_u \xi^*(\omega) = \gamma_u \Big(\omega - a_{q^*} \mathbb{1}(\omega \geq a_{q^*}) \Big). \tag{7}$$

Notice that, since $1 \leq a_{q^*} \leq 2$, we have $0 \leq \xi^*(\omega) \leq 1$ 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 of the economy involves no trading. To show this, we first argue that, in any equilibrium, there is no trade of 1-records, i.e., $q^*(1) = 0$. Indeed, type-1 consumers always earn a zero trading surplus when they sell their records. Moreover, by Equation (7), the price $p^*(1)$ they receive in return can be no higher than γ_u. Therefore, their net payoff is no higher than γ_u, which by assumption is strictly smaller than $\bar{r}$. This implies that type-1 consumers do not sell their data to the platform in any equilibrium, i.e., $q^*(1) = 0$. 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 (7) 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.[12]

Why are equilibria inefficient in this case? 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 low-type consumer needs to be paid a market price $p^*(1)$ that is higher than her reservation utility $\bar{r}$. Unfortunately, the platform is unwilling to pay such a high price, since the value of a 1-record from its perspective is, at most, $\gamma_u < \bar{r}$. △

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

[12] This inefficiency is foreshadowed by Corollary 1. Indeed, this example satisfies the sufficient conditions of this result, since $\gamma_u < \bar{r} < \frac{1+\gamma_u}{2} < 1 + \gamma_u$.

a payoff of γ_u, which would exceed the price it paid for the two records. This implies that at least one between $p^*(1)$ and $p^*(2)$ must exceed $\frac{\gamma_u}{2}$. We argue that $p^*(2) \leq \frac{\gamma_u}{2}$. Indeed, were this not the case, all high-type consumers would strictly prefer to sell, leading to an uninformed merchant price of $a_{q^*} = 2$, which by Equation (7) implies $p^*(2) = 0$, a contradiction. Therefore, let $p^*(1) \geq \frac{\gamma_u}{2}$. In this case, the low-type consumers strictly prefer to sell and, thus, $q^*(1) = \bar{q}(1)$. Just as in the constrained efficient allocation, the optimal mechanism given q^* involves setting $x^*(1|2) = \min\{\bar{q}(1), q^*(2)\}/q^*(2)$. That is, the platform creates a segment that includes all the low-type consumers and as many high-type ones as possible subject to inducing the lowest 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 (7), $p^*(2) = 0$. Thus, $q^*(2) = \min\{\bar{q}(2), \frac{\bar{q}(1)}{\bar{r}}\}$. We conclude that the unique equilibrium of this economy is inefficient since its induced welfare is $\mathcal{W}(q^*, x^*) = (1+\gamma_u)\bar{q}(1) + \max\{0, \bar{r} - \bar{q}(1)(1+\bar{r})\} < W^\circ$.

In this equilibrium, too many high-type consumers sell their record to the platform relative to what is efficient, i.e., $q^*(2) > q^\circ(2)$. 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. Specifically, she undermines the chances that other high-type consumers can buy the merchant's product at the lowest fee. More formally, notice that, on the one hand, such a consumer individually gains from selling her record to the platform, since $\sum_a x^*(a|2)u(a,2) = q^*(1)/q^*(2) \geq \bar{r}$. On the other hand, however, she does not help increase the aggregate trading surplus earned by all consumers. Indeed, $\xi^*(2) = 0$ (by Equation (6)). Therefore, this consumer's individual gain must come at the expense of the gains of other consumers.[13] △

We relegate the third parametric case, i.e., $\frac{\gamma_u}{2} \leq \bar{r} \leq \gamma_u$, to Appendix C. This case shares similar intuitions to the two cases already discussed. In a nutshell, it features multiple equilibria, inducing either inefficiently low or inefficiently high trade. In conclusion, our analysis shows that all equilibria of this simple economy are inefficient. We summarize these results in

[13] It is natural to wonder whether a negative price for 2-records could correct this inefficiency, as it would disincentivize the sale of these records. Such a negative price would be incompatible with equilibrium behavior, as the platform would then demand an infinite quantity of 2-records. More generally, we refer to Corollary A.1 in Appendix A for sufficient conditions in an arbitrary economy that trigger a similar kind of inefficiency.

Figure 2: Equilibrium Inefficiency in the Example of Section 3.1

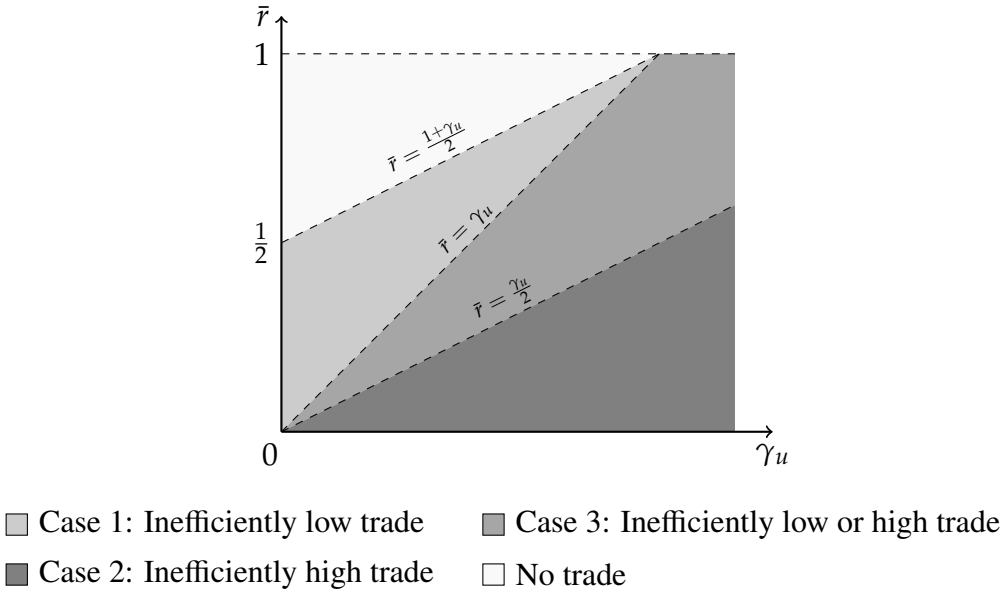

Figure 2.

4 Remedies to the Inefficiency

This section discusses alternative market designs that provide a remedy to the inefficiency of the competitive economy. Specifically, we propose three solutions. The first consists of establishing a "data union" that manages consumers' data records on their behalf. The second introduces data tax as a way to decentralize the data union. The third consists of making data markets "more complete" in which, in the spirit of Arrow (1969), the consumers and the platform can trade the externalities they create on each other.

4.1 The Role of a Data Union

We begin by discussing the consequences of introducing a new intermediary in the economy, which we call a *data union*.[14] A data union represents consumers and manages their data records on their behalf. That is, it collects data records from participating consumers, sells some or all of them to the platform, and distributes the proceeds back to the consumers, as

[14]The potential policy role of data unions has been discussed by Posner and Weyl (2018) and Seim et al. (2022).

compensation. The data union coordinates consumer actions, by unilaterally deciding which records should be sold to the platform and how consumers should be compensated. By doing so, the data union substitutes, in part, the competitive market. We show that a data union can implement allocations that are constrained efficient.

More specifically, we consider a data union that operates as follows. Consumers voluntarily decide whether to become members of the data union. If they do, they retain their reservation utility $r(\omega)$ unless the data union decides to sell their records to the platform. If they do not, they cannot unilaterally sell their record to the platform and, thus, just obtain their reservation utility $r(\omega)$. The union collects a database $\hat{q}$ and then sells a subset of it, $q \leq \hat{q}$, to the platform, at a price that extracts all the platform's expected payoff, namely, $V(q)$. Finally, the union distributes the proceeds $V(q)$ to its members. That is, the union chooses $p \in \mathbb{R}^{\Omega}$ such that $\sum_{\omega} p(\omega)\bar{q}(\omega) = V(q)$. Note that payment to a consumer can be negative and can depend on her type. The union maximizes the welfare of its members. Without loss of generality, we can assume that $\hat{q} = \bar{q}$, namely, the union acquires the data records of all consumers. We can write the problem of the data union as follows:

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

The second-to-last constraint requires that the data union makes no profits. The last constraint, instead, ensures that each type-ω consumer has no incentive to leave the union. By participating in the union, this consumer receives a compensation of $p(\omega)$. Additionally, with probability $\frac{q(\omega)}{\bar{q}(\omega)}$, her record is sold to the platform, in which case she obtains a payoff of $\sum_a u(a,\omega)x(a|\omega)$. With remaining probability, instead, her data record is not sold, and the consumer preserves 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.

The next result shows that, unlike equilibria of the competitive economy discussed earlier, the data union induces allocations that are constrained efficient.

Proposition 3. *Let (p^*, q^*, x^*) be a solution to the data union's problem. The allocation (q^*, x^*) is constrained efficient and, thus, maximizes consumers' welfare. 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 discussed in Section 3. First, while consumers can decide to leave the union, they have no say in whether their data records is sold to the platform. This allows the data union to coordinate consumers in a way that competitive markets cannot. Second, the data union has bargaining power, that is, prices p are chosen by the union rather than determined by market clearing. Thanks to this, the data union both internalizes the externalities discussed in the previous section and can properly compensate consumers for their participation, thus leading to efficient allocations that maximize consumers' welfare.

We conclude by noting that the fact that the data union's objective is to maximize consumers' welfare is not essential to its ability to induce constrained efficiency allocations. To see this, consider the data economy from Section 3 but assume the platform has bargaining power, i.e., it chooses p rather than taking it as given. It can be shown that such a platform induces allocations that are constrained efficient. Unlike the data union, this "monopsonist" platform has no incentives to distribute the proceeds of its activity back to the consumers. Therefore, while allocations in this setting are constrained efficient, consumers' welfare is minimized, i.e. it equals R.

4.2 Data Taxes

While the data union helps achieve constrained-efficient allocations, it requires bargaining power, thus abandoning the competitive nature of the economy we have analyzed so far. It is then natural to wonder whether constrained-efficient allocations can be induced in the context of a competitive economy? In this section, we do so by introducing data taxes, levied on consumers, and show that they can restore the efficiency of the competitive equilibria.

More specifically, we enrich our competitive economy by assuming that whenever a type-ω consumer sells her record to the platform, she pays a "data tax" $\tau(\omega) \in \mathbb{R}$ to the government. When $\tau(\omega) \leq 0$, this tax is interpreted as a subsidy paid by the government. To make the

problem interesting, let us assume the government cannot run a deficit.[15] Besides taxes, all other components of the model are unchanged relative to Section 2, including the definition of equilibrium. The next result shows that any constrained efficient allocation can be supported as an equilibrium of the economy with taxation.

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

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

$$\tau^*(\omega) \triangleq p^*(\omega) + \sum_a x^\circ(a|\omega)u(a,\omega) - r(\omega). \tag{8}$$

Additionally, define $\zeta^*(\omega) \triangleq q^\circ(\omega)/\bar{q}(\omega)$, for all ω. It is straightforward to check that $(p^*, \zeta^*, q^\circ, x^\circ)$ 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. Indeed, notice that 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 ζ^*, data markets clear. Therefore, $(p^*, \zeta^*, q^\circ, x^\circ)$ is an equilibrium that supports the constrained efficient allocation (q°, x°). In this equilibrium, consumer welfare equals R 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.

Example of a Data Tax. The data tax corrects the inefficiency of the competitive economy by using taxation to make 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. For instance, consider Case 1 discussed in Section 3.1. In that case, 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. Taxation restores efficiency by subsidizing these consumers just enough to make

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

them indifferent between selling or not. Specifically, $\tau^*(1) = \gamma_u - \bar{r} < 0$. This subsidy is financed by taxing the high-type consumers, namely $\tau^*(2) = 1 - \bar{r}$. 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 Making Data Markets More Complete

In this last subsection, we show how the inefficiency of the competitive economy can be corrected by enriching the price system, i.e., by letting the price of a data record depend on more than its type ω. We do so by allowing consumers to trade "the way" their records are used by the platform. In other words, at the time of trade, the platform and the consumer must agree on how the record will be used by the platform—i.e., which fee a it will recommend to the merchant.[16] This approach follows standard ways of modeling competitive economies with externalities (e.g., Arrow (1969) and Laffont (1976)). We refer to this setting as the Lindahl economy.

More formally, we define an economy featuring one market for each pair (a, ω). In such markets, ω-records can be traded for use a at a price $p(a, \omega)$. A type-ω consumer decides in which market to sell her ω-record, if any. That is, for all a, she chooses $\zeta(a, \omega) \in [0, 1]$, i.e., the probability of selling her record to the platform for use a. 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 needs to deliver on this promise. That is, the platform's problem can be written as:

$$\begin{aligned} \max_{q,x} \quad & \sum_{a,\omega} \Big(v(a,\omega) - p(a,\omega) \Big) x(a|\omega) q(\omega) \\ \text{such that} \quad & \sum_{\omega} \big(\pi(a,\omega) - \pi(a',\omega) \big) x(a|\omega) q(\omega) \geq 0 \qquad \forall\, a, a' \in A \end{aligned} \tag{9}$$

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

[16] 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)).

Lindahl economy has richer markets, with prices that depend on a and not just on ω.[17]

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

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

(a). Given p^*, (q^*, x^*) solves the platform's problem in (9).

(b). Given p^*, ζ^* solves the consumers problem. That is, for all ω,

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

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

Before presenting the main result of this section, we introduce a more demanding efficiency benchmark than the one we considered so far.

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 with the notion of constrained efficiency (Definition 2), an unconstrained efficient allocation (q, x) does not require x to solve $\mathcal{P}_q$ but, more simply, that x is obedient given q. That is, x does not need to be sequentially optimal for the platform given q. Therefore, the welfare induced by an unconstrained efficient allocation is weakly higher than that induced by 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^*, ζ^*, q^*, x^*) be an equilibrium of the Lindahl economy. The equilibrium allocation (q^*, x^*) is unconstrained efficient. Therefore, 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]

[17] In particular, the timing of the two economies is the same and the platform has no more commitment power in one or the other. To see this, note that we could have equivalently written conditions (a) and (b) in Definition 1 as Equation (9), with the additional restriction that $p(a, \omega) = p(\omega)$ for all (a, ω).

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

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. Since the economy is competitive, the platform still earns zero profit and, thus, consumer welfare is $W^\dagger$. 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 unconstrained-efficient allocations and constrained-efficient ones coincide, leading to a welfare of $W^\circ = W^\dagger = \bar{r} + \bar{q}(1)(1 + \gamma_u - 2\bar{r})$. Moreover, just as in Section 3.1, the unconstrained-efficient allocation $(q^\dagger, x^\dagger)$ is unique and is 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. To contrast this, we now discuss an equilibrium of the Lindahl economy and show it is unconstrained efficient. Let (p^*, ζ^*, 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 $\zeta^*(1, \omega) = \frac{\bar{q}(1)}{\bar{q}(\omega)}$ and $\zeta^*(2, \omega) = 0$, so that ζ^* 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 $p^*(a = 1, \omega = 2) = -(1 - \bar{r})$. We argue that this is an equilibrium of the Lindahl economy. To see this, 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^*.

Notice the crucial role of the price system in inducing an efficient allocation. 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 $p^*(1, \omega = 2) < 0$. Despite the negative price, the platform does not have an incentive to acquire an arbitrary quantity of such records because it needs to guarantee the merchant is willing to charge a low fee $a = 1$ to all of them. $\triangle$

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*.

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 (2022): "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. 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.

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.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp (2019): "Big Data and Firm Dynamics," *AEA Papers and Proceedings*, 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.

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.

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

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

By focusing on the linear specification $v = \gamma_u u + \gamma_\pi \pi$, we can explicitly compute V and W and thus characterize equilibrium prices and Ψ_q. We start by providing such characterizations. From Figure 1, we can explicitly compute

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

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

where $w(q) := \sum_\omega \omega q(\omega)$ is the market value and $\pi_m(q) := \max_a \sum_\omega \pi(a,\omega) q(\omega)$ is the merchant's profit when there is no information intermediation. Note that $w(q)$ is linear and $\pi_m(q)$ is convex, so V and W are concave functions and their supergradients are well-defined.

Next, we characterize the supergradients of V and W. Fix any $q \geq 0$. Let A_q be the set of maximizers of $\max_a \sum_\omega \pi(a,\omega) q(\omega)$. Let $\phi_a \in \mathbb{R}^\Omega$ be defined as $\phi_a(\omega) := \gamma_u \omega + (\gamma_\pi - \gamma_u) a \mathbb{1}(\omega \geq a)$. Define:

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

where *cov* refers to convex hull, and define:

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

$$\text{and } \phi(\omega) \geq \tilde{\phi}(\omega), \text{ if } q(\omega) = 0\}.$$

Then we have the following result.

Lemma A.1. *1. The set of supergradients of $V(q)$ at q is $\Phi_q(v)$.*

2. $\Psi_q = \Phi_q(v)$ *when* $\gamma_\pi > \gamma_u$; $\Psi_q = \Phi_q(v+u)$ *when* $\gamma_\pi \leq \gamma_u$.

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 case of $\gamma_\pi > \gamma_u$.

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 simply 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.

The second claim follows since from (A.1), when $\gamma_\pi > \gamma_u$, $W(q) = V(q)$; when $\gamma_\pi \leq \gamma_u$, Ψ_q can be derived in exactly the same way by replacing γ_u with $1 + \gamma_u$. □

Note that Φ_q and Ψ_q are generically unique. Moreover, Theorem 5.2 of Bertsimas and Tsitsiklis (1997) states that the set of supergradients of the value function is the same as the optimal solutions of the dual problem. Therefore, $\Phi_q(v)$ is also the set of solutions to the following 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.

With these notions, next we show Proposition 1 and Lemma 1.

Proof of Proposition 1. We first discuss the case where $\gamma_\pi > \gamma_u$. It is easy to see that in this case the planner's solution to ($\mathcal{SB}$) 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.

Next we discuss the case where $\gamma_\pi \leq \gamma_u$. In this case the constraint for x to be sequentially rational can be relaxed, and the planner's problem $\mathcal{SB}$ becomes $\mathcal{FB}$. A data allocation (q, x) is 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. Firstly, (μ, λ) achieves the minimum of $\mathcal{P}'_q(v + u)$, so $\mu \in \Phi_q(v + u)$. Secondly, $\mu(\omega) = r(\omega)$ whenever $q(\omega) < \bar{q}(\omega)$. Taking $\psi_q = \mu$, we conclude the "only if" direction.

Next we show 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:

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

In other words, the platform's optimal payoff is zero if (A.2) is feasible, and infinity 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^*, ζ^*, q^*, x^*), we must have $q^* \leq \bar{q}$ and q^* solves (1) given p^*.

Next, we turn to analyze the efficiency of the equilibria of the competitive data economy.

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^*, ζ^*, 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. □

Next we discuss two sufficient conditions under which no equilibrium can be constrained efficient. The two conditions correspond to the low-trade case and high-trade case in the example of Section 3.1. The first condition is stated as Corollary 1 and the second is stated as Corollary A.1.

Proof of Corollary 1. Let (p^*, ζ^*, 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 type-$\underline{\omega}$ will get a zero trading surplus on the platform, this means they will strictly prefer not to sell their data records. However, this is inefficient because given $q^* \neq 0$, by Lemma A.1 we know 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 sufficient condition in spirit of the high-trade case. The conditions are slightly convoluted, but they can be broken down as follows. Define ω_1 and the highest type and ω_2 as the second highest type. Firstly, we assume that as long as all ω_1 consumers sell their data, the merchant will strictly prefer to set the fee at ω_1. This helps us pin down the value of data, and can be guaranteed by the requirement that $\bar{q}(\omega_1)\omega_1 > \omega \sum_{\omega' \geq \omega} \bar{q}(\omega)$ for all $\omega < \omega_1$. Moreover, we assume that if the monopoly price is ω_1, then the platform will want to pool type-ω_2 consumers with type-ω_1 consumers to lower the fee for them. This is guaranteed by the requirement that $r(\omega_2) < \gamma_u \omega_2$. In such environments, we show that for a range of $r(\omega_1)$, type-ω_1 consumers will get an average surplus higher than their marginal contribution,

resulting in a congestion problem. This corresponds to the high-trade case of Section 3.1. We summarize and illustrate below.

Corollary A.1. *Consider $\gamma_\pi \le \gamma_u$. Assume $\bar{q}(\omega_1)\omega_1 > \omega \sum_{\omega' \ge \omega} \bar{q}(\omega)$ 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^*, ζ^*, q^*, x^*) be a competitive equilibrium. Next we prove it cannot be constrained efficient by discussing several cases. Note that since $\gamma_\pi \le \gamma_u$, 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.

- Case 1: $q^*(\omega_2) < \bar{q}(\omega_2), q^*(\omega_1) < \bar{q}(\omega_1)$. This is inefficient. To see this, 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. This means the social welfare strictly increases by introducing this market, so q^* is inefficient.

- Case 2: $q^*(\omega_2) < \bar{q}(\omega_2), q^*(\omega_1) = \bar{q}(\omega_1)$. In this case, by assumption, the unique monopoly price for market q^* is ω_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 q^* is inefficient.

- Case 3: $q^*(\omega_2) = \bar{q}(\omega_2), q^*(\omega_1) < \bar{q}(\omega_1)$. We make an observation that we must have $x^*(\omega_1|\omega) = 0$ for all $\omega < \omega_1$, otherwise some ω consumers will sell their data while left with nothing, which violates social optimality. Next we argue that (q^*, x^*) is not efficient. Firstly, 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 we know q^* cannot be efficient. Secondly, 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 ω_1 consumers from the platform, their welfare increases by $r(\omega_1) > \gamma_\pi \omega_1$ while all other markets (with fee smaller than ω_1) are not affected, which means social welfare strictly increases. Therefore, (q^*, x^*) cannot be efficient in this case. Thirdly, 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 the case that $q^*(\omega_1) = \bar{q}(\omega_1)$,

which is a contradiction. Therefore, we conclude that any such equilibrium cannot be constrained efficient.

- Case 4: $q^*(\omega_2) = \bar{q}(\omega_2), q^*(\omega_1) = \bar{q}(\omega_1)$. In this case, by our assumption, the unique monopoly price is ω_1. Therefore, we have $\psi_{q^*}(\omega_1) = \gamma_\pi \omega_1 < r(\omega_1)$. By Proposition 1 we know q^* is inefficient.

To sum up, under these assumptions no competitive equilibrium is constrained efficient. □

A.2 Proofs for Section 4

In this section we provide proofs for Section 4, in particular, Proposition 3 and Proposition 5. The arguments for Proposition 4 is right after the statement, so we skip it here. In this section, we do not assume the linear specification and consider general platform payoff v.

Proof of Proposition 3. **Only If.** Let (q^*, x^*) be a solution to the planner's problem $\mathcal{SB}$. First, we argue that (q^*, x^*) is a solution of a relaxed version of the data union's problem. We first discard the consumer's participation constraint from the data union's problem. Then, let us substitute the constraint $\sum_\omega \hat{q}(\omega)p(\omega) = V(q)$ into the data union's objective. By doing so, prices p do not appear in the relaxed problem. Summing up, the relaxed problem is

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

This problem is exactly the planner's problem. Since (q^*, x^*) is a solution to the relaxed problem, it must yield a value that is weakly higher than the value of the data union's problem, the proof is complete if we find prices p^* such that the participation constraints are satisfied and the data union's budget is balanced, given (q^*, x^*).

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

$$G(q^*, x^*) \;=\; V(q^*) - \sum_{\omega} \bar{q}(\omega)\tilde{p}(\omega)$$

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

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 must be no smaller than $\sum_\omega \bar{q}(\omega) r(\omega)$ because it is always feasible for the planner.

Since the union cannot earn a profit, we redistribute $G(q^*, x^*)$ back to the consumers in a uniform manner. Specifically, we let $t(\omega) = G(q^*, x^*)$ (recall that $\sum_\omega \bar{q}(\omega) = 1$). Therefore, if $\zeta^*(\omega) = 1$ was optimal under $\tilde{p}(\omega)$, it is still optimal under $p^*(\omega) \geq \tilde{p}(\omega)$.

We thus constructed a profile (p^*, q^*, x^*) that is feasible for the data union. Moreover, since (q^*, x^*) solves the relaxed problem, it also solves the data union's problem.

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

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

By the "only if" direction, we know there exist p° such that $(p^\circ, q^\circ, x^\circ)$ is feasible for the data union. (A.3) implies that:

$$\begin{aligned} & \textstyle\sum_\omega p^*(\omega) \bar{q}(\omega) + \sum_{a,\omega} u(a,\omega) x^*(a|\omega) q^*(\omega) + \sum_\omega (\bar{q}(\omega) - q^*(\omega)) r(\omega) \\ < \; & \textstyle\sum_\omega p^\circ(\omega) \bar{q}(\omega) + \sum_{a,\omega} u(a,\omega) x^\circ(a|\omega) q^\circ(\omega) + \sum_\omega (\bar{q}(\omega) - q^\circ(\omega)) r(\omega) \end{aligned}$$

This contradicts (p^*, q^*, x^*) being a solution to the data union's problem, so we conclude it must be constrained efficient. □

Proof of Proposition 5. **Step 1:** Let (p^*, q^*, x^*, ζ^*) 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

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

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

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

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^*,ζ^*) such that (p^*,q^*,x^*,ζ^*) 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 $\zeta(\cdot,\omega)$ and we can therefore assume to choose ζ^* such that $\zeta^*(\cdot,\omega)\bar{q}(\omega) = x^*(\cdot|\omega)q^*(\omega)$.

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

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

Since (q^*,x^*) is a first-best efficient allocation, we know $\chi^*(a,\omega) := x^*(a|\omega)q^*(\omega)$ solves $\mathcal{FB}'$. Define $\zeta^*(a,\omega) = \chi^*(a,\omega)/\bar{q}(\omega)$. 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^*, $\zeta^*(\omega)$ 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 with any a, so $zeta^*(\cdot,\omega)$ is optimal. When $\mu^*(\omega) > 0$, by complementary slackness, we have that $\sum_a \zeta^*(a,\omega) = 1$. Therefore, no type-$\omega$ consumer keeps the data. Since selling the data with any a gives the consumer a payoff of $\mu^*(\omega) + r(\omega)$. They are indifferent between different a and thus $\zeta^*(\omega)$ 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 conclusion is given by Theorem 2 of 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 $x \in x^*(q)$. □

With 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 both p' and p'', q solves the platform's first-stage problem, then by Lemma A.2 we know both

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

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 $(p', q', x') \in F(p, q, x)$. □

With 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 $p^*(2) = 0$; $\zeta^*(1) = 1$ and $\zeta^*(2) = \min\{1, \frac{\bar{q}(1)}{\bar{r}\bar{q}(2)}\}$; $q^*(1) = \bar{q}(1)$ 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

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

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 $0 \leq q^*(1) \leq \bar{q}(1)$ 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)). We take this stance because we want to focus on the trade between the platform and the consumers. Since we do not consider transfers between the merchant and the platform, it is natural that inefficiency can arise if the merchant's profit is taken into account.[20] That said, we show in this section that an analogous result to Proposition 2 holds if we define the efficiency notion to incorporate the merchant's profit.

Specifically, define the social welfare to be:

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

Define the notion of social efficiency as follows:

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

$$\max_{q,x} \quad SW(q,x)$$

[20] In the case where the platform can charge a service fee to the merchant, the platform can extract all the merchant's profit. Therefore, the platform's payoff essentially becomes $v(a,\omega) + \pi(a,\omega)$, which is equal to $\gamma_u u(a,\omega) + (1+\gamma_\pi)\pi(a,\omega)$. This is is a special case of our model (where $\gamma_u' = \gamma_u$ and $\gamma_\pi' = 1+\gamma_\pi$). In this case, the notion of welfare introduced in Equation (2) coincides with the social welfare (Equation (D.1)).

$$\begin{aligned} &\textit{such that} \quad q \leq \bar{q}, \\ &\textit{and} \quad\quad x \textit{ solves } \mathcal{P}_q. \end{aligned}$$

We have the following result, which extends Proposition 2 to this alternative notion of efficiency.

Proposition D.1. *Let (p^*, ζ^*, 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 the sufficiency of the proposition here. Let (p^*, ζ^*, 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,

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

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 still the pooling externality. When $\gamma_\pi > \gamma_u$, the only optimal mechanism for the platform given any q is full disclosure, as argued in the proof of Proposition 2, so in this case 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 elaborate the two sources of externality 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.

E Expropriation Economy

The expropriation economy is one where the property right of data records belongs to the platform. Our analysis can be directly applied to show that this economy is in general inefficient.

Specifically, the following result says that the expropriation economy is constrained efficient if and only if consumers' reservation values are low so that they do not care whether they are expropriated.

Corollary E.1. *The expropriation economy is constrained efficient if and only if there exists* $\psi_{\bar{q}}$ *such that* $r(\omega) \leq \psi_{\bar{q}}(\omega)$ *for all* ω.

Proof. The expropriation economy is one where the database of the platform is $\bar{q}$. The result then follows directly from Proposition 1. □

Therefore, whenever there exists some consumer whose reservation utility is large, the expropriation economy will not be constrained efficient.