The Economics of Data Externalities

Shota Ichihashi*

May 12, 2020

Abstract

I study a model in which a firm buys data from consumers. There are data externalities, whereby data of some consumers reveal information about others. Using an information design approach, I characterize the kind of data externality that maximizes or minimizes consumer surplus and the firm profit. I apply the results to solve an information-design problem in which the firm chooses what information to collect from consumers, taking into account the impact of the externalities on the cost of sourcing data.

*Bank of Canada, 234 Wellington Street West, Ottawa, ON K1A 0G9, Canada. Email: shotaichihashi@gmail.com. Website: http://shotaichi.weebly.com/. The opinions expressed in this article are the author's own and do not reflect the views of the Bank of Canada.

1 Introduction

Online platforms, such as Facebook and Google, collect data from users, learn their characteristics, and personalize services and advertising. Biotechnology companies collect genetic information about individuals to assess their health risks. Carmakers collect driving data through vehicles to learn how the human driver behaves. What is common in these examples is that a firm collects data from individuals to learn about some uncertain state of the world.

I capture these examples in the following model: A firm aims to learn about an uncertain state of the world. Consumers have some information about the state. The information corresponds to consumer data. The game consists of two stages. First, the firm sets prices to buy data. Second, each consumer decides whether to sell her data. The payoffs of the firm and consumers depend on monetary transfers and what the firm learns about the state. The firm may set negative prices (i.e., charge positive fees) when consumers benefit from the firm learning the state.

The key feature of the model is the presence of *data externalities*, whereby data of some consumers reveal information about others. The recent works on data markets find that a data externality can lead to low prices of data and excessive data collection (Acemoglu et al., 2019; Bergemann et al., 2019; Choi et al., 2019). The novelty of this paper is to study more general welfare implications by considering a richer class of data externalities and consumer preferences.

To focus on data externalities, I consider the following problem: Fix any Blackwell experiment μ_0 about the state. μ_0 represents the aggregate data in the economy. A profile of experiments $(\mu_1, \dots, \mu_n)$ for n consumers is called an *allocation of data*. An allocation of data $(\mu_1, \dots, \mu_n)$ is *feasible* if it contains the same information as μ_0, i.e., μ_0 and $(\mu_1, \dots, \mu_n)$ induce the same distribution of posteriors. For any μ_0, I ask which feasible allocation of data maximizes or minimizes consumer surplus and the firm's profit. Since μ_0 is fixed, I can consider various data externalities without changing the total amount of information.

Section 4 considers monotone consumer preferences. First, I assume that consumers are worse off if the firm learns more about the state. I show that consumer welfare is minimized and the firm's profit is maximized when consumers hold "substitutable" data, where the marginal value of individual data is negligible. This is consistent with the finding of the literature. In contrast, consumer surplus is maximized and the firm's profit is minimized if consumer data are "comple-

mentary," whereby the marginal value of individual data is high when many consumers provide their data. The paper constructs these allocations of data for any μ_0. In particular, constructing complementary data employs the *secrete sharing algorithm* of Shamir (1979).

Second, I assume that consumers are better off if the firm learns more about the state. In this case, the firm may charge positive fees to extract surplus from consumers. I show that a data externality can protect consumers. Specifically, consumer surplus is maximized when data are substituable, and it is minimized when data are complementary. Thus, the kind of data externality identified in the literature (i.e., substitutable data) can improve consumer welfare when the firm's data usage benefits consumers.

Section 5 considers two applications of the above results. The first application is the firm's information design problem: The firm chooses what information to collect from consumers to maximize profits. The firm can flexibly "design" a data externality by requesting consumers to share correlated information. Under arbitrary consumer preferences, I characterize the firm-optimal information structure. The second application is a monopoly pricing problem, where the firm uses data for third-degree price discrimination. The market outcome depends not only on what data the firm uses in the product market, but also on how data are initially allocated in the data market. I consider all allocations of data and characterize all possible pairs of the firm's profit and consumer surplus. The analysis highlights a beneficial role of data externalities for consumers.

This paper is closely related to recent works on data markets with data externalities (Easley et al., 2018; Acemoglu et al., 2019; Bergemann et al., 2019; Choi et al., 2019).[1] In particular, Acemoglu et al. (2019) and Bergemann et al. (2019) consider models in which a firm buys information from consumers to learn about their types, and consumer welfare is decreasing in the amount of collected information. There are two main differences between these papers and the current paper. First, these papers assume that the types and signals follow normal distributions. This assumption implies that the marginal value of individual data (and thus the marginal incentive of an individual to protect privacy) is decreasing in the total amount of information acquired by the firm. In contrast, I consider arbitrary information structures of consumer data. This generality enables me to consider a new kind of data externality, whereby individual data become more valuable as the firm

[1] Earlier works that consider externalities in information sharing include MacCarthy (2010) and Fairfield and Engel (2015).

collects more data. As the analysis shows, this data externality is important for deriving the kind of data externality that maximizes consumer welfare or the firm's profit in a more general setting. Second, while Acemoglu et al. (2019) and Bergemann et al. (2019) assume that data collection hurts consumers, I also consider beneficial data collection (Subsection 4.2) and general consumer preferences (Section 5). By doing so, I can examine how the impact of data externality depends on consumer preferences. One new insight from this analysis is that data externalities may protect consumers from the firm's market power. Overall, the paper complements the literature by providing a richer understanding of data externalities.

The paper is also related to the broad literature on information markets. One branch of this literature considers the optimal collection and sales of personal data (e.g., Admati and Pfleiderer 1986, Taylor 2004, Calzolari and Pavan 2006, Eső and Szentes 2007, Babaioff et al. 2012, Bergemann et al. 2015a, Hörner and Skrzypacz 2016, Bergemann et al. 2018, Agarwal et al. 2019, Ichihashi 2019). Another branch considers the optimal use of data such as price discrimination and targeting (e.g., Conitzer et al. 2012, De Corniere and De Nijs 2016, Ali et al. 2019, Madio et al. 2019, Montes et al. 2019, Bonatti and Cisternas 2020, De Corniere and Taylor 2020). Relative to this literature, I simplify the mechanism of data collection and data usage, and take a data externality as a key variable.

2 Model

There are $n \geq 1$ consumers represented by the set $\mathcal{N} = \{1, \ldots, n\}$. A firm buys data from consumers to learn about the state of the world $X \in \mathcal{X}$. $\mathcal{X}$ is finite, and all players share a common prior belief about X. Given a finite set $\mathcal{S}$ of realizations, I call any function $\mu : \mathcal{X} \to \Delta(\mathcal{S})$ an *experiment*.[2] Let $\mathbf{\Sigma}$ denote the set of all experiments with finite realization spaces. Given any $\mu \in \mathbf{\Sigma}$, let $\langle \mu \rangle \in \Delta(\Delta(\mathcal{X}))$ denote the distribution of posteriors induced by the prior and μ.

Let $\succeq$ denote the Blackwell (1953) order on experiments. That is, $\mu \succeq \mu'$ if $\langle \mu \rangle$ is a mean preserving spread of $\langle \mu' \rangle$, in which case we say that μ is more informative than μ'.

The *aggregate data* in the economy is denoted by an experiment μ_0. An *allocation of data* is a profile of n experiments $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) : \mathcal{X} \to \Delta(\mathcal{S}^{\mathcal{N}})$. $\mathcal{S}$ is a finite set of signal

[2]Given a set $\mathcal{S}$, $\Delta(\mathcal{S})$ denotes the set of all probability distributions over $\mathcal{S}$.

realizations, and μ_i represents the data held by consumer i. Given $X \in \mathcal{X}$, realizations from $(\mu_1(X), \ldots, \mu_n(X))$ may be correlated. An allocation of data $\boldsymbol{\mu}$ is *feasible* if $\langle \boldsymbol{\mu} \rangle = \langle \mu_0 \rangle$. Let $\mathcal{F}(\mu_0)$ denote the set of all feasible allocations of data. Below, I describe the game taking the allocation $\boldsymbol{\mu}$ as exogenously given, but the main focus is how the equilibrium depends on $\boldsymbol{\mu}$.

The game consists of two stages. In the first stage, the firm chooses a price vector $\mathbf{p} = (p_1, \ldots, p_n) \in \mathbb{R}^n$, where p_i is payment to consumer i. Each consumer i privately observes p_i. A negative price $p_i < 0$ is relevant when consumers benefit from data collection. In the second stage, all consumers simultaneously decide whether to sell their data. Specifically, let $a_i \in \{0, 1\}$ denote the data sharing decision of consumer i with $a_i = 1$ corresponding to sharing. Denote the profile of sharing decisions by $\mathbf{a} = (a_1, \ldots, a_n)$. Let $\mathcal{N}_{\mathbf{a}} = \{i \in \mathcal{N} : a_i = 1\}$ denote the set of consumers who sell their data. Then, the firm's data is given by the experiment $\boldsymbol{\mu}_{\mathbf{a}} = (\mu_i)_{i \in \mathcal{N}_{\mathbf{a}}} : \mathcal{X} \to \Delta(\mathcal{S}^{\mathcal{N}_{\mathbf{a}}})$. All players make decisions before X is realized. Thus, no player has private information.

A profile of data sharing decisions other than consumer i is denoted by $\mathbf{a}_{-i} \in \{0, 1\}^{n-1}$. If $\mathbf{a}_{-i}$ is such that any consumer $j \neq i$ shares her data, then it is written as $\mathbf{1}_{-i}$. For $a \in \{0, 1\}$, $(a, \mathbf{a}_{-i})$ denotes the profile of data sharing actions such that consumer i chooses a and other consumers choose $\mathbf{a}_{-i}$. Finally, $\boldsymbol{\mu}_{-i}$ denotes $\boldsymbol{\mu}_{(0, \mathbf{1}_{-i})}$.

If the firm collects data $\boldsymbol{\mu}_{\mathbf{a}}$, then it obtains a payoff of $\pi(\langle \boldsymbol{\mu}_{\mathbf{a}} \rangle) - \sum_{i \in \mathcal{N}} a_i p_i$, and consumer i obtains a payoff of $u_i(\langle \boldsymbol{\mu}_{\mathbf{a}} \rangle) + a_i p_i$. $\pi(\cdot)$ and each $u_i(\cdot)$ are defined on $\Delta(\Delta(\mathcal{X}))$. For simplicity, write $\pi(\langle \boldsymbol{\mu} \rangle)$ and $u_i(\langle \boldsymbol{\mu} \rangle)$ as $\pi(\boldsymbol{\mu})$ and $u_i(\boldsymbol{\mu})$, respectively.

The firm prefers more data: If $\mu \succeq \mu'$, then $\pi(\mu) \geq \pi(\mu')$. I impose more structures on $(u_i(\cdot))_{i \in \mathcal{N}}$ later. Normalize $\pi(\mu_\emptyset) = u_i(\mu_\emptyset) = 0$, where $\mu_\emptyset$ denotes an uninformative experiment, i.e., $\langle \mu_\emptyset \rangle$ is degenerate at the prior.

The solution concept is pure-strategy perfect Bayesian equilibrium (PBE) such that consumers hold passive beliefs, that is, each consumer i does not change her belief about $\mathbf{p}_{-i}$ after observing the firm's deviation that affects p_i. Hereafter, "equilibrium" refers to PBE with passive beliefs. Given an allocation of data $\boldsymbol{\mu}$, let $\mathcal{E}(\boldsymbol{\mu})$ denote the set of all equilibrium prices and data sharing decisions.

The main focus of the paper is how consumer surplus (i.e., the sum of the payoffs of all consumers) and the firm's profit depend on the allocation of data. The following notions simplify exposition.

Definition 1. Fix any experiment $\mu_0 \in \mathbf{\Sigma}$.

- An allocation of data $\boldsymbol{\mu}^*$ *maximizes consumer surplus with respect to* μ_0 if $\boldsymbol{\mu}^* \in \mathcal{F}(\mu_0)$, and there is $(\mathbf{a}^*, \mathbf{p}^*) \in \mathcal{E}(\boldsymbol{\mu}^*)$ such that for any $\boldsymbol{\mu} \in \mathcal{F}(\mu_0)$ and any $(\mathbf{a}, \mathbf{p}) \in \mathcal{E}(\boldsymbol{\mu})$,

$$\sum_{i \in \mathcal{N}} u_i(\boldsymbol{\mu}^*_{\mathbf{a}^*}) + a_i^* p_i^* \geq \sum_{i \in \mathcal{N}} u_i(\boldsymbol{\mu}_{\mathbf{a}}) + a_i p_i. \tag{1}$$

- An allocation of data $\boldsymbol{\mu}^*$ *minimizes consumer surplus with respect to* μ_0 if $\boldsymbol{\mu}^* \in \mathcal{F}(\mu_0)$, and there is $(\mathbf{a}^*, \mathbf{p}^*) \in \mathcal{E}(\boldsymbol{\mu}^*)$ such that for any $\boldsymbol{\mu} \in \mathcal{F}(\mu_0)$ and any $(\mathbf{a}, \mathbf{p}) \in \mathcal{E}(\boldsymbol{\mu})$,

$$\sum_{i \in \mathcal{N}} u_i(\boldsymbol{\mu}^*_{\mathbf{a}^*}) + a_i^* p_i^* \leq \sum_{i \in \mathcal{N}} u_i(\boldsymbol{\mu}_{\mathbf{a}}) + a_i p_i. \tag{2}$$

Analogously, I define an allocation of data that maximizes or minimizes the firm's profit. Definition 1 considers maximization or minimization given a fixed μ_0. Fixing μ_0 enables us to compare two economies that have the same information in aggregate but differ in how information is distributed across consumers. By doing so, I can pose the welfare implications of data externalities as a well-defined question. The following is the benchmark result for a single consumer.[3] Without data externalities, the outcome is efficient but consumer surplus is zero.

Claim 1. *Suppose* $n = 1$ *and the consumer holds data* μ_0*. In any equilibrium, consumer surplus is zero, and the firm obtains profit* $\max\{0, \pi(\mu_0) + u_1(\mu_0)\}$.

3 Substitutable and Complementary Allocations of Data

I introduce two allocations of data that are useful for describing the main results.

Definition 2. An allocation of data $\boldsymbol{\mu}$ is *perfectly substitutable* if, for any $i \in \mathcal{N}$, $\langle \boldsymbol{\mu} \rangle = \langle \boldsymbol{\mu}_{-i} \rangle$.

Definition 3. An allocation of data $\boldsymbol{\mu}$ is *perfectly complementary* if, for any $i \in \mathcal{N}$, $\langle \boldsymbol{\mu}_{-i} \rangle = \langle \boldsymbol{\mu}_{\emptyset} \rangle$, where $\mu_{\emptyset}$ is an uninformative experiment.

[3] The result follows from the standard argument of monopoly pricing with perfectly inelastic demand.

An allocation of data is perfectly substitutable if the marginal value of individual data is zero. One example of such an allocation is that all consumers hold identical data. A perfectly substitutable allocation captures an extreme version of a situation in which a firm can learn about a consumer from the data of other consumers. In contrast, a perfectly complementary allocation of data is such that the marginal value of individual data equals the value of the entire dataset. In other words, the dataset is valueless if the data of any single consumer is missing. This is an extreme version of "increasing returns to scale," whereby the data of some consumers increase the marginal value of data on other consumers.[4] If $n = 2$, the above definitions satisfy the complementarity and substitutability of two experiments studied by Börgers et al. (2013). For any aggregate data μ_0, there is a feasible allocation of data satisfying one of the above conditions.

Lemma 1. *Suppose $n \geq 2$, and take any experiment $\mu_0 \in \mathbf{\Sigma}$ as the aggregate data.*

1. *There is a feasible and perfectly substitutable allocation of data.*

2. *There is a feasible and perfectly complementary allocation of data.*

Proof. Take any experiment $\mu_0 : \mathcal{X} \to \Delta(\mathcal{Y})$. For Point 1, take an allocation of data $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)$ such that $\mu_i^* = \mu_0$ for all $i \in \mathcal{N}$. We have $\langle \boldsymbol{\mu}^* \rangle = \langle \boldsymbol{\mu}_{-i}^* \rangle = \langle \mu_0 \rangle$, because the firm observes the same realization $Y \in \mathcal{Y}$ across all μ_i^*'s with probability 1.

To show Point 2, I employ the secret sharing algorithm of Shamir (1979). The algorithm implies that there is a set $\mathcal{S}$ and a function $\boldsymbol{\nu} : \mathcal{Y} \to \Delta(\mathcal{S}^n)$ such that for any distribution over $\mathcal{Y}$, $\boldsymbol{\nu}_{-i}$ is uninformative about a realized $Y \in \mathcal{Y}$ for any $i \in \mathcal{N}$, but $\boldsymbol{\nu}$ is perfectly informative about Y (i.e., $\boldsymbol{\nu}_{-i}$ is the experiment created by $\boldsymbol{\nu}$ by omitting the i-th experiment). Define $\boldsymbol{\mu}^* : \mathcal{X} \to \Delta(\mathcal{S}^n)$ as a composite of μ_0 and $\boldsymbol{\nu}$: For any $X \in \mathcal{X}$, $\boldsymbol{\mu}^*(X)$ first draws $Y \in \mathcal{Y}$ according to $\mu_0(X) \in \Delta(\mathcal{Y})$, and then draws $(S_1, \ldots, S_n) \in \mathcal{S}^n$ according to $\boldsymbol{\nu}(Y) \in \Delta(\mathcal{S}^n)$. $\boldsymbol{\mu}^*$ is perfectly complementary and satisfies $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0 \rangle$. □

[4] In practice, data seem to exhibit increasing or decreasing returns to scale depending on the contexts. Arrieta-Ibarra et al. (2018) offer an insightful discussion on this point.

4 Welfare Implications of Data Externalities

I show that if consumer preferences are monotone, then a perfectly substitutable or complementary allocation of data consists of the best or the worst outcome.

4.1 Harmful Data Collection

Assume that data collection is harmful to consumers: For any experiments $\mu, \mu' \in \mathbf{\Sigma}$ such that $\mu \succeq \mu'$, $u(\mu) \leq u(\mu') \leq 0$.

Proposition 1. *For any aggregate data $\mu_0 \in \mathbf{\Sigma}$, a perfectly substitutable allocation of data $\boldsymbol{\mu}^*$ minimizes consumer surplus and maximizes the firm's profit with respect to μ_0. Given $\boldsymbol{\mu}^*$, the firm's profit is $\pi(\mu_0)$, consumer surplus is $\sum_{i\in\mathcal{N}} u_i(\mu_0)$, and the prices of data are zero.*

Proof. The existence of a feasible $\boldsymbol{\mu}^*$ follows from Point 1 of Lemma 1. Suppose that the allocation of data is $\boldsymbol{\mu}^*$ and the firm chooses $p_i = 0$ for all $i \in \mathcal{N}$. Then, there is an equilibrium in which all consumers sell their data. Any consumer is indifferent between selling and not selling whenever all other consumers sell their data, because $\boldsymbol{\mu}^*$ is perfectly substitutable. The firm does not benefit from decreasing p_i because consumer i rejects it. This equilibrium leads to the firm's profit $\pi(\mu_0)$ and consumer surplus $\sum_{i\in\mathcal{N}} u_i(\mu_0)$.

The above equilibrium is best for the firm and worst for consumers: Take any feasible allocation of data and any equilibrium. The firm's equilibrium profit is at most $\pi(\mu_0)$ because the firm cannot charge a positive fee when data collection lowers consumers' payoffs. Also, the equilibrium payoff of each consumer i is at least $u_i(\mu_0)$, because she can choose to not sell her data. □

Proposition 1 illustrates the data externality studied in the literature. When the data of some consumers reveal information about consumer i, the private loss of i from sharing her data decreases as other consumers share their data. The firm can exploit this externality to collect a large amount of data at low prices. The perfectly substitutable allocation captures this intuition in an extreme way. The next result presents an allocation of data that is best for consumers and worst for the firm.[5]

[5] Proposition 2 assumes $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0) \geq 0$. If $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0) < 0$, then there may be no pure-strategy PBE with passive beliefs under complementary data. For example, suppose additionally $\pi(\mu_0) + u_i(\mu_0) > 0$ for all $i \in \mathcal{N}$. If there is an equilibrium such that $k \geq 2$ consumers reject to sell data, the firm can deviate and buy data

Proposition 2. *For any aggregate data μ_0 such that $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0) \geq 0$, a perfectly complementary allocation of data $\boldsymbol{\mu}^*$ maximizes consumer surplus and minimizes the firm's profit with respect to μ_0. Under $\boldsymbol{\mu}^*$, the firm pays $-u_i(\mu_0) \geq 0$ to each consumer i. Thus, consumer surplus is zero, and the firm's profit is $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0)$.*

Proof. The existence of a feasible $\boldsymbol{\mu}^*$ follows from Point 2 of Lemma 1. Consider an equilibrium in which the firm pays each consumer i a price of $-u_i(\mu_0)$, which is the loss that i incurs by sharing her data conditional on that other $n-1$ consumers sell their data. By the similar argument as the proof of Proposition 1, we can verify that this is an equilibrium. Each consumer obtains a payoff of zero, and the firm obtains a payoff of $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0)$. This maximizes the payoff of each consumer, who never obtains a positive payoff.[6] This outcome also minimizes the firm's payoff, because for any $\boldsymbol{\mu} \in \mathcal{F}(\mu_0)$, the firm can obtain a payoff of at least

$$\pi(\boldsymbol{\mu}) - \sum_{i\in\mathcal{N}} [u_i(\boldsymbol{\mu}_{-i}) - u_i(\boldsymbol{\mu})] \geq \pi(\mu_0) - \sum_{i\in\mathcal{N}} [0 - u_i(\mu_0)] = \pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0).$$

Therefore, $\boldsymbol{\mu}^*$ maximizes consumer surplus and minimizes the firm's profit. □

Proposition 2 extends the "de-correlation scheme" proposed by Acemoglu et al. (2019). They consider a trusted mediator who (i) collects data from consumers, (ii) computes transformed variables for each consumer by removing the correlation with the information of other consumers, and then (iii) sells the transformed data of those who are willing to sell their data. They assume a Gaussian information structure, and thus de-correlation in (ii) corresponds to a linear transformation of signals. In contrast, the transformation based on Proposition 2 is non-parametric, and an explicit construction is known (e.g., Shamir 1979).

from everyone by setting a small but positive price, as each consumer believes her data contribution is not pivotal. In an equilibrium where only consumer i rejects to sell data, the firm can deviate and obtain data at price $-u_i(\mu_0)$ from i. However, there is also no equilibrium in which the firm buys data from all consumers, because the firm then earns a negative profit. If the firm's offers are public, then there is a consumer-best and firm-worst equilibrium with no data collection.

[6] Suppose to the contrary that there is an equilibrium in which consumer i obtains a positive payoff. It means that the firm pays a positive price, and i strictly prefers to share her data. Since i holds a passive belief, the firm can slightly lower the price to buy data, which is a contradiction.

4.2 Beneficial Data Collection

Assume that consumers are better off if the firm has more data: For any $\mu, \mu' \in \mathbf{\Sigma}$ such that $\mu \succeq \mu'$, $u(\mu) \geq u(\mu') \geq 0$. In this case, a consumer can always retain her data and secure a non-negative payoff. The best and worst allocations of data are the mirror images of those under harmful data collection.

Proposition 3. *For any aggregate data $\mu_0 \in \mathbf{\Sigma}$, a perfectly complementary allocation of data $\boldsymbol{\mu}^*$ minimizes consumer surplus and maximizes the firm's profit with respect to μ_0. Under $\boldsymbol{\mu}^*$, each consumer i pays $u_i(\mu_0) \geq 0$, and the firm extracts full surplus $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0)$.*

Proof. Under $\boldsymbol{\mu}^*$, there is an equilibrium in which the firm sets $p_i^* = -u_i(\mu_0) \leq 0$ for all $i \in \mathcal{N}$ and all consumers share their data. Given $\mathbf{p}^*$, each consumer is indifferent between sharing and not sharing her data conditional on that all other consumers share their data. This is an equilibrium that maximizes the firm's profit, because the firm extracts the efficient total surplus while giving consumers the lowest possible payoff of zero. □

Proposition 4. *For any aggregate data $\mu_0 \in \mathbf{\Sigma}$, a perfectly substitutable allocation of data $\boldsymbol{\mu}^*$ maximizes consumer surplus and minimizes the firm's profit with respect to μ_0. Under $\boldsymbol{\mu}^*$, the firm collects data at a price of zero. Consumer surplus is $\sum_{i \in \mathcal{N}} u_i(\mu_0)$ and the firm's profit is $\pi(\mu_0)$.*

Proof. The same argument as the proof of Proposition 3 implies that there is an equilibrium in which the firm sets $p_i^* = 0$ for all $i \in \mathcal{N}$ and all consumers sell their data. This is an equilibrium that maximizes consumer surplus, because there is no equilibrium in which the firm pays a positive transfer given data collection is beneficial.[7] This, in turn, implies that the equilibrium minimizes the firm's profit. □

The intuition for Proposition 4 is similar to the free-rider problem. When the allocation of data is highly substitutable, a consumer has a low willingness to pay for having her data collected, provided that other consumers sell their data. If prices were exogenously given, this free-rider problem would inefficiently lower the level of data provision. However, when prices are endogenous, each

[7] This follows from the same argument as Footnote 6.

consumer's incentive to free-ride forces the firm to lower fees. Thus, the data externality protects consumers from the monopolist firm.[8]

The equilibrium in Proposition 4 seems consistent with the observation that data collection which potentially benefits consumers (e.g., collection of location data to improve web-mapping services) often involves no monetary transfer from consumers to firms. One potential explanation is that a firm cannot charge for data collection because the marginal contribution of individual data to improving the quality of the service or the product is negligible.

Remark 1 (Multiplicity of Equilibria). In Definition 1, I select an equilibrium $(\mathbf{a}^*, \mathbf{p}^*) \in \mathcal{E}(\boldsymbol{\mu}^*)$ that achieves a higher consumer surplus than in any equilibrium under any other feasible allocations of data. This definition leaves the possibility that, under $\boldsymbol{\mu}^*$, there is another equilibrium that leads to a low consumer surplus. For example, Proposition 4 selects an equilibrium under $\boldsymbol{\mu}^*$ in which the firm sets $p_i = 0$ for all i. However, there is also an equilibrium in which the firm charges a negative price $-u_i(\mu_0)$ to i and a price of zero to others.

Nonetheless, the results are not sensitive to the equilibrium selection in the following sense. If data collection is beneficial, then I can construct a sequence $(\boldsymbol{\mu}^k)_{k\in\mathbb{N}}$ of feasible allocations of data such that the sequence converges to $\boldsymbol{\mu}^*$ (in Proposition 3 or Proposition 4) and there is a unique equilibrium under $\boldsymbol{\mu}^k$ for each k. Thus, we can always approximate the consumer or the firm-optimal allocation with an allocation of data that has a unique equilibrium (Appendix A proves these claims).

If data collection is harmful, the equilibrium is unique under the consumer-optimal allocation of data studied in Proposition 2. Namely, if $\pi(\mu_0)+\sum_{i\in\mathcal{N}} u_i(\mu_0) > 0$, then the unique equilibrium is such that the firm collects all data by paying $-u_i(\mu_0)$ to each i.

In contrast, under the consumer-worst equilibrium in Proposition 2, there can be multiple equilibria. However, if $\pi(\mu_0) + u_i(\mu_0) > 0$ holds for some i, then the firm collects data from at least one consumer in any equilibrium. Moreover, if one consumer sells her data, other consumers are willing to sell data for free, given the perfectly substitutable allocation. Thus, in any equilibrium, the firm collects data at a price of zero.

[8]The logic is somewhat similar to how free-riding by shareholders prevents the raider from capturing surplus in corporate takeover (cf. Tirole 2010).

5 Applications

This section consists of two parts. First, I consider the information design problem of a firm that can choose any allocation of data. Second, I consider a firm that uses data to price discriminate in the product market. These applications illustrate that the results in the previous section are useful for solving problems with a broader class of consumer preferences.

5.1 Firm's Information Design Problem

This section considers a profit maximization problem of a firm that can request any data from any consumers. While the result relies on a strong assumption that the firm can potentially source any information, we may think of the problem as the first-best benchmark for the firm. Moreover, the unconstrained problem admits a clean characterization under arbitrary consumer preferences. The firm's problem is equivalent to finding an allocation of data that maximizes profit without the feasibility constraint:

Definition 4. An allocation of data $\boldsymbol{\mu}^*$ *globally maximizes the firm's profit* if there is $(\mathbf{a}^*, \mathbf{p}^*) \in \mathcal{E}(\boldsymbol{\mu}^*)$ such that for any allocation of data $\boldsymbol{\mu}$ and $(\mathbf{a}, \mathbf{p}) \in \mathcal{E}(\boldsymbol{\mu})$,

$$\pi(\boldsymbol{\mu}^*_{\mathbf{a}^*}) - \sum_{i \in \mathcal{N}} a_i^* p_i^* \geq \pi(\boldsymbol{\mu}_{\mathbf{a}}) - \sum_{i \in \mathcal{N}} a_i p_i. \tag{3}$$

If $n = 1$, then any $\mu^* \in \arg\max_{\mu \in \mathbf{\Sigma}} \pi(\mu) + u_1(\mu)$ globally maximizes the firm's profit (see Claim 1). The following result characterizes the firm-optimal allocation of data for $n \geq 2$.[9]

Proposition 5. *Suppose $n \geq 2$ and take any $(u_i(\cdot))_{i \in \mathcal{N}}$. Let $\mu_0^* \in \mathbf{\Sigma}$ satisfy*

$$\mu_0^* \in \arg\max_{\mu \in \mathbf{\Sigma}} \left(\pi(\mu) + \sum_{i \in \mathcal{N}} u_i(\mu) - \sum_{i \in \mathcal{N}} \min_{\mu' \preceq \mu} u_i(\mu') \right). \tag{4}$$

Then, there is an allocation $\boldsymbol{\mu}^*$ that satisfies $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0^* \rangle$ and globally maximizes the firm's profit.

[9] I restrict attention to $\pi(\cdot)$ and $(u_i(\cdot))_{i \in \mathcal{N}}$ such that the maximization and minimization problems in (4) have solutions.

Proof. Take any allocation $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and an equilibrium under $\boldsymbol{\mu}$. Without loss of generality, suppose that the firm buys μ_i from each consumer $i \in \mathcal{N}$ (otherwise, we can replace μ_i with $\mu_\emptyset$). Given passive beliefs, the equilibrium price for consumer i is $p_i = -u_i(\boldsymbol{\mu}) + u_i(\boldsymbol{\mu}_{-i})$. Thus, the firm's profit is

$$\begin{aligned}
&\pi(\boldsymbol{\mu}) + \sum_{i \in \mathcal{N}} [u_i(\boldsymbol{\mu}) - u_i(\boldsymbol{\mu}_{-i})] \\
\leq &\pi(\boldsymbol{\mu}) + \sum_{i \in \mathcal{N}} u_i(\boldsymbol{\mu}) - \sum_{i \in \mathcal{N}} \min_{\mu' \preceq \boldsymbol{\mu}} u_i(\mu') \qquad (5) \\
\leq &\max_{\mu} \left(\pi(\mu) + \sum_{i \in \mathcal{N}} u_i(\mu) - \sum_{i \in \mathcal{N}} \min_{\mu' \preceq \mu} u_i(\mu') \right). \qquad (6)
\end{aligned}$$

The first inequality (5) holds because $\boldsymbol{\mu}_{-i} \preceq \boldsymbol{\mu}$ for each i. Take a maximizer μ_0^* of (6). I construct an allocation $\boldsymbol{\mu}^*$ such that $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0^* \rangle$ and the firm attains profit (6) in some equilibrium under $\boldsymbol{\mu}^*$. The construction is as follows. For each $i \in \mathcal{N}$, pick an experiment $\mu_i^{MIN} \in \arg\min_{\mu \preceq \mu_0^*} u_i(\mu)$. Let $\boldsymbol{\nu}^i_{-i} = (\nu^i_j)_{j \in \mathcal{N} \setminus \{i\}}$ denote a perfectly complementary allocation of data for consumers in $\mathcal{N} \setminus \{i\}$ such that $\langle \boldsymbol{\nu}^i_{-i} \rangle = \mu_i^{MIN}$. Also, let $\boldsymbol{\nu}^*$ denote a perfectly complementary allocation of data for n consumers such that $\langle \boldsymbol{\nu}^* \rangle = \langle \mu_0^* \rangle$. Consider the allocation of data $\boldsymbol{\mu}^*$ such that each consumer j has $(\nu^i_j)_{i \in \mathcal{N} \setminus \{j\}}$ and ν^*_j. Consider the strategy profile in which the firm sets a price of $p_i^* := -u_i(\mu_0^*) + u_i(\mu_i^{MIN})$ to consumer i, and all consumers sell their data. It is optimal for consumer i to sell her data: If i does not sell data, her payoff is $u_i(\mu_i^{MIN})$ because other $n-1$ consumers sell data and the firm obtains $\boldsymbol{\nu}^i_{-i}$. If i sells data, her gross utility is $u_i(\mu_0^*)$. Thus, p_i^* is the maximum amount that i is willing to pay. There is no profitable deviation for the firm regarding prices, because $p_i^* \leq 0$ holds for all i, and the firm cannot lower the price. In this equilibrium, the firm attains profit (6). □

The intuition is follows. Suppose the firm collects data μ^* in aggregate. The maximum amount that consumer i is willing to pay for data collection is $u_i(\mu^*) - u_i(\mu^*_{-i})$, where μ^*_{-i} is the information that the firm would collect without i's data. Since μ^* is more informative than μ^*_{-i}, $u_i(\mu^*) - \min_{\mu \preceq \mu_0^*} u_i(\mu)$ is the upper bound of the transfer. In the proof, I construct an allocation of data such that $n-1$ consumers other than i jointly hold data $\arg\min_{\mu \preceq \mu_0^*} u_i(\mu)$, so that i's gross utility from not selling data is indeed $\min_{\mu \preceq \mu_0^*} u_i(\mu)$.

The objective (4) captures the distortion in the firm's incentive to collect data. The sum of the first two terms $\pi(\mu^*) + \sum_{i\in\mathcal{N}} u_i(\mu^*)$ is total surplus: If a piece of data increases total surplus, then the firm prefers to collect it because the firm can extract the welfare gain as a monopolist. If $n = 1$, the firm chooses μ^* to maximize total surplus. The third term $\min_{\mu\preceq\mu_0^*} u_i(\mu)$, which is specific to $n \geq 2$, captures the firm's extra incentive to collect data. By collecting more data with a carefully chosen correlation structure, the firm can lower the utility of consumer i from refusing to sell data. This reduces the price of data and increases the firm's profit. The result shows that this intuition holds for any consumer preferences. The following result is a directly corollary.

Corollary 1. *The firm-optimal allocation of data in Proposition 5 has the following properties.*

1. *If each $u_i(\cdot)$ is monotone in $\succeq$, then under the firm-optimal allocation of data, the firm fully learns the state.*

2. *Any efficient experiment $\mu_E^* \in \arg\max_\mu \pi(\mu) + \sum_{i\in\mathcal{N}} u_i(\mu)$ cannot be strictly more informative than any firm-optimal allocation of data $\boldsymbol{\mu}^*$.*

Proof. Take any $i \in \mathcal{N}$. If $u_i(\cdot)$ is increasing in $\succeq$, then $\min_{\mu'\preceq\mu} u_i(\mu') = u_i(\mu_\emptyset) = 0$. If $u_i(\cdot)$ is decreasing in $\succeq$, then $\min_{\mu'\preceq\mu} u_i(\mu') = u_i(\mu)$. Thus, (3) reduces to $\Pi(\mu) := \pi(\mu) + \sum_{i\in\mathcal{N}_+} u_i(\mu)$, where $\mathcal{N}_+$ is the set of i's such that $u_i(\cdot)$ is increasing. Since $\Pi(\cdot)$ is increasing in $\succeq$, the fully informative experiment maximizes (3), which completes the proof of Point 1.

For Point 2, suppose there is a firm-optimal allocation of data $\boldsymbol{\mu}^*$ that is strictly less informative than μ_E^*. Since μ_E^* is efficient, we have

$$\pi(\boldsymbol{\mu}^*) + \sum_{i\in\mathcal{N}} u_i(\boldsymbol{\mu}^*) \leq \pi(\mu_E^*) + \sum_{i\in\mathcal{N}} u_i(\mu_E^*).$$

Since $\mu_E^* \succeq \boldsymbol{\mu}^*$, we have $\min_{\mu'\preceq\mu_E^*} u_i(\mu') \leq \min_{\mu'\preceq\boldsymbol{\mu}^*} u_i(\mu')$. Combining these inequalities, we obtain

$$\pi(\boldsymbol{\mu}^*) + \sum_{i\in\mathcal{N}} u_i(\boldsymbol{\mu}^*) - \sum_{i\in\mathcal{N}} \min_{\mu'\preceq\boldsymbol{\mu}^*} u_i(\mu') \leq \pi(\mu_E^*) + \sum_{i\in\mathcal{N}} u_i(\mu_E^*) - \sum_{i\in\mathcal{N}} \min_{\mu'\preceq\mu_E^*} u_i(\mu'). \tag{7}$$

Thus, μ_E^* is also the firm-optimal information, which completes the proof. □

The next result uses the proposition to characterize the consumer-worst outcome. Despite the potential heterogeneity of consumer preferences, the worst outcome involves full information acquisition, and each consumer gets the lowest payoff according to her $u_i(\cdot)$.

Corollary 2. *Among all allocations of data and all equilibria, the lowest consumer surplus is* $\sum_{i\in\mathcal{N}} \min_{\mu\preceq\mu_{FULL}} u_i(\mu)$*, where* μ_{FULL} *is the fully informative signal.*

Proof. By replacing μ_0^* in the proof of Proposition 5 with μ_{FULL}, we can construct an allocation $\boldsymbol{\mu}^*$ such that $\langle\boldsymbol{\mu}^*\rangle = \langle\mu_{FULL}\rangle$ and the firm pays $p_i^* = -u_i(\mu_{FULL}) + \min_{\mu\preceq\mu_{FULL}} u_i(\mu) \leq 0$ to each i. The resulting consumer surplus is $\sum_{i\in\mathcal{N}} u_i(\mu_{FULL}) + p_i^* = \sum_{i\in\mathcal{N}} \min_{\mu\preceq\mu_{FULL}} u_i(\mu)$. This is the lowest consumer surplus because i can always secure a payoff of at least $\min_{\mu\preceq\mu_{FULL}} u_i(\mu)$ by refusing to sell data. □

5.2 Monopoly Pricing

This subsection assumes that the firm uses data to price discriminate. Suppose that the firm sells a good to consumers, each of whom demands one unit. The production cost is zero. n consumers have a common value of X to the good.[10] X has a finite support and is positive with probability 1.

Initially, the firm only knows the prior of X. However, the firm can buy data and use it for pricing. Formally, given the allocation of data $\boldsymbol{\mu}$, the firm and consumers play the following game: First, in the *data market*, as in the previous section, the firm chooses a price vector $\mathbf{p}$, and each consumer i decides whether to sell data μ_i. Second, in the *product market*, the firm updates its belief about X based on collected data, and then sets a product price t. Finally, consumers make (identical) purchase decisions. It is without loss of generality that the firm sets the same price across all consumers.

Suppose that the firm pays the total amount of p for data and sets a product price of t, and m consumers buy goods. In ex post terms, the firm's profit and consumer surplus are $mt - p$ and $m(X-t)+p$, respectively. The average firm profit and the average consumer surplus are $\frac{1}{n}(mt-p)$ and $\frac{1}{n}[m(X-t)+p]$, respectively.

Let $\bar{w} := \mathbb{E}[X]$ denote the average total surplus under the efficient outcome.[11] Let $u_\emptyset$ and $\pi_\emptyset$ denote the average expected consumer surplus and the average firm profit, when the firm buys no

[10] The common value assumption simplifies exposition. The same result holds for independent and private values.

[11] Since X is almost surely positive, the efficient outcome is that all consumers buy goods with probability 1.

information and all players behave optimally in the product market. For simplicity, assume that the optimal product price given no information is unique, so that $u_\emptyset$ is unique.

I characterize all possible outcomes across all allocations of data. The setting is similar to Bergemann et al. (2015b), which considers possible pairs of consumer surplus and the seller profit across all information structures available to the seller. The difference is that the firm in my model needs to buy information from consumers.

If there is one consumer ($n = 1$), then the firm can set a price of collecting data to make the consumer indifferent between selling and not selling data. If she does not sell data, her payoff in the product market is $u_\emptyset$. Thus, the consumer's net equilibrium payoff is $u_\emptyset$.

Claim 2. *If there is a single consumer, then the following two conditions are equivalent.*

1. *There is an allocation of data (which is equal to the aggregate data) such that the equilibrium payoffs of the firm and the consumer are π^* and u^*, respectively.*
2. *$u^* = u_\emptyset$ and $\pi_\emptyset \leq \pi^* \leq \bar{w} - u_\emptyset$.*

Proof. Suppose Point 1 holds. $u^* \geq u_\emptyset$ holds because the consumer can secure $u_\emptyset$ by not sharing data. $u^* > u_\emptyset$ means that the consumer shares data (say) μ^*. Let u_{μ^*} denote the consumer's payoff in the product market given data μ^*. Let p_1^* denote the equilibrium transfer from the firm to the consumer in exchange for data. Since $u^* = u_{\mu^*} + p_1^* > u_\emptyset$, the firm can slightly lower p_1^* to strictly increase its profit while collecting μ^*. This is a contradiction, and thus $u^* = u_\emptyset$. The sum of the payoffs of the consumer and the firm is at most $\bar{w}$, and the firm can always choose to not collect data. Thus, $\pi_\emptyset \leq \pi^* \leq \bar{w} - u_\emptyset$.

Suppose that Point 2 holds. Let π_{μ^*} denote the firm profit in the product market given μ^*. Bergemann et al. (2015b) shows that there is an experiment μ^* such that $u_{\mu^*} = u_\emptyset$ and $\pi_{\mu^*} = \pi^*$. Consider the following strategy profile: The seller sets $p_1 = 0$ and the consumer sells data. Regardless of whether the consumer sells data, the firm sets a price optimally to achieve $(\pi^*, u_\emptyset)$ in the product market. This consists of an equilibrium. In particular, the consumer is willing to share data because it does not change her payoff in the product market. □

To state the main result, define the surplus triangle:

$$\Delta := \left\{(\pi, u) \in \mathbb{R}^2 : \pi + u \leq \bar{w}, u \geq 0, \pi \geq \pi_\emptyset\right\}. \tag{8}$$

Whenever there are multiple consumers, any outcome in the surplus triangle can arise. Thus, data externalities drastically expand the set of possible outcomes.

Proposition 6. *Suppose $n \geq 2$. A pair (π, u) of the average profit and the average consumer surplus can arise in some equilibrium given some allocation of data if and only if $(\pi, u) \in \Delta$.*

Proof. Suppose $n \geq 2$. To show the "if" part, take any $(\pi^*, u^*) \in \Delta$. Bergemann et al. (2015b) show that there is $\mu^* \in \mathbf{\Sigma}$ such that if the firm has μ^*, then the resulting average outcome in the product market is (π^*, u^*). Suppose that $u^* < u_\emptyset$ (resp. $u^* \geq u_\emptyset$). Proposition 1 (resp. Proposition 4) guarantees that there is an allocation of data such that the firm collects data μ^* at a price of zero. In this equilibrium, (π^*, u^*) arises as the net (average) payoffs of the firm and consumers. The "only if" part holds because consumers can secure zero payoffs by selling no data and buying nothing, and the firm can secure $\pi_\emptyset$ by obtaining no data and set an optimal price given the prior. □

Figure 1 depicts the result. The surplus triangle Δ corresponds to AEC. EC represents the firm's profit from no data, and AC describes the total surplus from the efficient allocation (all values are in terms of the average across consumers). If the market consists of a single consumer, then the possible outcomes correspond to the line BD. Thus, the consumer never benefits from data. In contrast, if the market consists of multiple consumers, then any outcome in AEC can arise depending on the allocation of data.

The result suggests that a social planner who cares about consumers should consider not only what inference the firm can make from the aggregate data, but also how the data are initially allocated to consumers. To see this, compare the following two scenarios. First, suppose that the aggregate data enable the firm to perfectly price discriminate. Then, consumer surplus is zero in the product market. However, the net consumer surplus can be positive if the data held by different consumers are complementary so that they exhibit increasing returns to scale. This is because the firm pays compensation to collect data. Second, suppose that the aggregate data correspond to a "consumer-optimal segmentation" in Bergemann et al. (2015b). In this case, consumer surplus in the product market is high. However, if consumer data are complementary, then the firm can charge a fee in the data market to extract the surplus accruing to consumers in the product market.

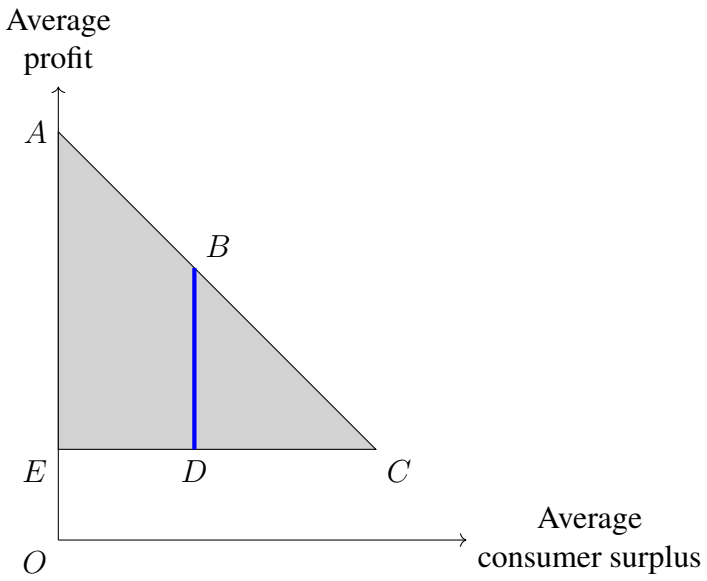

Figure 1: The set of possible outcomes for $n = 1$ (blue line, BD) and $n \geq 2$ (gray triangle, AEC).

6 Conclusion

How the initial allocation of resources affects the market outcome is an important question, and this paper applies the question to markets for data. For any aggregate data μ_0 in the economy, there are perfectly complementary and substitutable allocations of data consistent with μ_0. These allocations of data maximize or minimize the prices of data, consumer welfare, and the firm's profit in different environments. These results emphasize the importance of considering not only how the firm uses collected data but also how data are initially allocated across consumers.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar (2019), "Too much data: Prices and inefficiencies in data markets." Working Paper 26296, National Bureau of Economic Research.

Admati, Anat R and Paul Pfleiderer (1986), "A monopolistic market for information." *Journal of Economic Theory*, 39, 400–438.

Agarwal, Anish, Munther Dahleh, and Tuhin Sarkar (2019), "A marketplace for data: An algorithmic solution." In *Proceedings of the 2019 ACM Conference on Economics and Computation*, 701–726.

Ali, S Nageeb, Gregory Lewis, and Shoshana Vasserman (2019), "Voluntary disclosure and personalized pricing." Technical report, National Bureau of Economic Research.

Arrieta-Ibarra, Imanol, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E Glen Weyl (2018), "Should we treat data as labor? Moving beyond "Free"." In *AEA Papers and Proceedings*, volume 108, 38–42.

Babaioff, Moshe, Robert Kleinberg, and Renato Paes Leme (2012), "Optimal mechanisms for selling information." In *Proceedings of the 13th ACM Conference on Electronic Commerce*, 92–109, ACM.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2019), "The economics of social data." Cowles foundation discussion paper 2203, Yale University.

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin (2015a), "Selling experiments." Technical report, Mimeo, MIT and Yale.

Bergemann, Dirk, Alessandro Bonatti, and Alex Smolin (2018), "The design and price of information." *American Economic Review*, 108, 1–48.

Bergemann, Dirk, Benjamin Brooks, and Stephen Morris (2015b), "The limits of price discrimination." *The American Economic Review*, 105, 921–957.

Blackwell, David (1953), "Equivalent comparisons of experiments." *The annals of mathematical statistics*, 24, 265–272.

Bonatti, Alessandro and Gonzalo Cisternas (2020), "Consumer scores and price discrimination." *The Review of Economic Studies*, 87, 750–791.

Börgers, Tilman, Angel Hernando-Veciana, and Daniel Krähmer (2013), "When are signals complements or substitutes?" *Journal of Economic Theory*, 148, 165–195.

Calzolari, Giacomo and Alessandro Pavan (2006), "On the optimality of privacy in sequential contracting." *Journal of Economic theory*, 130, 168–204.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim (2019), "Privacy and personal data collection with information externalities." *Journal of Public Economics*, 173, 113–124.

Conitzer, Vincent, Curtis R Taylor, and Liad Wagman (2012), "Hide and seek: Costly consumer privacy in a market with repeat purchases." *Marketing Science*, 31, 277–292.

De Corniere, Alexandre and Romain De Nijs (2016), "Online advertising and privacy." *The RAND Journal of Economics*, 47, 48–72.

De Corniere, Alexandre and Greg Taylor (2020), "Data and competition: a general framework with applications to mergers, market structure, and privacy policy."

Easley, David, Shiyang Huang, Liyan Yang, and Zhuo Zhong (2018), "The economics of data." *Available at SSRN 3252870*.

Eső, Péter and Balazs Szentes (2007), "Optimal information disclosure in auctions and the handicap auction." *The Review of Economic Studies*, 74, 705–731.

Fairfield, Joshua AT and Christoph Engel (2015), "Privacy as a public good." *Duke LJ*, 65, 385.

Hörner, Johannes and Andrzej Skrzypacz (2016), "Selling information." *Journal of Political Economy*, 124, 1515–1562.

Ichihashi, Shota (2019), "Non-competing data intermediaries."

MacCarthy, Mark (2010), "New directions in privacy: Disclosure, unfairness and externalities." *ISJLP*, 6, 425.

Madio, Leonardo, Yiquan Gu, and Carlo Reggiani (2019), "Exclusive data, price manipulation and market leadership."

Montes, Rodrigo, Wilfried Sand-Zantman, and Tommaso Valletti (2019), "The value of personal information in online markets with endogenous privacy." *Management Science*, 65, 1342–1362.

Shamir, Adi (1979), "How to share a secret." *Communications of the ACM*, 22, 612–613.

Taylor, Curtis R (2004), "Consumer privacy and the market for customer information." *RAND Journal of Economics*, 631–650.

Tirole, Jean (2010), *The theory of corporate finance*. Princeton University Press.

Appendix

A Omitted Proofs for Remark 1: Equilibrium Multiplicity

I impose the following mild restriction on consumer preferences.

Assumption 1. For any $\mu, \mu' \in \mathbf{\Sigma}$, any $\alpha \in [0,1]$, and any $i \in \mathcal{N}$, $u_i(\alpha\langle\mu\rangle + (1-\alpha)\langle\mu\rangle) = \alpha u_i(\langle\mu\rangle) + (1-\alpha)u_i(\langle\mu\rangle)$.

Assumption 1 holds if each consumer i has some underlying payoff $u_i(a, X)$, which depends on the firm's (unmodeled) action a and a realized state X. Denoting the firm's action at $b \in \Delta(\mathcal{X})$ by $a(b)$, we can write $u_i(\mu) = \int_{\Delta(\mathcal{X})} \int_{\mathcal{X}} u_i(a(b), X) db(X) d\langle\mu\rangle(b)$. Since $u_i(\mu)$ is linear, it satisfies Assumption 1.

A.1 "Approximation result" for Proposition 3

Proposition 7. *Fix any aggregate data $\mu_0 \in \mathbf{\Sigma}$ such that $\pi(\mu_0) > 0$. There is a sequence of feasible allocations of data $(\boldsymbol{\mu}^k)_{k\in\mathbb{N}}$ that satisfies the following.*

1. *For each $\boldsymbol{\mu}^k$, there is a unique equilibrium.*
2. *As $k \to +\infty$, the equilibrium consumer surplus converges to 0, and the firm's profit converges to $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0)$.*
3. *For each $i \in \mathcal{N}$, as $k \to +\infty$, $\langle\mu_i^k\rangle$ weakly converges to $\langle\mu_\emptyset\rangle$.*

Proof. Take any perfectly complementary allocation of data $\boldsymbol{\mu}^C = (\mu_1^C, \ldots, \mu_n^C)$ such that $\langle \boldsymbol{\mu}^C \rangle = \mu_0$. For each $k \in \mathbb{N}$, consider the following allocation of data $(\mu_i^k)_{i \in \mathcal{N}}$: (i) with probability $1 - \frac{1}{k}$, $(\mu_i^k)_{i \in \mathcal{N}} = \boldsymbol{\mu}^C$, and (ii) with probability $\frac{1}{k}$, $\mu_i^k = \mu_0$ or $\mu_i^k = \mu_\emptyset$ holds. Specifically, conditional on (ii), one of n consumers (say i) is randomly picked with probability $\frac{1}{n}$, and $\mu_i^k = \mu_0$ holds. For any other consumer j, $\mu_j^k = \mu_\emptyset$ holds. Thus, given (ii), there is exactly one consumer with $\mu_i^k = \mu_0$.

Take any equilibrium, and suppose to the contrary that the firm does not collect data from (say) consumer 1. Suppose that the firm deviates and offers to pay $p_1 = \varepsilon$ with a sufficiently small $\varepsilon > 0$. Consumer 1 then strictly prefers to share data because she earns $\varepsilon > 0$ and benefits from the firm's learning. Consider the positive probability event where (ii) is realized and only i's data is μ_0. On this event, the firm's profit strictly increases by collecting data μ_1^k. Thus, the above deviation is profitable for a small $\varepsilon > 0$, leading to a contradiction. Given that the firm collects all data, the maximum price the firm can charge to consumer i is $p_i^k = -u_i(\mu_{1,\mathbf{1}_{-i}}^k) + u_i(\mu_{0,\mathbf{1}_{-i}}^k) = \left(1 - \frac{1}{k}\right) u_i(\mu_0) + \frac{1}{nk} u_i(\mu_0)$. In the unique equilibrium, the firm charges each consumer i a price of p_i^k, which converges to $u_i(\mu_0)$ as $k \to +\infty$. Points 2 and 3 directly follows from these observations. □

A.2 "Approximation result" for Proposition 4

Proposition 8. *Fix any aggregate data $\mu_0 \in \boldsymbol{\Sigma}$. There is a sequence of feasible allocations of data $(\boldsymbol{\mu}^k)_{k \in \mathbb{N}}$ that satisfies the following.*

1. *For each $\boldsymbol{\mu}^k$, there is a unique equilibrium.*
2. *As $k \to +\infty$, the equilibrium consumer surplus converges to $\sum_{i \in \mathcal{N}} u_i(\mu_0)$, and the firm's profit converges to $\pi(\mu_0)$.*
3. *For each $i \in \mathcal{N}$, as $k \to +\infty$, $\langle \mu_i^k \rangle$ weakly converges to $\langle \mu_0 \rangle$.*

Proof. For each $k \in \mathbb{N}$, consider the following allocation of data $(\mu_i^k)_{i \in \mathcal{N}}$: For each realized $X \in \mathcal{X}$, (i) with probability $1 - \frac{1}{k}$, $\mu_i^k = \mu_0$ for all $i \in \mathcal{N}$; (ii) with probability $\frac{1}{k}$, only one of n consumers, say i, has $\mu_i^k = \mu_0$, and any other consumer j has $\mu_j^k = \mu_\emptyset$ (i.e., this part is the same as (ii) in the proof of Proposition 7). By the same argument as the proof of Proposition 7, the firm collects data from all consumers in any equilibrium. Given that the firm collects all data,

the maximum price the firm can charge to consumer i is $p_i^k = u_i(\mu^k_{1,\mathbf{1}_{-i}}) - u_i(\mu^k_{0,\mathbf{1}_{-i}}) = \frac{1}{nk} u_i(\mu_0)$. In the unique equilibrium, the firm charges each consumer i a price of p_i^k, which converges to 0 as $k \to +\infty$. Points 2 and 3 directly follow from the above observations. $\square$