The Economics of Data Externalities

Shota Ichihashi*

December 15, 2020

Abstract

I study a model of a firm that buys data from consumers. There are data externalities, whereby data of some consumers reveal information about others. I characterize data externalities that maximize or minimize consumer surplus and the firm's profit. I use 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 data externalities on the cost of buying data. The firm collects no less information than the efficient amount. In some cases, we can solve the firm's problem using a two-step concavification method.

Keywords: externalities; data markets; privacy; information design

*Bank of Canada, 234 Wellington Street West, Ottawa, ON K1A 0G9, Canada. Email: shotaichihashi@gmail.com.
I thanks seminar participants at the Bank of Canada. 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.

Motivated by these examples, I study the following model: A firm aims to learn about an uncertain state of the world, and consumers have some information about it. The information of each consumer corresponds to her 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 fees for collecting data) when consumers benefit from the firm learning about 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 work on data markets finds that a data externality can lower the prices of data and cause excessive data collection (Acemoglu et al., 2019; Bergemann et al., 2019; Choi et al., 2019). This paper studies 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, which represents the aggregate data in the economy. A profile of experiments $(\mu_1, \ldots, \mu_n)$ for n consumers is called an *allocation of data*. An allocation $(\mu_1, \ldots, \mu_n)$ is *feasible* if it contains the same information as μ_0, i.e., they induce the same distribution of posteriors on the state. For any μ_0, I ask which feasible allocation of data maximizes or minimizes consumer surplus and the firm's profit. Because μ_0 is fixed, I can study the welfare implications of data externalities without changing the total amount of information.

Section 4 identifies the best and worst outcomes in a few instances. 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 zero. This result 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 "complementary," 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. The firm may then charge positive fees to extract surplus. I show that a data externality can protect consumers. Specifically, consumer surplus is maximized when data are substitutable, 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.

While Section 4 mainly studies monotone consumer preferences, I extend the result on the consumer-worst outcome to arbitrary preferences. In general, the worst outcome for consumer i is that other consumers hold the information (contained in the aggregate data μ_0) that harms consumer i, and her data provision is pivotal in enabling the firm to learn the beneficial component of μ_0. By combining the idea of substitutable and complementary data, we can construct an allocation of data that creates the worst situation simultaneously for all consumers.

Section 5 considers two applications of the above results. The first application is the firm's data collection problem: The firm chooses what information to collect from consumers to maximize profits. The firm can flexibly design data externalities by requesting consumers to share correlated information. For arbitrary consumer preferences, I characterize the firm-optimal information structure. Under a certain condition, we can solve the firm's problem with a two-step concavification method: The first step identifies the information the firm collects when a consumer refuses to sell her data, and the second step identifies the information the firm collects on the equilibrium path.

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. In the spirit of Bergemann et al. (2015b), 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 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) mainly consider the case in which data collection lowers consumer welfare, 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.

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

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, $X \in \mathcal{X}$. The set $\mathcal{X}$ 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. We say that μ is more informative than μ' if $\langle \mu \rangle$ is a mean preserving spread of $\langle \mu' \rangle$, and write it as $\mu \succeq \mu'$ (Blackwell, 1953)

The *aggregate data* in the economy is denoted by an experiment μ_0. An *allocation of data* is a profile of n experiments $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) : \mathcal{X} \to \Delta(\mathcal{S}^{\mathcal{N}})$. The set $\mathcal{S}$ is a finite set of signal realizations, and μ_i represents consumer i's data. 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 $\langle \boldsymbol{\mu} \rangle = \langle \mu_0 \rangle$. For any $\mu_0 \in \mathbf{\Sigma}$, 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 exogenous. However, our focus is on 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. The firm may choose a negative price $p_i < 0$ 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 move before X is realized. As a result, no player has private information about X.

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 all consumers $j \neq i$ sell their data, 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 $\boldsymbol{\mu}_{(0, \mathbf{1}_{-i})}$.

All players maximize their expected payoffs, and ex post payoffs are as follows. If the firm

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

collects data $\boldsymbol{\mu}_{\mathbf{a}}$, it obtains a payoff of $\pi(\langle\boldsymbol{\mu}_{\mathbf{a}}\rangle) - \sum_{i\in\mathcal{N}} a_i p_i$, and each consumer i obtains a payoff of $u_i(\langle\boldsymbol{\mu}_{\mathbf{a}}\rangle) + a_i p_i$. The functions $\pi(\cdot)$ and $(u_i(\cdot))_{i\in\mathcal{N}}$ 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 $\pi(\mu) \geq \pi(\mu')$. Also, normalize $\pi(\mu_\emptyset) = u_i(\mu_\emptyset) = 0$, where $\mu_\emptyset$ is an uninformative experiment (i.e., $\langle\mu_\emptyset\rangle$ is degenerate at the prior). I will later impose more structures on $(u_i(\cdot))_{i\in\mathcal{N}}$.

The solution concept is perfect Bayesian equilibrium (PBE) in which consumers hold passive beliefs—i.e., 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 (Section 4.4 discusses the role of passive beliefs).

The following notions simplify exposition. Below, consumer surplus refers to the sum of the expected payoffs of all consumers. Also, given an allocation of data $\boldsymbol{\mu}$, let $\mathcal{E}(\boldsymbol{\mu})$ denote the set of all equilibria.

Definition 1. Fix any experiment $\mu_0 \in \mathbf{\Sigma}$. An allocation of data $\boldsymbol{\mu}^*$ *maximizes (respectively, minimizes) consumer surplus with respect to* μ_0 if $\boldsymbol{\mu}^* \in \mathcal{F}(\mu_0)$, and there is an equilibrium $E^* \in \mathcal{E}(\boldsymbol{\mu}^*)$ such that for any $\boldsymbol{\mu} \in \mathcal{F}(\mu_0)$ and any $E \in \mathcal{E}(\boldsymbol{\mu})$, consumer surplus under E^* is weakly greater (respectively, smaller) than the one under E.

Analogously, I define an allocation of data that maximizes or minimizes the firm's expected payoff. Definition 1 considers maximization or minimization under a fixed μ_0. Thus, we can compare two economies that have the same aggregate data but differ in how data are distributed across consumers.

Definition 1 admits the possibility that although $\boldsymbol{\mu}^*$ maximizes consumer surplus, $\mathcal{E}(\boldsymbol{\mu}^*)$ contains another equilibrium that gives a lower consumer surplus than some equilibrium under another allocation. Section 4.4 discusses the multiplicity of equilibria and addresses this concern.

The following is the benchmark result for a single consumer. Without data externalities, the equilibrium is typically efficient and consumer surplus is zero.[3]

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

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.

An allocation of data is perfectly substitutable if the marginal value of individual data is zero. One example is when all consumers hold the same 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. Perfect complementarity 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 experiments in 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 $\mu_0 \in \boldsymbol{\Sigma}$ 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 $\mu_0 : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$. For Point 1, take an allocation of data $(\mu_1^*, \ldots, \mu_n^*)$ 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, we use the secret sharing algorithm of Shamir (1979). The algorithm provides a set $\mathcal{S}$ and a function $\boldsymbol{\nu} : \mathcal{Y} \rightarrow \Delta(\mathcal{S}^n)$ 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 (here, $\boldsymbol{\nu}_{-i}$ is the experiment created by $\boldsymbol{\nu}$ by omitting the i-th experiment). Define $\boldsymbol{\mu}^* : \mathcal{X} \rightarrow \Delta(\mathcal{S}^n)$ as a composite

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

of μ_0 and $\boldsymbol{\nu}$: For any $X \in \mathcal{X}$, $\boldsymbol{\mu}^*(X)$ first draws Y according to $\mu_0(X) \in \Delta(\mathcal{Y})$, then draws $(S_1, \ldots, S_n) \in \mathcal{S}^n$ according to $\boldsymbol{\nu}(Y) \in \Delta(\mathcal{S}^n)$. The experiment $\boldsymbol{\mu}^*$ is perfectly complementary and satisfies $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0 \rangle$. □

If $n = 2$, the construction of complementary signals is simple.[5] Take any $\mu_0 : \mathcal{X} \to \Delta(\mathcal{Y})$ with $\mathcal{Y} = \{1, 2, \ldots, m\}$. We decompose μ_0 into two experiments, μ_0^1 and μ_0^2. First, independently of the state X, μ_0^1 draws a permutation α of $1, 2, \ldots, m$ uniformly randomly from the set of all the permutations. Second, given the realized permutation α, μ_0^2 draws signal realization $\alpha(Y)$ whenever μ_0 draws $Y \in \mathcal{Y}$. The allocation (μ_0^1, μ_0^2) is perfectly complementary and contains the same information as μ_0.[6]

An economic example of a complementary allocation is as follows: Suppose consumer 1 is a seller, whose type θ_S is -1 or 1. Consumer 2 is a buyer, whose type θ_B is -1 or 1. For example, θ_S is the horizontal characteristics of the seller's product, and θ_B is the buyer's taste. The firm (e.g., a platform) tries to learn the value of the match, $X = \theta_S \cdot \theta_B$. Suppose θ_S and θ_B are independently distributed with the prior $\mathbf{P}(\theta_S = 1) = \mathbf{P}(\theta_B = 1) = 0.5$. Then, if the buyer and the seller have signals that respectively reveal θ_B and θ_S, it is a perfectly complementary allocation. Indeed, knowing θ_B or θ_S alone provides no information about X.

4 Welfare Implications of Data Externalities

Lemma 1 helps us find an allocation of data that maximizes or minimizes consumer surplus and the firm's profit. I first consider monotone consumer preferences, then extend a part of the results for general preferences. Throughout the section, we assume $n \geq 2$.

4.1 Harmful Data Collection

Assume that data collection harms consumers: For any experiments $\mu, \mu' \in \mathbf{\Sigma}$ such that $\mu \succeq \mu'$, $u_i(\mu) \leq u_i(\mu') \leq 0$ for all $i \in \mathcal{N}$.

[5] Shamir's algorithm uses the polynomial interpolation.

[6] In particular, suppose we only have μ_0^2 and observe realization $Y \in \mathcal{Y}$. Then, the conditional probability of X given Y is $\mathbf{P}^{\mathcal{X}}(X|Y) = \frac{\mathbf{P}^{\mathcal{X}}(X)\cdot\left[\frac{1}{m}\mu^0(1|X)+\cdots+\frac{1}{m}\mu^0(m|X)\right]}{\frac{1}{m}\mu^0(1)+\cdots+\frac{1}{m}\mu^0(m)} = \mathbf{P}^{\mathcal{X}}(X)$. Here, $\mu^0(Y)$ is the ex ante probability of Y under μ^0, and $\mathbf{P}^{\mathcal{X}}(X)$ is the ex ante probability of X under the prior.

Proposition 1. *For any aggregate data $\mu_0 \in \Sigma$, 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 Lemma 1. Suppose that the allocation of data is $\boldsymbol{\mu}^*$ and the firm chooses $p_i = 0$ for all i. Then, there is an equilibrium in which all consumers sell their data: Because $\boldsymbol{\mu}^*$ is perfectly substitutable, each consumer is indifferent between selling and not selling her data, whenever all other consumers sell their data. The firm does not benefit from decreasing p_i because consumer i will reject it. This equilibrium leads to the firm's profit $\pi(\mu_0)$ and consumer surplus $\sum_{i\in\mathcal{N}} u_i(\mu_0)$.

To show the welfare implications, take any feasible allocation and any (possibly mixed-strategy) equilibrium. The firm's profit is at most $\pi(\mu_0)$ because it cannot charge negative prices when data collection harms consumers. Also, the equilibrium payoff of each consumer i is at least $u_i(\mu_0)$, because she can refuse to sell her data. A substitutable allocation $\boldsymbol{\mu}^*$ attains these bounds. □

Proposition 1 illustrates the kind of a 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 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.

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 Lemma 1. Suppose that the firm offers each consumer i a price of $-u_i(\mu_0)$ and all consumers sell their data. Because of the complementarity, price $-u_i(\mu_0)$ is consumer i's loss of selling her data conditional on that other consumers sell their data. A similar argument as Proposition 1 implies this is an equilibrium. All consumers receive zero payoffs, and the firm obtains $\pi(\mu_0) + \sum_{i\in\mathcal{N}} u_i(\mu_0)$. This outcome maximizes the payoff of each consumer, who never obtains a positive payoff.[7] Also, it minimizes the firm's payoff, because for any $\boldsymbol{\mu} \in \mathcal{F}(\mu_0)$, the firm can offer each consumer a price of $-u_i(\mu_0)$ to secure $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0)$. As a result, $\boldsymbol{\mu}^*$ maximizes consumer surplus and minimizes the firm's profit. □

Proposition 2 does not consider the case in which collecting μ_0 is inefficient relative to collecting no data, i.e., $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0) < 0$. Appendix A shows that in this case, there is no equilibrium with passive beliefs, but there is a perfect Bayesian equilibrium with non-passive beliefs, in which the firm collects no data and all players obtain zero payoffs.

Proposition 2 extends the "de-correlation scheme" in 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, then (iii) sells the transformed data of those who are willing to sell their data. They assume a Gaussian signal structure, so that Step (ii) is 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).

4.2 Beneficial Data Collection

Assume now that consumers are better off if the firm has more data: For any $\mu, \mu' \in \boldsymbol{\Sigma}$ such that $\mu \succeq \mu'$, $u_i(\mu) \geq u_i(\mu') \geq 0$ for each $i \in \mathcal{N}$. When data collection is beneficial, 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 $\mu_0 \in \boldsymbol{\Sigma}$, 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

[7] 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 consumer i strictly prefers to share her data. Because she holds a passive belief, the firm can lower the price to buy the same data, which is a contradiction.

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 $\mu_0 \in \Sigma$, 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 and all consumers sell their data. This equilibrium maximizes consumer surplus, because there is no equilibrium in which the firm pays a positive price when data collection is beneficial.[8] This 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 exogenous, this free-rider problem would inefficiently lower the level of data provision. However, when prices are endogenous, this free-ride incentive forces the firm to lower fees. As a result, the data externality protects consumers from the monopolist firm.[9]

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

Remark 1 (The set of possible payoffs). The above results characterize the highest and lowest payoffs of the players across all feasible allocations of data. Appendix D shows that any payoff between the best and worst payoffs can arise under some allocation of data.

4.3 The Consumer-Worst Outcome Under General Utilities

If consumers hold non-monotone preferences, a complementary or substitutable allocation of data may no longer provide surplus bounds. However, a combination of these allocations attains the

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

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

consumer-worst outcome under arbitrary preferences.

Proposition 5. *Take any $(u_i(\cdot))_{i\in\mathcal{N}}$ and any aggregate data μ_0. There is an allocation of data $\boldsymbol{\mu}^*$ that minimizes consumer surplus with respect to μ_0. Under $\boldsymbol{\mu}^*$, each consumer i receives payoff $\min_{\mu\preceq\mu_0} u_i(\mu)$.*

Proof. Because $\min_{\mu\preceq\mu_0} u_i(\mu)$ is a lower bound of the payoffs consumer i can secure by refusing to sell her data, it suffices to construct $\boldsymbol{\mu}^*$ that achieves $\min_{\mu\preceq\mu_0} u_i(\mu)$ as an equilibrium payoff of i. The construction is as follows. For each $i\in\mathcal{N}$, pick an experiment $\mu_i^{MIN}\in\arg\min_{\mu\preceq\mu_0} u_i(\mu)$. Also, 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 $\langle\boldsymbol{\nu}^i_{-i}\rangle=\mu_i^{MIN}$. Let $\boldsymbol{\nu}^*$ denote a perfectly complementary allocation 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 offers $p_i^*:=-u_i(\mu_0)+u_i(\mu_i^{MIN})$ to each 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. The firm has no profitable deviation, because $p_i^*\le 0$ holds for all i, and the firm cannot lower the price. □

Intuitively, the aggregate data μ_0 contains information about X that benefits consumer i and information that harms i. In such a case, the worst outcome for consumer i is that other consumers already hold information that decreases i's welfare, but her data contribution is pivotal in realizing the beneficial component of μ_0. The firm can then collect data at negative prices, because each consumer benefits a lot by providing her data. The above result shows that we can indeed construct an allocation of data that attains such a situation simultaneously for all consumers.

We can also find an outcome that minimizes consumers' payoffs across all allocations of data without feasibility constraints. Although consumer preferences can be heterogeneous, the allocation of data that fully reveals the state minimizes the payoffs of all consumers simultaneously.

Corollary 1. *There is an all allocation of data $\boldsymbol{\mu}$ and an equilibrium $E\in\mathcal{E}(\boldsymbol{\mu})$ in which each consumer receives a payoff of $\min_{\mu\preceq\mu_{FULL}} u_i(\mu)$, where μ_{FULL} is the fully informative signal. In E, each consumer's payoff is minimized across all allocations and all equilibria.*

Proof. By replacing μ_0 with μ_{FULL} in the proof of Proposition 5, we can construct an allocation $\boldsymbol{\mu}^*$ such that $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_{FULL} \rangle$ and consumer i receives $\min_{\mu \preceq \mu_{FULL}} u_i(\mu)$, which is the lowest value because i can secure at least $\min_{\mu \preceq \mu_{FULL}} u_i(\mu)$ by not selling data. □

4.4 Discussion on the Multiplicity of Equilibria

Definition 1 selects 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. This definition leaves the possibility that another equilibrium under $\boldsymbol{\mu}^*$ leads to a low consumer surplus. For example, Proposition 4 selects an equilibrium 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.

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

If data collection is harmful, the equilibrium is unique under the consumer-optimal allocation of data studied in Proposition 2. Under the consumer-worst allocation in Proposition 2, the equilibrium is unique if $\pi(\mu_0) + u_i(\mu_0) > 0$ for some i. Given this inequality, 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.

Another restriction in Definition 1 is that I compare equilibria with passive beliefs. Allowing non-passive beliefs is likely to introduce other equilibria, which the above selection argument may not eliminate. For example, consumers can sustain high prices with a belief system such that if the firm decreases the price, each consumer i believes that the firm collects data from other consumers in a way that increases i's cost of selling data. However, allowing non-passive beliefs does not change the surplus bounds in the above results, except for the consumer-optimal allocation under harmful data collection. Appendix C provides details.

5 Applications

This section provides two applications of the above results. First, I study the firm's data collection problem. Second, I consider a firm that uses data to price discriminate consumers in the product market. These applications show that we can use the results in the previous section to solve problems in which the aggregate data is not fixed.

5.1 Optimal Data Collection

I study a profit maximization problem of a firm that can request any data from consumers. Although it is a strong assumption that the firm can potentially source any information, we may view the problem as the first-best benchmark for the firm. Also, the unconstrained problem admits a clean characterization. The firm's problem is equivalent to finding an allocation of data that maximizes a profit without any feasibility constraint:

Definition 4. An allocation of data $\boldsymbol{\mu}^*$ *globally maximizes the firm's profit* if there is an equilibrium $E^* \in \mathcal{E}(\boldsymbol{\mu}^*)$ such that for any allocation of data $\boldsymbol{\mu}$ and $E \in \mathcal{E}(\boldsymbol{\mu})$, the firm's expected payoff in E^* is higher than the one in E.

If $n = 1$, any welfare-maximizing experiment $\mu^* \in \arg\max_{\mu \in \boldsymbol{\Sigma}} \pi(\mu) + u_1(\mu)$ also globally maximizes the firm's profit (see Claim 1). The following result characterizes the firm-optimal allocation of data for $n \geq 2$ (Remark 2 and Appendix F provide conditions under which the problem (1) has a solution; hereafter, we focus on the payoff functions that satisfy those conditions).

Proposition 6. *Suppose $n \geq 2$ and take any $(u_i(\cdot))_{i \in \mathcal{N}}$. Let $\mu_0^* \in \boldsymbol{\Sigma}$ solve*

$$\max_{\mu \in \boldsymbol{\Sigma}} \left(\pi(\mu) + \sum_{i \in \mathcal{N}} u_i(\mu) - \sum_{i \in \mathcal{N}} \min_{\mu_i \preceq \mu} u_i(\mu_i) \right). \tag{1}$$

Then, there is an allocation $\boldsymbol{\mu}^*$ that satisfies $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0^* \rangle$ and globally maximizes the firm's profit. The maximum payoff of the firm associated with $\boldsymbol{\mu}^*$ is (1).

Proof. Take a maximizer μ_0^* of (1). Proposition 5 implies that there is an allocation of data $\boldsymbol{\mu}^*$ under which each consumer i receives a payoff of $\min_{\mu \preceq \mu_0^*} u_i(\mu)$. Under this allocation, the firm's profit is $\pi(\mu_0^*) + \sum_{i\in\mathcal{N}} u_i(\mu_0^*) - \sum_{i\in\mathcal{N}} \min_{\mu_i \preceq \mu_0^*} u_i(\mu_i)$. Take any allocation of data $\boldsymbol{\mu}$ and any equilibrium $E \in \mathcal{E}(\boldsymbol{\mu})$. Let $\boldsymbol{\mu}' = (\mu_1', \dots, \mu_n')$ denote the data the firm collects in E (we may have $\mu_i' = \mu_\emptyset$ if the firm does not collect data from consumer i). Because the consumer can secure a payoff of $u_i(\boldsymbol{\mu}'_{-i})$ by not selling her data, the firm's profit in E is at most

$$\begin{aligned}
&\left(\pi(\boldsymbol{\mu}') + \sum_{i\in\mathcal{N}} u_i(\boldsymbol{\mu}') - \sum_{i\in\mathcal{N}} u_i(\boldsymbol{\mu}'_{-i})\right) \\
\leq &\left(\pi(\boldsymbol{\mu}') + \sum_{i\in\mathcal{N}} u_i(\boldsymbol{\mu}') - \sum_{i\in\mathcal{N}} \min_{\mu_i \preceq \boldsymbol{\mu}'} u_i(\mu_i)\right) \\
\leq &\left(\pi(\mu_0^*) + \sum_{i\in\mathcal{N}} u_i(\mu_0^*) - \sum_{i\in\mathcal{N}} \min_{\mu_i \preceq \mu_0^*} u_i(\mu_i)\right).
\end{aligned}$$

Here, the first inequality is from $\boldsymbol{\mu}'_{-i} \preceq \boldsymbol{\mu}'$, and the second inequality is from (1). Therefore, $\boldsymbol{\mu}^*$ globally maximizes the firm's profit, and $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0^* \rangle$. □

The objective (1) 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. The third term $\min_{\mu_i \preceq \mu} u_i(\mu_i)$ 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. The firm can then collect i's data at a lower price. The following result is a direct corollary.

Corollary 2. *The firm-optimal allocation of data in Proposition 6 has the following properties.*

1. *If each $u_i(\cdot)$ is monotone in $\succeq$, the firm fully learns the state.*
2. *Any efficient experiment $\mu_E^* \in \arg\max_{\mu \in \boldsymbol{\Sigma}} \pi(\mu) + \sum_{i\in\mathcal{N}} u_i(\mu)$ cannot be strictly more informative than any firm-optimal allocation of data $\boldsymbol{\mu}^*$.*

Proof. To prove Point 1, take any $i \in \mathcal{N}$. If $u_i(\cdot)$ is increasing in $\succeq$, $\min_{\mu_i \preceq \mu} u_i(\mu_i) = u_i(\mu_\emptyset) = 0$. If $u_i(\cdot)$ is decreasing in $\succeq$, $\min_{\mu_i \preceq \mu} u_i(\mu_i) = u_i(\mu)$. Thus, the maximand in (1) becomes $\Pi(\mu) := \pi(\mu) + \sum_{i\in\mathcal{N}_+} u_i(\mu)$, where $\mathcal{N}_+$ is the set of i's such that $u_i(\cdot)$ is increasing. Because $\Pi(\cdot)$ is increasing in $\succeq$, the fully informative experiment solves the maximization problem (1).

For Point 2, suppose there is a firm-optimal allocation $\boldsymbol{\mu}^*$ that is strictly less informative than μ_E^*. Because μ_E^* maximizes total surplus, we have

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

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

$$\pi(\boldsymbol{\mu}^*) + \sum_{i \in \mathcal{N}} u_i(\boldsymbol{\mu}^*) - \sum_{i \in \mathcal{N}} \min_{\mu' \preceq \boldsymbol{\mu}^*} u_i(\mu') \leq \pi(\mu_E^*) + \sum_{i \in \mathcal{N}} u_i(\mu_E^*) - \sum_{i \in \mathcal{N}} \min_{\mu' \preceq \mu_E^*} u_i(\mu'). \tag{2}$$

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

With a stronger assumption, we can solve the firm's problem (1) using concavification (e.g., Aumann and Maschler 1995 and Kamenica and Gentzkow 2011). To state the next result, we assume that there are functions $\hat{\pi} : \Delta(\mathcal{X}) \to \mathbb{R}$ and $\hat{u} : \Delta(\mathcal{X}) \to \mathbb{R}$ such that for each $\mu \in \boldsymbol{\Sigma}$, we have $\pi(\mu) = \int_{\Delta(\mathcal{X})} \hat{\pi}(b) d\langle\mu\rangle(b)$ and $u(\mu) = \int_{\Delta(\mathcal{X})} \hat{u}(b) d\langle\mu\rangle(b)$.[10] For simplicity, we identify π and u with $\hat{\pi}$ and $\hat{u}$, respectively. Also, for any function $f : \Delta(\mathcal{X}) \to \mathbb{R}$, let $\mathcal{V}[f]$ denote a concavification of f, and let $\mathcal{V}[f](b)$ denote the concavification evaluated at $b \in \Delta(\mathcal{X})$.[11] If the consumers share the same utilities, we can solve the firm's problem with a two-step concavification method.

Corollary 3. *Assume all of the $n \geq 2$ consumers have the same utility function, $\frac{1}{n}u(\cdot)$. Given the common prior $b_0 \in \Delta(\mathcal{X})$, the firm's payoff* (1) *under the optimal allocation of data is*

$$\mathcal{V}\left[\mathcal{V}[\pi + u] - u\right](b_0). \tag{3}$$

Proof. Because consumers share the same utility $\frac{1}{n}u(\cdot)$, we can write the firm's problem (1) as

$$\max_{\nu \in \boldsymbol{\Sigma}} \left\{ \max_{\mu \succeq \nu} \left(\pi(\mu) + u(\mu) - u(\nu)\right) \right\}. \tag{4}$$

[10] Functions $\hat{\pi}$ and $\hat{u}$ exist, for example, if the firm chooses some action a after learning about X from the information collected, and the ex post payoff of each player depends only on (a, X). Appendix F provides details.

[11] A concavification of a function f is the smallest concave function that is everywhere weakly greater than f. For details, see, e.g., Kamenica and Gentzkow (2011).

Fix any $\nu \in \mathbf{\Sigma}$, and consider the problem $\max_{\mu \succeq \nu} (\pi(\mu) + u(\mu) - u(\nu))$, which is an information design problem in which the designer chooses μ that is more informative than ν to maximize $\pi(\mu) + u(\mu) - u(\nu)$. Appendix E shows

$$\max_{\mu \succeq \nu} (\pi(\mu) + u(\mu) - u(\nu)) = \int_{\Delta(\mathcal{X})} \{\mathcal{V}[\pi + u](b) - u(b)\}\, d\langle \nu \rangle(b). \tag{5}$$

As a result, (4) is written as

$$\max_{\nu \in \mathbf{\Sigma}} \int_{\Delta(\mathcal{X})} \{\mathcal{V}[\pi + u](b) - u(b)\}\, d\langle \nu \rangle(b), \tag{6}$$

which equals (3). □

We now apply Corollary 3 to an example.

Example 1. Suppose $\mathcal{X} = \{0, 1\}$, so that we can identify $\Delta(\mathcal{X})$ with $[0, 1]$, where $b \in [0, 1]$ is the probability of $X = 1$. Assume $\pi(\cdot) \equiv 0$, and consumers share the same utility $\frac{1}{n} u(b)$. In Figure 1, the black and red solid lines depict $u(b)$ and $\mathcal{V}[u](b)$, respectively. Figure 2 depicts $\mathcal{V}[u](b) - u(b)$ and its concavification, $\mathcal{V}[\mathcal{V}[u] - u](b)$.

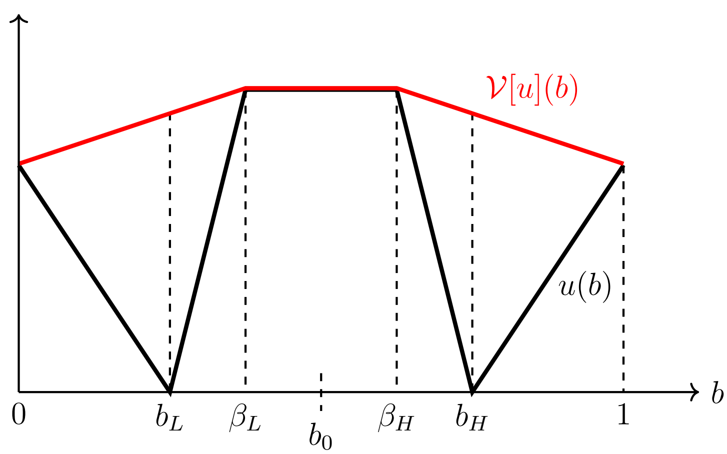

Figure 1: Utility u and its concavification $\mathcal{V}[u]$

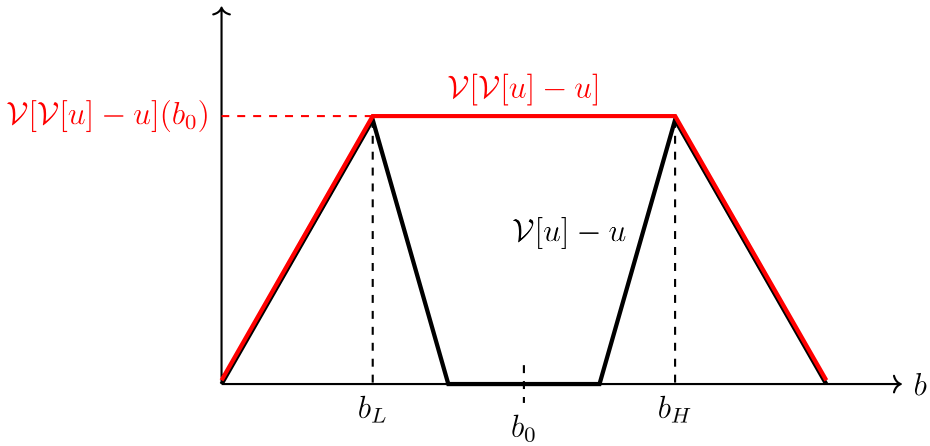

Figure 2: Function $\mathcal{V}[u] - u$ and its concavification

Corollary 3 states that across all allocations of data and equilibria, the maximum payoff of the firm given prior b_0 is $\mathcal{V}[\mathcal{V}[u] - u](b_0)$ (see Figure 2). We can also derive the firm-optimal signal (i.e., μ_0^* in Proposition 6) from the two-step concavification. First, the concavification in Figure 2 splits the prior b_0 into b_L and b_H. This step corresponds to the concavification of $\mathcal{V}[u] - u$. Second, Figure 1 splits b_L into 0 and β_L, and splits b_H into β_H and 1. This step corresponds to the concavification of u. As a result, the firm-optimal signal generates four posteriors, $0, \beta_L, \beta_H$, and 1. Under the firm-optimal allocation, the data of any $n - 1$ consumers induce posteriors b_L and b_H. These posteriors do not arise on the equilibrium path, but increase the firm's profit by reducing the outside option of consumers from refusing to sell their data. Finally, the firm-optimal signal is strictly more informative than the one that maximizes total surplus, which induces posteriors in $[\beta_L, \beta_H]$. This observation conforms to Corollary 2.

Remark 2. Proposition 6 does not by itself tell us that a solution of (1) exists, but Appendix F provides sufficient conditions under which a solution exists. In particular, it studies a setting in which the firm learns about the state from collected information, then takes an action that affects the payoffs of all players. In this setting, $\pi(\mu)$ and $(u_i(\mu))_{i\in\mathcal{N}}$ are the equilibrium payoffs of the subgame in which the firm has collected μ. Then, the problem (1) has a solution, provided the firm has finitely many actions and breaks ties in a certain way.

5.2 Monopoly Price Discrimination

This subsection assumes that the firm uses data to price discriminate. 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.[12] X has a finite support and is positive with probability 1.

Given the allocation of data $\boldsymbol{\mu}$ about X, 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 observe X and 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.[13] Let $u_\emptyset$ and $\pi_\emptyset$ denote the average expected consumer surplus and the average firm profit, when the firm buys no 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 characterizes possible pairs of consumer surplus and the seller profit across all information structures. 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. *Suppose $n = 1$. 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.*

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

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

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 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 Point 2 holds. I write π_μ for the firm profit in the product market given $\mu \in \mathbf{\Sigma}$. Bergemann et al. (2015b) shows that there is a $\mu^* \in \mathbf{\Sigma}$ 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 doing so does not change her payoff in the product market. □

To state the main result, define the surplus triangle:

$$\Delta := \left\{(\pi, u) \in \mathbb{R}^2 : \pi + u \leq \bar{w}, u \geq 0, \pi \geq \pi_\emptyset\right\}. \tag{7}$$

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

Proposition 7. *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 a $\mu^* \in \mathbf{\Sigma}$ such that if the firm has μ^*, the resulting average outcome in the product market is (π^*, u^*). Suppose $u^* < u_\emptyset$ (resp. $u^* \geq u_\emptyset$). Proposition 1 (resp. Proposition 4) implies that there is an allocation of data such that the firm collects μ^* at a price of zero. In the 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 3 depicts the possible outcomes for $n = 1$ and $n \geq 2$. 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 for some 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, because the firm compensates consumers for providing data if they hold complementary 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.

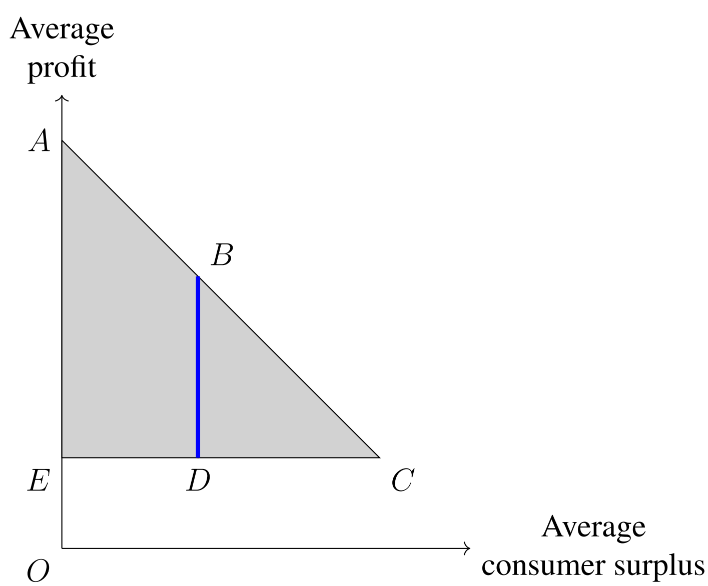

Figure 3: The set of possible outcomes for $n = 1$ (blue line, BD) and $n \geq 2$ (gray triangle, AEC).

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, we can find 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. The 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.

Aliprantis, Charalambos D and Kim Border (2006), *Infinite dimensional analysis: a hitchhiker's guide*. Springer Science & Business Media.

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.

Aumann, Robert J and Michael Maschler (1995), *Repeated games with incomplete information*. MIT press.

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."

Kamenica, Emir and Matthew Gentzkow (2011), "Bayesian persuasion." *American Economic Review*, 101, 2590–2615.

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.

Online Appendix

A Appendix for Proposition 2

Proposition 2 shows that a perfectly complementary allocation of data maximizes consumer surplus and minimizes the firm's profit if $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0) \geq 0$. We now assume $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0) < 0$.

Claim 3. *If $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0) < 0$ and $\pi(\mu_0) > 0$, under a perfectly complementary allocation of data, there is no perfect Bayesian equilibrium with passive beliefs.*

Proof. Suppose to the contrary that there is an equilibrium. First, suppose that some consumer i accepts a positive price with a positive probability. On this event, the firm collects data from all other consumers; otherwise, the firm would collect an uninformative signal and make a negative profit from i (with a positive probability conditional on the event), but it could then profitably deviate by offering a negative price to i. As a result, whenever a consumer observes a positive price, she correctly anticipates that the firm will collect data from all other consumers. Thus, the firm has to offer a price of at least $-u(\mu_0)$ to each consumer i whenever it collects data. However, such an offer leads to a negative profit, which is a contradiction. Second, suppose the firm collects no data with probability 1. Because consumers have passive beliefs, the firm can then deviate and collect all data at an arbitrarily small positive price. This is a contradiction. □

Claim 4. *If $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0) < 0$, under a perfectly complementary allocation of data, there is a perfect Bayesian equilibrium in which the firm collects no data and all players obtain zero payoffs.*

Proof. Consider the following strategy profile: On the equilibrium path, the firm offers a price of zero to all consumers, who refuse to sell their data. If the firm deviates and increases a price to consumer i, then i believes that the firm offers a greater price than $-u_j(\mu_0)$ and buys data from all other consumers $j \neq i$. This is an equilibrium: Given the consumers' non-passive beliefs, the firm can acquire data μ_0 only by paying $-u_i(\mu_0)$ or more to each consumer i, but such an offer will lead to a negative profit. As a result, the firm has no incentive to deviate. □

B Proofs for Remark 4.4: Multiplicity of Equilibria

I impose the following restriction on consumer preferences.

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

Assumption 1 is weak. It holds if each consumer i has some underlying payoff $u_i(a, X)$ that depends on the firm's (unmodeled) action a and a realized state X. Denoting the firm's action at a posterior $b \in \Delta(\mathcal{X})$ 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.

B.1 "Approximation result" for Proposition 3

Proposition 8. *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 $\boldsymbol{\mu}^C = (\mu_1^C, \dots, \mu_n^C)$ such that $\langle\boldsymbol{\mu}^C\rangle = \mu_0$. For each $k \in \mathbb{N}$, consider the following allocation $(\mu_i^k)_{i\in\mathcal{N}}$: (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 $p_1 = \varepsilon > 0$. Consumer 1 then strictly prefers to sell data because she earns ε and benefits from the firm's learning. Consider the positive probability event in which (ii) is realized and $\mu_1^k = \mu_0$. 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 i is $p_i^k = -u_i(\mu^k_{1,\mathbf{1}_{-i}}) + u_i(\mu^k_{0,\mathbf{1}_{-i}}) = -\left(1 - \frac{1}{k} + \frac{1}{nk}\right) u_i(\mu_0)$. In the unique equilibrium, the firm charges each consumer i a non-positive 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. □

B.2 "Approximation result" for Proposition 4

Proposition 9. *Fix any aggregate data $\mu_0 \in \Sigma$ 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 $\sum_{i\in\mathcal{N}} u_i(\mu_0)$, 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., (ii) is the same as the one in the proof of Proposition 8). By the same argument as the proof of Proposition 8, 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. □

C Equilibria with Non-Passive Beliefs

This subsection provides three observations. First, I present an example in which an equilibrium with non-passive beliefs exists under an extreme allocation of data identified in the main analysis. Second, I present an example in which non-passive beliefs can increase the surplus bound for the consumer-optimal outcome under harmful data collection. Finally, I show that other surplus bounds do not change even if we allow non-passive beliefs.

C.1 Example of an Equilibrium with Non-Passive Beliefs

Consider harmful data collection with a perfectly substitutable allocation of data. For simplicity, suppose all consumers have the same $u_i(\cdot)$. Proposition 1 shows that there is an equilibrium in which each consumer i receives a payoff of $u_i(\mu_0) \leq 0$, which minimizes consumer surplus. Under the same allocation, there is a PBE in which one consumer (say 1) receives a payoff of zero, and other consumers continue to receive $u_i(\mu_0)$. To see this, consider the following strategy profile: The firm offers price $-u_1(\mu_0)$ to consumer 1, who sell her data, and it offers a price of zero to other consumers, who do not sell their data. Whenever the firm deviates and offers a different price to consumer i, she believes that the firm offers negative prices to all consumers $j \neq i$ and they refuse to sell their data. If $\pi(\mu_0) + u_1(\mu_0) \geq 0$, this is an equilibrium. In particular, unlike the case of passive beliefs, the firm can no longer collect data for free from consumer $j \neq 1$, because any consumer who detects a deviation believes that her data provision is pivotal. In this equilibrium, consumer surplus is greater than the equilibrium (with passive beliefs) in which all consumers provide their data for free.

C.2 Non-Passive Beliefs: Consumer-Optimal Allocation Under Harmful Data Collection

Suppose data collection is harmful. Proposition 2 shows that under a certain condition, a perfectly complementary allocation attains the maximum consumer surplus of zero. I show that a PBE with non-passive beliefs can attain a positive consumer surplus.

To see this, suppose the state is two-dimensional: $X = (X_1, X_2)$. For each $j \in \{1, 2\}$, let $\mu(j)$ denote the experiment that fully reveals the realization of X_j and is uninformative about the other dimension. Also, let $\mu(12)$ denote the experiment that full reveals the state.

Consider the following payoffs: $u(\mu(1)) = u(\mu(2)) = -1$, $u(\mu(12)) = -3$, and $\pi(\mu(1)) = \pi(\mu(2)) = \pi(\mu(12)) = 4$. The setting satisfies the assumption in Proposition 2. Suppose there are two consumers, and consumers 1 and 2 hold $\mu(1)$ and $\mu(2)$, respectively. Consider the following equilibrium: The firm collects data only from consumer 1 at price 2 and offers a price of zero to consumer 2. Consumer 2 has the passive belief. Consumer 1, in contrast, believes that if the firm deviates, it collects data from consumer 2 at price (say) $+\infty$. The firm has no profitable deviation: If the firm decreases the price to consumer 1, she refuses to sell her data, because she believes that consumer 2 sells her data, so consumer 1's loss of selling her data is now $u(\mu(12)) - u(\mu(2)) = -2$ as opposed to -1. Also, the firm does not benefit from buying data from consumer 2, because she requires a price of at least 2, and the marginal value of the second unit of data is zero to the firm. In this equilibrium, consumer surplus is 1, which is strictly greater than the maximum value under passive beliefs.

C.3 Non-Passive Beliefs: The Robustness of Other Surplus Bounds

Except for the consumer-optimal allocation under harmful data collection, allowing non-passive beliefs does not change the surplus bounds. First, the consumer-worst outcomes remain the same, because I derive the results based on the observations that consumers can secure a payoff of $u_i(\mu_0)$ (under harmful data collection) or 0 (under beneficial data collection) by refusing to sell data. Because this payoff is a lower bound of a consumer's outside option for any beliefs, non-passive beliefs do not change the consumer-worst outcome (by the same argument, Proposition 5 extends). The rest of the results continue to hold, because they only use the upper and lower bounds of possible equilibrium prices that do not depend on consumers' beliefs. The firm-best and the firm-worst outcomes under harmful data collection remain the same, because the result only uses the fact that the firm cannot charge negative prices but can always pay $-u_i(\mu_0)$ to collect data from each consumer i. Also, the firm-best and the consumer-worst outcomes under beneficial data collection remain the same, because the firm can always collect data at a price arbitrarily close to zero. Finally, the firm-worst outcome under beneficial data collection remains the same, because the firm cannot charge more than $u_i(\mu_0)$ on each consumer i.

D The Set of Equilibrium Payoffs

This appendix provides results on the set of all equilibrium payoffs across all feasible allocations of data. Throughout the appendix, we fix the aggregate data, μ_0.

Claim 5. *Consider harmful data collection with* $\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0) \geq 0$. *For each player, any payoff between the worst and the best payoffs can arise under some data allocation: For any* $\pi \in [\pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0), \pi(\mu_0)]$, *there is some feasible allocation of data and the corresponding equilibrium in which the firm receives a payoff of* π. *Similarly, for any* $U \in [\sum_{i \in \mathcal{N}} u_i(\mu_0), 0]$, *there is some feasible allocation of data and the corresponding equilibrium in which the consumer surplus is U.*

Proof. Take any $\alpha \in [0,1]$. Take a perfectly complementary allocation of data $\boldsymbol{\mu}^C = (\mu_1^C, \ldots, \mu_n^C)$ such that $\langle \boldsymbol{\mu}^C \rangle = \mu_0$. Consider the following allocation $(\mu_i^*)_{i \in \mathcal{N}}$: (i) with probability α, $(\mu_i^*)_{i \in \mathcal{N}} = \boldsymbol{\mu}^C$, and (ii) with probability $1-\alpha$, $\mu_i^* = \mu_0$ for all $i \in \mathcal{N}$. Conditional on that all consumers $j \neq i$ sells their data, consumer i's utility from refusing to sell data is $(1-\alpha)u_i(\mu_0)$. If she sells her data, her utility is $u_i(\mu_0)$. As a result, the minimum price consumer i is willing to accept to sell her data is $p_i^* = u_i(\mu_0) - (1-\alpha)u_i(\mu_0) = \alpha u_i(\mu_0)$. Now, there is an equilibrium in which the firm offers price $\alpha u_i(\mu_0)$ to each consumer i and all consumers sell their data. In particular, if the firm buys data only from consumer i, the firm loses the gross revenue of $\alpha\pi(\mu_0)$ and saves the total price of $-\alpha \sum_{j \neq i} u_j(\mu_0)$. Because $\pi(\mu_0) + \sum_{j \neq i} u_j(\mu_0) \geq \pi(\mu_0) + \sum_{j \in \mathcal{N}} u_j(\mu_0) \geq 0$, the firm does not benefit from such a deviation. A similar argument implies that the firm has no profitable deviation, and neither do consumers. In this equilibrium, the firm's payoff is $\pi(\mu_0) + \alpha \sum_{i \in \mathcal{N}} u_i(\mu_0)$, and the consumer surplus is $(1-\alpha)\sum_{i \in \mathcal{N}} u_i(\mu_0)$. By moving α from 0 to 1, we obtain the result. □

By the same argument as above, we obtain the following result.

Claim 6. *Consider beneficial data collection. For each player, any payoff between the worst and the best payoffs can arise under some data allocation: For any $\pi \in [\pi(\mu_0), \pi(\mu_0) + \sum_{i \in \mathcal{N}} u_i(\mu_0)]$, there is some feasible allocation of data and the corresponding equilibrium in which the firm receives a payoff of π. Similarly, for any $U \in [0, \sum_{i \in \mathcal{N}} u_i(\mu_0)]$, there is some feasible allocation of data and the corresponding equilibrium in which the consumer surplus is U.*

E Appendix for the Proof of Corollary 3

In this appendix, we identify μ with $\langle \mu \rangle$. We prove equation (5), which is equivalent to

$$\max_{\mu \succeq \nu} \int_{\Delta(\mathcal{X})} f(b) d\mu(b) = \int_{\Delta(\mathcal{X})} \mathcal{V}[f](b) d\nu(b), \tag{8}$$

where $f = \pi + u$. First, by the definition of $\mathcal{V}[f]$, we can find a mean-preserving transition kernel $Q : b \mapsto Q(\cdot|b) \in \Delta(\Delta(\mathcal{X}))$ such that[14]

$$\int_{\Delta(\mathcal{X})} \mathcal{V}[f](b)d\nu(b) = \int_{\Delta(\mathcal{X})} \int_{\Delta(\mathcal{X})} f(\beta)dQ(\beta|b)d\nu(b) = \int_{\Delta(\mathcal{X})} f(b)d\mu'(b),$$

where $\mu'(\cdot) := \int_{\Delta(\mathcal{X})} Q(\cdot|b)d\nu(b)$. Because $\mu' \succeq \nu$, we have

$$\max_{\mu \succeq \nu} \int_{\Delta(\mathcal{X})} f(b)d\mu(b) \geq \int_{\Delta(\mathcal{X})} \mathcal{V}[f](b)d\nu(b). \tag{9}$$

To show the converse, take any $\mu \succeq \nu$. Given a mean-preserving transition kernel Q that satisfies $\mu(\cdot) = \int_{\Delta(\mathcal{X})} Q(\cdot|b)d\nu(b)$, we have

$$\int_{\Delta(\mathcal{X})} f(b)d\mu(b) = \int_{\Delta(\mathcal{X})} \int_{\Delta(\mathcal{X})} f(\beta)dQ(\beta|b)d\nu(b) \leq \int_{\Delta(\mathcal{X})} \mathcal{V}[f](b)d\nu(b).$$

Thus we obtain (8). □

F The Existence of a Solution in Problem (1)

I provide conditions under which the firm's problem (1), which is written as

$$\max_{\mu \in \mathbf{\Sigma}} \left\{ \pi(\mu) + \sum_{i \in \mathcal{N}} u_i(\mu) + \sum_{i \in \mathcal{N}} \max_{\mu' \in G(\mu)} (-u_i(\mu')) \right\}, \quad \text{where} \tag{10}$$

$$G(\mu) = \{\mu' \in \mathbf{\Sigma} : \mu' \preceq \mu\}, \tag{11}$$

has a solution. Throughout the appendix, I view $\mu \in \mathbf{\Sigma}$ as a Bayes plausible element of $\Delta(\Delta(\mathcal{X}))$, and use weak* topology on $\Delta(\Delta(X))$.

First, if $\pi(\cdot)$ and $(u_i(\cdot))_{i \in \mathcal{N}}$ satisfy a certain continuity assumption, a version of Berge Maximum Theorem ensures the existence of a solution.

Claim 7. *Suppose $\pi(\cdot)$ is upper semicontinuous and $(u_i(\cdot))_{i \in \mathcal{N}}$ is continuous on $\Delta(\Delta(\mathcal{X}))$ with respect to weak* topology. Then, the problem (1) has a solution.*

[14] A function $Q : \Delta(\mathcal{X}) \to \Delta(\Delta(\mathcal{X}))$ is a mean-preserving transition kernel if, for all $b \in \Delta(\mathcal{X})$, $\int_{\Delta(\mathcal{X})} \beta Q(d\beta|b) = b$. Because I only use probability distributions with finite supports, I omit the formal measure-theoretic description, such as a Borel σ-algebra for $\Delta(\mathcal{X})$.

Proof. If the correspondence $G(\cdot)$ is compact-valued and upper hemicontinuous, Lemma 17.30 of Aliprantis and Border (2006) implies that $\sum_{i \in \mathcal{N}} \max_{\mu' \in G(\mu)} (-u_i(\mu'))$ is upper semicontinuous in μ. Then, the maximand of (10) becomes upper semicontinuous. As a result, Theorem 2.43 of Aliprantis and Border (2006) implies that a solution exists.

Thus, it suffices to show that $G(\cdot)$ is compact-valued and upper hemicontinuous. Let $\Phi \subset \mathbb{R}^{\Delta(\mathcal{X})}$ denote the set of all continuous convex functions. By Theorem 7 of Blackwell (1953), we can write $G(\mu)$ as

$$
\begin{aligned}
G(\mu) &= \bigcap_{\phi \in \Phi} \left\{ \mu' \in \Delta(\Delta(\mathcal{X})) : \int_{\Delta(\mathcal{X})} \phi d\mu \geq \int_{\Delta(\mathcal{X})} \phi d\mu' \right\} \\
&= \bigcap_{\phi \in \Phi} G(\phi, \tau),
\end{aligned}
$$

where $G(\phi, \mu) = \left\{ \mu' \in \Delta(\Delta(\mathcal{X})) : \int_{\Delta(\mathcal{X})} \phi d\mu \geq \int_{\Delta(\mathcal{X})} \phi d\mu' \right\}$. The set $G(\phi, \mu)$ is compact, because the set of probability distributions on $\Delta(\mathcal{X})$ is compact, and $G(\phi, \mu)$ is a closed subset of it. Indeed, if $\{\mu_n\} \subset G(\phi, \mu)$ and $\mu_n \to \mu_0$, then $\int \phi \mu_n \to \int \phi \mu_0$ for any $\phi \in \Phi$, because ϕ is a continuous function defined on a compact set $\Delta(\mathcal{X})$ (and thus bounded). Thus, if $\mu_n \to \mu_0$ and $\int \phi d\mu \geq \int \phi d\mu_n$, then $\int \phi d\mu \geq \int \phi d\mu_0$. This implies $\mu_0 \in G(\phi, \mu)$, and thus $G(\phi, \mu)$ is closed.

The upper hemicontinuity of $G(\phi, \cdot)$ is shown as follows. Take $\hat{\mu}_n \to \hat{\mu}$ and $\mu_n \to \mu$ such that $\int \phi d\hat{\mu}_n \geq \int \phi d\mu_n$ for any n. Then, by the similar argument as above, we get $\int \phi d\hat{\mu} \geq \int \phi d\mu$, and thus $\mu \in G(\phi, \hat{\mu})$. Since $G(\phi, \cdot)$ has a closed graph and is closed-valued, it is upper hemicontinuous. Because $G(\mu)$ is the intersection of compact-valued upper hemicontinuous correspondences, Point 2 of Theorem 17.25 of Aliprantis and Border (2006) implies that it is upper hemicontinuous. □

While Claim 7 provides a condition for the existence, it might be unclear whether the condition holds in applications. For example, suppose that the firm uses a signal to learn about X and then takes a payoff-relevant action. In such a case, $\pi(\mu)$ and $(u_i(\cdot))_{i \in \mathcal{N}}$ are not the primitives but the equilibrium payoffs from the firm's optimal behavior.

We now provide clearer conditions for the existence when the firm chooses an action. Consider the following three-stage game. First, the firm sets prices to buy data. Second, consumers decide whether to sell their data. Finally, the firm chooses an action a from a compact subset $\mathcal{A}$ of $\mathbb{R}^m$.

Let $\pi(a, X)$ and $u_i(a, X)$ denote the ex post payoffs of the firm and consumer i if the firm chooses $a \in \mathcal{A}$ and the realized state is X. The solution concept continues to perfect Bayesian equilibrium.

Abusing terminology, we say that an allocation of data $\boldsymbol{\mu}^*$ globally maximizes the firm's profit if $\boldsymbol{\mu}^*$ satisfies Definition 4, where "equilibrium" in Definition 4 now refers to equilibrium of this extended game.

We provide two conditions that guarantee the existence.

Claim 8. *Suppose that $\pi(\cdot, \cdot)$ and $(u_i(\cdot, \cdot))_{i \in \mathcal{N}}$ are continuous, and the firm has a unique best response $a(b) := \arg\max_{a \in \mathcal{A}} \int_{\mathcal{X}} \pi(a, X) db(X)$ for each $b \in \mathcal{X}$. For each $\mu \in \Delta(\Delta(\mathcal{X}))$, define $\pi(\mu) = \int_{\Delta(\mathcal{X})} \pi(a(b), b) d\mu$ and $u_i(\mu) = \int_{\Delta(\mathcal{X})} u_i(a(b), b) d\mu$ for each $i \in \mathcal{N}$. Then, (10) has a solution, μ_0^*. Moreover, there is an allocation of data $\boldsymbol{\mu}^*$ such that $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0^* \rangle$ and $\boldsymbol{\mu}^*$ globally maximizes the firm's profit.*

Proof. Note that $a(b)$ and $\pi(a, X)$ are continuous, and $\mathcal{A}$ is compact and $\mathcal{X}$ is finite. As a result, $\pi(a(b), b) := \int_{\mathcal{X}} \pi(a(b), X) db(X)$ is continuous and bounded. Corollary 15.7 of Aliprantis and Border (2006) implies that $\pi(\mu) = \int_{\Delta(\mathcal{X})} \pi(a(b), b) d\mu$ is (weak*) continuous in μ. By the same argument, $(u_i(\mu))_{i \in \mathcal{N}}$ is continuous in μ. By Claim 7, (10) has a solution. By the construction of $\pi(\mu)$ and $(u_i(\mu))_{i \in \mathcal{N}}$, any solution of (10) has the associated allocation of data that globally maximizes the firm's profit. $\square$

Claim 8 exclude the case in which the firm has finitely many actions and is indifferent between two actions at some posterior. The next result guarantees the existence of the firm-optimal allocation of data in such a case. We prepare some notations. Assume now that $\mathcal{A}$ is finite. For each posterior $b \in \mathcal{X}$, let $a_i(b)$ denote the firm's best response that breaks ties to minimize consumer i's expected payoff. Also, let $a^{\mathcal{N}}(b)$ denote the firm's best response that breaks ties to maximize the sum of the expected payoffs of all players. Define $\pi^{\mathcal{N}}(\mu) := \int_{\Delta(\mathcal{X})} \pi(a^{\mathcal{N}}(b), b) d\mu$, $u_i^{\mathcal{N}}(\mu) := \int_{\Delta(\mathcal{X})} u_i(a^{\mathcal{N}}(b), b) d\mu$, and $u_i^i(\mu) := \int_{\Delta(\mathcal{X})} u_i(a_i(b), b) d\mu$. Give these tie-breaking rules of the firm, the problem similar to (10) has a solution and provides the firm-optimal data collection.

Claim 9. *The problem*

$$\max_{\mu \in \boldsymbol{\Sigma}} \left\{ \pi^{\mathcal{N}}(\mu) + \sum_{i \in \mathcal{N}} u_i^{\mathcal{N}}(\mu) + \sum_{i \in \mathcal{N}} \max_{\mu' \in G(\mu)} \left(-u_i^i(\mu') \right) \right\} \tag{12}$$

has a solution, μ_0^**. There is an allocation of data* $\boldsymbol{\mu}^*$ *such that* $\langle \boldsymbol{\mu}^* \rangle = \langle \mu_0^* \rangle$ *and* $\boldsymbol{\mu}^*$ *globally maximizes the firm's profit.*

Proof. First, we show the following result: Take any function $v(a, X)$. Let $a^v(b)$ denote the firm's best response at each posterior $b \in \mathcal{X}$ such that the firm breaks ties to maximize $v(a, b) := \int_{\mathcal{X}} v(a, X) db(X)$. Then, $v(a(b), b)$ is upper semicontinuous. To show this result, suppose to the contrary that there is an $\varepsilon > 0$ and $\{b_n\} \subset \Delta(\mathcal{X})$ such that $b_n \to b$ but $v(a(b_n), b_n) \geq v(a(b), b) + \varepsilon$ for all $n \in \mathbb{N}$. Because $\mathcal{A}$ is finite, we can choose a subsequence $\left\{b_{m(n)}\right\}$ so that for some $a' \in A$, $a(b_{m(n)}) = a'$ for all n. This implies that

$$v(a(b), b) \geq v(a', b) \geq v(a(b), b) + \varepsilon, \tag{13}$$

which is a contradiction. (The first inequality comes from the tie braking rule in favor of $v(a, X)$.) Thus, $v(a(b), b)$ is upper semi-continuous on $\Delta(\mathcal{X})$.

The above observations implies that $\pi^{\mathcal{N}}(\mu) + \sum_{i \in \mathcal{N}} u_i^{\mathcal{N}}(\mu)$ is upper semicontinuous in μ. Also, because the tie breaking that minimizes consumer i's payoff is equivalent to the one that maximizes $-u_i(a, X)$, $-u_i^i(\mu')$ is upper semicontinuous in μ'. Lemma 17.30 of Aliprantis and Border (2006) then implies that $\max_{\mu' \in G(\mu)} (-u_i^i(\mu'))$ is upper semicontinuous in μ. Therefore, the maximand of (12) is upper semicontinuous, and the problem has a solution.

By the construction of $\pi^{\mathcal{N}}(\mu)$ and $(u_i^{\mathcal{N}}(\mu), u_i^i(\mu))_{i \in \mathcal{N}}$ and the same argument as Proposition 6, any solution of (12) has the associated allocation of data $\boldsymbol{\mu}^*$ that gives the firm a payoff equal to (12). In the (firm-optimal) equilibrium under $\boldsymbol{\mu}^*$, the firm breaks tie to minimize consumer i's payoff whenever she refuses to sell her data, and the firm breaks tie to maximize the total surplus on the equilibrium path.

The allocation $\boldsymbol{\mu}^*$ globally maximizes the firm's profit. To see this, take any allocation $\boldsymbol{\mu}$ and equilibrium $E \in \mathcal{E}(\boldsymbol{\mu})$. We modify E so that upon choosing a, the firm breaks ties to minimize the sum of the payoffs of the consumers who have refused to sell data; if there is no such consumer, the firm breaks ties to maximize the sum of the payoffs of all players. Correspondingly, the firm offers a weakly lower price so that each consumer behaves in the same way as in E but is now indifferent between sharing and not sharing her data. This modification creates a new equilibrium $E' \in \mathcal{E}(\boldsymbol{\mu})$ that gives the firm a weakly greater payoff, because it now collects the same data at lower prices. The firm's payoff in E' is the maximand of (12) evaluated at $\boldsymbol{\mu}$. Therefore, the firm is better off under an equilibrium of $\boldsymbol{\mu}^*$ than any $E \in \mathcal{E}(\boldsymbol{\mu})$. $\square$