Non-competing Data Intermediaries

Shota Ichihashi*

May 28, 2019

Abstract

I study competition among data intermediaries—technology companies and data brokers that collect consumer data and sell them to downstream firms. When firms use consumer data to extract rents, intermediaries have to compensate consumers for personal data. I show that competition among intermediaries fails: If they offer high compensation, consumers share their data with multiple intermediaries, which lowers the price of data in the downstream market and hurts intermediaries. This leads to multiple equilibria with different allocations of data among intermediaries: There is a monopoly equilibrium, and an equilibrium with greater data concentration benefits intermediaries and hurts consumers.

Keywords: information markets, intermediaries, personal data, privacy

*Bank of Canada, 234 Wellington Street West, Ottawa, ON K1A 0G9, Canada. Email: shotaichihashi@gmail.com.
This paper was formerly circulated under the title "Natural Monopoly for Data Intermediaries." I thank Jason Allen, Sitian Liu, and seminar participants at the Bank of Canada, Decentralization Conference 2019, and Yokohama National University. The opinions expressed in this article are the author's own and do not reflect the views of the Bank of Canada.

1 Introduction

This paper studies competition among data intermediaries, which collect and distribute personal data between consumers and firms. Data brokers, such as LiveRamp and Nielsen, collect consumer data and sell them to firms such as retailers and advertisers. Technology companies, such as Google and Facebook, collect user data and share them indirectly through targeted advertising spaces. I regard these companies as data intermediaries and study how they compete for personal data.

Specifically, consider the following situation. Consumers decide whether to join online platforms and provide personal data. Platforms share collected data with third parties, which can earn revenue from data. However, the use of data by third parties could hurt consumers through price discrimination, spam, and further data leakage. Thus, online platforms have to provide consumers valuable services and rewards (e.g., social media) in order to obtain personal data.

I model such a situation as a two-sided market for personal data. The main focus is on the price-setting behavior of data intermediaries. On the one side, they set "prices" to obtain consumer data. The prices represent the quality of online services or rewards that consumers can enjoy in exchange for providing their data. On the other side, intermediaries set prices to sell data to third parties.

The main question is whether competition among intermediaries dissipates profits. The question is important for understanding how the surplus created by data is allocated among economic agents. In traditional markets, the answer to this question is often yes: The idea reminiscent of Demsetz (1968) suggests that intermediaries compete in the upstream market to have market power in the downstream market, and this drives their profits to zero. However, in the market for data, this may not be the case.

The model consists of two stages. In the upstream market, intermediaries make offers to consumers. An offer consists of a set of data and compensation. Then, each consumer decides whether to accept each offer. The benefit for a consumer of accepting an offer is that she can earn compensation. The cost is that an intermediary may sell her data to downstream firms, whose use of data hurts the consumer. Consumers' decisions determine the allocation of data, which specifies whose and what data each intermediary holds. In the downstream market, intermediaries post prices to sell the data to firms.

The key idea of the paper is that intermediaries prefer to *not* compete for data because the competition will lower the price of data. To see this, suppose that intermediaries offer high compensations to obtain more personal data. Consumers then share their data with *multiple* intermediaries. This intensifies price competition and lowers the price of data in the downstream market. The economic force is driven by the non-rivalry of data: Unlike conventional economic goods, the same data can be simultaneously obtained and sold by any number of intermediaries.

I show that this economic force leads to multiple equilibria with different allocations of data among intermediaries. There are two main findings. First, there is a monopoly equilibrium, in which a single intermediary extracts the maximum possible surplus. Other intermediaries have no incentive to compensate consumers for data, because consumers will then share their data with multiple intermediaries. Thus, competition among data intermediaries may not dissipate profits.

Second, I focus on *data concentration*, which refers to a situation where a small number of intermediaries obtain a large amount of data. There are equilibria with different degrees of concentration. I show that data concentration benefits intermediaries and hurts consumers, if consumers have the increasing marginal costs of sharing data and firms have decreasing marginal revenues of using data. This is because large intermediaries can compensate consumers based on the infra-marginal cost and charge firms for the infra-marginal value of data.

Moreover, I study the different margins of data concentration. Data concentration at the *intensive* margin occurs, for example, if an intermediary acquires another intermediary that holds different data on the same group of consumers (e.g., one intermediary has location data and another intermediary has health data on consumers in the US). Data concentration at the *extensive* margin occurs, for example, if an intermediary acquires another intermediary that has the same kind of data on different consumers (e.g., one intermediary has health data on consumers in the US and another intermediary has health data on consumers in the EU). I show that the different margins of data concentration can have different welfare implications.

The paper helps us understand two issues of the data economy. One is why consumers do not seem to be compensated for providing their data (Arrieta-Ibarra et al., 2018; Carrascal et al., 2013). I show that the market for data could fail to reward consumers as suppliers of personal data. This explanation, which does not depend on consumer unawareness or the lack of transparency, could be important, because there has been increasing awareness of data sharing practices, and regulators

have tried to ensure consumers' control over data (e.g., The EU General Data Protection Regulation). The other issue is data concentration in the hands of major Internet platforms (e.g., Sokol and Comerford, 2015). I show that data concentration can arise even though data are non-rivalrous and the model excludes network externalities or returns to scale. Moreover, data concentration can hurt consumers by lowering compensation for data. This result has a potential implication on regulating dominant online platforms.

The rest of the paper is organized as follows. Section 2 discusses related works and Section 3 describes the model. Section 4 considers two benchmarks: One is the case of a monopoly intermediary, and the other is when data are rivalrous. In Section 5, I begin with the case where each consumer has a single piece of data, and then I consider the general case. Section 6 studies the welfare impact of data concentration. Section 7 provides extensions, and Section 8 concludes.

2 Literature Review

This paper relates to two strands of literature. First, it relates to the recent literature on markets for data. Bergemann and Bonatti (2019) consider a monopoly data intermediary and study under what condition the intermediation of data can be profitable. They assume that a downstream firm uses data for price discrimination that hurts consumers. In contrast, I assume that the intermediation is profitable and focus on the issues of competition and data concentration. Moreover, rather than microfound how downstream firms use consumer data, I define preferences directly over the sets of data firms acquire.

More broadly, this paper relates to works on markets for data beyond the context of price discrimination. Bergemann et al. (2018) consider a model of data provision and data pricing. Jones et al. (2018) study, among other things, how different property rights of data affect economic outcomes in a semi-endogenous growth model. Choi et al. (2018) consider consumers' privacy choices in the presence of an information externality. Gu et al. (2018) study data brokers' incentives to merge data. They consider both the submodular and supermodular functions as the revenue function of a downstream firm. In contrast, I assume supermodularity but endogenize the allocation of data among intermediaries. Kim (2018) considers a model of a monopoly advertising platform and studies consumers' privacy concerns, market competition, and vertical integration between the

platform and sellers.

Second, the paper relates to the literature on platform competition in two-sided markets. Relative to this literature, the novelty of this paper is that I consider the combination of negative cross-side externality and multi-homing (which is captured by the non-rivalry of data). The literature typically assumes that the interaction of the two sides is mutually beneficial (e.g., Armstrong (2006); Caillaud and Jullien (2003); Rochet and Tirole (2003)). This is natural in many applications such as video games (consumers and game developers) and credit cards (cardholders and merchants). In this case, platform competition involves undercutting prices charged to at least one side, which is sustainable even if multi-homing is possible. In contrast, I consider the case where one side (i.e., firms) benefits whereas the other side (i.e., consumers) loses from the interactions (i.e., transfer of personal data). In this case, competition involves raising compensation for consumers, which cannot be sustained. Caillaud and Jullien (2003) show that intermediaries have an incentive to make their services non-exclusive to relax price competition. Their result is logically distinct from mine. In their model, the profits of intermediaries that offer non-exclusive services still go to zero as matching frictions disappear. Negative cross-side externalities also appear in models of advertising platforms, such as Anderson and Coate (2005) and Reisinger (2012). There, the presence of advertisers imposes negative externalities on users due to nuisance costs.

3 Model

There are $N \in \mathbb{N}$ consumers, $K \in \mathbb{N}$ data intermediaries, and a single downstream firm. (Section 7 considers multiple downstream firms.) Where it does not cause confusion, N and K denote the sets of consumers and intermediaries, respectively. Figure 1 depicts the game: Intermediaries obtain data in the upstream market and sell them in the downstream market. Below, I describe the model in detail.

Upstream Market

Each consumer $i \in N$ has a finite set of data, $\mathcal{D}_i$. Each element of $\mathcal{D}_i$ is an indivisible and non-rivalrous good. $\mathcal{D} := \cup_{i \in N} \mathcal{D}_i$ denotes the set of all data.

At the beginning of the game, each intermediary $k \in K$ simultaneously makes an *offer* $(D_i^k, \tau_i^k)_{i \in N}$. $\tau_i^k \in \mathbb{R}$ is the amount of compensation that intermediary k is willing to pay for

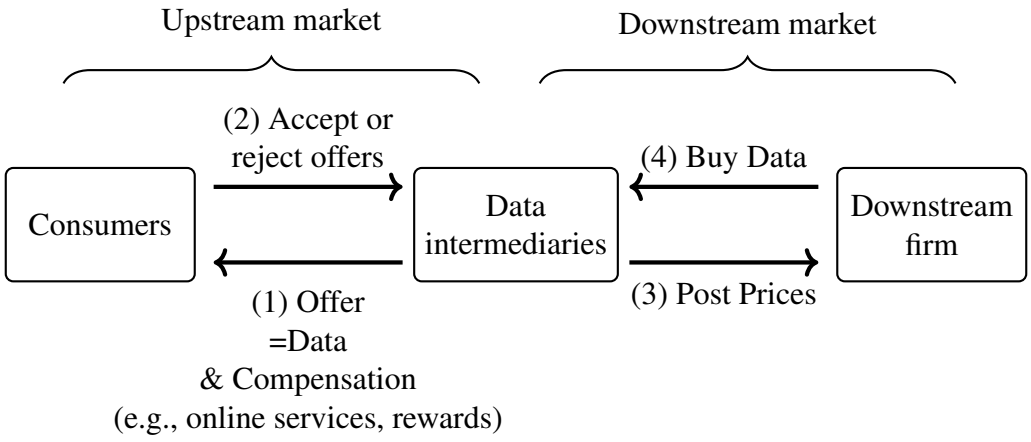

Figure 1: Timing of Moves

data $D_i^k \subset \mathcal{D}_i$. Compensation represents the quality of online services and monetary rewards. If $D_i^k \neq \emptyset$, I call (D_i^k, τ_i^k) a *non-empty offer*. Consumer i does not observe offers to other consumers, $(D_j^k, \tau_j^k)_{k \in K, j \in N \setminus \{i\}}$. This assumption simplifies the analysis by restricting coordination among consumers.

After observing $(D_i^k, \tau_i^k)_{k \in K}$, each consumer i decides which offers to accept. Motivated by the non-rivalry of data, I impose no restriction on the number of offers consumers can accept. Formally, each consumer i simultaneously chooses $K_i \subset K$, where $k \in K_i$ means that consumer i provides data D_i^k to intermediary k and earns τ_i^k. These decisions determine intermediary k's data $D^k = \cup_{i \in N^k} D_i^k$, where $N^k := \{i \in N : k \in K_i\}$ is the set of consumers who accept the offers from intermediary k. All intermediaries and the firm publicly observe $(D^1, \ldots, D^K)$, which I call the *allocation of data*. Given any $D^k \subset \mathcal{D}$, let $D_i^k := D^k \cap \mathcal{D}_i$ denote intermediary k's data on consumer i.

Downstream Market

Given the allocation of data $(D^1, \ldots, D^K)$, each intermediary k simultaneously posts a price $p^k \in \mathbb{R}$ for its data. The firm then chooses the set $K' \subset K$ of intermediaries from which it buys data. As a result, the firm obtains data $D := \cup_{k \in K'} D^k$. Note that the firm obtains consumer i's data $d_i \in \mathcal{D}_i$ if and only if there is $k \in K$ such that $d_i \in D_i^k$ and $k \in K_i \cap K'$. $d_i \in D_i^k$ means that intermediary k asks for d_i, and $k \in K_i \cap K'$ means that consumer i accepts the offer of intermediary k and the firm buys data from k.

Preferences

All players maximize expected payoffs, and their ex post payoffs are as follows. The payoff of each intermediary is revenue minus compensation: Suppose that intermediary k pays compensation τ_i^k to each consumer $i \in N^k$ and posts a price of p_k, and the firm buys data from a set K' of intermediaries. Then, intermediary k obtains a payoff of $\mathbf{1}_{\{k \in K'\}} p_k - \sum_{i \in N^k} \tau_i^k$, where $\mathbf{1}_{\{x \in X\}}$ is the indicator function, which is 1 or 0 if $x \in X$ or $x \notin X$, respectively.

The payoff of each consumer is as follows. Suppose that consumer i earns a compensation of τ_i^k from each intermediary in K_i, and the firm obtains $D_i \subset \mathcal{D}_i$. Then, i's payoff is $\sum_{k \in K_i} \tau_i^k - C_i(D_i)$. The first term is the total compensation. The second term $C_i(D_i)$ is consumer i's cost of sharing data D_i with the downstream firm. I assume that consumers incur the increasing and convex costs of sharing data.

Assumption 1. For each $i \in N$, $C_i : 2^{\mathcal{D}_i} \to \mathbb{R}_+$ satisfies the following.

1. $C_i(D_i)$ is increasing in D_i: For any $X, Y \subset \mathcal{D}_i$ such that $X \subset Y$, $C_i(Y) \geq C_i(X)$.

2. $C_i(D_i)$ is supermodular in D_i: For any $X, Y \subset \mathcal{D}_i$ with $X \subset Y$ and $d \in \mathcal{D}_i \setminus Y$, it holds that

$$C_i(Y \cup \{d\}) - C_i(Y) \geq C_i(X \cup \{d\}) - C_i(X). \tag{1}$$

3. $C_i(\emptyset) = 0$ and $C_i(\{d\}) > 0$ for all $d \in \mathcal{D}_i$.

The assumption holds if $\mathcal{D}_i$ is a singleton. C_i should be thought of as a reduced form capturing a consumer's loss from, say, price discrimination, privacy concern, and intrusive marketing compaign.

The payoff of the downstream firm is as follows. If the firm obtains data $D \subset \mathcal{D}$ and the total payment to intermediaries is p, the firm obtains a payoff of $\Pi(D) - p$. The first term is the firm's *revenue* from data D. I assume that the firm benefits from data but the marginal revenue is decreasing:

Assumption 2. $\Pi : 2^{\mathcal{D}} \to \mathbb{R}_+$ satisfies the following.

1. $\Pi(D)$ is increasing in D: For any $X, Y \subset \mathcal{D}$ such that $X \subset Y$, $\Pi(Y) \geq \Pi(X)$.

2. $\Pi(D)$ is submodular in D: For any $X, Y \subset \mathcal{D}$ with $X \subset Y$ and $d \in \mathcal{D} \setminus Y$, it holds that

$$\Pi(X \cup \{d\}) - \Pi(x) \geq \Pi(Y \cup \{d\}) - \Pi(Y). \tag{2}$$

(If inequality (2) is strict for any $X \subsetneq Y$, Π_i is *strictly* submodular.)

3. $\Pi(\emptyset) = 0$.

Submodularity is motivated by the idea that data typically exhibits decreasing returns to scale (Varian, 2018).

Timing and Solution Concept

The timing of the game, depicted in Figure 1, is as follows. First, each intermediary simultaneously makes an offer to each consumer. After privately observing the offers, each consumer simultaneously decides the set of offers to accept. The decision of each consumer determines the allocation of data. Then, each intermediary simultaneously posts a price to the firm. Finally, the firm chooses the set of intermediaries from which it buys data. I consider pure-strategy perfect Bayesian equilibrium.

3.1 Discussion of Modeling Assumptions and Applications

Before proceeding to the analysis, I discuss modeling assumptions and applications.

Consumers' Benefit and Loss from the Use of Data

The model assumes that the use of data by downstream firms hurts consumers. A motivation for this assumption is that the harmful use of personal data by third parties has been actively discussed by policymakers as a key issue of online privacy problems (Federal Trade Commission, 2014). However, in reality, consumers may share data and enjoy benefits such as social media, search engines, and coupons. The model incorporates these benefits as the endogenous choice of compensation by intermediaries.

Another implicit assumption is that, although intermediaries can choose compensation, they cannot affect consumers' loss C_i from the firm's use of data. This reflects the difficulty of writing

a fully contingent contract over how and which third parties can use personal information. The lack of commitment over the sharing and use of data plays an important role in other models of data economy, such as (Huck and Weizsacker, 2016; Jones et al., 2018).

Finally, it is not crucial that consumers do not incur a loss from sharing data with intermediaries: As discussed in Section 7, even if consumers incur an exogenous loss of sharing data with intermediaries, the main insights continue to hold.

Applications

Online Platforms: We may view data intermediaries as online platforms such as Google and Facebook. The model captures the following situation: Platforms provide online services to consumers in exchange for their data. D_i^k represents the set of data that consumers need to provide to use platform k, and τ_i^k represents the quality of k's service. Platforms may share data with third parties, such as advertisers, retailers, and political consulting firms. Data sharing with each third party can benefit (e.g., better targeting) or hurt (e.g., price discrimination and privacy concern) a data subject, but the aggregate impact is negative and lowers a consumer's payoff by $C_i(D_i^k)$.

Three remarks are in order. First, the model formulates compensation as one-to-one transfer. This is mainly to simplify the analysis, and the results continue to hold as long as compensation is costly for intermediaries. This is natural if an intermediary needs to invest to improve the quality of its service.

Second, in the model, the benefit for consumer i of sharing data with intermediary k depends only on τ_i^k. However, if we interpret intermediaries as online platforms, we may think that the benefit should also depend on how much data other consumers provide (e.g., social media). I exclude such a situation deliberately to clarify that the results are not driven by network externalities or returns to scale.

Finally, the model abstracts from the institutional details of online advertising platforms. For instance, they distribute personal data indirectly through sponsored search or targeted display advertising. For another instance, they compete for not only data but also the attention of consumers. Nonetheless, by regarding these platforms as pure data intermediaries, I can isolate a novel economic mechanism potentially relevant to their competition.

Data Brokers: We can interpret intermediaries as data brokers such as LiveRamp, Nielsen, and Oracle. The business model of these firms is to collect personal information from online and offline sources, and resell or share that information with others such as retailers and advertisers (Federal Trade Commission, 2014).

Some data brokers obtain data from consumers in exchange for monetary compensation (e.g., Nielsen Home Scan). However, data brokers commonly obtain personal information without interacting with consumers. The model could also fit such a situation. To see this, suppose that data brokers obtain individual purchase history from retailers. Consider the following chain of transactions: Retailers compensate customers and record their purchases. For example, retailers may offer discounts to customers who sign up for loyalty cards. Retailers then sell these records to data brokers, which resell the data to downstream firms. We can regard retailers in this example as consumers in the model. The cost C_i represents the compensation that retailer i has to pay to its customers.

Alternatively, the model can be useful for understanding how the incentives of data brokers would look like if they had to source data directly from consumers. The question is of growing importance, as awareness of data sharing practices increases and policymakers try to ensure that consumers have control over their data (e.g., The EU General Data Protection Regulation and California Consumer Privacy Act).

4 Preliminary Analysis

I provide two benchmarks, which will be compared with the main specification.

4.1 Monopoly Intermediary

Consider a monopoly data intermediary. For any data $D \subset \mathcal{D}$, I write consumer i's cost $C_i(D \cap \mathcal{D}_i)$ as $C_i(D)$. If the intermediary obtains and sells data D, consumer i requires compensation of at least $C_i(D)$, and the firm is willing to pay up to $\Pi(D)$. This leads to the following result.

Claim 1. *In any equilibrium, the monopoly intermediary obtains and sells data $D^M \subset \mathcal{D}$ that*

satisfies

$$D^M \in \arg\max_{D \subset \mathcal{D}} \Pi(D) - \sum_{i \in N} C_i(D). \tag{3}$$

All consumers and the firm obtain zero payoffs.

Later, I use D^M to describe equilibria with multiple intermediaries. If the right hand side of (3) has multiple maximizers, I use one of them arbitrarily as D^M and conduct the analysis.

4.2 Competing Intermediaries for Rivalrous Goods

Suppose that data are rivalrous—each consumer can provide each piece of data to *at most one* intermediary.[1] Such a model captures the market for conventional economic goods. In this case, competition among intermediaries dissipates profits and enables consumers to extract full surplus. Intuitively, if one intermediary earns a positive profit by obtaining data D^k, another intermediary can offer consumers slightly higher compensation to *exclusively* obtain D^k. This implies that no intermediary can earn a positive profit. The proof is in Appendix A.

Claim 2. *Suppose that data are rivalrous and there are multiple intermediaries. In any equilibrium, all intermediaries and the firm obtain zero payoffs. If Π is strictly supermodular, in any equilibrium, there is at most one intermediary that obtains non-empty data.*

5 Main Analysis

This section considers the main specification: Multiple intermediaries buy and sell non-rivalrous data.

5.1 Equilibrium in the Downstream Market

The following lemma shows that the equilibrium revenue of each intermediary in the downstream market is equal to the marginal contribution of its data to the firm's revenue. The result relies on

[1] Formally, I assume that each consumer i can accept a collection of offers $(D_i^k, \tau_i^k)_{k \in K_i}$ if and only if $D_i^k \cap D_i^j = \emptyset$ for any distinct $j, k \in K_i$.

the submodularity of the firm's revenue function Π. The proof of the uniqueness is involved and relegated to Appendix B.[2]

Lemma 1. *Suppose that each intermediary k holds data D^k. In any (subgame perfect) equilibrium of the downstream market, intermediary k obtains a revenue of*

$$\Pi^k := \Pi\left(\bigcup_{j\in K} D^j\right) - \Pi\left(\bigcup_{j\in K\setminus\{k\}} D^j\right), \tag{4}$$

and the firm obtains data $\cup_{k\in K} D^k$.

Proof. I show that there is an equilibrium (of the downstream market) in which each intermediary k posts a price of Π^k, and the firm buys all data. First, the submodularity of Π implies that $\Pi(\cup_{k\in K'\cup\{j\}} D^j) - \Pi(\cup_{k\in K'} D^j) \geq \Pi^j$ for all $K' \subset K$. Thus, if each intermediary k sets a price of Π^k, the firm prefers to buy all data. Second, if intermediary k strictly increases its price, the firm strictly prefers buying data from intermediaries in $K \setminus \{k\}$ to buying data from a set of intermediaries containing k. Finally, if an intermediary lowers the price, it decreases its revenue. Thus, no intermediary has a profitable deviation. □

Lemma 1 has two implications. First, on and off the equilibrium paths, consumers anticipate that any data they share with intermediaries will be sold to the downstream firm. Second, intermediaries earn zero revenue in the downstream market if they hold the same data. This is similar to Bertrand competition with homogeneous products. More generally, the revenue of an intermediary depends only on the part of the data that other intermediaries do not hold.

Corollary 1. *Suppose that each intermediary $j \neq k$ holds data D^j. The equilibrium revenue of intermediary k in the downstream market is identical between when it holds D^k and $D^k \cup D'$, where D' is any subset of $\cup_{j\neq k} D^j$.*

[2] Lerner and Tirole (2004) focus on a symmetric environment but do not assume the submodularity. Gu et al. (2018) assume $K = 2$ and consider both submodularity and supermodularity. The uniqueness of the equilibrium revenue is new in the literature.

5.2 Consumers with Single Unit Data

To capture the main idea simply, I begin with the case where each consumer has a single piece of data:

Assumption 3. For each $i \in N$, $\mathcal{D}_i = \{d_i\}$.

The following notion simplifies the exposition.

Definition 1. An allocation of data $(D^1, \ldots, D^K)$ is *partitional* if $D^k \cap D^j = \emptyset$ for all $k, j \in K$ with $k \neq j$.

The following result states that although data are non-rivalrous, no two intermediaries obtain the same piece of data along the equilibrium path. Intuitively, if two intermediaries obtained the same data, one of them could profitably deviate by not collecting the data and reducing compensation to the data subjects. The proof is in Appendix C.

Proposition 1. *In any equilibrium, the allocation of data is partitional.*

Proposition 1 resembles product differentiation; however, products in this model are data, which come from consumers. Thus, intermediaries' motive for product differentiation affects consumer surplus. The following result illustrates this point: It presents equilibria in which data provided by consumers and their surplus are equal to the monopoly outcome. The proof is in Appendix D.

Theorem 1. *Take any partitional allocation of data* $(D^1, \ldots, D^K)$ *with* $\cup_{k \in K} D^k = D^M$. *Then, there is an equilibrium with the following properties.*

1. *The equilibrium allocation of data is* $(D^1, \ldots, D^K)$.
2. *Consumer surplus is zero: In the upstream market, intermediary* k *pays consumer* i *a compensation of* $\mathbf{1}_{\{d_i \in D^k\}} C_i(d_i)$.
3. *In the downstream market, each intermediary* k *obtains a revenue of*

$$\Pi(D^M) - \Pi(D^M \setminus D^k).$$

The intuition is as follows. Suppose that intermediary j deviates and compensates consumers for providing data in D^k with $k \neq j$. Then, these consumers will provide their data with not only intermediary j but also k. Indeed, when consumers share data with one intermediary, they also prefer to accept offers from other intermediaries: By doing so, consumers can earn higher total compensation without increasing the loss from the firm's use of data. However, once consumers share data with both k and j, intermediary j cannot charge the firm a positive price for the data. Anticipating this, intermediary j prefers to not compensate for any data in D^k. Since each intermediary faces no competing offers, it can acquire data at the monopsony price, $C_i(d_i)$. Also, no intermediary has an incentive to acquire data in $\mathcal{D} \setminus D^M$, because consumers ask for greater compensation than the price of their data in the downstream market.

The non-rivalry of data is important not only for consumers' receiving zero surplus (Point 2) but also for the multiplicity of equilibrium allocations. Indeed, if data are rivalrous, under a mild condition, at most one intermediary acquires data (Claim 2).

One implication of the theorem is that there is a monopoly equilibrium. Thus, the presence of multiple intermediaries may not dissipate their profits.

Theorem 2. *There is an equilibrium in which a single intermediary acts as a monopolist described in Claim 1.*

Proof. Take $D^k = D^M$ and $D^j = \emptyset$ for all $j \neq k$ in Theorem 1. □

The comparison of Theorem 2 with Claim 2 indicates that the non-rivalry of data relaxes competition among intermediaries, and it can transfer all surplus from consumers to intermediaries. This argument has the following policy implication: The EU's General Data Protection Regulation aims to ensure *data portability*, under which consumers have "the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided." Let us interpret the current model and the one for Claim 2 as the economy with and without data portability, respectively. Then, these results illustrate that data portability could relax ex ante competition for data and lower consumer surplus.

A natural question is whether there are equilibria other than those described in Theorem 1. The answer is yes: As the following example shows, both compensation and the set of data shared by consumers can be different from the monopoly outcome.

Example 1. There is an equilibrium in which no data are shared. Consider a single consumer and two intermediaries. On the path of play, both intermediaries offer $(\{d_1\}, 0)$ and the consumer rejects them. If an intermediary unilaterally deviates and offer $(\{d_1\}, \tau)$ with $\tau \geq C_1(d_1)$, the consumer accepts offers of *both* intermediaries. This consists of an equilibrium. In particular, no intermediary has an incentive to obtain data, because the consumer will then share her data with all intermediaries, following which the price of the data is zero.

There is also an equilibrium where a consumer extracts full surplus $\Pi(d_1) - C_1(d_1)$: One intermediary, say 1, offers $(\{d_1\}, \Pi(d_1))$, and the other intermediary offers $(\{d_1\}, 0)$. On the path of play, the consumer accepts only $(\{d_1\}, \Pi(d_1))$. If intermediary 1 unilaterally deviates and *lowers* compensation to τ_1^1 such that $C_1(d_1) < \tau_1^1 < \Pi(\{d_1\})$, the consumer accepts offers of both intermediaries. This consists of an equilibrium. In particular, intermediary 1 has no incentive to lower compensation, because the consumer will then share her data with all intermediaries.

A common feature of these equilibria is that a consumer punishes a deviating intermediary 1 by disseminating her data d_1 with multiple intermediaries. This is possible because intermediary 2 also asks for d_1 at compensation between zero and $C_1(d_1)$. Note that the consumer is indifferent between accepting and rejecting the offer from intermediary 2, conditional on her accepting the offer from intermediary 1. However, on the equilibrium path, the consumer has to reject the offer from intermediary 2: If she accepted, intermediary 1 would prefer to not compensate for data d_1 that would be valueless in the downstream market.

However, I argue that these equilibria, in which intermediaries make offers that are rejected on the equilibrium path, are not robust for two reasons. First, such an equilibrium disappears if intermediaries incur costs to enter the market or costs to make offers to consumers, no matter how small the costs are.[3] Second, such an equilibrium disappears if consumers accept offers whenever they are indifferent. In the above example, if the consumer adopts such a tie-breaking rule, she accepts intermediary 2's offer, and thus the equilibrium cannot be sustained. This idea motivates the following notion.

Definition 2. An *acceptance equilibrium* is an equilibrium in which consumers accept all non-

[3]Consider an additional stage in which intermediaries simultaneously decide whether to enter the market at cost, after which those that have entered the market play the current model. Then, it cannot be an equilibrium that an intermediary enters the market but makes an offer that is rejected for sure.

empty offers on the equilibrium path.

Proposition 2. *Acceptance equilibrium has the following properties.*

1. *All equilibria in Theorem 1 are acceptance equilibria.*
2. *In any acceptance equilibrium, consumer surplus is zero.*
3. *In any acceptance equilibrium, if the firm buys data D, then $D \subsetneq D^M$ never holds. Thus, if $D^M = \mathcal{D}$, the set of all acceptance equilibria consists of the equilibria in Theorem 1.*

Proof. Point 1 follows from the proof of Theorem 1. To show Point 2, take any acceptance equilibrium. Let $(D^1, \ldots, D^K)$ denote the allocation of data, which is partitional by Proposition 1. Suppose to the contrary that intermediary k pays consumer i a strictly greater compensation than $C_i(d_i)$. If another intermediary offers $(\{d_i\}, \tau)$ with $\tau \geq 0$, consumer i accepts it by definition of acceptance equilibrium. Then, the price of i's data is zero in the downstream market (Corollary 1). Thus, intermediary k can profitably deviate by not obtaining d_i. This implies that no intermediary other than k makes a non-empty offer to consumer i. This, in turn, implies that intermediary k can profitably deviate by slightly lowering the compensation. This is a contradiction. Thus, in any acceptance equilibrium, each consumer receives either no compensation or zero compensation.

To show Point 3, suppose $D \subsetneq D^M$. Then, intermediary k can profitably deviate by additionally offering $(\{d_i\}, C_i(d_i) + \varepsilon)$ for a small $\varepsilon > 0$ to any consumer i with $d_i \in D^M \setminus D$. Note that intermediary k can obtain data d_i at (nearly) $C_i(d_i)$, because other intermediaries make empty offers. This strictly increases intermediary k's payoff, because the increment of k's payoff is equal to the monopolist's gain from acquiring data $D^M \setminus D$, which is positive by definition of D^M. □

Point 2 of Proposition 2 implies that if intermediaries incur even small costs to make offers to consumers, any equilibrium predicts that consumer surplus is zero. Point 3 implies that if the value of data for the firm is high relative to the negative impact on consumers, the set of acceptance equilibria is equal to the equilibria in Theorem 1.

5.3 Consumers with Multidimensional Data

This section assumes that each consumer i has any finite set $\mathcal{D}_i$ of data. This setting involves a new challenge: On the one hand, acceptance equilibrium may not exist. On the other hand, if I

consider equilibrium other than acceptance equilibrium, I can sustain almost any profile of offers as equilibrium using the same logic as Example 1. To avoid this difficulty, I continue to focus on acceptance equilibrium but impose a restriction under which an acceptance equilibrium exists.

Assumption 4. The monopoly intermediary acquires all data in equilibrium, i.e., $D^M = \mathcal{D}$.

The assumption holds if the firm's marginal revenue from data is high relative to consumers' marginal costs of sharing the data. For example, for any Π and $(C_i)_{i\in N}$, the assumption holds if I scale up Π to $\alpha\Pi$ for a large $\alpha > 1$. Under Assumption 4, I obtain a similar result to Theorem 1. The proof is in Appendix E.

Theorem 3. *Take any partitional allocation of data $(D^1, \ldots, D^K)$ with $\cup_{k\in K} D^k = D^M$. Then, there is an acceptance equilibrium with the following properties.*

1. *The equilibrium allocation of data is $(D^1, \ldots, D^K)$.*
2. *In the upstream market, intermediary k pays consumer i a compensation of*

$$\hat{\tau}_i^k := C(\mathcal{D}_i) - C(\mathcal{D}_i \setminus D_i^k). \tag{5}$$

3. *In the downstream market, each intermediary k obtains a revenue of*

$$\hat{p}^k := \Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D^k). \tag{6}$$

A key difference from the case of single unit data (Theorem 1) is the equilibrium compensation. Point 2 of Theorem 3 shows that intermediary k compensates consumer i according to the additional loss that consumer i incurs by sharing D_i^k conditional on sharing data to other intermediaries $j \neq k$. Unless C_i is additively separable, this creates a wedge between the total compensation $\sum_{k\in K} \hat{\tau}_i^k$ and the cost $C_i(\mathcal{D}_i)$. I will exploit this property in the next section.

6 Welfare Implications of Data Concentration

In any equilibrium described in Theorem 1 and Theorem 3, the allocation of data is a partition of D^M. This enables us to compare equilibria from the perspective of data concentration. I say that

one allocation of data is more concentrated than another if the former is coarser than the latter as a partition. The allocation becomes more concentrated, for example, if one intermediary acquires another.

Definition 3. Take two partitional allocations of data, (D^k) and $(\hat{D}^k)$. We say that $(\hat{D}^k)$ is *more concentrated than* (D^k) if (i) $\cup_k D^k = \cup_k \hat{D}^k$ and (ii) for each $k \in K$, there is $\ell \in K$ such that $D^k \subset \hat{D}^\ell$.

The following result summarizes the impact of data concentration on consumers and intermediaries. The proof is in Appendix F.

Theorem 4.

- *Consider equilibria in Theorem 1. Intermediaries' total profit is higher in an equilibrium with a more concentrated allocation of data.*
- *Consider equilibria in Theorem 3. Consumer surplus is lower and intermediaries' total profit is higher in an equilibrium with a more concentrated allocation of data.*

The intuition is as follows. As in Lemma 1, the price of data D^k is $\Pi(\cup_{j\in K} D^j) - \Pi(\cup_{j\in K\setminus\{k\}} D^j)$, the additional revenue the firm can earn from D^k conditional on having other data. If there are many intermediaries each of which has a small part of D^M, the contribution of each piece of data is close to the marginal revenue $\Pi(D^M) - \Pi(D^M \setminus \{d\})$. In contrast, if a few intermediaries hold D^M, the contribution of data held by each intermediary is large. Thus, intermediaries can set a higher price to extract the infra-marginal value of its data. Since $\Pi(\cdot)$ is submodular, the latter leads to a greater total revenue for intermediaries. Symmetrically, if a consumer's cost C_i is supermodular, data concentration hurts consumers. This is because a large intermediary can compensate consumers for their data on the basis of the infra-marginal cost.

6.1 Intensive and Extensive Margins of Data Concentration

The allocation of data can be more concentrated at the intensive or extensive margin. To see this, consider the following example. There are two intermediaries. The left block of Figure 2 depicts a situation in which intermediary 1 obtains location data on all consumers in the US and EU, and

intermediary 2 obtains purchase data on all consumers in the US and EU. The right block of Figure 2 is an alternative allocation where intermediary 1 holds location and purchase data on consumers in the US, and intermediary 2 obtains location and purchase data on consumers in the EU. The allocation of data in the right block is more concentrated at the *intensive margin*, because each intermediary has more data about each data subject.

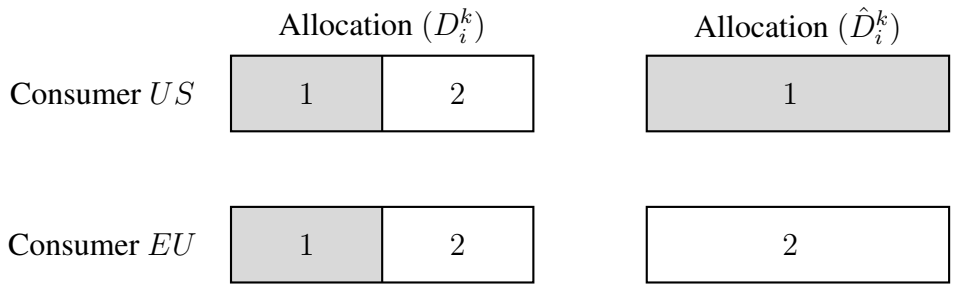

Figure 2: Data concentration at the intensive margin

Next, suppose that there are are four intermediaries. The left block of Figure 3 depicts a situation in which intermediaries 1 and 3 acquire location data on consumers in the US and EU, respectively, and intermediaries 2 and 4 acquire purchase data on consumers in the US and EU, respectively. Suppose that intermediaries 1 and 3 merge. The right block of Figure 3 depicts such a situation where the new intermediary is labeled as 1. After the merger, the allocation of data becomes more concentrated at the *extensive margin*, because the new intermediary has location data on wider population.

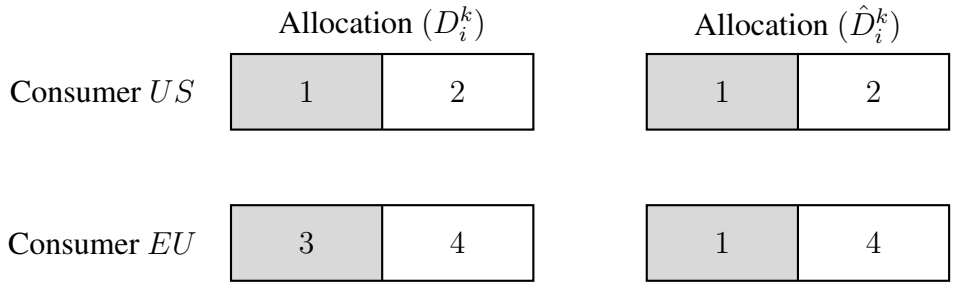

Figure 3: Data concentration at the extensive margin

The following definition generalizes these examples.

Definition 4. Let $(D^k)_k$ and $(\hat{D}^k)_k$ denote two partitional allocations of data with $\cup_k D^k = \cup_k \hat{D}^k$.

1. $(\hat{D}^k)_k$ is *more concentrated than* $(D^k)_k$ *at the intensive margin* if for any given $i \in N$ and any $k \in K$, there is $\ell \in K$ such that $D_i^k \subset \hat{D}_i^\ell$.

2. $(\hat{D}^k)_k$ is *more concentrated than* $(D^k)_k$ *at the extensive margin* if $(\hat{D}^k)_k$ is more concentrated than $(D^k)_k$, and for any $i \in N$ and any $k \in K$, there is ℓ such that $D_i^k = \hat{D}_i^\ell$.

Proposition 3. *Consider equilibria described in Theorem 3.*

1. *Data concentration at the intensive margin lowers consumer surplus. The impact on intermediaries' total profit is ambiguous.*

2. *Data concentration at the extensive margin increases intermediaries' total profit, and it does not affect consumer surplus.*

Note that data concentration at the intensive margin does not necessarily imply data concentration Definition 3. In Figure 2, $(\hat{D}_i^k)$ is more concentrated at the intensive margin, however, intermediaries' total profits could be higher or lower at $(\hat{D}_i^k)$ than (D_i^k) depending on the shape of Π. For example, suppose that the firm's revenue function Π is separable across consumers, and the revenue function for each consumer is submodular. Then, concentration at the intensive margin leads to higher profits of intermediaries. In contrast, if the firm's revenue function Π is separable across data, then concentration at the intensive margin might reduce intermediaries' profits. In the example of location and purchase data, intermediaries' profits may decrease if the firm's revenue is the sum of revenue from location data and revenue from purchase data, each of which is a submodular set function.

7 Extensions

This section uses the formulation of single unit data (Subsection 5.2), but a similar extension applies to multidimensional data (Subsection 5.3).

7.1 Beneficial Use of Data

So far, I have assumed that a downstream firm's use of personal data hurts consumers. In reality, consumers may also benefit from sharing their data with third parties. To capture such a situation,

assume $C_i(d_i) < 0$ for each i. That is, each consumer gains $B_i := -C_i(d_i) > 0$ if the firm acquires her data.

In this case, intermediaries may offer negative compensation as a fee to transfer consumers' data. Thus, the relevant question is whether competing intermediaries have an incentive to lower fees. The proof of the following result is in Appendix G.

Proposition 4. *Suppose that the firm's use of data benefits consumers. For any $K \geq 1$, almost all consumers share their data with the firm. Moreover, competition among intermediaries benefits consumers:*

1. *If there is a monopoly intermediary, all consumers obtain zero payoffs.*
2. *If there are at least two intermediaries, in any equilibrium, all consumers obtain payoffs of at least $B_i > 0$.*

The intuition is as follows. For simplicity, suppose that intermediaries are restricted to offering non-negative fees. The key observation is that if multiple intermediaries offer positive fees to consumers, each consumer accepts at most one offer. This is because consumers can enjoy the benefit $B_i > 0$ as long as one intermediary transfer their data to the firm. Since consumers share their data with at most one intermediary, each intermediary tries to undercut the fees of other intermediaries in order to obtain data. This lowers the equilibrium fees down to zero. The result contrasts with the main specification: When the firm's use of data hurts consumers, competition among intermediaries could benefit consumers by raising compensation. However, this does not work because consumers will then share data with multiple intermediaries.

7.2 Multiple Downstream Firms

The model can readily take into account multiple downstream firms if they do not interact with each other. Suppose that there are L firms, where firm $\ell \in L$ has revenue function Π^ℓ that depends only on data available to ℓ. Each consumer i's cost of sharing data is $\sum_{\ell \in L} C_i^\ell$, where each C_i^ℓ can depend on what data of consumer i firm ℓ has.

This setting is equivalent to the one with a single firm. First, Lemma 1 implies that each intermediary k posts a price of $\Pi_\ell(\cup_k D^k) - \Pi_\ell(\cup_{j\neq k} D^k)$ to firm ℓ in the downstream market. Note that I implicitly assume that intermediaries can price discriminate firms.

Given the pricing rule, the revenue of intermediary k given the allocation of data $(D^k)_k$ is $\sum_{\ell\in L}[\Pi^\ell(\cup_k D^k) - \Pi^\ell(\cup_{j\neq k} D^k)]$. By setting $\Pi := \sum_{\ell\in L}\Pi^\ell$, we can calculate the equilibrium revenue of each intermediary in the downstream market as in Lemma 1.

Second, intermediaries cannot commit to not sell data to downstream firms. Thus, once a consumer shares her data with one intermediary, the data is sold to all firms. This means that in equilibrium, each consumer i decides which offers to accept in order to maximize total compensation minus the cost $\sum_{\ell\in L} C_i^\ell(D_i)$. Therefore, by setting $C_i := \sum_{\ell\in L} C_i^\ell$, we can apply the same analysis as before. Note that this extension can accommodate the case where some firms impose loss ($C_i^\ell > 0$) and some impose benefit ($C_i^\ell < 0$) on consumers, because Section 3 only requires that $\sum_{\ell\in L} C_i^\ell > 0$.

7.3 Privacy Concern Toward Data Intermediaries

Consumers may incur the (exogenous) cost of sharing data not only with downstream firms but also with data intermediaries. I can incorporate this by assuming that each consumer incurs a loss of ρK_i by sharing her data with K_i intermediaries.

This does not change the result qualitatively. If $\rho > 0$, intermediaries obtain less data than the original model, because it has to pay a compensation of at least $c + \rho$ to each consumer. Any equilibrium allocation of data is partitional, and there are multiple equilibria one of which is a monopoly equilibrium.

8 Conclusion

This paper studies competition among data intermediaries, which obtain data from consumers and sell them to downstream firms. The model incorporates two key features of personal data. One is that data are non-rivalrous, and the other is that the use of data by third parties could negatively affect consumers. These features drastically change the nature of competition relative to the intermediation of physical goods: Data intermediaries may secure monopoly profit in some equi-

librium, and the equilibrium allocation of data across intermediaries is not unique. This enables me to compare equilibria with different degrees of data concentration. Under a certain condition, an equilibrium with greater data concentration is associated with higher profits of intermediaries and lower consumer welfare.

References

Anderson, Simon P and Stephen Coate (2005), "Market provision of broadcasting: A welfare analysis." *The Review of Economic studies*, 72, 947–972.

Armstrong, Mark (2006), "Competition in two-sided markets." *The RAND Journal of Economics*, 37, 668–691.

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.

Bergemann, Dirk and Alessandro Bonatti (2019), "Markets for information: An introduction." *Annual Review of Economics*, 11, 1–23.

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

Caillaud, Bernard and Bruno Jullien (2003), "Chicken & egg: Competition among intermediation service providers." *RAND journal of Economics*, 309–328.

Carrascal, Juan Pablo, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira (2013), "Your browsing behavior for a big mac: Economics of personal information online." In *Proceedings of the 22nd international conference on World Wide Web*, 189–200, ACM.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim (2018), "Privacy and personal data collection with information externalities."

Demsetz, Harold (1968), "Why regulate utilities?" *The Journal of Law and Economics*, 11, 55–65.

Federal Trade Commission (2014), "Data brokers: A call for transparency and accountability." *Washington, DC*.

Gu, Yiquan, Leonardo Madio, and Carlo Reggiani (2018), "Data brokers co-opetition." *Available at SSRN 3308384*.

Huck, Steffen and Georg Weizsacker (2016), "Markets for leaked information." *Available at SSRN 2684769*.

Jones, Charles, Christopher Tonetti, et al. (2018), "Nonrivalry and the economics of data." In *2018 Meeting Papers*, 477, Society for Economic Dynamics.

Kim, Soo Jin (2018), "Privacy, information acquisition, and market competition."

Lerner, Josh and Jean Tirole (2004), "Efficient patent pools." *American Economic Review*, 94, 691–711.

Reisinger, Markus (2012), "Platform competition for advertisers and users in media markets." *International Journal of Industrial Organization*, 30, 243–252.

Rochet, Jean-Charles and Jean Tirole (2003), "Platform competition in two-sided markets." *Journal of the european economic association*, 1, 990–1029.

Sokol, D Daniel and Roisin Comerford (2015), "Antitrust and regulating big data." *Geo. Mason L. Rev.*, 23, 1129.

Varian, Hal (2018), "Artificial intelligence, economics, and industrial organization." In *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press.

Appendix

A Proof of Claim 2

Below, I write $X - Y$ to mean $X \setminus Y$, and $X - Y - Z$ to mean $(X \setminus Y) \setminus Z$. Take any $K \geq 2$ and suppose to the contrary that there is an equilibrium in which one intermediary, say 1, obtains

a positive payoff. Suppose that each intermediary k obtains data D_i^k from consumer $i \in N^k$ at compensation τ_i^k. Define $D^* := \cup_k D^k$. Suppose that intermediary 2 deviates and offers each consumer $i \in N^1$ an offer of $(D_i^1 \cup D_i^2, \tau_i^1 + \tau_i^2 + \varepsilon)$. Then, all consumers in N^1 accept the offer of *only* intermediary 2. In the downstream market, the revenue of intermediary 2 increases from $\Pi(D^*) - \Pi(D^* - D^2)$ to $\Pi(D^*) - \Pi(D^* - D^1 - D^2)$, which yields a net gain of $\Pi(D^* - D^2) - \Pi(D^* - D^1 - D^2)$. By Assumption 2, $\Pi(D^* - D^2) - \Pi(D^* - D^1 - D^2) \geq \Pi(D^*) - \Pi(D^* - D^1)$. Since intermediary 1 obtains a positive payoff if intermediary 2 did not deviate, it holds that $\Pi(D^*) - \Pi(D^* - D^1) - \sum_{i \in N^1} \tau_i^1 > 0$, which implies $\Pi(D^* - D^2) - \Pi(D^* - D^1 - D^2) - \sum_{i \in N_1} (\tau_i^1 + \varepsilon) > 0$ for a small $\varepsilon > 0$. Thus, intermediary 2 has a profitable deviation, which is a contradiction.

Second, suppose to the contrary that there is an equilibrium where the firm obtains a positive payoff. This means that multiple intermediaries obtain different non-empty data. If $\Pi(\cup_k D^k) = \sum_{k \in K} \Pi(D^k)$, then the firm's payoff would be zero. Thus, $\Pi(\cup_k D^k) > \sum_{k \in K} \Pi(D^k)$ holds. This implies that, in the upstream market, an intermediary can unilaterally deviate and increase its payoff by offering slightly higher compensation to consumers in order to obtain $\cup_{k \in K} D^k$. This is a contradiction, and thus the firm obtains a payoff of zero. This argument also implies that, if Π is strictly supermodular, in any equilibrium, there is at most one intermediary that obtains non-empty data.

B Proof of Lemma 1

Proof. Take any allocation of data $(D^1, \ldots, D^K)$. I show that the equilibrium revenue of each intermediary k is at most Π^k. Suppose to the contrary that (without loss of generality) intermediary 1 obtains a strictly greater revenue than Π^1. Let $K' \ni 1$ denote the set of intermediaries from which the firm buys data.

First, in equilibrium, $\Pi(\cup_{k \in K'} D^k) = \Pi(\cup_{k \in K} D^k)$. To see this, note that if $\Pi(\cup_{k \in K'} D^k) < \Pi(\cup_{k \in K} D^k)$, then there is some $\ell \in K$ such that $\Pi(\cup_{k \in K'} D^k) < \Pi(\cup_{k \in K' \cup \{\ell\}} D^k)$. Such intermediary ℓ can profitably deviate by setting a sufficiently low positive price, because the firm then buys data D^ℓ. This is a contradiction.

Second, define $K^* := \left\{\ell \in K : \ell \notin K', p^\ell = 0\right\} \cup K'$. Note that K^* satisfies $\Pi(\cup_{k \in K'} D^k) = \Pi(\cup_{k \in K} D^k) = \Pi(\cup_{k \in K^*} D^k)$, $\sum_{k \in K'} p^k = \sum_{k \in K^*} p^k$, and $p^j > 0$ for all $j \notin K^*$.

It holds that

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} p^k = \max_{J\subset K\setminus\{1\}} \left(\Pi(\cup_{k\in J} D^k) - \sum_{k\in J} p^k\right). \tag{7}$$

To see this, suppose one side is greater than the other. If the left hand side is strictly greater, then intermediary 1 can profitably deviate by slightly increasing its price. If the right hand side is strictly greater, then the firm would not buy D^1. In either case, we obtain a contradiction.

Let J^* denote a solution of the right hand side of (7). I consider two cases. First, suppose that there exists some $j \in J^* \setminus K^*$. By the construction of K^*, $p^j > 0$. Then, intermediary j can profitably deviate by slightly lowering p^j. To see this, note that

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} \hat{p}^k < \Pi(\cup_{k\in J^*} D^k) - \sum_{k\in J^*} \hat{p}^k, \tag{8}$$

where $\hat{p}^k = p^k$ for all $k \neq j$ and $\hat{p}^j = p^j - \varepsilon > 0$ for a small $\varepsilon > 0$. This implies that after the deviation by intermediary j, the firm buys data D^j. This is because the left hand side of (8) is the maximum revenue that the firm can obtain if it cannot buy data D^j, and the right hand side is the lower bound of the revenue that the firm can achieve by buying D^j. Thus, the firm always buy data D^j, which is a contradiction.

Second, suppose that $J^*\setminus K^* = \emptyset$, i.e., $J^* \subset K^*$. This implies that the right hand side of (7) can be maximized by $J^* = K^*\setminus\{1\}$, because Π is submodular and $\Pi(\cup_{k\in K^*} D^k) - \Pi(\cup_{k\in K^*\setminus\{\ell\}} D^k) \geq p^\ell$ for all $\ell \in K^*$. Plugging $J^* = K^* \setminus \{1\}$, we obtain

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} p^k = \Pi(\cup_{k\in K^*\setminus\{1\}} D^k) - \sum_{k\in K^*\setminus\{1\}} p^k. \tag{9}$$

I show that there is $j \notin K^*$ such that

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D^k) < \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D^k). \tag{10}$$

Suppose to the contrary that for all $j \notin K^*$,

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D^k) = \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D^k). \tag{11}$$

By submodularity, this implies that

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D^k) = \Pi(\cup_{k\in K\setminus\{1\}} D^k).$$

Then, we can write (9) as

$$\Pi(\cup_{k\in K} D^k) - \sum_{k\in K^*} p^k = \Pi(\cup_{k\in K\setminus\{1\}} D^k) - \sum_{k\in K^*\setminus\{1\}} p^k$$

which implies $\Pi^1 = p^1$, a contradiction. Thus, there must be $j \notin K^*$ such that (10) holds. Such intermediary j can again profitably deviate by lowering its price, which is a contradiction. Therefore, intermediary k's revenue is at most Π^k.

Finally, I show that in equilibrium, each intermediary k gets a revenue of at least Π^k. This follows from the submodularity if Π: If intermediary k sets a price of $\Pi^k - \varepsilon$, the firm buys D^k no matter what prices other intermediaries set. Thus, intermediary k must obtain a payoff of at least Π^k in equilibrium. Combining this with the previous part, we can conclude that in any equilibrium, each intermediary k obtains a revenue of Π^k. □

C Proof of Proposition 1

Proof. Suppose to the contrary that there is an equilibrium in which multiple intermediaries, say 1 and 2, obtain the same piece of data d_i. Since consumer i prefers to share her data, the sum of compensations from intermediaries 1 and 2 is at least $C_i(d_i) > 0$. This implies that at least one intermediary, say 1, pays a positive compensation to consumer i. However, intermediary 1 can increase its payoff by offering $(\emptyset, 0)$ to consumer i. By Corollary 1, this does not reduce intermediary 1's revenue in the downstream market. Moreover, it reduces intermediary 1's expense in the upstream market. This is a contradiction. □

D Proof of Theorem 1

Proof. Take any disjoint allocation of data $(D^1, \ldots, D^K)$ with $\cup_{k\in K} D^k = D^M$. Let N^k denote the set of consumers from whom intermediary k obtains data. Consider the following strategy profile:

If $d_i \in D^k$, intermediary k offers $(d_i, C_i(d_i))$ to consumer i. Otherwise, it offers $(\emptyset, 0)$. In the downstream market, intermediaries set prices according to Lemma 1. The off-path behaviors of consumers are as follows. Suppose that a consumer detects a deviation by any intermediary. Then, the consumer accepts a set of offers to maximize her payoff, but here, the consumer accepts an offer if she is indifferent between accepting and rejecting it.

First, all consumers are indifferent between accepting and rejecting the offers, and thus it is optimal for them to accept all non-empty offers. Second, intermediaries and the firm have no profitable deviation in the downstream market by Lemma 1. Third, suppose that intermediary k unilaterally deviates in the upstream market and offers (D_i^k, τ_i^k) to each consumer i. Note that we can without loss of generality focus on offers such that $(D_i^k, \tau_i^k) = (\emptyset, 0)$ for all $i \in \cup_{j \neq k} N^j$. Indeed, if k pays a positive compensation to consumer $i \in N^j$, consumer i also accepts the offer of intermediary j. By Corollary 1, this does not increase intermediary k's revenue. Let $D^{-k} := \cup_{j \neq k} D^j$ denote the data held by intermediaries other than k. Let $\hat{D}^k \subset \mathcal{D} \setminus D^{-k}$ denote the data that intermediary k obtains as a result of the deviation. If this deviation is strictly profitable for k, it holds that $\Pi(\hat{D}^k \cup D^{-k}) - \Pi(D^{-k}) - \sum_{d \in \hat{D}^k} C_i(d) > \Pi(D^k \cup D^{-k}) - \Pi(D^{-k}) - \sum_{d \in D^k} C_i(d)$. However, this never holds because the monopolist could then earn strictly higher revenue from $\hat{D}^k \cup D^{-k}$ than D^M, which is a contradiction. □

E Proof of Theorem 3

Proof. Suppose that each intermediary k offers $(D_i^k, \hat{\tau}_i^k)$ to each consumer i and sets a price of data following Lemma 1. I show that this strategy profile is an equilibrium. First, Lemma 1 implies that there is no profitable deviation in the downstream market. Second, suppose that intermediary k deviates and offers $(\tilde{D}_i^k, \tilde{\tau}_i^k)$ to each consumer i. Without loss of generality, we can assume that $\tilde{D}_i^k \subset D_i^k$. The reason is as follows. If consumer i rejects $(\tilde{D}_i^k, \tilde{\tau}_i^k)$, intermediary k replace it with $(\tilde{D}_i^k, \tilde{\tau}_i^k) = (\emptyset, 0)$. If consumer i accepts $(\tilde{D}_i^k, \tilde{\tau}_i^k)$ but $\tilde{D}_i^k \subsetneq D_i^k$, it means that intermediary k obtains some data $d \in \tilde{D}_i^k \setminus D_i^k$. Because $\cup_k D^k = D^M = \mathcal{D}$, there is another intermediary that obtains data d. By Corollary 1, intermediary k is indifferent between offering $(\tilde{D}_i^k \setminus \{d\}, \tilde{\tau}_i^k)$ and offering $(\tilde{D}_i^k, \tilde{\tau}_i^k)$. Let $D^- := D^k \setminus \tilde{D}_i^k$ denote the set of data that are not acquired by the firm as a result of intermediary k's deviation. If intermediary k deviates in this way, its revenue in

the downstream market decreases by $\Pi(D^M) - \Pi(D^M \setminus D^k) - [\Pi(D^M \setminus D^-) - \Pi(D^M \setminus D^k)] = \Pi(D^M) - \Pi(D^M - D^-)$. In the upstream market, if consumer i provides data $\tilde{D}_i^k$ to intermediary k, then it is optimal for consumer i to accept other offers from non-deviating intermediaries, because C_i is supermodular. This implies that the minimum compensation that intermediary k has to pay is $C_i(\mathcal{D}_i \setminus D_i^-) - C_i(\mathcal{D}_i \setminus D_i^k)$. Thus, intermediary k's compensation to consumer i in the upstream market decreases by $C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^k) - [C_i(\mathcal{D}_i \setminus D_i^-) - C_i(\mathcal{D}_i \setminus D_i^k)] = C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^-)$. Thus, k's total compensation decreases by $\sum_{i \in N} \left[C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^-)\right]$. Because $D^M = \mathcal{D}$ is the optimal choice of the monopolist, it holds that $\Pi(D^M) - \Pi(D^M \setminus D^-) - \sum_{i \in N}[C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^-)] > 0$. Therefore, the deviation does not increase intermediary k's payoff. □

F Proof of Theorem 4

Proof. Let $(\hat{D}_k)_{k \in K}$ and $(D_k)_{k \in K}$ denote two partitional allocations of data such that the former is more concentrated than the latter. Without loss of generality, assume that $\cup_k \hat{D}^k = \cup_k D^k = \mathcal{D}$. Note that in general, for any set $S_0 \subset S$ and a partition $(S_1, \ldots, S_K)$ of S_0, we have

$$\begin{aligned}
&\Pi(S) - \Pi(S - S_0) \\
=&\Pi(S) - \Pi(S - S_1) + \Pi(S - S_1) - \Pi(S - S_1 - S_2) + \cdots \\
&+ \Pi(S - S_1 - S_2 - \cdots - S_{K-1}) - \Pi(S - S_1 - S_2 - \cdots - S_K) \\
\geq &\sum_{k \in K} \left[\Pi(S) - \Pi(S - S_k)\right],
\end{aligned}$$

where the last inequality follows from the submodularity of Π. For any $\ell \in K$, let $K(\ell) \subset K$ satisfy $\hat{D}^\ell = \sum_{k \in K(\ell)} D^k$. The above inequality implies

$$\begin{aligned}
\Pi(\mathcal{D}) - \Pi(\mathcal{D} - \hat{D}^\ell) &\geq \sum_{k \in K(\ell)} \left[\Pi(\mathcal{D}) - \Pi(\mathcal{D} - D^k)\right], \forall \ell \in K \\
\Rightarrow \sum_{\ell \in K} \left[\Pi(\mathcal{D}) - \Pi(\mathcal{D} - \hat{D}^\ell)\right] &\geq \sum_{\ell \in K} \sum_{k \in K(\ell)} \left[\Pi(\mathcal{D}) - \Pi(\mathcal{D} - D^k)\right].
\end{aligned}$$

In the last inequality, the left and the right hand sides are the total revenue for intermediaries in the downstream market under $(\hat{D}^k)$ and (D^k), respectively. We can prove the result on consumer surplus by replacing Π with $-C_i$. Note that if $(\hat{D}^k)$ is more concentrated than (D^k), $(\hat{D}_i^k)$ is more

concentrated than (D_i^k). □

G Proof of Proposition 4

Proof. Take any equilibrium, and suppose to the contrary that consumers in $D \subset N$ do not share their data. This means that any offer that these consumers face contains a fee of weakly greater than B_i. Then, intermediary (say) 1 can weakly increase its payoff by offering a compensation of 0 to consumers in D. Note that following this deviation, consumers in D accept offers of only intermediary 1. This increases the net payoff of intermediary 1 by $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D) > 0$. This shows that in any equilibrium, all consumers shares her data.

To show Point 1, suppose that the market consists of a monopoly intermediary. Now, if consumers obtain strictly positive payoffs, the intermediary can strictly increase its payoff by slightly increasing the fees offered to those consumers, which is a contradiction. Thus, in any equilibrium, each consumer i shares her data and pays a fee of B_i. Finally, it is indeed an equilibrium that the intermediary offers a fee of B_i to all consumers, all of whom accept the offer.

To prove Point 2, I first show that there is an equilibrium where all consumers share their data. Consider the strategy profile where all intermediaries offer zero fee to all consumers, who accept all offers; if intermediary k unilaterally deviates, consumers who are affected by the deviation share their data with all intermediaries $j \neq k$, and they share their data with k if and only if k offers a non-positive fee. First, the strategy of each consumer is optimal both on and off the equilibrium paths because accepting zero fee increases her payoffs by $B_i > 0$. Second, no intermediary has an incentive to deviate, because it either obtains no data or obtains data that other intermediaries hold. Therefore, the proposed strategy profile is an equilibrium.

Next, suppose to the contrary that there is an equilibrium in which all consumers share data but some consumers obtain payoffs strictly lower than B_i. Let D denote the set of those consumers. Without loss of generality, suppose that intermediary 2 charges positive fees to all consumers in D. Then, if intermediary 1 can deviate and offers them a fee of zero, consumers in D share their data *only* with 1. This strictly benefits intermediary 1 because its payoff in the downstream market increases by at least $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D) > 0$. Therefore, each consumer i shares her data with zero or lower fees and obtains a payoff of at least B_i. □