Non-competing Data Intermediaries

Click here to download the latest version.

Shota Ichihashi*

September 30, 2019

Abstract

I consider a model of markets for personal data, where data intermediaries (e.g., online platforms and data brokers) buy data from consumers and sell them to downstream firms. Competition among intermediaries has a limited impact on improving consumer welfare: If intermediaries offer high prices for data, consumers share data with multiple intermediaries, which lowers the downstream price of data and hurts intermediaries. This leads to multiple equilibria. There is a monopoly equilibrium, and an equilibrium with greater data concentration benefits intermediaries and hurts consumers. I generalize the results to arbitrary consumer preferences and study information design by data intermediaries.

Keywords: information markets, intermediaries, personal data, privacy

*Bank of Canada, 234 Wellington Street West, Ottawa, ON K1A 0G9, Canada. Email: shotaichihashi@gmail.com.
I thank Jason Allen, Itay Fainmesser, Matthew Gentzkow, Sitian Liu, Paul Milgrom, Shunya Noda, Makoto Watanabe, and seminar participants at the Bank of Canada, CEA Conference 2019, Decentralization Conference 2019, Yokohama National University, the 30th Stony Brook Game Theory Conference, and EARIE 2019. The opinions expressed in this article are the author's own and do not reflect the views of Bank of Canada.

1 Introduction

I consider a model of markets for personal data, in which data intermediaries collect and distribute personal data between consumers and downstream firms. For instance, online platforms, such as Google and Facebook, collect user data and share them indirectly through targeted advertising spaces. For another instance, data brokers, such as Acxiom and Nielsen, collect consumer data and sell them to retailers and advertisers (Federal Trade Commission, 2014).[1] The model clarifies how the interaction among data intermediaries shapes the creation and distribution of surplus from consumer data.

To make this concrete, consider online platforms that collect consumer data and share them with retailers and advertisers. The use of data by these third parties may hurt consumers through intrusive marketing campaigns, price discrimination, and spam. If so, platforms need to compensate consumers for sharing their data. Compensation might be monetary transfers or non-monetary benefits such as better quality of online services (e.g., social media and web mapping services).

The main question is whether competition among data intermediaries benefits consumers. Specifically, does competition incentivize data intermediaries to offer consumers better services and greater rewards? Does competition benefit consumers by changing the amounts and kinds of data that downstream firms acquire? This is a key question in recent policy debates on competition in digital markets (Crémer et al., 2019; Furman et al., 2019; Morton et al., 2019).

The model consists of consumers, data intermediaries, and downstream firms. Each consumer has a finite set of data (or data labels), say, email address, location, and browsing histories. In the upstream market, each intermediary decides what data to request from each consumer and how much compensation to offer. Each consumer then decides whether to accept each offer, balancing compensation she can earn and the expected benefit or loss she will experience when an intermediary sells her data to downstream firms. Each intermediary then learns what data other intermediaries have collected.[2] Finally, in the downstream market, intermediaries post prices and sell collected data to downstream firms.

A key idea of the paper is that competition may *not* increase compensation.[3] To see this,

[1] Section 3 discusses these applications in detail.

[2] Subsection 3.1 motivates this assumption.

[3] This may contrast with casual intuition. For instance, Furman et al. (2019) state that "it might have been that with more competition consumers would have given up less in terms of privacy or might even have been paid for their

consider an equilibrium in which an intermediary, say 1, collects location data. If another intermediary, say 2, offers positive compensation for the same data, then consumers will share the data with *both* intermediaries. This intensifies price competition and lowers the price of location data in the downstream market. Anticipating this, intermediary 2 prefers to not make a competing offer. This enables intermediary 1 to act as a monopoly of location data. The economic force is driven by the non-rivalry of data: The same data can be simultaneously obtained and sold by multiple intermediaries.

The above economic force leads to equilibria with the following two properties. First, intermediaries collect mutually exclusive sets of data, and the aggregate set of data bought by downstream firms is the same as under monopoly. Thus, competition does not affect what data consumers give up to downstream firms. Second, intermediaries act as local monopsonies in the upstream market: To collect data, an intermediary pays each consumer just enough compensation to cover her loss from downstream firms' use of the data. This limits the extent to which competition benefits consumers through greater compensation.

I show that the above equilibria have different degrees of *data concentration*. In a less concentrated equilibrium, many intermediaries collect small sets of data and earn low profits. In some cases, a lower concentration transfers surplus from intermediaries to firms, not to consumers. However, I also provide a condition on consumer preferences under which lower concentration benefits consumers. I connect this result with the welfare impact of "breaking up platforms."[4]

Some of the above results assume that downstream firms' use of data negatively affects consumers. However, the main insight holds even when consumers' payoffs depend arbitrarily on what data downstream firms acquire. In this setting, I characterize an equilibrium that (under a weak assumption) maximizes intermediary surplus and minimizes consumer surplus among all equilibria. The analysis shows that competition among data intermediaries generally has a limited impact on improving consumer welfare.

Finally, I use this general setting to study information design by competing intermediaries. A downstream firm uses data for price discrimination and product recommendation. Intermediaries

data." For another instance, Morton et al. (2019) state that "an easy method to pay consumers, combined with price competition for those consumers, might significantly erode the high profits of many incumbent platforms."

[4]See, e.g., *Elizabeth Warren on Breaking Up Big Tech,* N.Y. TIMES (June 26, 2019), www.nytimes.com/2019/06/26/us/politics/elizabeth-warren-break-up-amazon-facebook.html

can potentially obtain any informative signals (i.e., Blackwell experiments) about consumers' willingness to pay. In the equilibrium described above, a single intermediary obtains a fully informative signal. The resulting consumer surplus is equal to the one under hypothetical Bayesian persuasion in which consumers directly disclose information to the firm.

The contribution of the paper is two-fold. First, it contributes to the recent discussion of competition in digital markets. The model identifies an economic force that relaxes competition among data intermediaries. The result helps us understand why consumers do not seem to be compensated properly for their data provision (Arrieta-Ibarra et al., 2018). The model also explains data concentration as an equilibrium and clarifies how it hurts consumers. My explanation does not depend on network externalities or returns to scale. Second, the paper connects information design with markets for information, the two areas that currently do not have much overlap in the literature.[5]

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. Section 5 describes unique equilibrium payoffs in the downstream market. Section 6 assumes that consumers incur loss of sharing data with downstream firms. I show that there are multiple non-competitive equilibria. This section also studies the welfare impacts of data concentration. Section 7 generalizes these results by allowing general consumer preferences. This section also studies information design by competing intermediaries. Section 8 provides extensions, and Section 9 concludes.

2 Literature Review

This paper relates to two strands of literature. First, it relates to a growing literature on markets for data. Bergemann and Bonatti (2019b) consider a monopoly data intermediary and assume that a downstream firm uses data for price discrimination that hurts consumers. They show that the intermediary could earn a positive profit even when intermediation lowers total surplus. In contrast, I assume that the intermediation of data is profitable and focus on competition and data concentration. The economic mechanism of my paper is amenable to but independent of data externality, which is a key component in the model of Bergemann and Bonatti (2019b).

[5] Bergemann and Bonatti (2019b) is one of the initial attempts to establish such a connection.

More broadly, this paper relates to works on markets for data beyond the context of price discrimination. De Corniere and De Nijs (2016) study the design of an online advertising auction where a platform can use consumer data to improve the quality of match between consumers and advertisements. Gu et al. (2018) study data brokers' incentives to merge data. While I mainly assume that a downstream firm's revenue is a submodular function of datasets, they consider supermodularity as well.[6] In contrast to their work, I endogenize intermediaries' data collection in the upstream market, which enables me to conduct consumer welfare analysis. Bergemann et al. (2018) consider a model of data provision and data pricing. Bonatti and Cisternas (Forthcoming) study the aggregation of consumers' purchase histories and study how data aggregation and transparency impact a strategic consumer's incentives. 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. 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 two-sided markets (e.g., Armstrong 2006; Caillaud and Jullien 2003; Carrillo and Tan 2015; Galeotti and Moraga-González 2009; Hagiu and Wright 2014; Rhodes et al. 2018; Rochet and Tirole 2003). My paper differs from this literature in two ways. One is that my results are not driven by network externalities. Indeed, all results hold even when a market consists of a single consumer. The other is more substantive. The literature often assumes that a transaction between two sides is mutually beneficial.[7] This is natural in many applications such as video games (consumers and game developers) and credit cards (cardholders and merchants). When a transaction is mutually beneficial, platform competition involves undercutting prices charged to at least one side, which is sustainable even if multi-homing is possible. In contrast, I assume that a transaction (i.e., a downstream firm's acquiring data) benefits one side (i.e., a firm) but may benefit or hurt the other side (i.e., a consumer). When the use of data hurts consumers, intermediaries may compete for consumer data by raising compensation. I show that such competition does not occur due to the nonrivalry of data.

[6]However, Proposition 1 shows that the main insight holds regardless of the shape of a firm's revenue function.

[7]Exceptions are advertising platforms. For example, Anderson and Coate (2005) and Reisinger (2012) consider models where the presence of advertisers imposes negative externalities on viewers due to nuisance costs.

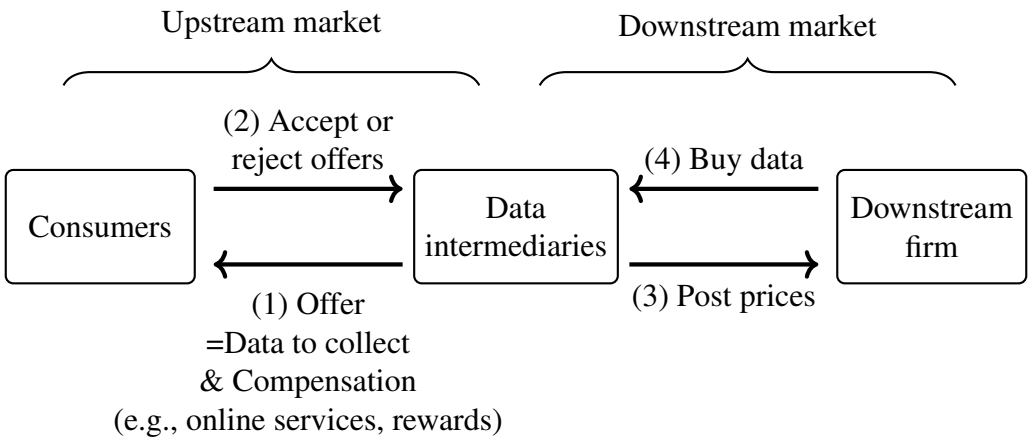

Figure 1: Timing of Moves

In my model, the nonrivalry of data relaxes competition among data intermediaries. This echoes the findings of the literature that multihoming by one side relaxes platform competition for that side (e.g., Caillaud and Jullien 2003). However, there are two key differences. First, in the literature, multihoming by one side not only relaxes competition for that side but also intensifies competition for the other side. In contrast, in my model, the nonrivalry of data may enable a single intermediary to secure a monopoly profit. Second, many of my results—such as the analysis of data concentration, general consumer preferences, and information design—have no counterpart in the literature.

3 Model

There are $N \in \mathbb{N}$ consumers, $K \in \mathbb{N}$ data intermediaries, and a single downstream firm.[8] Where it does not cause confusion, N and K denote the sets of consumers and intermediaries, respectively. Figure 1 depicts the game: Intermediaries obtain consumer data in the upstream market and then sell them in the downstream market. The detail is as follows.

Upstream Market

Each consumer $i \in N$ has a finite set $\mathcal{D}_i$ of data. Each element of $\mathcal{D}_i$ represents a data label such as i's email address, location, or browsing history, and it is treated as an individual and non-

[8] As I show in Section 8, this is equivalent to a model with multiple downstream firms that do not interact with each other.

rivalrous good. $\mathcal{D} := \cup_{i \in N} \mathcal{D}_i$ denotes the set of all data in the economy.

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 i's data $D_i^k \subset \mathcal{D}_i$. Compensation τ_i^k represents the quality of online services and the amount of monetary rewards. $\tau_i^k < 0$ is interpreted as a fee to transfer data. If $D_i^k \neq \emptyset$, I call (D_i^k, τ_i^k) a *non-empty offer*.

After observing offers, each consumer i simultaneously decides which offers to accept. Motivated by the non-rivalry of data, I assume that consumers can accept any number of offers. Formally, each consumer i 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. I call $(D^1, \ldots, D^K)$ 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

Intermediaries and the firm observe the allocation of data $(D^1, \ldots, D^K)$. Then, 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 the firm buys data $D := \cup_{k \in K'} D^k$ at total price $\sum_{k \in K'} p^k$. 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. $k \in K_i \cap K'$ 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 that 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 her data $D_i \subset \mathcal{D}_i$. Then, i's payoff is $\sum_{k \in K_i} \tau_i^k + U_i(D_i)$. The first term is the total compensation from intermediaries. The second

term $U_i(D_i)$ is consumer i's gross payoff when the firm acquires her data D_i from intermediaries. For example, U_i is a decreasing (set) function if the firm's use of data lowers consumer welfare. I normalize $U_i(\emptyset) = 0$ and impose more structures later. Note that U_i is independent of what data the downstream firm has on other consumers $j \neq i$. The results do not rely on this assumption (see Subsection 8.3 for the detail).

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

Assumption 1. $\Pi : 2^{\mathcal{D}} \to \mathbb{R}_+$ satisfies the following.

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

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

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

   (If inequality (1) is strict for any $X \subsetneq Y$, Π_i is *strictly* submodular.)

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

Point 2 (submodularity) simplifies the equilibrium pricing in the downstream market. However, Section 7 shows that some of the insights continue to hold without Point 2. Point 3 is normalization.

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. Second, each consumer simultaneously decides the set of offers to accept. After observing the allocation of data, each intermediary simultaneously posts a price to the firm. Finally, the firm chooses the set of intermediaries from which it buys data. The solution concept is pure-strategy subgame perfect equilibrium.

3.1 Discussion of Assumptions

I comment on two important modeling assumptions.

Observable allocation of data

It is crucial that intermediaries observe what data others collected before setting downstream prices. I assume this for two reasons. First, in practice, some data intermediaries disclose what kind of data they collect. For example, a data broker CoreLogic states that it holds property data covering more than 99.9% of U.S. property records.[9] Also, if an intermediary collects data directly from consumers, it needs to communicate what data it collects (e.g., Nielsen Homescan). Although there may be a verifiability problem and intermediaries' incentives to over- or understate what data they collect, it would be a reasonable starting point to assume that the allocation of data is observable.

Second, intermediaries have an incentive to make the allocation of data observable, because it often makes them better off in the Pareto sense. To see this, suppose that each intermediary privately observes what data it collects. Consider an equilibrium where intermediary k pays a positive compensation to consumers and sells their data at a positive price. Then, intermediary k can profitably deviate by collecting no data and charges the same price to the downstream firm. In particular, the firm cannot detect this deviation because it does not observe what data intermediary k has collected. This argument implies that there is no equilibrium in which intermediaries pay positive compensation to consumers. If U_i only takes negative values for all i, then only equilibrium involves no data sharing. Relative to such a situation, intermediaries are better off when the allocation of data is publicly observable.

Finally, although somewhat unnatural, another possible specification is that the data held by each intermediary is observable to the downstream firm but not to other intermediaries. Such a setting is intractable because there is no pure-strategy equilibrium.[10]

Timing

I assume that intermediaries set prices after observing the allocation of data. The idea is similar to models of endogenous product differentiation such as the one in d'Aspremont et al. (1979),

[9] https://www.corelogic.com/about-us/our-company.aspx (accessed July 11, 2019)

[10] Consider a strategy profile in which the firm acquires non-empty data D at a positive price. Then, some intermediary k can profitably deviate by obtaining data D for free from consumers and setting a slightly lower price than how much the firm would pay without k's deviation. Note that other intermediaries cannot detect such a deviation. In the setting of Section 6, this implies that there is no pure-strategy equilibrium.

where sellers set prices after observing their choices of product design. What data an intermediary collects (i.e., offer) is often a part of platform design or a company's policy. For example, we may interpret an offer that collects location data as a web mapping service (i.e., Google maps). In contrast, after collecting data, intermediaries will have many opportunities to share the data. Then, it is reasonable to assume that intermediaries can adjust downstream prices of data more quickly than adjusting what data they collect.

3.2 Applications

I present several interpretations of data intermediaries and motivate other assumptions not discussed in the previous subsection.

Online Platforms

The model can capture competition for data among online platforms such as Google and Facebook. Given an offer (D_i^k, τ_i^k), 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. For example, a web mapping service such as Google maps corresponds to D_i^k that contains i's location data but not, say, her political preferences. Platforms may share data with advertisers, retailers, and political consulting firms, which benefits or hurts data subjects (e.g., beneficial targeting or harmful price discrimination). The net effect is summarized by $U_i(D_i)$.

Several remarks are in order. First, $U_i(\cdot)$ is exogenous, that is, intermediaries cannot influence how the firm's use of data affects consumers. 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 markets for data such as Huck and Weizsacker (2016) and Jones et al. (2018).

Second, compensation is modeled as one-to-one transfer. This is to simplify the analysis. The results hold even if the cost of compensating consumers is non-linear. The assumption of costly compensation is natural if compensation is monetary transfer or an intermediary needs to invest to improve the quality of its service.

Third, the benefit for consumer i of sharing data with intermediary k depends only on τ_i^k. If we

interpret intermediaries as online platforms, we may think that the benefit should increase if other consumers provide more data (e.g., social media). However, I exclude such a situation to clarify that the results are not driven by network externalities or returns to scale.

Finally, this paper abstracts from competition for consumer *attention*, which is relevant to advertising platforms. Competition for attention is different from that for data because attention is a scarce resource. If consumers need to visit platforms to generate data but multi-homing is prohibitively costly due to scarce attention, then the non-rivalry of data may not hold.

Data Brokers

Intermediaries can be interpreted as data brokers such as LiveRamp, Nielsen, and Oracle. Data brokers collect personal data from online and offline sources, and resell or share that data 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, it is common that data brokers obtain personal data without interacting with consumers. The model could fit such a situation. For example, suppose that data brokers obtain individual purchase records from retailers. Consider the following chain of transactions: Retailers compensate customers and record their purchases, say, by offering discounts to customers who sign up for loyalty cards. Retailers then sell these records to data brokers, which resell the data to third parties. We can regard retailers in this example as consumers in the model.

The model can also 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's GDPR and California Consumer Privacy Act).

Mobile Application Industry

Kummer and Schulte (2019) empirically show that mobile application developers trade greater access to personal information for lower app prices, and consumers choose between lower prices and greater privacy when they decide which apps to install. Moreover, app developers share collected data with third parties for direct monetary benefit (see Kummer and Schulte 2019 and references therein). The model captures such economic interactions as a two-sided market for consumer data.

4 Two Benchmarks

I begin with two benchmarks, which I will compare with the main specification.

4.1 Monopoly Intermediary

Consider a monopoly intermediary ($K = 1$). For any set of data $D \subset \mathcal{D}$, I write $U_i(D \cap \mathcal{D}_i)$ as $U_i(D)$. Suppose that the intermediary obtains and sells data D. If $U_i(D) < 0$, the intermediary can obtain consumer i's data at compensation $-U_i(D)$. If $U_i(D) > 0$, the intermediary can charge a fee of $U_i(D)$ to transfer i's data. In the downstream market, the intermediary can set a price of $\Pi(D)$ to extract full surplus from the firm. Thus, I obtain the following result.

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

$$D^M \in \arg\max_{D \subset \mathcal{D}} \Pi(D) + \sum_{i \in N} U_i(D). \tag{2}$$

All consumers and the firm obtain zero payoffs.

Later, I use D^M to describe equilibria with multiple intermediaries. If (2) has multiple maximizers, I pick any one of them as D^M and conduct the analysis.

4.2 Competition for Rivalrous Goods

Suppose that data are rivalrous—each consumer can provide each piece of data to *at most one* intermediary.[11] Such a model corresponds to the market for physical goods.[12] In this case, competition among intermediaries dissipates profits and enables consumers to extract full surplus (see Appendix A for the proof).

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.*

[11] 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 $j, k \in K_i$.

[12] This model is similar to Stahl (1988), who shows that competition among intermediaries for physical goods can lead to a Walrasian outcome.

Intermediaries make zero profit due to Bertrand competition in the upstream market: If one intermediary earned a positive profit by obtaining data D^k, then another intermediary could profitably deviate by offering consumers slightly higher compensation to *exclusively* obtain D^k. For such a deviation to be unprofitable, the equilibrium payoffs of all intermediaries have to be zero.

5 Equilibrium Analysis: Downstream Market

Hereafter, I consider multiple intermediaries with non-rivalrous data. First, I show that the equilibrium revenue of each intermediary k in the downstream market is unique and equal to the firm's marginal revenue from k's data. The result relies on the submodularity of the firm's revenue function Π.[13]

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

$$\Pi^k := \Pi\left(\bigcup_{j\in K} D^j\right) - \Pi\left(\bigcup_{j\in K\setminus\{k\}} D^j\right). \tag{3}$$

Lemma 1 implies that intermediaries earn zero revenue if they hold the same data. This is similar to Bertrand competition with homogeneous products. More generally, the revenue of an intermediary is determined by the part of its 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 $D^k \cup D'$ for any $D' \subset \cup_{j\neq k} D^j$.*

[13] Lerner and Tirole (2004) focus on a symmetric environment but do not assume submodularity. Gu et al. (2018) assume $K = 2$ and consider both submodularity and supermodularity. To the best of my knowledge, the uniqueness of the equilibrium revenue for any K is a new result.

6 Equilibrium with Costly Data Sharing

Given the unique equilibrium outcome in the downstream market (Lemma 1), I solve equilibrium compensation and data sharing decision in the upstream market. I begin with a simple setup and later consider more general settings.

6.1 Single Unit Data

First, assume that each consumer i has a single piece of data and she incurs a loss of C_i if the firm acquires her data.

Assumption 2. For each $i \in N$, $\mathcal{D}_i = \{d_i\}$ and $C_i := -U(\{d_i\}) > 0$.

A motivation for this assumption is that the harmful use of personal data by third parties has been discussed by policymakers as a key issue of online privacy problems (Federal Trade Commission, 2014). C_i should be thought of as a reduced form capturing a consumer's (expected) loss from, say, price discrimination, privacy concern, and intrusive marketing compaign. The following notion simplifies the exposition.

Definition 1. The allocation of data $(D^1, \ldots, D^K)$ is *partitional* if no two intermediaries obtain the same piece of data: $D^k \cap D^j = \emptyset$ for all $k, j \in K$ with $k \neq j$.

The following result presents equilibria that are equivalent to the monopoly equilibrium in terms of compensation and the set of data consumers give up to the downstream firms. Thus, competition may not increase compensation or privacy. Recall that D^M is the set of data that a monopoly intermediary would acquire (see Appendix C for the proof).

Theorem 1. *Competition may not increase compensation or privacy: 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: Intermediary k pays consumer i a compensation of $\mathbf{1}_{\{d_i \in D^k\}} C_i$.*

The theorem states that any partition of D^M can arise as an equilibrium allocation of data. Thus, intermediaries collect mutually exclusive sets of data, and the aggregate data collected is equal to the one under monopoly.[14] Across these equilibria, consumer surplus is zero (monopoly level). Thus, the equilibria in Theorem 1 differ only in how intermediaries and the firm divide the surplus created by D^M (Section 6.3 investigates this point).

The intuition for Theorem 1 is as follows. Take any equilibrium described above. Suppose that intermediary 2 deviates and offers positive compensation to consumers for data D^1, which intermediary 1 is going to acquire. Then, these consumers will share D^1 with not only intermediary 2 but 1. Indeed, when consumers share data with one intermediary, they also prefer to share data with other intermediaries that offer positive compensation: By doing so, consumers can earn higher total compensation without increasing the loss from the firm's use of data.[15] However, if consumers share D^1 with intermediaries 1 and 2, these intermediaries have to set a downstream price of zero for D^1 (Lemma 1). Anticipating this, intermediary 2 prefers to not compensate for D^1. Since each intermediary faces no competing offers, it can collect data at the monopsony price C_i. This also implies that intermediaries face the same marginal costs and benefits of collecting data as a monopolist. Thus, competition does not change the aggregate data collected relative to monopoly.

The non-rivalry of data is important not only for consumers obtaining zero surplus (Point 2) but also for the multiplicity of allocations of data: If data were rivalrous, a mild condition guarantees that at most one intermediary acquires non-empty data (Claim 2).

Theorem 1 implies that there is a monopoly equilibrium. Thus, the presence of multiple homogeneous intermediaries may have no impact on the outcome.

Theorem 2. *For any number of intermediaries in the market, there is an equilibrium in which a single intermediary acts as a monopolist described in Claim 1.*

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

The results have several implications. First, competition among data intermediaries may not occur (Theorem 2). Moreover, even if competition occurs, it does not benefit consumers. This

[14] Indeed, in any equilibrium, the allocation of data is partitional. If two intermediaries k and j acquires the same data, then one of them can profitably deviate not collecting the data. The deviating intermediary can save compensation to consumers without losing revenue in the downstream market (Corollary 1).

[15] As I show in Section 8, this argument holds even if consumers incur (exogenous) losses from sharing data with each intermediary.

is captured by non-monopoly equilibria in Theorem 1. In these equilibria, intermediaries obtain small sets of data (relative to monopoly) in the upstream market. This intensifies price competition in the downstream market because different sets of data are imperfect substitutes from the firm's perspective. However, in the upstream market, each intermediary k acts as a monopsony of data D^k. Thus, competition among intermediaries benefits the downstream firm, not consumers (Subsection 6.3 formalizes this). The observation contrasts with the case of rivalrous goods, where competition occurs only in the upstream market (Claim 2).

Second, the results are driven by consumers' ability to share data with multiple intermediaries. This observation connects my results to *data portability* under the EU's GDPR. Data portability states that data controller, such as online platforms, must allow consumers to transfer their data across competitors. Let us interpret the models with non-rivalrous and rivalrous data as the economy with and without data portability, respectively. Then, Theorem 1 and Claim 2 imply that data portability may relax ex ante competition for data and transfer surplus from consumers to intermediaries.[16]

Third, Theorem 2 gives a rationale to the frequently used assumption in the literature that the market consists of a monopoly data seller.[17] We can justify the assumption as a subgame of the extended game in which multiple data sellers first acquire information at cost and then sell collected data.

The results are robust to various extensions. For example, consumers could incur exogenous costs of sharing data with intermediaries (e.g., privacy concern against data intermediaries); U_i could depend on what data the firm holds on consumer $j \neq i$ (e.g., downstream firms use consumer j's data to predict the characteristics of consumer i); intermediaries could incur heterogeneous costs of processing and storing data. Section 7 and Section 8 discuss some of them in detail.

Remark 1. Are there equilibria other than those in Theorem 1? The answer is yes. To see this, consider a single consumer and two intermediaries. There is an equilibrium in which the consumer extracts full surplus $\Pi(d_1) - C_1$: One intermediary, say 1, offers $(\{d_1\}, \Pi(d_1))$, and the other

[16] It would be interesting to examine the welfare impact of data portability by incorporating this potential downside and the intended benefit of preventing consumer lock-in, which the current model does not capture. Krämer and Stüdlein (2019) study a model in which consumers' switching costs depend on data portability.

[17] See, for example, Babaioff et al. (2012), Bergemann et al. (2018), Bergemann and Bonatti (2019b), Bimpikis et al. (2019), and references therein. Sarvary and Parker (1997) is one of the early works that study competition between information sellers.

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 $C_1 < \tau_1^1 < \Pi(d_1)$, then the consumer accepts offers of both intermediaries. This consists of an equilibrium. Intermediary 1 has no incentive to lower compensation because the consumer will then share her data with both intermediaries, following which the price of the data is zero.

There is also an equilibrium in which no data are shared. 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 $\tau \geq C_1$, 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 both intermediaries.

I do not focus on these equilibria for the following reason. In terms of intermediaries' payoffs, the equilibria in Example 1 are Pareto dominated by those in Theorem 1. To study the non-competitive nature of the market for data, it would be reasonable to exclude the former.[18] The equilibria in Theorem 1 are also suitable for studying how the surplus created by data is divided, because they have the same total surplus.

6.2 Multidimensional Data

I now relax assumptions on consumer preferences. Assume that each consumer i has a finite set $\mathcal{D}_i$ of data and incurs increasing convex costs of sharing data with the firm.

Assumption 3. For each $i \in N$, the cost of sharing data $C_i := -U_i$ satisfies the following.

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

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

$$C_i(Y \cup \{d\}) - C_i(Y) \geq C_i(X \cup \{d\}) - C_i(X). \tag{4}$$

This setting involves a new challenge: The equilibria in Theorem 1 have a simple and nice property that each intermediary k asks consumer i for data $d_i \in D_i^k$ and consumers accept all

[18] If C_i is constant across $i \in N$ and $\Pi(D)$ depends only on the cardinality of D, then Theorem 1 corresponds to the set of all equilibria that are Pareto undominated from intermediaries' perspective.

non-empty offers. In contrast, the current setting may not have such an equilibrium.[19] To avoid this difficulty, I impose the following assumption.

Assumption 4. $(U_i)_{i\in N}$ and Π are such that a monopoly intermediary obtains and sells all data, i.e., $D^M = \mathcal{D}$.[20]

Assumption 4 naturally holds in the following two settings. One is when the downstream firm is a seller that uses data for price discrimination. If the firm can perfectly price discriminate consumers using all data $\mathcal{D}$, then the assumption holds. Subsection 7.2 microfounds U_i and Π using this interpretation. The other is when there is an informational externality among consumers, under which a monopoly intermediary can source data cheaply from consumers. To formally examine this, I need to extend the model so that U_i can depend on other consumers' data. Such an extension is discussed in Subsection 8.3.

In terms of primitives, Assumption 4 holds if the firm's marginal revenue from data is high relative to consumers' marginal costs of sharing the data.[21] Under Assumption 4, Theorem 1 extends (see Appendix D for the proof).

Theorem 3. *Take any partitional allocation of data $(D^1, \dots, D^K)$ 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 $(D^1, \dots, D^K)$.*
2. *Intermediary k collects consumer i's data D_i^k at compensation $\hat{\tau}_i^k$, which is i's marginal cost of sharing D_i^k:*

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

In particular, there is an equilibrium in which a single intermediary acts as a monopolist.

A key difference from the case of single unit data (Theorem 1) is the equilibrium compensation (5). Intermediary k now compensates consumer i according to the additional loss that she incurs

[19] For example, suppose that $N = 1$, $\mathcal{D}_i = \{a, b\}$, $C_i(a) = C_i(b) = 0$, $C_i(\{a, b\}) = +\infty$, and $\Pi(a) = \Pi(b) > 0$. A monopolist collects either a or b at zero compensation. If $K = 2$, in any pure-strategy equilibrium, one intermediary offers $(\{a\}, \Pi(a))$, the other intermediary offers $(\{b\}, \Pi(b))$, and the consumer accepts only one of them. Thus, the consumer extracts full surplus if there are multiple intermediaries.

[20] In the current setting, this is equivalent to the assumption that (A) total surplus is maximized when the firm acquires D^M. If there are informational externalities among consumers, then (A) is different from Assumption 4. In that case, my results continue hold under Assumption 4. See Subsection 8.3 for the detail.

[21] For any Π and $(U_i)_{i\in N}$, the assumption holds if the firm's revenue function is $\alpha\Pi$ with a large $\alpha > 1$.

by sharing D_i^k conditional on sharing data with 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 $C_i(\mathcal{D}_i)$. To have a better intuition, consider the following example.

Example 1 (Breaking up data intermediaries). Each consumer has her location and financial data. The downstream firm profits from data but there is a risk of data leakage. Each consumer incurs an expected loss of \$20 from this potential data leakage if only if the firm holds *both* location and financial data (otherwise, she incurs no loss).

Suppose that the market consists of a monopoly intermediary. Then, the intermediary obtains both location and financial data and pays \$20, leaving zero surplus to consumers. For example, the intermediary may operate an online service that requires consumers to provide these data.

Now, suppose that a regulator breaks up the monopolist into two intermediaries, 1 and 2. Theorem 3 implies that in one of the equilibria, intermediaries 1 and 2 collect location and financial data, respectively, and each intermediary pays a compensation of \$20. For example, two intermediaries may operate mobile applications that collect different data, and each application delivers the value of \$20 to consumers. In this equilibrium, each consumer obtains a net surplus of \$20. Thus, breaking up a monopolist may change the equilibrium allocation of data, increase compensation, and benefit consumers.[22] The following subsection generalizes this observation.

6.3 Data Concentration

Theorems 1 and 3 state that any partition of D^M can arise as an equilibrium allocation of data. We can interpret an equilibrium corresponding to a coarser partition as an equilibrium with a greater data concentration among intermediaries. The following definition formalizes this idea:

Definition 2. 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) $\cup_{k \in K} D^k = \cup_{k \in K} \hat{D}^k$ 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 impacts of data concentration on consumers and intermediaries (see Appendix E for the proof).

[22] However, there is also an equilibrium in which a single intermediary acts as a monopolist. This paper does not explore which equilibrium is more likely to arise.

Theorem 4. *Data concentration benefits intermediaries and may hurt consumers and the downstream firm:*

1. *Consider equilibria in Theorem 1. Intermediaries' total profit is higher and the firm's profit is lower in an equilibrium with a more concentrated allocation of data.*

2. *Consider equilibria in Theorem 3. Consumer surplus and the firm's profit are 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 downstream price of data D^k is the firm's marginal revenue $\Pi(\cup_{j\in K} D^j) - \Pi(\cup_{j\in K\setminus\{k\}} D^j)$ from D^k. If there are many intermediaries each of which has a small subset of D^M, then the contribution of each piece of data is close to $\Pi(D^M) - \Pi(D^M \setminus \{d\})$. In contrast, if a few intermediaries jointly hold D^M, each of them can charge a high 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 base compensation on the infra-marginal cost of sharing data.

The nonrivalry of data is important for conducting the meaningful welfare analysis of data concentration. Indeed, if data are rivalrous as in Claim 2, then under a mild condition, only one intermediary obtains data.

7 Equilibrium with General Preferences

So far, I have assumed that consumers incur losses when the downstream firm obtains their data. In reality, firms' use of data may also benefit consumers. For example, a downstream firm may be a financial institution that uses consumer data for fraud detection (e.g., Federal Trade Commission 2014). More generally, the benefit or loss for a consumer of giving up her data to downstream firms should depend on the amount and kind of data.

Motivated by this observation, I allow consumers to have *any* preferences on the downstream firm's use of data. I present a natural extension of a monopoly equilibrium, which captures the

non-competitive feature of markets for personal data. I use this result to study information design by data intermediaries.

7.1 Partially Monopolistic Equilibrium

The following result generalizes Theorem 3 (see Appendix F for the proof).

Proposition 1 (Partially Monopolistic Equilibrium (PME)). *Suppose that U_i is any set function for each i, and Π is any increasing set function. If $K \geq 2$, under Assumption 4, there is an equilibrium in which a single intermediary obtains all data and pays each consumer i a compensation of $\max_{D \subset \mathcal{D}_i} U_i(D) - U_i(\mathcal{D}_i)$. Thus, consumer i obtains an equilibrium payoff of $\max_{D \subset \mathcal{D}_i} U_i(D)$.*

If $U_i(D_i)$ is decreasing for each i, the PME reduces to a monopoly equilibrium. In contrast, suppose that $\max_{D \subset \mathcal{D}_i} U_i(D) > U_i(\emptyset) = 0$, that is, consumer i prefers to share some data with the downstream firm for free. Proposition 1 implies that consumer surplus in the PME is then greater than under monopoly (Claim 1) but lower than in the market with rivalrous goods (Claim 2).

To see why competition benefits consumers when $U_i^* := \max_{D \subset \mathcal{D}_i} U_i(D) > 0$, consider the extreme case where consumer i prefers to share all data for free, i.e., $U_i^* = U_i(\mathcal{D}_i)$. A monopoly intermediary extracts full surplus from consumer i by charging a fee of $U_i^* > 0$. In contrast, if there are multiple intermediaries and intermediary k charges a positive fee, then another intermediary $j \neq k$ can offer a slightly lower fee to *exclusively* obtain data from consumer i. Indeed, consumer i has no incentive to accept the offer of intermediary k, because she can enjoy a benefit of U_i^* as long as intermediary j transfers her data. This restores Bertrand competition, which drives down the equilibrium fees to zero. However, competition does not force intermediaries to offer positive compensation (i.e., negative fees). Due to the non-rivalry of data, once intermediaries offer positive compensation, consumers share data with all of them, which will hurt intermediaries.

Proposition 1 states that the above intuition applies to arbitrary preferences. Figure 2 assumes $N = 1$ and depicts U_i and Π as functions of the amount of data that the firm has on i. U_i is non-monotone, and Π now exhibits increasing returns to scale. First, the monopoly intermediary obtains all data at a compensation of $-U_i(\mathcal{D}_i)$ (short red dotted arrow). Let us decompose the monopoly compensation $-U_i(\mathcal{D}_i)$ into two parts: The monopolist extracts surplus created by $D_i^* \in \arg\max_{D \subset \mathcal{D}_i} U_i(D)$ from consumer i by charging $U_i(D_i^*) > 0$, and it obtains additional

data $\mathcal{D}_i \setminus D_i^*$ at the minimum compensation $U_i(D_i^*) - U_i(\mathcal{D}_i)$ (long blue dotted arrow). In contrast, when there are multiple intermediaries, competition prevents intermediaries from extracting surplus $U_i(D_i^*)$. This guarantees that each consumer i obtains a payoff of at least $U_i(D_i^*)$. However, competition does not increase compensation for data $\mathcal{D}_i \setminus D_i^*$, the sharing of which hurts consumer i. Thus, in the partially monopolistic equilibrium, a single intermediary acquires all data and compensates consumers according to the loss $U_i(D_i^*) - U_i(\mathcal{D}_i)$ of sharing $\mathcal{D}_i \setminus D_i^*$. Finally, the compensation in the PME is lower than $\Pi(\mathcal{D}_i)$, which is the compensation that the consumer would have received in the market for physical goods (black dashed arrow).

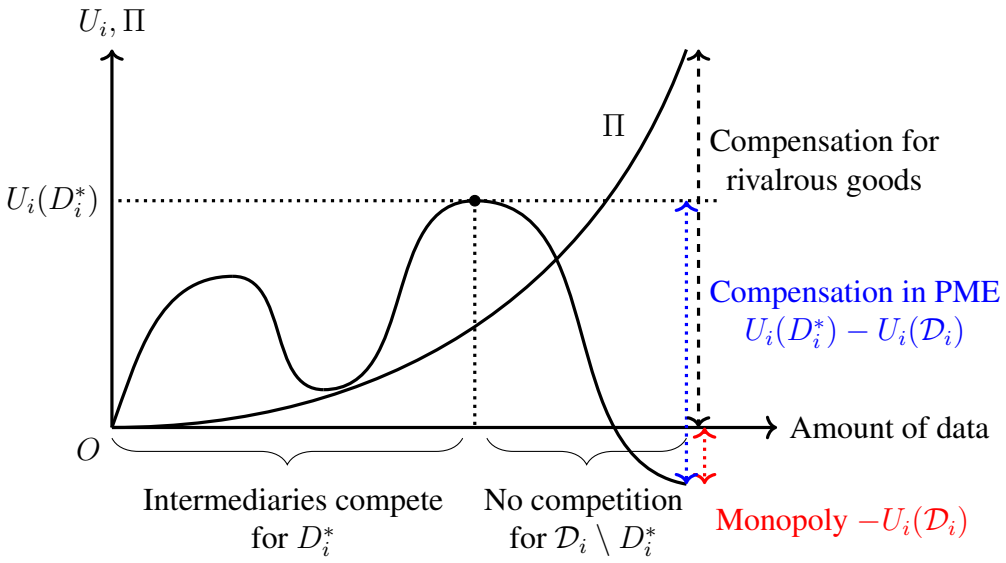

Figure 2: Partially monopolistic equilibrium

The next result shows that if the market consists of many intermediaries, then the PME minimizes consumer surplus and maximizes intermediary surplus across all equilibria (see Appendix G for the proof). This result corroborates the claim that the PME is a natural extension of a monopoly equilibrium. To state the result, let $CS_i(K)$ denote the set of all possible equilibrium payoffs of consumer i when the market consists of K intermediaries.

Proposition 2. *As the number K of intermediaries grows large, the worst consumer surplus and the best intermediary surplus in equilibrium converge to those in the partially monopolistic equilibrium. Formally, the following holds.*

1. *For each $i \in N$, $\lim_{K\to\infty} (\inf CS_i(K)) = \max_{D\subset\mathcal{D}} U_i(D)$. The result holds even when $\mathcal{D}$ is infinite*

as long as the right hand side is well-defined.

2. *Suppose that $\mathcal{D}$ is finite and Π is strictly increasing. There is a $K^* \in \mathbb{N}$ such that for any $K \geq K^*$ and $i \in N$,* $\min CS_i(K) = \max_{D \subset \mathcal{D}} U_i(D)$.

The intuition is as follows. Suppose that there are K intermediaries and in some equilibrium, consumer i obtains a payoff of $U_i(D_i^*) - \delta_K$ with $\delta_K > 0$. If an intermediary offers (D_i^*, ε) with $\varepsilon < \delta_K$, consumer i prefers to accept it. Because any intermediary can always deviate and offer (D_i^*, ε), each intermediary obtains a payoff of at least δ_K. This implies that intermediary surplus is at least $K \cdot \delta_K$. However, intermediary surplus is bounded from above by the total surplus at the efficient outcome, which is finite. Thus, $\delta_K \to 0$ as K grows large, i.e., the worst consumer surplus converges to $U_i(D_i^*)$ as the number of intermediaries grows large. Point 2 shows that under a stronger assumption, $U_i(D_i^*)$ is exactly the lowest equilibrium payoff of consumer i for a sufficiently large but finite K. Finally, in the PME, consumer surplus is $\sum_{i \in N} U_i(D_i^*)$ and intermediaries obtain the remaining surplus from the efficient outcome. Thus, the PME is (approximately) an intermediary-optimal outcome for a large K.

A takeaway of this section is that *the benefit for consumers from competition among intermediaries is typically lower than in traditional markets for physical goods.* Competition eliminates fees that consumers would have to pay in a monopoly market. However, due to the nonrivalry of data, competing intermediaries have no incentive to offer positive compensation. The latter creates a gap between consumer benefits from competition in the market for data and in traditional markets.

7.2 Information Design by Data Intermediaries

I use the above results to study information design by data intermediaries. I assume that a downstream firm is a seller that uses data for product recommendation and price discrimination. Each piece of data is an informative signal about consumers' willingness to pay, and intermediaries can potentially collect any signals.

The formal description is as follows. Assume for simplicity that there is a single consumer (thus, omit subscript i). A firm is a seller that provides $M \in \mathbb{N}$ products $1, \ldots, M$. The consumer has a unit demand, and her values for products, $\mathbf{u} := (u_1, \ldots, u_M)$, are independently

and identically distributed according to a cumulative distribution function F with a finite support $V \subset (0, +\infty)$.[23]

The consumer has a set of data $\mathcal{D}$, where each $d \in \mathcal{D}$ is a signal (Blackwell experiment) from which the seller can learn about $\mathbf{u}$. I assume that $\mathcal{D}$ consists of all signals with finite realization spaces and that intermediaries can ask consumers for any finite set of signals.[24]

After buying a set of data $D \subset \mathcal{D}$ from intermediaries, the seller learns about $\mathbf{u}$ from signals in D. Then, the seller sets a price and recommends one of M products to the consumer. Finally, the consumer observes the value and the price of the recommended product, and she decides whether or not to buy it.[25] A recommendation could be an advertiser displaying a targeted advertisement or an online retailer showing a product as a personalized recommendation. If the consumer buys product m at price p, her payoff from this transaction is $u_m - p$. Otherwise, her payoff is zero. The seller's payoff is its revenue. In any subgame where the seller has obtained data D, I consider pure-strategy perfect Bayesian equilibrium such that the seller calculates its posterior beliefs based on the prior F and signals in D on and off the equilibrium paths.[26]

An important observation is that Assumption 4 holds, i.e., a monopoly intermediary collects all data $\mathcal{D}$. Indeed, if the seller has all data, it can access a fully informative signal and perfectly learn $\mathbf{u}$. Then, the seller can recommend the highest value product and perfectly price discriminate the consumer, which maximizes total surplus. Thus, a monopoly intermediary, which can extract total surplus, collects and sells all data in equilibrium.

To simplify exposition, I prepare several notations. Given a set D of signals, let $U(D)$ and $\Pi(D)$ denote the expected payoffs of the consumer and the seller, respectively, when the seller that has D optimally sets a price and recommends a product, and the consumer makes an optimal purchase decision. Note that $\Pi(D)$ is increasing because a larger D corresponds to a more informative

[23] I define F as a left-continuous function. Thus, $1 - F(p)$ is the probability that the consumer's value for any given product is weakly greater than p at the prior.

[24] To close the model, I need to specify how realizations of different signals are correlated conditional on $\mathbf{u}$. One way is to use the formulation of Gentzkow and Kamenica (2017): Let X be a random variable that is independent of $\mathbf{u}$ and uniformly distributed on $[0, 1]$ with typical realization x. A signal d is a finite partition of $V^M \times [0, 1]$, and the seller observes a realization $s \in d$ if and only if $(\mathbf{u}, x) \in s$. However, the result does not rely on this particular formulation.

[25] The model assumes that the seller only recommends one product, and thus the consumer cannot buy non-recommended products. This captures the restriction on how many products can be marketed to a given consumer. See Ichihashi (Forthcoming) for a detailed discussion of the motivation behind this formulation.

[26] I assume that the seller breaks ties in favor of the consumer. The existence of an equilibrium is shown in Ichihashi (Forthcoming).

signal. Define $p(F) := \min(\arg\max_{p \in V} p[1 - F(p)])$. $p(F)$ is the lowest monopoly price given a value distribution F.

Consider a benchmark with a monopoly intermediary. The intermediary obtains the efficient amount of information (such as a fully informative signal) and extracts full surplus from the consumer and the seller. Thus, consumer surplus is $U(\emptyset)$, which is her payoff in a hypothetical scenario in which the seller recommends one of M products randomly at a price of $p(F)$.

If the market consists of multiple intermediaries, consumer surplus in the partially monopolistic equilibrium is equal to the one in a hypothetical scenario where the consumer directly discloses information to the seller. In other words, consumer surplus is equal to the one in Bayesian persuasion (see Appendix H for the proof).

Proposition 3. *Suppose that there are multiple intermediaries. In the partially monopolistic equilibrium, one intermediary (say* 1*) obtains a fully informative signal, and the consumer obtains a payoff of* $\max_{d \in \mathcal{D}} U(d)$*. Moreover, this equilibrium satisfies the following.*

1. *If the seller provides a single product (*$M = 1$*), all intermediaries earn zero payoffs. The consumer obtains payoff* $U(d^*)$*, where* d^* *is the consumer-optimal segmentation in Bergemann et al. (2015).*

2. *Suppose that the seller provides multiple products (*$M \geq 2$*). For a generic prior* F *satisfying* $p(F) > \min V > 0$*, intermediary* 1 *earns a positive payoff that is independent of the number of intermediaries.*[27]

The intuition is as follows. First, consider Point 1. Bergemann et al. (2015) show that there is a signal d^* such that (i) d^* maximizes the consumer's payoff, i.e., $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$, (ii) the seller is indifferent between obtaining d^* and nothing, i.e., $\Pi(d^*) = \Pi(\emptyset)$, and (iii) d^* maximizes total surplus $U(d) + \Pi(d)$. (i) implies that competing intermediaries cannot charge the consumer a positive fee for d^*. (ii) implies that they cannot charge the firm a positive price for d^*. Moreover, (iii) implies that intermediaries cannot make a profit by obtaining and selling additional information. Thus, in the PME, the consumer obtains a payoff of $U(d^*)$, and no intermediaries

[27] A generic F means that the statement holds for any probability distribution in $\Delta(V) \subset \Delta(\mathbb{R})$ satisfying $p(F) > \min V$, except for those that belong to some Lebesgue measure-zero subset of $\Delta(V)$.

can make a positive profit. In this case, competition among intermediaries yields the consumer all welfare gain from her information. Moreover, when K is large, this equilibrium (PME) is worst for the consumer. This implies when $M = 1$ and K is large, the equilibrium outcome is (almost) unique.

Second, consider Point 2. Ichihashi (Forthcoming) shows that if the prior F satisfies the condition in Point 2, then any consumer-optimal signal $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$ leads to inefficiency. Intuitively, d^* conceals some information about which product is most valuable to the consumer. This benefits the consumer by inducing the seller to lower prices, but it leads to inefficiency due to product mismatch. This inefficiency (under the hypothetical Bayesian persuasion) creates a room for competing intermediaries to earn a positive profit: An intermediary can additionally obtain information that enables the seller to perfectly learn the consumer's values. The consumer requires a positive compensation to share such information. This, in turn, implies that a single intermediary can act as a monopoly of the information. Thus, competition benefits the consumer relative to monopoly but it does not completely dissipate intermediaries' profits.

8 Extensions

8.1 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 utility of sharing data is $\sum_{\ell \in L} U_i^\ell$, where each U_i^ℓ depends on the set of i's data that firm ℓ obtains.

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 plus $\sum_{\ell \in L} U_i^\ell(D_i)$. Therefore, we can apply the same analysis as before by defining $U_i := \sum_{\ell \in L} U_i^\ell$.

8.2 Privacy Concern Toward Data Intermediaries

Consumers may incur exogenous costs of sharing data with not only downstream firms but also data intermediaries. I can incorporate this by assuming that consumer i incurs a loss of ρK_i by sharing her data with K_i intermediaries. For the case of single unit data (Subsection 6.1), the result does not change qualitatively. If $\rho > 0$, intermediaries obtain less data than the original model, because it has to pay a compensation of at least $C_i + \rho$ to each consumer. Any equilibrium allocation of data is partitional, and there are multiple equilibria, one of which is a monopoly equilibrium.

8.3 Informational Externality Among Consumers

So far, I have assumed that U_i depends only on D_i. That is, the payoff of consumer i does not depend on what data the downstream firm has on consumer $j \neq i$. This assumption might fail, for instance, if the firm uses data on consumer j to infer consumer i's willingness to pay and price discriminate i on that basis. For another instance, U_i could depend on data on other consumers if the firm chooses a single action such as a price or a product design based on the aggregate data.

The model can incorporate such dependency ("informational externality") by writing U_i as $U_i(D_i, D_{-i})$, where $D_i \subset \mathcal{D}_i$ and $D_{-i} \subset \cup_{j \in N \setminus \{i\}} \mathcal{D}_j$. Suppose that for any D_{-i}, $U_i(\cdot, D_{-i})$ satisfies assumptions in the previous sections such as submodularity. Then, all the results continue to hold under the additional assumption that each consumer does not observe offers made to other consumers. To see why we need this assumption, suppose that offers are publicly observable and intermediary k makes a deviating offer to consumer i. When U_j depends on what data the firm will have on consumer i, then this deviation may affect the data-sharing decision of consumer $j \neq i$ to intermediary $\ell \neq k$. In this case, intermediaries may not be able to sustain a monopoly outcome since each intermediary may fail to internalize how its deviation affect other intermediaries.

Intuitively, if there is an informational externality among a large number of consumers, As-

sumption 4 is more likely to hold. This is because an externality creates a gap between the gains from data that accrue to a monopoly intermediary and the marginal compensation received by consumers (Bergemann and Bonatti, 2019a).

9 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: Data are non-rivalrous, and the use of data by third parties could affect consumers. These features drastically change the nature of competition relative to the intermediation of physical goods: When firms' use of data hurts consumers, data intermediaries may secure monopoly profit in some equilibrium, 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. The main insights hold even when consumers have heterogeneous and arbitrary preferences over the firm's use of data: Intermediaries compete for data that consumers would voluntarily share with the firm, and a single intermediary acts as a monopsony of data for which consumers would require compensation.

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.

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 and Alessandro Bonatti (2019a), "The economics of social data."

Bergemann, Dirk and Alessandro Bonatti (2019b), "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.

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

Bimpikis, Kostas, Davide Crapis, and Alireza Tahbaz-Salehi (2019), "Information sale and competition." *Management Science*, 65, 2646–2664.

Bonatti, Alessandro and Gonzalo Cisternas (Forthcoming), "Consumer scores and price discrimination." *Review of Economic Studies*.

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

Carrillo, Juan and Guofu Tan (2015), "Platform competition with complementary products." Technical report, Working paper.

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

Crémer, Jacques, Yves-Alexandre de Montjoye, and Heike Schweitzer (2019), "Competition policy for the digital era." *Report for the European Commission*.

d'Aspremont, Claude, J Jaskold Gabszewicz, and J-F Thisse (1979), "On hotelling's" stability in competition"." *Econometrica: Journal of the Econometric Society*, 1145–1150.

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

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

Furman, Jason, D Coyle, A Fletcher, D McAules, and P Marsden (2019), "Unlocking digital competition: Report of the digital competition expert panel." *HM Treasury, United Kingdom*.

Galeotti, Andrea and José Luis Moraga-González (2009), "Platform intermediation in a market for differentiated products." *European Economic Review*, 53, 417–428.

Gentzkow, Matthew and Emir Kamenica (2017), "Bayesian persuasion with multiple senders and rich signal spaces." *Games and Economic Behavior*, 104, 411–429.

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

Hagiu, Andrei and Julian Wright (2014), "Marketplace or reseller?" *Management Science*, 61, 184–203.

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

Ichihashi, Shota (Forthcoming), "Online privacy and information disclosure by consumers." *American Economic Review*.

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."

Krämer, Jan and Nadine Stüdlein (2019), "Data portability, data disclosure and data-induced switching costs: Some unintended consequences of the general data protection regulation." *Economics Letters*, 181, 99–103.

Kummer, Michael and Patrick Schulte (2019), "When private information settles the bill: Money and privacy in googles market for smartphone applications." *Management Science*.

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

Morton, Fiona Scott, Theodore Nierenberg, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A Douglas Melamed, and Jamie Morgenstern (2019), "Report: Committee for the study of digital platforms-market structure and antitrust subcommittee." *George J. Stigler Center for the Study of the Economy and the State, The University of Chicago Booth School of Business*.

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

Rhodes, Andrew, Makoto Watanabe, and Jidong Zhou (2018), "Multiproduct intermediaries."

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

Sarvary, Miklos and Philip M Parker (1997), "Marketing information: A competitive analysis." *Marketing science*, 16, 24–38.

Stahl, Dale O (1988), "Bertrand competition for inputs and walrasian outcomes." *The American Economic Review*, 189–201.

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 $D^* := \cup_{k \in K} D^k$. 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 intermediary 2 but not 1. Lemma 1 implies that, 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 1, $\Pi(D^* - D^2) - \Pi(D^* - D^1 - D^2) \geq \Pi(D^*) - \Pi(D^* - D^1)$. 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 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 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 earns a lower revenue. Thus, no intermediary has a profitable deviation.

I next turn to proving uniqueness. 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 inter-

mediary 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

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} p^k = \max_{J\subset K\setminus\{1\}} \left(\Pi(\cup_{k\in J} D^k) - \sum_{k\in J} p^k\right). \tag{6}$$

To see this, suppose that 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 (6). 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

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} \hat{p}^k < \Pi(\cup_{k\in J^*} D^k) - \sum_{k\in J^*} \hat{p}^k, \tag{7}$$

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 (7) 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 (6) 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

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} p^k = \Pi(\cup_{k\in K^*\setminus\{1\}} D^k) - \sum_{k\in K^*\setminus\{1\}} p^k. \tag{8}$$

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

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D^k) < \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D^k). \tag{9}$$

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

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

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 (8) 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 (9) 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 Theorem 1

Proof. Take any partitional 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) 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 (or equivalently, the set of consumers) that intermediary k obtains as a result of the deviation. If this deviation is strictly profitable for k, it holds that $\Pi(\hat{D}^k \cup D^{-k}) - \Pi(D^{-k}) - \sum_{i \in \hat{D}^k} C_i > \Pi(D^k \cup D^{-k}) - \Pi(D^{-k}) - \sum_{i \in D^k} C_i$. However, this never holds because the monopolist could then earn strictly higher revenue by obtaining and selling $\hat{D}^k \cup D^{-k}$ instead of D^M, which is a contradiction. □

D 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 $\Pi(D^M) - \Pi(D^M \setminus D^k) - [\Pi(D^M \setminus D^-) - \Pi(D^M \setminus D^k)] = \Pi(D^M) - \Pi(D^M \setminus D^-)$. 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 an optimal choice of the monopolist, it holds that $\Pi(D^M) - \Pi(D^M \setminus D^-) - \sum_{i \in N}[C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^-)] \geq 0$. Therefore, the deviation does not increase intermediary k's payoff. □

E 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), then for each $i \in N$, $(\hat{D}_i^k)$ is more concentrated than (D_i^k). □

F Proof of Proposition 1

Proof. Consider the following strategy profile: In the upstream market, intermediary 1 offers $(\mathcal{D}_i, U(D_i^*) - U(\mathcal{D}_i))$ to each consumer i. Other intermediaries offer $(D_i^*, 0)$ to each consumer i. Consumers accept only the offer of intermediary 1. If an intermediary deviates, then consumers optimally decide which intermediaries to share data with, breaking ties in favor of sharing data. In the downstream market, if intermediary 1 does not deviate in the upstream market, then any intermediary $j \neq 1$ sets a price of zero, and intermediary 1 sets a price of $\Pi(\mathcal{D}) - \Pi(D^{-1})$, where D^{-1} is the set of data that intermediaries other than 1 hold. If intermediary 1 deviates in the upstream market, then assume that players play any equilibrium of the corresponding subgame.

I show that the suggested strategy profile consists of an equilibrium. First, I show that intermediary 1 has no incentive to deviate. Suppose that intermediary 1 deviates and obtains data D_i^1 from each consumer i. Let $\hat{D}_i$ denote the set of all data that consumer i shares as a result of 1's deviation ($D_i^1 \subsetneq \hat{D}_i$ if consumer i also shares data with some intermediary $j \neq 1$). The revenue of intermediary 1 in the downstream market is at most $\Pi(\cup_{i\in N}\hat{D}_i)$. The compensation to each consumer i has to be at least $\tau_i \geq U(D_i^*) - U(\hat{D}_i)$. To see this, suppose $U(D_i^*) > U(\hat{D}_i) + \tau_i$. The left hand side is the payoff that consumer i can attain by sharing data exclusively with intermediary $k > 1$. The right hand side is her maximum payoff conditional on sharing data with intermediary 1. Note that all intermediaries other than 1 offer zero compensation. Then, $U(D_i^*) > U(\hat{D}_i) + \tau_i$ implies that consumer i would strictly prefer to reject the offer from intermediary $k \neq 1$. Now, these bounds on revenue and cost imply that intermediary 1's payoff after the deviation is at most $\Pi(\cup_{i\in N}\hat{D}_i) - \sum_{i\in N}[U_i(D_i^*) - U_i(\hat{D}_i)] = \Pi(\cup_{i\in N}\hat{D}_i) + \sum_{i\in N} U_i(\hat{D}_i) - \sum_{i\in N} U_i(D_i^*)$. Since the efficient outcome involves full data sharing, this is at most $\Pi(\cup_{i\in N}\mathcal{D}_i) + \sum_{i\in N} U_i(\mathcal{D}_i) - \sum_{i\in N} U_i(D_i^*) = \Pi(\cup_{i\in N}\mathcal{D}_i) - \sum_{i\in N}[U_i(D_i^*) - U_i(\mathcal{D}_i)]$, which is intermediary 1's payoff without deviation. Thus, there is no profitable deviation for intermediary 1.

Second, suppose that intermediary 2 deviates and offers (D_i^2, τ_i^2) to each consumer i. Without loss of generality, assume that each consumer accepts the offer. Let D_i^{-1} denote the set of data that consumer i provides to intermediaries in $K \setminus \{1\}$ after the deviation. If the consumer accepts the offer of intermediary 1, her payoff increases by $U_i(\mathcal{D}_i) - U_i(D_i^{-1}) + U_i(D_i^*) - U_i(\mathcal{D}_i) \geq$

$U_i(\mathcal{D}_i) - U_i(D_i^*) + U_i(D_i^*) - U_i(\mathcal{D}_i) = 0$. The inequality follows from $U_i(D_i^*) \geq U_i(D_i^{-1})$. Thus, each consumer i prefers to accept the offer of intermediary 1. If $\tau_i^2 \geq 0$, this implies that intermediary 2's could be better off (relative to the deviation) by not collecting D_i^2, because it can save compensation without losing revenue in the downstream market. Indeed, intermediary 2's revenue in the downstream market is zero for any increasing Π. If $\tau_i^2 < 0$, consumer i strictly prefers sharing data with intermediary 1 to sharing data with intermediary 2. Overall, these imply that intermediary 2 does not benefit from the deviation. □

G Proof of Proposition 2

Proof. (Proof of Point 1) I prepare several notations. Define $U_i^* := \max_{D \subset \mathcal{D}} U_i(D)$, and $TS^* := \Pi(\mathcal{D}) + \sum_{i \in N} U_i(\mathcal{D}) > 0$ where $U_i(\mathcal{D}) := U_i(\mathcal{D}_i)$. Assumption 4 ensures that TS^* is the maximum total surplus.

As U_i^* is an equilibrium payoff in the PME, $\inf CS_i(K) \leq U_i^*$ holds for all $K \in \mathbb{N}$. Thus, we obtain $\limsup_{K \to \infty}(\inf CS_i(K)) \leq U_i^*$. To obtain the result, it suffices to show that

$$\liminf_{K \to \infty}(\inf CS_i(K)) \geq U_i^*.$$

Suppose to the contrary that $\liminf_{K \to \infty}(\inf CS_i(K)) < U_i^* - 3\delta$ for some δ. This implies that there exists a strictly increasing subsequence $\{K_n\} \subset \mathbb{N}$ such that $\inf CS_i(K_n) < \liminf_{K \to \infty}(\inf CS_i(K_n)) + \delta < U_i^* - 2\delta$. This implies that for each K_n, there exists an equilibrium E_n in which the payoff of consumer i, denoted by CS_i^n, satisfies $CS_i^n < U_i^* - \delta$.

I show that this leads to a contradiction. Take any K_n. Suppose that intermediary k deviates and offers (D_i^*, ε) with $\varepsilon \in (0, \delta)$ to consumer i. If consumer i rejects this deviating offer, her payoff is at most $CS_i(K_n)$. If she accepts the deviating offer and rejects all other offers, her payoff is $U_i^* - \varepsilon > U_i^* - \delta$. Thus, consumer i accepts the deviating offer. This implies that for each n, in equilibrium E_n, any intermediary earns a payoff of at least δ, which implies that the sum of payoffs of all intermediaries is at least $K_n\delta$. However, for a large K_n, we obtain $K_n\delta > TS^*$, which is a contradiction. Combining $\liminf_{K \to \infty}(\inf CS_i(K)) \geq U_i^*$ and $\limsup_{K \to \infty}(\inf CS_i(K)) \leq U_i^*$, we obtain $\lim_{K \to \infty}(\inf CS_i(K)) = U_i^*$.

(Proof of Point 2) Define $m := \min_{d \in \mathcal{D}, D \subset \mathcal{D}} \Pi(D) - \Pi(D \setminus \{d\}) > 0$. Let K^* satisfy $K^* >$

TS^*/m. Suppose that there are $K \geq K^*$ intermediaries and take any equilibrium. Suppose (to the contrary) that the payoff of consumer i is $U_i(D_i^*) - \delta$ with $\delta > 0$. I derive a contradiction by assuming that any intermediary obtains a payoff of at least m. Suppose to the contrary that intermediary k earns a strictly lower payoff than m. If intermediary k deviates and offers (D_i^*, ε) with $\varepsilon \in (0, \delta)$ to consumer i, then she accepts this offer. Let D_i^{-k} denote the data that consumer i shares with intermediaries in $K \setminus \{k\}$ as a result of k's deviation. Then, $D_i^* \setminus D_i^{-k} \neq \emptyset$ holds. To see this, suppose to the contrary that $D_i^* \subset D_i^{-k}$. Then, consumer i could be strictly better off by rejecting intermediary k's offer (D_i^*, ε) because $\varepsilon > 0$. However, conditional on rejecting k's deviating offer, the set of offers that consumer i faces shrinks relative to the original equilibrium. Thus, the maximum payoff the consumer can achieve by rejecting k's deviating offer is at most $U_i(D_i^*) - \delta < U_i(D_i^*) - \varepsilon$, which is a contradiction. Since consumer i accepts the offer of intermediary k and $D_i^* \setminus D_i^{-k} \neq \emptyset$, intermediary k can earn a profit arbitrarily close to m from consumer i. This implies that in the equilibrium, any intermediary earns a payoff of at least m; otherwise, an intermediary can profitably deviate by offering empty offers to all consumers in $N \setminus \{i\}$ and (D_i^*, ε) to consumer i. However, if each intermediary earns at least m, the sum of payoffs of all intermediaries is at least $Km > TS^*$. This implies that one of consumers and the firm obtains a negative payoff, which is contradiction. Therefore, in any equilibrium, any consumer obtains a payoff of at least $U_i(D_i^*)$. □

H Proof of Proposition 3

Proof. Note that Proposition 1 holds even when $\mathcal{D}$ is not finite. Let d_{FULL} denote a fully informative signal. I show Point 1. Assuming that there is a single product ($M = 1$), Bergemann et al. (2015) show that there is a signal d^* that satisfies the following conditions: $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$; $\Pi(d^*) = \Pi(\emptyset)$; d^* maximizes total surplus, i.e., $U(d^*) + \Pi(d^*) = U(d_{FULL}) + \Pi(d_{FULL})$. Namely, d^* simultaneously maximizes consumer surplus and total surplus without increasing the seller's revenue. These properties imply that intermediary 1's revenue in the downstream market is equal to the compensation it pays in the upstream market: $\Pi(d_{FULL}) - \Pi(\emptyset) = \Pi(d_{FULL}) - \Pi(d^*) = U(d^*) - U(d_{FULL})$. Thus, all intermediaries earn zero payoffs.

I show Point 2. Ichihashi (Forthcoming) shows that if $M = 2$, then for a generic F satisfying

$p(F) > \min V$, any signal $d^{**} \in \arg\max_{d \in \mathcal{D}} U(d)$ leads to an inefficient outcome. This implies $\Pi(d_{FULL}) + U(d_{FULL}) > \Pi(d^{**}) + U(d^{**}) \geq \Pi(\emptyset) + U(d^{**})$. Then, $\Pi(d_{FULL}) - \Pi(\emptyset) - [U(d^{**}) - U(d_{FULL})] > 0$. Thus, intermediary 1 earns a positive profit. $\square$