Non-competing Data Intermediaries

Click here to download the latest version.

Shota Ichihashi*

July 29, 2019

Abstract

I study competition among data intermediaries—technology companies and data brokers that collect consumer data and sell them to downstream firms. When firms' use of data hurts consumers, intermediaries need to compensate consumers for collecting their data. However, competition may not increase compensation: If intermediaries offer high compensation, consumers share data with multiple intermediaries, which lowers the price of data in the downstream market 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 use the results to solve 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, Sitian Liu, Shunya Noda, and seminar participants at the Bank of Canada, CEA Conference 2019, Decentralization Conference 2019, Yokohama National University, and the 30th Stony Brook Game Theory Conference. The opinions expressed in this article are the author's own and do not reflect the views of Bank of Canada.

1 Introduction

This paper studies competition among data intermediaries, which collect and distribute personal data between consumers and firms: Data brokers, such as LiveRamp and Nielsen, collect consumer data and sell them to retailers and advertisers (Federal Trade Commission, 2014). Technology companies, such as Google and Facebook, collect user data and share them indirectly through targeted advertising spaces. Mobile application developers collect user data and share them with third parties (Kummer and Schulte, 2019). I regard these companies as data intermediaries and study their competition and its welfare implication.

For example, consider consumers deciding whether to join online platforms, which collect personal data and share them with third parties. The use of data by third parties may benefit or hurt consumers. For instance, it may hurt consumers through price discrimination, spam, and further data leakage. In this case, platforms need to provide consumers valuable services and rewards (e.g., social media) to obtain their data.

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

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

The model consists of consumers, data intermediaries, and downstream firms. In the upstream market, intermediaries collect data from consumers in exchange for compensation. In the downstream market, intermediaries post prices and sell collected data to firms. What data each intermediary acquires in the upstream market depends on consumers' data-sharing decisions: Each consumer decides what data to share with each intermediary, balancing compensation it offers and the expected benefit or loss she will experience when downstream firms acquire her data.

We may think that competition among intermediaries increases equilibrium compensation to consumers. The key idea of the paper is that this does not occur: Suppose that intermediaries offer consumers high compensations to obtain more data. Consumers then share their data with multiple intermediaries. This intensifies price competition and lowers the price of data in the downstream market, which hurts intermediaries. Thus, *paying more for consumer data makes the data worth less in the downstream market*. The economic force is driven by the non-rivalry of data: Unlike physical goods, the same data can be simultaneously obtained by any number of intermediaries.

I show that this economic force leads to multiple equilibria that differ in the set of data that each intermediary acquires. There are three main findings. First, when the use of data by downstream firms hurts consumers, there is a monopoly equilibrium in which one intermediary extracts the maximum possible surplus. Other intermediaries do not compensate consumers, because if they do, consumers will then share their data with multiple intermediaries. Thus, competition among data intermediaries may not dissipate their profits.

The second main finding is on *data concentration*, which refers to a small number of intermediaries acquiring a large amount of data. There are equilibria with different degrees of concentration. I show that if consumers incur the increasing marginal loss of sharing data with firms, then data concentration benefits intermediaries and hurts consumers. This is because large intermediaries compensate consumers based on their infra-marginal loss. I also study the intensive and extensive margins of data concentration and explore their welfare implications.

The third main finding concerns a general case in which downstream firms' use of data may hurt or benefit consumers, depending on the amounts and types of data that firms acquire. For any quasilinear consumer preferences, I identify an equilibrium that (under a weak assumption) maximizes intermediary surplus and minimizes consumer surplus. In this *partially monopolistic equilibrium,* intermediaries compete for some data but one intermediary acts as a monopolist for other data. As a result, intermediary surplus and consumer surplus fall between those under monopoly and a hypothetical setting in which data are treated as rivalrous goods.

I use this result to study information design by competing intermediaries. I assume that downstream firms use consumer data for price discrimination and product recommendation. Intermediaries can potentially obtain and sell any informative signals (Blackwell experiments) about consumers' willingness to pay. In the partially monopolistic equilibrium, a single intermediary obtains

a fully informative signal, with which firms can perfectly price discriminate and recommend the highest-value products to consumers. The equilibrium consumer surplus is equal to the one in a hypothetical Bayesian persuasion where consumers directly disclose information to firms.

The paper relates to two issues of the data economy. One is why consumers do not seem to be compensated for providing their data (Arrieta-Ibarra et al., 2018; Carrascal et al., 2013). My paper helps us understand this issue by clarifying a mechanism by which compensation to consumers fails to reflect the value of their data to downstream firms. The explanation does not depend on consumer unawareness or the lack of transparency. I argue that such an explanation is important because there has been increasing awareness of data sharing practices, and regulators have tried to ensure consumers' control over data (e.g., the EU General Data Protection Regulation). The other issue is data concentration in the hands of major Internet platforms (e.g., Sokol and Comerford, 2015). I show that data concentration can arise even though data are non-rivalrous and the model excludes network externalities. The welfare implications of data concentration are new in the literature.

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 a unique equilibrium payoff 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 the recent literature on markets for data. Bergemann and Bonatti (2019b) study under what condition a data intermediary can earn a positive profit. They consider a monopoly data intermediary and assume that a downstream firm uses data for price discrimination that hurts consumers. In contrast, I assume that the intermediation of data is profitable and focus on competition and data concentration. My results give a

rationale to the assumption of a monopolistic data seller (see also the discussion after Theorem 2).

More broadly, this paper relates to works on markets for data beyond the context of price discrimination. 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 data set, they consider supermodularity as well.[1] In contrast to their work, I endogenize how intermediaries collect consumer data in the upstream market. This 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 platform competition in two-sided markets. The literature typically assumes that a transaction between two sides is mutually beneficial (e.g., Armstrong (2006); Caillaud and Jullien (2003); Rochet and Tirole (2003)). This is natural in many applications such as video games (consumers and game developers) and credit cards (cardholders and merchants). 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 transaction hurts one side—that is, when downstream firms use data to extract rents from consumers— competition among intermediaries involves raising compensation for consumers. I show that such competition does not occur when multi-homing is possible, which is captured by the nonrivalry of data.

Caillaud and Jullien (2003) show that intermediaries have an incentive to make their services non-exclusive in order to relax price competition. Their result is logically distinct from mine. For instance, in their model, intermediaries earn positive (but below monopoly) profits only if matching technology is costly and imperfect. In my model, intermediaries can earn a monopoly profit without these frictions. Negative cross-side externalities also appear in models of advertising plat-

[1] However, Proposition 3 shows that the main insight holds regardless of the shape of a firm's revenue function.

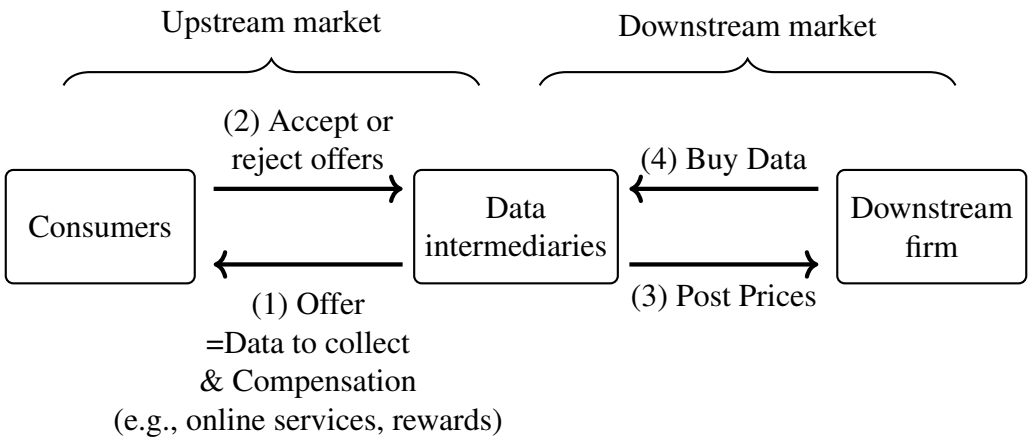

Figure 1: Timing of Moves

forms, such as Anderson and Coate (2005) and Reisinger (2012). There, the presence of advertisers imposes negative externalities on users due to nuisance costs.

3 Model

There are $N \in \mathbb{N}$ consumers, $K \in \mathbb{N}$ data intermediaries, and a single downstream firm.[2] 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 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 $\mathcal{D}_i$ is an indivisible and *non-rivalrous* good.[3] $\mathcal{D} := \cup_{i \in N} \mathcal{D}_i$ 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 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*.

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

[3] For example, $\mathcal{D}_i = \{i$'s age, i's email address, i's income$\}$.

After observing offers, each consumer i decides which offers to accept. Motivated by the non-rivalry of data, I impose no restriction on the number of offers consumers can accept. Formally, each consumer i simultaneously chooses $K_i \subset K$, where $k \in K_i$ means that consumer i provides data D_i^k to intermediary k and earns τ_i^k. These decisions determine intermediary k's data $D^k = \cup_{i \in N^k} D_i^k$, where $N^k := \{i \in N : k \in K_i\}$ is the set of consumers who accept the offers from intermediary k. 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

All intermediaries and the firm publicly 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 uses data to extract rents from consumers. I normalize $U_i(\emptyset) = 0$ and impose more structures later. Note that U_i does not depend on what data the downstream firm has on other consumers $j \neq i$. However, 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. $\Pi(D)$ is increasing in D: For any $X, Y \subset \mathcal{D}$ such that $X \subset Y$, $\Pi(Y) \geq \Pi(X)$.

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

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

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

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

Submodularity is motivated by the idea that data typically exhibit decreasing returns to scale (Varian, 2018). The main insight continues to hold only with Point 1 (increasing Π), as Section 7 shows.

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. The decision of each consumer determines the allocation of data. Then, each intermediary simultaneously posts a price to the firm. Finally, the firm chooses the set of intermediaries from which it buys data. The solution concept is pure-strategy subgame perfect equilibrium in which each consumer's data-sharing decision depends only on offers to her and not on offers to other consumers. This restriction ensures that equilibria I consider are not sensitive to each consumer's information about other consumers' offers.

3.1 Discussion of Assumptions

I comment on two important assumptions of the model and elaborate on the role they play in the analysis that follows.

Observability of the allocation of data

It is crucial for my analysis that the allocation of data is publicly observable. I assume this for two reasons. First, in practice, data intermediaries seem to 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.[4] Also, if an intermediary collects data directly from consumers, then it needs to reveal what data it collects (e.g., Nielsen Homescan). Although there could be a verifiability problem and intermediaries' strategic incentives to over- or understate what data they have, it would be a reasonable starting point to assume that each intermediary observes what kind of data other intermediaries collect.

Second, theoretically, 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 to the firm at a positive price. Then, intermediary k can profitably deviate by collecting no data and charges the same price to the firm. This argument implies that if each intermediary privately observes its data, then the market may break down and intermediaries may obtain zero profits in equilibrium.[5]

Finally, another possible specification would be that the data held by each intermediary is observable to the firm but not to other intermediaries. However, such a setting is intractable because there is no pure-strategy equilibrium.[6]

Timing

I assume that it is after observing the allocation of data that intermediaries set prices. There are two motivations for this assumption. First, what data an intermediary collects (i.e., offer) is a part of platform design or a company's policy. In contrast, after collecting data, intermediaries will have many opportunities to share the data. Then, it is reasonable to assume that intermediaries can adjust prices of data in the downstream market more quickly than adjusting what data they collect.

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

[5] This is the case if consumers incur loss from the firm's use of data as in Section 6.

[6] 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.

Second, If intermediaries collect data and set prices at the same time, then as I argue above, there may be no pure-strategy equilibrium.

A consequence of these assumptions is that intermediaries obtain different sets of data to avoid price competition in the downstream market.

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 the following situation: Online platforms such as Google and Facebook provide services to consumers in exchange for their data. D_i^k represents the set of data that consumers need to provide to use platform k, and τ_i^k represents the quality of k's service. (For example, a web mapping service such as Google maps would correspond to an offer such that D_i^k contains i's location data but not, say, i's political preference.) Platforms may share data with third parties, such as advertisers, retailers, and political consulting firms. Data sharing with each third party can benefit (e.g., better targeting) or hurt (e.g., price discrimination and privacy concern) a data subject. The aggregate impact of these effects is summarized by $U_i(D_i)$.

Several remarks are in order. First, $U_i(\cdot)$ is assumed to be exogenous, that is, intermediaries cannot directly 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, the model formulates compensation as one-to-one transfer. This is mainly to simplify the analysis; the results continue to 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, I assume that the benefit for consumer i of sharing data with intermediary k depends only on τ_i^k. However, if we interpret intermediaries as online platforms, we may think that the

benefit should increase if other consumers provide more data (e.g., social media). I deliberately exclude such a situation to clarify that the results are not driven by network externalities or returns to scale.

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

Data Brokers

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

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

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

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. We may think that competition encourages app developers to "pay" more for consumer data in the form of lower app prices and higher qualities. I will show that this may not be the case because paying more for consumer data makes data worth less in the downstream market.

4 Preliminary Analysis

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 the right hand side of (2) has multiple maximizers, I pick one of them arbitrarily 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.[7] Such a model corresponds to the market for physical goods.[8] 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.*

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, no intermediary can earn a positive profit in equilibrium.

5 Equilibrium Analysis: Downstream Market

Hereafter, I consider the main specification: Multiple intermediaries buy and sell non-rivalrous data. First, I show that the equilibrium revenue of each intermediary in the downstream market is unique and equal to the marginal contribution of its data to the firm's revenue. The result relies on the submodularity of the firm's revenue function Π.[9]

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

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

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

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

[9] 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.

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

Proof. I show that there is an equilibrium (of the downstream market) in which each intermediary k posts a price of Π^k and the firm buys all data. First, the submodularity of Π implies that $\Pi(\cup_{k\in K'\cup\{j\}} D^j) - \Pi(\cup_{k\in K'} D^j) \geq \Pi^j$ for all $K' \subset K$. Thus, if each intermediary k sets a price of Π^k, the firm prefers to buy all data. Second, if intermediary k 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. The uniqueness of the equilibrium revenue is relegated to Appendix B. □

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

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

6 Equilibrium with Costly Data Sharing

Given Lemma 1, I describe equilibria of the entire game. To illustrate the main idea in a simply way, this section assumes that $U_i(D_i)$ is decreasing in D_i, that is, each consumer obtains a lower payoff if she shares a greater set of data with the firm.

6.1 Single Unit Data

I begin with the simplest case in which each consumer i has single unit data and her payoff decreases by 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 actively 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 states that although data are non-rivalrous, no two intermediaries obtain the same piece of data on the equilibrium path (see Appendix C for the proof).

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

Intuitively, if two intermediaries obtained the same data, then one of them could increase a profit by not collecting the data: The deviation does not change its revenue in the downstream market (Lemma 1) but reduces compensation to consumers.

Proposition 1 resembles product differentiation.[10] As products in this model are consumer data, intermediaries' incentives for product differentiation affect consumer surplus. The following result illustrates this point: It presents equilibria in which consumer surplus and total surplus are equal to the monopoly outcome. Recall that D^M is the set of data that a monopoly intermediary acquires (see Appendix D for the proof).

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

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

[10] While the author is not aware of empirical evidence in support of the equilibrium prediction, the IT research and advisory company Gartner states "data brokers tend to specialize in certain industries in order to gain a competitive advantage." (https://www.gartner.com/smarterwithgartner/how-to-choose-a-data-broker/)

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

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

The theorem states that any partition of D^M can arise as an equilibrium allocation of data. Thus, different intermediaries collect different (possibly empty) data, and the aggregate data collection leads to the monopoly set of data transferred. Importantly, consumer surplus is zero (monopoly level) across all of these equilibria. Thus, the equilibria 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, and consider an intermediary's incentive to deviate. For example, suppose that intermediary 1 deviates and offers positive compensation to consumers for data D^2, which intermediary 2 is going to acquire. Then, these consumers will provide their data to not only intermediary 1 but 2. 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.[11] However, if consumers share data with intermediaries 1 and 2, these intermediaries have to set a price of zero for D^2 in the downstream market. Anticipating this, intermediary 1 prefers to not compensate for any data in D^2. Since each intermediary faces no competing offers, it can acquire data at the monopsony price C_i. Also, intermediaries have no incentive to acquire data in $\mathcal{D} \setminus D^M$ because consumers ask for greater compensation than the price of their data in the downstream market.

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

Theorem 1 implies that there is a monopoly equilibrium. Thus, the presence of multiple intermediaries may not dissipate their profits:

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.*

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

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

The results have several implications. First, *data portability* under the European General Data Protection Regulations might lower consumer welfare. Data portability states that data controller, such as online platforms, must allow consumers to withdraw and transfer their data across competitors. Let us interpret the models of non-rivalrous and rivalrous data as the economy with and without data portability, respectively. Then, the comparison of Theorem 2 with Claim 2 implies that data portability could relax ex ante competition for data and transfer surplus from consumers to intermediaries.[12]

Second, market concentration might not be a right measure of market power that data intermediaries hold against consumers. Theorem 1 implies that many intermediaries obtaining small pieces of data are consistent with consumers obtaining zero surplus. Indeed, such a situation transfers surplus from intermediaries to downstream firms, not to consumers. This is because, for any allocation of data, each intermediary acts as a monopsony of each piece of data.

Third, Theorem 2 gives a rationale to the frequently used assumption in the literature that the market consists of a monopolistic data seller.[13] We can view a model of a monopoly data seller as a subgame of the extended game in which multiple data sellers first acquire information at cost and then sell collected data.

Remark 1. The existence of the monopoly equilibrium (Theorem 2) is robust to a variety of extensions. For example, consumers could incur exogenous costs of sharing data with intermediaries (e.g., privacy concern); U_i could depend on what data other consumers share with the firm (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 2. Are there equilibria other than those in Theorem 1? The answer is yes: The following example shows that both consumer and total surplus can be different from the monopoly outcome.

[12] 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.

[13] 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.

Example 1. 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 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, which are Pareto undominated. Since I aim to understand the non-competitive nature of the market for data, it would be reasonable to focus on the latter.[14] Pareto undominated equilibria also let me focus on the issue of how the surplus created by data is divided, because all Pareto undominated equilibria have the same total surplus.

6.2 Multidimensional Data

I now generalize Theorems 1 and 2 by relaxing assumptions on the sets of data and consumer preferences. I assume that each consumer i has any finite set $\mathcal{D}_i$ of data and that consumers incur increasing convex costs of sharing data.

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

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

[14] Indeed, Theorem 1 characterizes all equilibria that are Pareto undominated in terms of the intermediaries' profits. The same statement applies to Theorem 3.

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

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

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^k and consumers accept all non-empty offers. In contrast, the current setting may not have such an equilibrium.[15] To avoid this difficulty, I focus on an environment where Theorem 1 naturally extends.

Assumption 4. A monopoly intermediary obtains and sells all data, i.e., $D^M = \mathcal{D}$.[16]

There are two settings in which Assumption 4 naturally holds. One is when the downstream firm is a seller who uses data to learn about consumers' values. If the firm can perfectly price discriminate consumers with 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 the primitives, Assumption 4 holds if the firm's marginal revenue from data is high relative to consumers' marginal costs of sharing the data.[17] Under Assumption 4, Theorem 1 extends (see Appendix E for the proof).

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

1. *The equilibrium allocation of data is $(D^1, \ldots, D^K)$.*
2. *In the upstream market, intermediary k acquires consumer i's data D_i^k at compensation $\hat{\tau}_i^k$, which is the marginal cost of sharing D_i^k:*

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

[15] Formally, this general setting may not have an equilibrium that reduces to the one in Theorem 1 when I assume that each consumer has single unit data.

[16] 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.

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

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

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

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. Point 2 of Theorem 3 states that intermediary k compensates consumer i according to the additional loss that consumer i incurs 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. 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). Theorem 3 implies that there are at least two equilibria: In one, intermediaries 1 and 2 acquire location and financial data, respectively, and each intermediary pays each consumer 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. In the other equilibrium, intermediary 1 acquires both location and financial data and pays \$20, leaving zero surplus to consumers. Thus, consumer surplus is lower in the latter equilibrium. The following subsection generalizes this observation.

6.3 Welfare Impacts of 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 F for the proof).

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

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

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

The intuition is as follows. As in Lemma 1, the price of data D^k is $\Pi(\cup_{j\in K} D^j) - \Pi(\cup_{j\in K\setminus\{k\}} D^j)$, the additional revenue the firm can earn from D^k conditional on having other data. If there are many intermediaries each of which has a small part of D^M, then the contribution of each piece of data is close to the marginal revenue $\Pi(D^M) - \Pi(D^M \setminus \{d\})$. In contrast, if a few intermediaries hold a large fraction of D^M, the contribution of each data set is large. Thus, each intermediary can set 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 compensate consumers for their data according to the infra-marginal cost.

6.4 Intensive and Extensive Margins of Data Concentration

The allocation of data can be more concentrated at the intensive or extensive margin. To see this, consider the following example. There are two intermediaries. The left block of Figure 2 depicts a situation in which intermediary 1 obtains location data on all consumers in the US and EU, and intermediary 2 obtains financial data on all consumers in the US and EU. The right block of Figure 2 is an alternative allocation where intermediary 1 holds location and financial data on consumers in the US, and intermediary 2 obtains location and financial data on consumers in the EU. I say that the allocation of data in the right block is more concentrated at the *intensive margin*, because each intermediary knows more about consumers in its dataset.

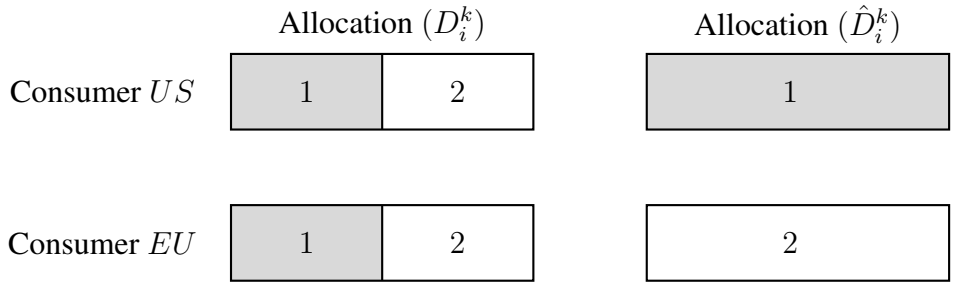

Figure 2: Data concentration at the intensive margin

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

Figure 3: Data concentration at the extensive margin

The following definition generalizes these examples.

Definition 3. Let $(D^k)_k$ and $(\hat{D}^k)_k$ denote two partitional allocations of data with $\cup_k D^k = \cup_k \hat{D}^k$.

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

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

Proposition 2. *Consider equilibria described in Theorem 3.*

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

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

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

7 Equilibrium with General Preferences

This section substantively relaxes the assumptions on preferences. First, I allow consumers to have any U_i and the firm to have any increasing Π. I show that there is a *partially monopolistic equilibrium*, which generalizes the monopoly equilibrium in Theorem 2. Second, if U_i is any (possibly non-monotone) submodular function, then any partition of D^M can arise as an equilibrium allocation of data, which generalizes Theorem 3. Third, I use the results to study information design by competing intermediaries.

7.1 Partially Monopolistic Equilibrium

I continue to assume that there are multiple intermediaries and maintain Assumption 4. The following result generalizes Theorem 1 (see Appendix G for the proof).

Proposition 3 (Partially Monopolistic Equilibrium (PME)). *Suppose that U_i is any set function for each i, and Π is any increasing set function. There is an equilibrium in which a single intermediary obtains all data and pays a compensation of $\max_{D\subset\mathcal{D}_i} U_i(D) - U_i(\mathcal{D}_i)$ to each consumer i, who obtains an equilibrium payoff of $\max_{D\subset\mathcal{D}_i} U_i(D)$.*

If $U_i(D_i)$ is decreasing for each i, then the partially monopolistic equilibrium reduces to the monopoly equilibrium, in which a single intermediary extracts full surplus from consumers and the firm. In contrast, suppose that consumer i prefers to share some data with the firm for free, i.e., $\max_{D\subset\mathcal{D}_i} U_i(D) > U_i(\emptyset) = 0$. Proposition 3 implies that consumer surplus is then greater than in the monopoly market (Claim 1) but lower than in the market with rivalrous goods (Claim 2). Thus, in general, *competition among data intermediaries benefits consumers relative to monopoly, however, the benefit is typically lower than in traditional markets with rivalrous goods.*

To see why competition benefits consumers when $U_i^* := \max_{D\subset\mathcal{D}_i} U_i(D) > 0$, suppose that 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. Thus, if intermediaries offer non-negative fees, consumer i can credibly share her data with at most one intermediary. Then, Bertrand competition lowers the equilibrium fees down 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 3 states that the above intuition applies to arbitrary preferences. Figure 4 depicts U_i as a function of the amount of data on i that the firm acquires. First, the monopoly intermediary obtains all data at a compensation of $-U_i(\mathcal{D}_i)$ (short dotted arrow). We can interpret the monopoly compensation $-U_i(\mathcal{D}_i)$ as the sum of two terms: 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 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 $\mathcal{D}_i \setminus D_i^*$.

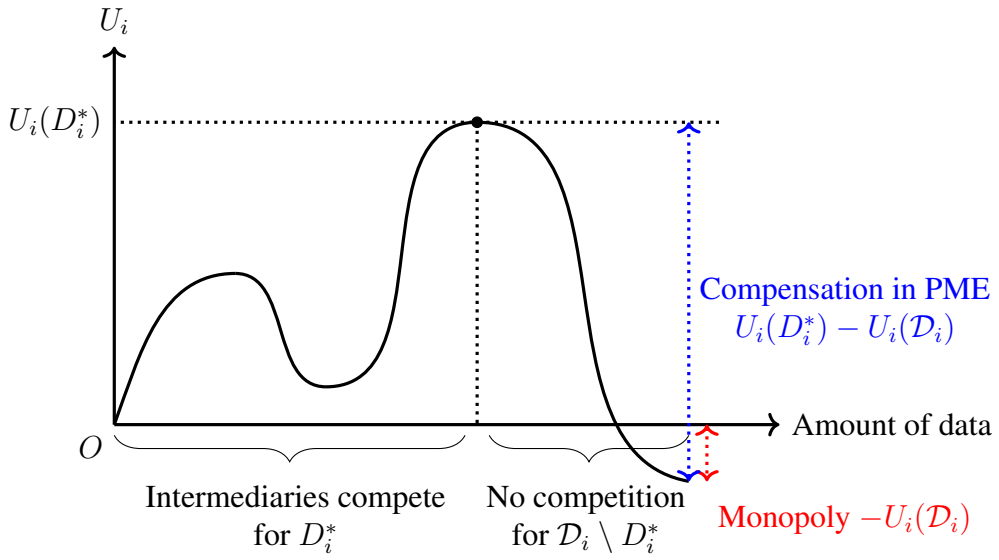

Figure 4: Partially monopolistic equilibrium

The next result shows that, if the market consists of many intermediaries, then the partially monopolistic equilibrium minimizes consumer surplus and maximizes the joint profits of all intermediaries. This observation justifies the claim that the partially monopolistic equilibrium is a natural extension of the monopoly equilibrium (see Appendix H for the proof):

Proposition 4. *Suppose that Π is strictly increasing. For a sufficiently large number K of intermediaries, the partially monopolistic equilibrium in Proposition 3 minimizes consumer surplus and maximizes intermediary surplus across all equilibria.*

The intuition is as follows. Suppose that in equilibrium, consumer i obtains a payoff of $U_i(D_i^*) - \delta$ with $\delta > 0$. Then, any intermediary can exclusively obtain a (potentially subset of) data D_i^* by offering (D_i^*, ε) with $\varepsilon < \delta$. The possibility of such a deviation guarantees that each intermediary earns a positive payoff in equilibrium. I prove that the payoff from the deviation is bounded from below by a positive value that is independent of the number K of intermediaries. This is a contradiction because as K grows large, the payoff of some intermediary has to go to zero.

We turn our focus to Theorem 3, which shows that any partition of $D^M (= \mathcal{D})$ can arise as an equilibrium allocation of data. The following result generalizes the result for any submodular U_i (see Appendix I for the proof).

Proposition 5. *Suppose that each U_i is submodular and there are $K = 2J$ intermediaries. Take any partitional allocation of data $(D^1, \ldots, D^J)$ for J intermediaries with $\cup_{k=1}^{J} D^k = \mathcal{D}$. There is an equilibrium in which intermediary $k \leq J$ obtains D^k and pays a compensation of*

$$L_i(D_i^k) := \max \left\{ \max_{D \subset \mathcal{D}_i} U_i(D) - U_i(D \cup D_i^k), 0 \right\}. \tag{7}$$

to each consumer i.

In this equilibrium, intermediary k compensates consumer i for D_i^k according to the maximum loss that she could incur by sharing D_i^k, where the maximum is taken across all possible data other than D_i^k that she could share with the firm. If U_i is decreasing, then submodularity implies $L_i(D_i^k) = U_i(\mathcal{D} \setminus D_i^k) - U_i(\mathcal{D})$, which is equal to the equilibrium compensation in Theorem 3. If U_i is non-monotone, then $L_i(D_i^k)$ could be different from the actual loss that i incurs by sharing D_i^k. If $D^k = \mathcal{D}$ for a unique k and $D^j = \emptyset$ for any other $j \neq k$, then the equilibrium coincides with the partially monopolistic equilibrium.

Finally, the submodularity of U_i implies that data concentration hurts consumers and benefits intermediaries, which generalizes Proposition 4.

Proposition 6. *Consider equilibria in Proposition 5. An equilibrium with a greater data concentration (i.e., coarser partition) is associated with lower consumer surplus and greater intermediary surplus.*

Proof. Take any disjoint sets $X, Y \subset \mathcal{D}_i$. Let $D^* \in \arg\max_{D \subset \mathcal{D}_i} U_i(D) - U_i(D \cup X \cup Y)$. The submodularity of U_i implies that

$$\begin{aligned}
&U_i(D^* \cup X \cup Y) + U_i(D^*) \leq U_i(D^* \cup X) + U_i(D^* \cup Y) \\
\iff &U_i(D^*) - U_i(D^* \cup X \cup Y) \leq U_i(D^* \cup X) - U_i(D^* \cup X \cup Y) + U_i(D^* \cup Y) - U_i(D^* \cup X \cup Y) \\
\iff &U_i(D^*) - U_i(D^* \cup X \cup Y) \leq L_i(Y) + L_i(X) \\
\iff &L_i(X \cup Y) \leq L_i(X) + L_i(Y).
\end{aligned}$$

The last inequality and the submodularity of Π imply that the total compensation is lower and the total revenue of intermediaries is greater in an equilibrium corresponding to a coarser partition. As all equilibria in Proposition 5 have the same total surplus, we obtain the desired result. □

7.2 Information Design by Competing Data Intermediaries

I use Proposition 3 to study information design by competing data intermediaries. I assume that a downstream firm is a seller that uses consumer data for product recommendation and price discrimination. Each piece of data $d \in \mathcal{D}_i$ is a signal (Blackwell experiment) about consumer i's willingness to pay. Intermediaries can potentially obtain any signals from consumers.

The formal description is as follows. Assume for simplicity that there is a single consumer (thus, omit subscript i). A downstream 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, $u := (u_1, \ldots, u_M)$, are independently and identically distributed according to a cumulative distribution function F with a finite support $V \subset (0, +\infty)$.[18]

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 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.[19]

After buying data $D \subset \mathcal{D}$ from intermediaries, the seller learns about values 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.[20] A recommendation could be an advertiser displaying a targeted advertisement or an online retailer showing a product as a personalized recommendation. The consumer's payoff from this transaction is $u_m - p$ if she buys product m at price 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

[18] 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.

[19] To close the model, I need to specify how realizations of different signals are correlated conditional on u. One way is to use the formulation of Gentzkow and Kamenica (2017): Let X be a random variable that is independent of 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 $(v, x) \in s$. However, the result does not rely on this particular formulation.

[20] 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 (2019) for a detailed discussion of the motivation behind this formulation.

pure-strategy perfect Bayesian equilibrium such that the seller calculates its posterior belief based on the prior F and signals in D on and off the equilibrium paths.[21]

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

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 signal. Define

$$p(F) := \min(\arg\max_{p \in V} p[1 - F(p)])$$

as 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 $U(\emptyset)$, the 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 J for the proof).

Proposition 7. *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 sells 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).*

[21] I assume that the seller breaks ties in favor of the consumer. The existence of an equilibrium is shown in Ichihashi (2019).

2. *Suppose that the seller sells 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.*[22]

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 profits by obtaining and selling additional information. Thus, in an equilibrium where d^* is obtained and sold, no intermediaries can make a positive profit.[23] In this case, competition among intermediaries yields the consumer all welfare gain from her information.

Second, consider Point 2. Ichihashi (2019) shows that if the prior F satisfies the condition in Point 2, then any signal d^* that maximizes the consumer's payoff $U(d)$ leads to inefficiency. Intuitively, d^* garbles the 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 an inaccurate product recommendation. 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.

Finally, we can apply a similar analysis to any model of Bayesian persuasion in which full disclosure is efficient. For example, Haghpanah and Siegel (2019) consider a consumer-optimal market segmentation in the context of multi-dimensional screening. Embedding their model into my model of competing intermediaries leads to a similar conclusion as Point 2 of Proposition 7.

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

[23] This equilibrium is equivalent to the partially monopolistic equilibrium in terms of the expected payoff of each player.

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, Assumption 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.

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

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

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

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

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*.

Haghpanah, Nima and Ron Siegel (2019), "Consumer-optimal market segmentation." *Available at SSRN 3333940*.

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

Ichihashi, Shota (2019), "Online privacy and information disclosure by consumers." *Available at SSRN 3112905*.

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.

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

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

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

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

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

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

Appendix

A Proof of Claim 2

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

a positive payoff. Suppose that each intermediary k obtains data D_i^k from consumer $i \in N^k$ at compensation τ_i^k. Define $D^* := \cup_k D^k$. Suppose that intermediary 2 deviates and offers each consumer $i \in N^1$ an offer of $(D_i^1 \cup D_i^2, \tau_i^1 + \tau_i^2 + \varepsilon)$. Then, all consumers in N^1 accept the offer of intermediary 2 but not 1. 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 the equilibrium revenue of each intermediary k is at most Π^k. Suppose to the contrary that (without loss of generality) intermediary 1 obtains a strictly greater revenue than Π^1. Let $K' \ni 1$ denote the set of intermediaries from which the firm buys data.

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

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

It holds that

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

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

Let J^* denote a solution of the right hand side of (8). 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

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} \hat{p}^k < \Pi(\cup_{k\in J^*} D^k) - \sum_{k\in J^*} \hat{p}^k, \tag{9}$$

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 (9) 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 (8) 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

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

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

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

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

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D^k) = \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D^k). \tag{12}$$

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 (10) 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 (11) holds. Such intermediary j can again profitably deviate by lowering its price, which is a contradiction. Therefore, intermediary k's revenue is at most Π^k.

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

C Proof of Proposition 1

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

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

E Proof of Theorem 3

Proof. Suppose that each intermediary k offers $(D_i^k, \hat{\tau}_i^k)$ to each consumer i and sets a price of data following Lemma 1. I show that this strategy profile is an equilibrium. First, Lemma 1 implies that there is no profitable deviation in the downstream market. Second, suppose that intermediary k deviates and offers $(\tilde{D}_i^k, \tilde{\tau}_i^k)$ to each consumer i. Without loss of generality, we can assume that $\tilde{D}_i^k \subset D_i^k$. The reason is as follows. If consumer i rejects $(\tilde{D}_i^k, \tilde{\tau}_i^k)$, intermediary k replace it with $(\tilde{D}_i^k, \tilde{\tau}_i^k) = (\emptyset, 0)$. If consumer i accepts $(\tilde{D}_i^k, \tilde{\tau}_i^k)$ but $\tilde{D}_i^k \subsetneq D_i^k$, it means that intermediary k obtains some data $d \in \tilde{D}_i^k \setminus D_i^k$. Because $\cup_k D^k = D^M = \mathcal{D}$, there is another intermediary

that obtains data d. By Corollary 1, intermediary k is indifferent between offering $(\tilde{D}_i^k \setminus \{d\}, \tilde{\tau}_i^k)$ and offering $(\tilde{D}_i^k, \tilde{\tau}_i^k)$. Let $D^- := D^k \setminus \tilde{D}_i^k$ denote the set of data that are not acquired by the firm as a result of intermediary k's deviation. If intermediary k deviates in this way, its revenue in the downstream market decreases by $\Pi(D^M) - \Pi(D^M \setminus D^k) - [\Pi(D^M \setminus D^-) - \Pi(D^M \setminus D^k)] = \Pi(D^M) - \Pi(D^M - D^-)$. In the upstream market, if consumer i provides data $\tilde{D}_i^k$ to intermediary k, then it is optimal for consumer i to accept other offers from non-deviating intermediaries, because C_i is supermodular. This implies that the minimum compensation that intermediary k has to pay is $C_i(\mathcal{D}_i \setminus D_i^-) - C_i(\mathcal{D}_i \setminus D_i^k)$. Thus, intermediary k's compensation to consumer i in the upstream market decreases by $C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^k) - [C_i(\mathcal{D}_i \setminus D_i^-) - C_i(\mathcal{D}_i \setminus D_i^k)] = C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^-)$. Thus, k's total compensation decreases by $\sum_{i \in N} \left[C_i(\mathcal{D}_i) - C_i(\mathcal{D}_i \setminus D_i^-)\right]$. Because $D^M = \mathcal{D}$ is 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. □

F Proof of Theorem 4

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

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

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

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

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

G Proof of Proposition 3

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 accept an offer of 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 inter-

mediary 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. □

H Proof of Proposition 4

Proof. In the equilibrium of Proposition 3, each consumer obtains a payoff of $U_i(D_i^*)$, intermediary surplus is $\Pi(\mathcal{D}) - \sum_{i \in N} U_i(D_i^*)$, and the firm's payoff is zero. Thus, it suffices to show that in any equilibrium, each consumer i obtains a payoff of at least $U_i(D_i^*)$. Define $m := \min_{d \in \mathcal{D}, D \subset \mathcal{D}} \Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus \{d\}) > 0$ and $TS^* := \Pi(\mathcal{D}) + \sum_{i \in N} U_i(\mathcal{D}) > 0$ where $U_i(D) := U_i(D \cap \mathcal{D}_i)$. 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 $U_i(D_i^*)$. □

I Proof of Proposition 5

Proof. $L_i(D_i^k)$ in (7) is the maximum loss that consumer i incurs by sharing D_i^k with the firm, where the maximum is taken across all sets of data she shares through other intermediaries. Define $M_i(D_i^k)$ as follows. If $L_i(D_i^k) = 0$, let $M_i(D_i^k) = \mathcal{D}_i$, i.e., the set of all data on consumer i. If $L_i(D_i^k) > 0$, let $M_i(D_i^k) \in \arg\max_{D \subset \mathcal{D}_i} U_i(D) - U_i(D \cup D_i^k)$. If there are multiple maximizers, choose any maximizer D such that there is no other maximizer $\tilde{D}$ with $D \subsetneq \tilde{D}$. Such a D exists because $\mathcal{D}$ is finite. Note that $U_i(M_i(D_i^k)) - U_i(M_i(D_i^k) \cup D_i^k) = L_i(D_i^k)$. An important observation is that $M_i(D_i^k) \supset D_i^{-1}$ where $D_i^{-1} = \cup_{k=2}^J D_i^k$ is the set of data consumer i shares with intermediaries other than 1. If $L_i(D_i^k) = 0$, $M_i(D_i^k) \supset D_i^{-1}$ holds because $M_i(D_i^k) = \mathcal{D}_i$. This is also true if $L_i(D_i^k) > 0$. To see this, suppose to the contrary that $M_i(D_i^k) \supsetneq D_i^{-1}$. The submodularity of U_i implies that $U_i(M_i(D_i^k) \cup D') - U_i(M_i(D_i^k) \cup D' \cup D_i^k) \geq L_i(D_i^k)$ where $D' := D_i^{-1} \setminus M_i(D_i^k) \neq \emptyset$. This contradicts the construction of $M_i(D_i^k)$.

Consider the following strategy profile. Each intermediary $k \leq J$ offers $(D_i^k, L_i(D_i^k))$ to each consumer i. Each intermediary $k + \ell$ with $\ell \in \{1, \ldots, J\}$ offers $(M_i(D_i^k), 0)$ to consumer i. In the downstream market, the behavior of intermediaries and the firm follows Lemma 1. On the path of play, consumers accept offers from intermediaries $1, \ldots, K$ and reject others. Off the path of play, consumers optimally choose the set of offers to accept, breaking ties in favor of accepting an offer.

I show that the above strategy profile consists of an equilibrium. First, each consumer i prefers to accept offer $(D_i^k, L_i(D_i^K))$ on and off the equilibrium paths by the construction of $L_i(D_i^k)$. On the path of play, given that consumer i shares data with intermediaries $1, \ldots, J$, she does not strictly

benefit from accepting the offer of intermediary $k > J$. Thus, each consumer has no incentive to deviate.

Any intermediary $j > J$ has no incentive to deviate. To see this, note that consumers continue to accept offers from intermediaries $1, \dots, J$ after j's deviation. Since $\cup_{k=1}^{J} D^k = \mathcal{D}$, j's deviation does not strictly increase its payoff in the downstream market. Also, j cannot strictly increase its payoff in the upstream market because consumers, who share data $\mathcal{D}$ to intermediaries $1, \dots, K$, have no incentive to pay to intermediary j.

I show that intermediary $k \leq J$ has no incentive to deviate. Suppose that an intermediary (say 1) deviates and offers $(\hat{D}_i^1, \hat{\tau}_i^1)$ to each consumer i. Without loss of generality, assume that all consumers accept offers from intermediary 1. Note that consumers continue to accept offers from intermediaries $2, \dots, J$. Thus, we can assume that $\hat{D}_i^1 \subset \mathcal{D} \setminus D_i^{-1}$. Note that intermediary 1's revenue in the downstream market is $\Pi(D^{-1} \cup (\cup_{i \in N} \hat{D}_i^1)) - \Pi(D^{-1})$. The compensation $\hat{\tau}_i^1$ to consumer i satisfies $U(D_i^{-1} \cup \hat{D}_i^1) + \hat{\tau}_i^1 \geq U(M_i(D_i^k))$. Indeed, if the right hand is strictly greater, then consumer i strictly prefers sharing data with intermediaries $2, \dots, J, J+1$ to sharing data with intermediaries $1, \dots, J$ (here, I use $M_i(D_i^k) \supset \hat{D}_i^1$). Thus, intermediary 1's payoff from the deviation is at most $\Pi(D^{-1} \cup (\cup_{i \in N} \hat{D}_i^1)) - \Pi(D^{-1}) + \sum_{i \in N} \left[U(D_i^{-1} \cup \hat{D}_i^1) - U(M_i(D_i^k)) \right]$. The sum of the first and third terms, $\Pi(D^{-1} \cup (\cup_{i \in N} \hat{D}_i^1)) + \sum_{i \in N} U(D_i^{-1} \cup \hat{D}_i^1)$, is maximized at $\hat{D}_i^1 = \mathcal{D}_i$ by Assumption 4. Thus, the deviating payoff is bounded from above by

$$\Pi(\mathcal{D}) - \Pi(D^{-1}) + \sum_{i \in N} \left[U(\mathcal{D}_i) - U(M_i(D_i^k)) \right],$$

which is intermediary 1's payoff on the path of play. Thus, no intermediary has a profitable deviation. □

J Proof of Proposition 7

Proof. Note that Proposition 3 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 (2019) 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. □