Natural Monopoly for Data Intermediaries

Shota Ichihashi*

April 19, 2019

Abstract

I study competition among data intermediaries—data brokers and technology companies that collect consumer data and sell them to downstream firms. Under the assumption that firms use consumer data to extract rents, consumers have to be compensated for their personal information. I show that competition among intermediaries fails: If they offer high compensation, consumers would share their data with multiple intermediaries, which lowers the price of the data in the downstream market and hurts intermediaries. I show that this leads to multiple equilibria with different allocations of data among intermediaries. There is a monopoly equilibrium, and equilibrium with a more concentrated allocation of data benefits intermediaries and hurts consumers.

Keywords: information markets, intermediaries, personal data, privacy

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

1 Introduction

I study 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 downstream buyers such as retailers and advertisers. Technology companies, such as Google and Facebook, collect data from users and sell targeted advertising spaces. I regard these companies as data intermediaries and study how they compete for personal data.

Imagine the following example. Consumers can enjoy online services by sharing personal information with platforms (i.e., intermediaries). However, platforms might share the data with, say, retailers, advertisers, and consulting firms. Consumers could then incur a loss such as price discrimination, intrusive marketing and political campaign, and further data leakage. These costs for consumers of sharing personal data are driven by two fundamental characteristics of data: The same data can be simultaneously used for different purposes, and it is difficult to write a complete contract over who can obtain data and how to use them.

The key to my analysis is the price-setting behaviors of data intermediaries. On the one side, intermediaries set "prices" to buy data from consumers (which I call compensation). Prices capture the quality of online services or the amount of monetary transfer that consumers can enjoy in exchange for providing their data. On the other, intermediaries set prices to sell the data to downstream firms that can raise revenue from them. Prices in the downstream market affect the demand for data, which, in turn, affects consumers' behavior as suppliers of personal data.

The main question is whether competition among intermediaries dissipates profits. The question is important for understanding how the surplus generated by personal data is allocated among economic agents. Now, in traditional markets, the answer to this question is often yes because competition for inputs in the upstream market drives rents to zero—the idea reminiscent of Demsetz (1968). In the market for data, this may not be the case.

The model consists of consumers, firms, and data intermediaries. Intermediaries obtain data from consumers in the upstream market and sell the data to firms in the downstream market. Firms can raise revenue from the data but the use of data negatively affects consumers. Thus, intermediaries have to compensate consumers in order to obtain their data. The amount of compensation determines the allocation of data, which specifies the amount and kind of data that each intermediary holds. The allocation of data affects price competition among intermediaries in the downstream market.

The key idea of the paper is the following economic force that deters competition for data. Suppose that intermediaries offer high compensations to obtain more data from consumers. Then, consumers would share their data with multiple intermediaries, because data are non-rivalrous. This lowers the price of data in the downstream market, which, in turn, hurts intermediaries.

This economic force leads to multiple equilibria with different allocations of data among intermediaries. I characterize the set of equilibria and analyze its properties. There are two main findings. First, there exists a monopoly equilibrium, where a single intermediary acquires data at the minimum level of compensation and extracts full surplus from firms. Other intermediaries have no incentive to compensate consumers for their data, because consumers will then share their data with multiple intermediaries.

Second, there are mulitple equilibria with different allocations of data. I show that an equilibrium with a higher degree of data concentration leads to higher revenue of intermediaries. This is because a "large" intermediary can charge the firm for the data based on its average value instead of marginal value.

I also explore different margins of data concentration. Data concentration at the "intensive" margin refers to the situation where each consumer provides her data to a small number of intermediaries. Data concentration at the "extensive" margin refers to the situation where a small number of intermediaries source data from a large number of consumers. In particular, I provide a condition under which data concentration at the intensive margin leads to a lower consumer surplus, whereas concentration at the extensive margin may not. Thus, different margins of data concentration could have different welfare implications.

The paper helps us understand two issues of the digital economy. One is the recent discussion on why consumers do not seem to be compensated for providing their data (Arrieta-Ibarra et al., 2018; Carrascal et al., 2013). I illustrate an economic mechanism under which "competition for data" fails to reward consumers as suppliers of personal data. The other is data concentration in the hands of major internet platforms. I show that data concentration can arise as equilibrium even in the absence of network externalities or returns to scale.

Two modeling assumptions are crucial for these results. One is that data are non-rivalrous, and the other is that firms' use of data negatively affects consumers. If I alter either of these assumptions, the presence of multiple intermediaries dissipates profits and strictly benefits consumers. I will later argue that comparing the model of non-rivalrous data with that of rivalrous data sheds light on the potential welfare implications of "data portability" in the EU's General Data Protection Regulation.

The rest of the paper is organized as follows. Section 2 discusses related works and Section 3 describes the model. Section 4 characterizes equilibria. There, I show that there is a monopoly equilibrium and that data concentration benefits intermediaries. Section 5 allows consumers to have multiple pieces of data and conducts richer analysis of data concentration. Section 6 provides extensions, and Section 7 concludes.

2 Literature Review

This paper relates to two strands of literature. First, it relates to the recent literature on markets for data. Bergemann and Bonatti (2019) consider a model of a monopoly data intermediary and study under what conditions intermediation of data is profitable. They assume that a downstream firm uses data for price discrimination that hurts consumers. I assume that the intermediation is profitable, and focus on the issues of competition and data concentration. Also, rather than microfound how a downstream firms use consumer data, I abstract away from these details and use a simpler payoff structure.

More broadly, this paper relates to works on markets for data beyond the context of price discrimination. Bergemann et al. (2018) consider a model of data provision and data pricing. Choi et al. (2018) consider consumers' privacy choices in the presence of an information externality. Gu et al. (2018) study the incentives of data brokers to merge data, taking the allocation of data as exogenous. Kim (2018) considers a model of monopoly advertising platform and studies consumers' privacy concerns, market competition, and vertical integration between the platform and sellers.

Second, the paper relates to the literature on platform competition in two-sided markets. Relative to this literature, the novelty of this paper is that I consider the combination of negative cross-side externality and multi-homing (which is captured by the non-rivalry of data). The literature typically assumes that the interaction of the two sides is mutual beneficial (e.g., Armstrong

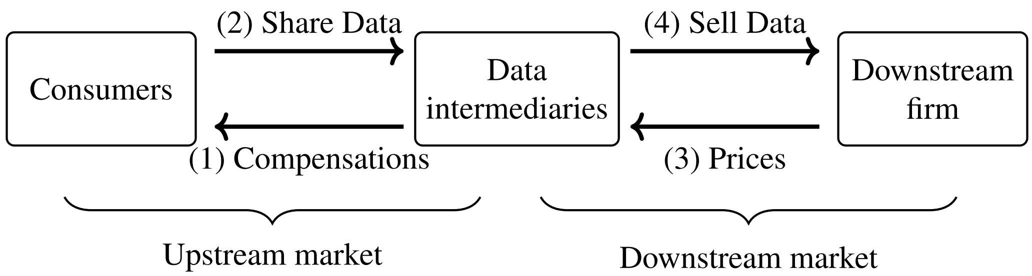

Figure 1: Timing of Moves

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

3 Model

There are a continuum of consumers uniformly distributed on $[0, 1]$, $K \in \mathbb{N}$ data intermediaries, and $L \in \mathbb{N}$ downstream firms. For now, I assume $L = 1$, relegating the general case to Section 6. Where it does not cause confusion, I use K to mean the set of all intermediaries. As depicted in Figure 1, the game consists of two parts—intermediaries buy data in the upstream market and sell them in the downstream market.

Upstream Market

Each consumer $i \in [0, 1]$ has some personal data, which I treat as indivisible and non-rivalrous goods. At the beginning of the game, each intermediary $k \in K$ simultaneously makes an *offer*

$(\tau_i^k)_{i\in[0,1]}$, where $\tau_i^k \in \mathbb{R}$ is measurable in $i \in [0,1]$. τ_i^k is the compensation that intermediary k pays consumer i in exchange for her data. Compensation could capture not only monetary transfer but also the costly provision of online services. I assume that each consumer i does not observe offers made to other consumers, i.e., $(\tau_m^k)_{k\in K, m\in[0,1]\setminus\{i\}}$. This assumption is not crucial but simplifies the analysis by restricting possible coordination among consumers.

After observing offers $(\tau_i^1, \ldots, \tau_i^K)$, each consumer i simultaneously chooses $K_i \subset K$, the set of intermediaries with which i shares her data. Hereafter, I say that consumer i *shares her data* whenever $K_i \neq \emptyset$. Each consumer's decision determines $D_k := \{i \in [0,1] : k \in K_i\}$, the set of consumers who share their data with intermediary k. I call D_k intermediary k's *data*. All intermediaries and the firm observe $(D_1, \ldots, D_K)$, which I call the *allocation of data*.

Downstream Market

Given the allocation of data $(D_1, \ldots, D_K)$, intermediaries sell their data to the firm. Namely, each intermediary k simultaneously posts a price $p_k \in \mathbb{R}$ as a take-it-or-leave-it offer. Then, the firm chooses the set $K' \subset K$ of intermediaries from which it buys data. The firm's decision determines $D := \cup_{k\in K'} D_k \subset [0,1]$, the set of all consumers whose data are acquired by the firm. Note that the firm obtains consumer i's data if and only if $K_i \cap K' \neq \emptyset$, i.e., the firm buys data from at least one intermediary with which consumer i has shared her data in the upstream market.

Preferences

All players maximize expected payoffs, and their ex post payoffs are as follows. The payoff of each intermediary is revenue minus compensation: If intermediary k obtains data D_k and the firm buys data from a set K' of intermediaries, then k obtains a payoff of $\mathbf{1}_{\{k\in K'\}} p_k - \int_{D_k} \tau_i^k di$, where $\mathbf{1}_{\{k\in K'\}}$ is equal to 1 or 0 if $k \in K'$ or $k \notin K'$, respectively.

The payoff of each consumer is as follows. Suppose that consumer i shares her data with intermediaries in K_i. If the firm obtains i's data, then her payoff is $\sum_{k\in K_i} \tau_i^k - c$. Otherwise, her payoff is $\sum_{k\in K_i} \tau_i^k$. Here, $c > 0$ is a consumer's disutility of sharing her data with the firm. I abstract from the origins of c. It should be thought of as a reduced form capturing a consumer's loss from price discrimination or intrusive targeted advertising.

The payoff of the downstream firm is as follows. Suppose that the firm obtains data from the set K' of intermediaries. Let $D = \cup_{k\in K'} D_k \subset [0,1]$ denote the resulting data. Then, the firm's payoff is $\Pi(|D|) - \sum_{k \in K'} p_k$, where $|D|$ is the Lebesgue measure of D. The first term, which I call *revenue*, depends only on the size of data.[1] The second term is the total expense to buy data from intermediaries in K'. I assume that Π satisfies the following.

Assumption 1. $\Pi : [0,1] \to \mathbb{R}$ is increasing, strictly concave, continuously differentiable, and $\Pi(0) = 0$. Moreover, there is $m^* \in (0,1]$ such that $\Pi'(m^*) = c$.

The assumption is stronger than necessary but greatly simplifies exposition.[2] The concavity of Π is motivated by the idea that "data typically exhibits decreasing returns to scale" (Varian, 2018). Concavity also simplifies the derivation of the equilibrium prices of data in the downstream market.

Timing

The timing of the game, depicted in Figure 1, is as follows. First, each intermediary simultaneously makes an offer to each consumer. After privately observing the offers made to her, 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.

Solution Concept

Throughout the paper, I consider strategy profiles such that for each k, the set D_k of consumers who share their data with intermediary k is measurable both on the path of play and after any unilateral deviation.

Among these strategy profiles, I consider pure-strategy perfect Bayesian equilibrium (PBE) such that the set of all consumers who share their data with at least one intermediary is written as $[0,m]$ for some $m \in [0,1]$.[3] From now on, I abbreviate PBE satisfying these conditions as "equilibrium."

[1] This excludes the case where the firm values the data of some consumers more highly than the data of other consumers. I can incorporate such a situation by assuming that $\Pi(\cdot)$ is a submodular set function.

[2] The main results hold if Π is increasing and concave.

[3] Any pure-strategy equilibrium is "equivalent" to an equilibrium where the set of consumers who share their data is written as $[0,m]$. The equivalence is up to the indices of consumers sharing their data and the outcomes associated with a measure-zero set of consumers.

3.1 Two Interpretations of Data Intermediaries

First, we can interpret intermediaries as "data brokers" such as LiveRamp (formerly known as Acxiom), Nielsen, and Oracle. The assumption that intermediaries compensate consumers seems to contradict the observation that data brokers usually do not interact with consumers. However, I argue that 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 increase 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).

Second, we may view data intermediaries as online platforms such as Google and Facebook. The model abstracts from many of their institutional details. For example, these technology companies distribute personal data indirectly through sponsored search or targeted display advertising. Nonetheless, by regarding these platforms as pure data intermediaries, we can shed light on a novel economic mechanism potentially relevant to their competition.

4 Equilibrium Analysis

4.1 Monopoly Data Intermediary

As a benchmark, consider a monopoly data intermediary ($K = 1$). Suppose that the intermediary tries to obtain and sell the data of consumers in $[0, m]$. In the upstream market, each consumer in $[0, m]$ requires a compensation of at least c, anticipating that her data will be sold to the firm. Thus, the intermediary has to pay a total compensation of mc. In the downstream market, the intermediary can extract full surplus from the firm by posting a price of $\Pi(m)$. In equilibrium, the intermediary chooses how much data to acquire by balancing the marginal revenue $\Pi'(m)$ and the marginal cost c. The proof of the following result is relegated to Appendix A.

Proposition 1. *Let m^* satisfy $\Pi'(m^*) = c$. In any equilibrium, the monopoly intermediary obtains data from consumers in $[0, m^*]$ at compensation c. Almost all consumers and the firm obtain zero payoffs.*

4.2 Multiple Data Intermediaries

Now, I assume that the market consists of multiple intermediaries ($K \geq 2$). I solve the game backwards. First, given the allocation $(D_1, \ldots, D_K)$ of data, I derive the equilibrium prices of data in the downstream market. Second, given the equilibrium pricing, I derive the equilibrium behavior in the upstream market. Hereafter, for any measurable set $D' \subset [0,1]$, I write $\Pi(D')$ to mean $\Pi(|D'|)$ where $|D'|$ is a measure of D'.

The following lemma shows that the equilibrium price of data is equal to their marginal contribution to the firm's revenue. The result depends on the concavity of $\Pi(\cdot)$.

Lemma 1. *Suppose that each intermediary k holds data D_k. In the downstream market, there is a subgame perfect equilibrium in which intermediary k posts a price of*

$$\Pi_k := \Pi\left(\bigcup_{j \in K} D_j\right) - \Pi\left(\bigcup_{j \in K \setminus \{k\}} D_j\right) \tag{1}$$

for its data D_k, and the firm purchases data from all intermediaries.

Lemma 1 implies that two intermediaries obtain zero revenue in the downstream market if they hold the same data, which 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. Hereafter, "equilibrium" refers to equilibrium of the entire game in which the prices of data are given by (1).[4]

Corollary 1. *Suppose that each intermediary $j \neq k$ holds data D_j. Take any $D_k \subset [0,1]$ and $D' \subset \cup_{j \neq k} D_j$. Then, the equilibrium revenue of intermediary k in the downstream market is identical between when it holds D_k and $D_k \cup D'$.*

Given the pricing rule in Lemma 1, how does equilibrium compensation and the allocations of data look like? The following lemma presents key properties of the equilibrium allocation and compensation.

Lemma 2. *In any equilibrium, for almost all $i \in [0,1]$, the following holds:*

[4] Gu et al. (2018) prove that (1) is a unique equilibrium price of data if $K = 2$.

1. *If consumer i shares her data, there is a unique intermediary (say k_i) that offers a positive compensation to i.*

2. *On the equilibrium path, consumer i shares her data only with intermediary k_i.*

Point 1 of Lemma 2 indicates the lack of competition among data intermediaries in the upstream market. Intuitively, conditional on a consumer sharing her data with one intermediary, the consumer prefers to share the data with any other intermediaries that offer positive compensation. This is because additional data sharing does not change her loss from the firm's use of data, whereas it increases the total compensation that a consumer can earn.[5] This in turn implies that any consumer who shares her data earns a positive compensation from only one intermediary: If multiple intermediaries offer positive compensation, then the consumer will share her data with all of them, following which the price of her data is zero.

Point 2 excludes the case where some consumers share their data with intermediaries that offer zero compensation. Thus, although data are non-rivalrous, on the equilibrium path, no two intermediaries obtain the same piece of data. This observation motivates the following notion, which simplifies the exposition of the next result.

Definition 1. An allocation of data $(D_1, \ldots, D_K)$ is *disjoint* if $D_k \cap D_j$ has measure zero for all $k, j \in K$ with $k \neq j$.

The following result characterizes all equilibria consistent with Lemma 1. The proof is in Appendix D.

Theorem 1. *Let $m^* \in (0,1)$ be the unique value satisfying $\Pi'(m^*) = c$. The following two conditions are equivalent.*

1. *There is an equilibrium such that the allocation of data is $(D_k)_k$ and each consumer $i \in [0,1]$ earns compensation τ_i.*

[5]This argument might give an impression that the result hinges on the assumption that consumers do not incur any disutility from sharing data with intermediaries. However, Section 6 shows that all the results continue to hold even if each consumer incurs the cost proportional to the number of intermediaries with which the consumer shares her data.

2. *The allocation $(D_k)_k$ is disjoint and $\cup_{k \in K} D_k = [0, m]$ with $m \leq m^*$. Moreover, for almost all $i \in [0,1]$, it holds that*

$$\tau_i \in [c, \Pi'(m)] \quad \textit{if} \quad i \leq m;$$
$$\tau_i = 0 \quad \textit{if} \quad i > m.$$

The theorem implies that there are multiple equilibria: Any partition of a subset of $[0, m^*]$ can arise as an equilibrium allocation of data, and the equilibrium compensation can be anywhere between disutility c and marginal revenue $\Pi'(m)$. Plugging $m = m^*$ and $\tau_i = c$ for $i \in [0, m]$ in Theorem 1, I obtain the following result.

Corollary 2. *Take any disjoint allocation $(D_k)_k$ with $|\cup_{k \in K} D_k| = m^*$. Then, there is an equilibrium where the allocation of data is given by $(D_k)_k$ and consumer surplus is at the monopoly level (zero). Therefore, there is an equilibrium in which a single intermediary acts as a monopolist.*

The intuition is as follows. First, if intermediary (say) 1 acts as a monopolist, then other intermediaries have no incentives to compete with it. Indeed, suppose that intermediary $j \neq 1$ tries to obtain data $[0, m^*]$ by offering a higher compensation. Then, these consumers will share their data with not only intermediary j but 1. However, this lowers the price of data in the downstream market and hurts intermediary j. Also, if intermediary j tries to obtain data from consumers in $(m^*, 1]$, then they demand strictly greater compensation than the price of their data. Therefore, any intermediary $j \neq 1$ has no incentive to deviate and obtain data.

Moreover, Corollary 2 states that any partition of $[0, m^*]$ can arise in an equilibrium with the monopoly level of consumer surplus. The intuition is similar to the above. Given any disjoint allocation of data (D_k), intermediary j has no incentive to obtain data D_ℓ with $\ell \neq j$, because consumers in D_ℓ will then share their data with both intermediaries j and ℓ. Therefore, each intermediary k becomes a monopsony of personal data of consumers in D_k.

Theorem 1 also implies that there are equilibria where consumers obtain positive payoffs. For example, there is an equilibrium where intermediary (say) 1 obtains data $[0, m]$ with $m < m^*$, pays a compensation of $\tau_i = \Pi'(m) > c$ to each consumer $i \in [0, m]$, and charges a price of $\Pi(m)$ to the firm. In this equilibrium, intermediary 1 has no incentive to deviate and *lower* compensation, because consumers will then share their data with other intermediaries, who offer zero compensation. However, I will later show that this kind of equilibrium disappears once we slightly perturb the model and assume that intermediaries incur small costs to interact with consumers.

The multiplicity of equilibria described in Theorem 1 motivates me to compare equilibria that differ in the degree of *data concentration*. To do so, I introduce the following notion.

Definition 2. Let $(D_k)_k$ and $(D'_k)_k$ denote two disjoint allocations of data. Then, we say that $(D'_k)_k$ is *more concentrated* than $(D_k)_k$ if $\cup_k D_k = \cup_k D'_k$, and for any k, there is ℓ such that $D_k \subset D'_\ell$.

Roughly, one allocation is more concentrated than the other if the former is coarser than the latter as a partition. The following result states that an equilibrium with a more concentrated allocation of data leads to higher industry revenue of intermediaries. See Appendix E for the proof.

Proposition 2. *Let $(D_k)_k$ and $(D'_k)_k$ denote two equilibrium allocations of data. If $(D'_k)_k$ is more concentrated than $(D_k)_k$, then in these equilibria, the total revenue of intermediaries in the downstream market is higher under $(D'_k)_k$ than under $(D_k)_k$.*

The intuition is simple. As in Lemma 1, the price of data held by each intermediary k is equal to the additional revenue that the firm can earn from the data, conditional on having data held by other intermediaries. Now, fix the total amount of data $[0, m]$. Suppose that there are many intermediaries, each of which has a small part of $[0, m]$. Then, the contribution of each piece of data is small and close to the marginal revenue $\Pi'(m)$. On the other hand, if a small number of intermediaries hold $[0, m]$, then they can set a price based on the infra-marginal value of the data. Since $\Pi(\cdot)$ is concave, the latter leads to a greater total revenue for intermediaries.

In the current model, there is no systematic relationship between consumer surplus and data concentration. In Section 5, I enrich the model and study such a relationship.

4.3 "Robustness" of Monopoly Consumer Surplus

Theorem 1 poses a question of which equilibria are likely to arise. This subsection addresses the question in a somewhat ad hoc way: I show that if it is costly for intermediaries to make offers to consumers, then any equilibrium has the monopoly level of consumer surplus, no matter how small the cost.

I modify the action space and preferences of each intermediary as follows. First, each intermediary can choose to *not* make an offer, which is now distinguished from offering non-positive compensation. Second, there is $\gamma > 0$ such that, if an intermediary makes an offer to mass m of consumers, it incurs a cost of $m\gamma$, which is subtracted from the original profit.

The following result states that small transaction costs eliminate equilibria in which either the mass of consumers sharing their data or compensation profile is different from monopoly. The proof is in Appendix G.

Proposition 3. *There is $\varepsilon > 0$ such that for any $\gamma \in (0, \varepsilon)$, in any equilibrium, consumer surplus is zero. The mass $m(\gamma)$ of consumers who share their data satisfies $\Pi'(m(\gamma)) = c+\gamma$, and converges to m^* as $\gamma \to 0$.*

Proposition 3 does *not* mean that any equilibrium given $\gamma > 0$ is the monopoly equilibrium where a single intermediary acquires $[0, m(\gamma)]$. As before, there are a continuum of equilibria that differ in how data $[0, m(\gamma)]$ is allocated across intermediaries. Thus, as $\gamma \to 0$, the set of all equilibria "converges" to the one described in Corollary 2.

The intuition of Proposition 3 is as follows. Consider an equilibrium where intermediary 1 offers compensation $\tau_i > c$ to a positive mass of consumers. In this equilibrium, each intermediary offers zero compensation to all consumers from whom it does not acquire their data. Consumers reject these zero compensation on the equilibrium path, but they accept the offers if intermediary 1 deviates an lowers compensations. (In the proof, I show that this is the only way to sustain strictly greater compensations than c in equilibrium.) In the presence of the transaction costs, zero compensation, which are rejected on the equilibrium path, cannot consist of an equilibrium.

4.4 Discussion of Modeling Assumptions

I discuss two modeling assumptions. First, I consider what happens if data are *rivalrous*. Second, I consider the case where the downstream firm's use of data *benefits* consumers.

Rivalrous Data

What if data are rivalrous? For example, an online platform might offer an "exclusive dealing" that prevents consumers from sharing their data with other platforms. In this case, it would be reasonable to treat data as rivalrous goods despite their physical property.

To see how the rivalry of data changes the result, suppose now that each consumer can share her data with *at most one* intermediary. The proof of the following result is in Appendix F, which also proves the existence of a consumer-optimal equilibrium in this alternative setting.

Claim 1. *Suppose that each consumer can share her data with at most one intermediary. If there are multiple intermediaries ($K \geq 2$), in any equilibrium, all intermediaries earn zero profits. Moreover, there are equilibria where consumers extract full surplus.*

If data were rivalrous, competition dissipates intermediaries' profits and strictly benefits consumers. Namely, if an intermediary earns a positive profit from data $[0, m]$, then another intermediary can offer a slightly higher compensation to exclusively obtain the data from consumers in $[0, m]$. Thus, all intermediaries obtain zero profits in equilibrium. This contrasts with the baseline model where intermediaries can secure a monopoly profit.

While it would need more work to derive a concrete policy implication, the result has the following takeaway. The General Data Protection Regulation aims to ensure *data portability*, under which consumers have "the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided." Now, let us interpret the previous model and the current one as markets with and without data portability, respectively. Then, the comparison of Claim 1 and Corollary 2 illustrates that data portability could relax ex ante competition for data and lower consumer surplus.

Beneficial Use of Data

So far, I have assumed that the firm's use of personal data negatively affects consumers. In reality, consumers may also benefit from sharing their data with a firm. For example, the firm may use data to offer personalized services and products or display more relevant ads. To capture such a situation, assume $c < 0$. That is, each consumer gains $b := -c > 0$ if the firm acquires her data.

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

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

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

The intuition is as follows. For simplicity, suppose that intermediaries are restricted to offering nonnegative fees. The key observation is that if multiple intermediaries offer positive fees to consumers, then each consumer accepts at most one offer. This is because consumers can obtain a benefit of $b > 0$ as long as their data are transferred to the firm through one intermediary. Now, since consumers share their data with at most one intermediary, each intermediary tries to undercut the fees of other intermediaries in order to obtain data. This lowers the equilibrium fees down to zero.[6] The result contrasts with the baseline specification where the firm's use of data hurts consumers. There, competition among intermediaries is about raising compensation, which does not work because consumers will then share data with multiple intermediaries.

5 Concentration of Data and Consumer Welfare

I enrich the model to allow each consumer to have multiple pieces of personal data. In this case, each intermediary has to decide which pieces of data to obtain from each consumer. I will show that, under a certain condition, an equilibrium with a more concentrated allocation of data leads to a lower consumer surplus.

[6] Proposition 4 suggests that there are multiple equilbria if $K \geq 2$. Indeed, by the same logic as Theorem 1, I can show that any compensation between 0 and $\Pi'(1)$ can be sustained in equilibrium. However, as in Proposition 3, only the equilibrium with zero fees remains if it is costly for intermediaries to make offers.

The formal description is as follows. For ease of exposition, assume now that there are $N \in \mathbb{N}$ consumers, $1, \ldots, N$. Abusing notation slightly, I use N to mean the set of all consumers. Each consumer $i \in N$ has the *set* of data $Q_i := \{d_i^q\}_{q \in [0,1]}$. Note that each piece of data is indexed by $q \in [0, 1]$. Where it does not cause confusion, I identify Q_i with $[0, 1]$, except when I need to clarify whose data I am referring to. Thus, "consumer i shares data $[0, q']$" means "consumer i shares data $\{d_i^q\}_{q \in [0,q']}$."

In the upstream market, each intermediary k makes an offer to each consumer i. An offer is now a pair (Q_i^k, τ_i^k). Here, $Q_i^k \subset Q_i$ is the set of data that intermediary k asks for from consumer i, and τ_i^k is the amount of compensation that k is willing to pay. As before, each consumer i simultaneously decides whether to accept each offer, where "accept" means that consumer i provides Q_i^k to intermediary k and earns τ_i^k. In the downstream market, each intermediary k simultaneously posts a price for its data. Finally, the firm decides from which intermediaries to buy data. The allocation of data $(Q_i^k)_{i,k}$ now specifies what data each intermediary holds for each consumer.

The payoffs of each player are as follows. First, the payoff of each intermediary continues to be revenue from the firm minus total compensation.

Next, suppose that the firm acquires data $Q_i' \subset Q_i$ from consumer i. Then, consumer i's payoff is the total compensation from intermediaries minus disutility of sharing data $c(q_i)$, where $q_i := |Q_i'|$ is the measure of Q_i' when I regard it as a subset of $[0, 1]$.

Assumption 2. $c(q)$ is differentiable, increasing, and strictly convex.

The payoff of the downstream firm is revenue minus total payments to intermediaries. If the firm acquires a set Q_i of data from each consumer i, then its revenue is $\Pi(q)$ where $q := \sum_{i=1}^{N} |Q_i|$.

Assumption 3. $\Pi(q)$ is differentiable, increasing, and concave. Moreover, there is a unique $q^* \in (0, 1]$ such that $\Pi'(Nq^*) = c'(q^*)$.

As before, I consider pure-strategy perfect Bayesian equilibrium in which each consumer i shares data $[0, q_i]$ for some $q_i \in [0, 1]$. Also, I restrict my attention to strategy profiles where all relevant sets of data are measurable.

5.1 Monopoly Data Intermediary

The following result, which parallels Proposition 1, describes the monopoly equilibrium. As before, the monopoly intermediary decides how much data to acquire from each consumer, balancing marginal revenue $\Pi'(q)$ and marginal cost $c'(q_i)$.

Proposition 5. *Let q^* satisfy $\Pi'(Nq^*) = c'(q^*)$. In any equilibrium, the monopoly intermediary obtains data $[0, q^*]$ from each consumer at compensation $c(q^*)$. All consumers and the firm obtain zero payoffs.*

Proof. Suppose that the monopoly intermediary obtains data $[0, q_i]$ from each consumer $i \in N$. Since the intermediary can extract full surplus from all consumers and the firm, its payoff is $R(q_1, \ldots, q_N) := \Pi(\sum_i q_i) - \sum_i c(q_i)$. $R(q_1, \ldots, q_N)$ is strictly concave because its Hessian is negative definite. Thus, the first order condition is sufficient for a unique global maximum. By construction, $(q^*, \ldots, q^*)$ satisfies the FOC. $\square$

5.2 Multiple Data Intermediaries

Suppose now that there are multiple intermediaries. As in the baseline setting, I could characterize equilibria that differ in allocations of data and compensations. However, I focus on a particular class of equilibria to study data concentration. Let $Q^M := \cup_{i=1}^N \{d_i^q\}_{q \in [0,q^*]}$ denote the set of all data that consumers share under the monopoly (Proposition 5). Also, given any allocation of data $(Q_i^k)_{i,k}$, let $Q^k := \cup_i Q_i^k$ denote the data that intermediary k holds. As before, $(Q_i^k)_{i,k}$ is said to be disjoint if $|Q^k \cap Q^j| = 0$ for any distinct k and j. The proof of the following result is in Appendix I.

Proposition 6. *Take any disjoint allocation of data $(Q_i^k)_{i,k}$ such that $\cup_{i,k} Q_i^k = Q^M$. Then, there is an equilibrium with the following properties:*

1. *The allocation of data is $(Q_i^k)_{i,k}$.*

2. *In the upstream market, each intermediary k offers a compensation of*

$$\tau_i^k := c\left(Q^M\right) - c\left(Q^M - Q_i^k\right) \tag{2}$$

to consumer i for data Q_i^k, and all consumers accept all offers.

3. *In the downstream market, each intermediary k posts a price of*

$$\Pi_k := \Pi\left(Q^M\right) - \Pi\left(Q^M - Q^k\right) \tag{3}$$

for its data Q^k, and the firm purchases data from all intermediaries.

Proposition 6 shows that any partition of Q^M—personal data shared under monopoly—can arise as an equilibrium allocation of data. In such an equilibrium, each intermediary sets a price equal to the contribution of its data to the firm's revenue. This is a straightforward extension of Lemma 1. In the upstream market, each intermediary k compensates each consumer i according to the "marginal" disutility that k's data acquisition imposes on i.

Focusing on equilibria described in Proposition 6, I show that a certain notion of data concentration leads to a lower consumer surplus.

Definition 3. Take two disjoint allocations of data (Q_i^k) and $(\hat{Q}_i^k)$. $(\hat{Q}_i^k)$ is *individually more concentrated* than (Q_i^k) if (i) $\cup_{i,k} Q_i^k = \cup_{i,k} \hat{Q}_i^k$, (ii) for each $i \in N$ and any $k \in K$, there is $\ell \in K$ such that $Q_i^k \subset \hat{Q}_i^\ell$, (iii) for some $(i, k, \ell) \in N \times K^2$, $Q_i^k \subsetneq \hat{Q}_i^\ell$.

Definition 3 captures the data concentration at the "intensive margin." For example, consider Figure 2. Under allocation (Q_i^k), both consumers 1 and 2 provide most (but not all) of their data to intermediary 1 (white). Under allocation $(\hat{Q}_i^k)$, consumers 1 and 2 provide all of their data to intermediaries 1 (white) and 2 (gray), respectively. While (Q_i^k) might look more concentrated, $(\hat{Q}_i^k)$ is individually more concentrated, because any given consumer provides data to a smaller number of intermediaries under $(\hat{Q}_i^k)$. In other words, under $(\hat{Q}_i^k)$, each intermediary knows a lot about a consumer, conditional on that it knows something about the consumer.

Definition 3 contrasts with data concentration at the "extensive margin." In Figure 3, data are (intuitively) more concentrated toward intermediary 1 under $(\hat{Q}_i^k)$, because a relatively small number of intermediaries own data of a larger number of consumers. However, it does not satisfy Definition 3, and consumers are indifferent between $(\hat{Q}_i^k)$ and (Q_i^k). In this case, although a smaller number of intermediaries have data of a larger number of consumers, the amount of data that an intermediary has per consumer has not changed.

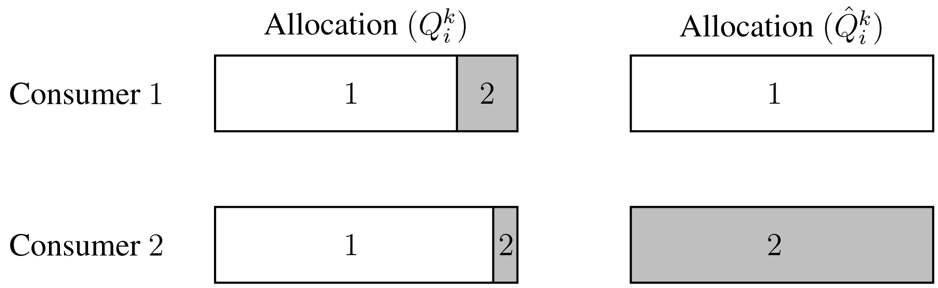

Figure 2: Data concentration at the intensive margin

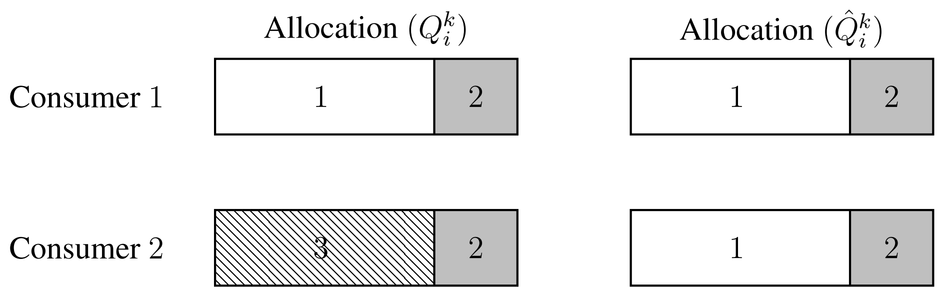

Figure 3: Data concentration at the extensive margin

Proposition 7. *Consider the class of equilibria described in Proposition 6. An equilibrium with individually more concentrated allocation of data is associated with lower consumer surplus.*

The intuition is similar to why data concentration leads to a greater revenue of intermediaries in Proposition 2. Consider an equilibrium with a more individually concentrated allocation of data. That is, each consumer shares a large amount of data with a small number of intermediaries. Point 2 of Proposition 6 implies that these intermediaries can compensate consumers according to the infra-marginal disutility of sharing data. As $c(\cdot)$ is convex, this leads to lower compensation compared to the case where each consumer shares small and different pieces of data with many intermediaries.

Beyond the specific welfare implication, one takeaway of Proposition 7 is that the "intensive" and "extensive" margins of data concentration can have different impacts on consumer welfare. Here, the intensive margin refers to how much an intermediary knows about a consumer, and the extensive margin refers to how many consumers' data an intermediary has. We could imagine data concentration in either direction, but the analysis of this section provides a sufficient condition under which the data concentration at the different margins affects consumer surplus differently:

Namely, concentration at the intensive margin leads to a lower consumer surplus, whereas concentration at the extensive margin does not.

6 Extensions

This section maintains the formulation of Section 3, but a similar extension applies to the one in Section 5.

6.1 Privacy Concern toward Data Intermediaries

Consumers may incur disutility of sharing data not only with downstream firms but also with data intermediaries. I can incorporate this by assuming that each consumer incurs a loss of ρK_i by sharing her data with K_i intermediaries, where $\rho > 0$.

In this case, intermediaries obtain a lower amount of data in equilibrium than before, because it has to pay a compensation of at least $c + \rho$. However, this does not change the main result. As Lemma 2, no two intermediaries offer positive compensations to the same set of consumers. Thus, there are again a continuum of equilibria, one of which is a monopoly equilibrium.

6.2 Vertically Differentiated Intermediaries

Suppose that intermediary k incurs a cost of $c_k \tau_i^k$ to pay a compensation of τ_i^k. The baseline model assumes $c_k = 1$ for all k, but c_k could be different across k. This extension is reasonable if compensation represents the costly provision of online services, and intermediaries incur different costs to provide the same quality of services. In this case, the main insight continues to hold: There is an equilibrium where the most efficient intermediary acts as a monopolist.

Proposition 8. *For any K, there is an equilibrium in which intermediary $k^* := \arg\min_{k \in K} c_k$ earns as high a profit as when the market consists only of itself.*

The intuition is the same as before: A less efficient intermediary has no incentive to compete, because either consumers share their data with multiple intermediaries, or the intermediary has to incur a greater cost of compensation than the price of the data.

6.3 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 the amount of data available to ℓ. Each consumer i's payoff (without transfer) is $\sum_{\ell \in L} u_\ell$, where $u_\ell = -c_\ell$ and $u_\ell = 0$ if firm ℓ does and does not obtain i's data,respectively.

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 whether to accept an offer by comparing $-\sum_\ell c_\ell$ with the compensation. Therefore, by setting $c := \sum_{\ell \in L} c_\ell$, we can apply the same analysis as before. Note that this extension can accommodate the case where some firms impose loss ($c_\ell > 0$) and some impose benefit ($c_\ell < 0$) on consumers, because Section 3 only requires that $\sum_{\ell \in L} c_\ell > 0$.

7 Conclusion

This paper provides a model of competing data intermediaries, which obtain data from consumers and sell them to downstream firms. The model incorporates two key features of personal data. One is that data are non-rivalrous, and the other is that firms' use of data could negatively affect consumers. These features drastically change the nature of competition, relative to the intermediation of physical goods. Namely, data intermediaries may secure monopoly profits in some equilibrium, and more importantly, the equilibrium allocation of personal data across intermediaries is not unique. This in turn enables me to compare equilibria with different degrees of data concentration. I show that, under a certain condition, data concentration benefits intermediaries but hurts consumers.

References

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

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

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

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

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

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

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

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

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

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

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

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.

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 Proposition 1

Proof. Consider the following strategy profile: The intermediary offers c to consumer $i \leq m^*$, and offers zero to consumer $i > m^*$. Also, it sets a price of $\Pi(m)$ at a node in which the intermediary has the data of size m. First, no consumer $i \leq m^*$ strictly benefits from unilaterally rejecting the offer. Also, no consumer $i > m^*$ benefits from accepting the offer of zero compensation.

Suppose that the intermediary deviates and buys data from the mass m of consumers. If $m > m^*$, the total compensation the intermediary has to pay increases by at least $\Delta c = \Delta \Pi'(m^*)$ where $\Delta = m - m^*$; in contrast, the revenue of the intermediary in the downstream market increases by $\Pi(m^* + \Delta) - \Pi(m^*)$. By the definition of m^*, such a deviation cannot be strictly profitable. If $m < m^*$, the total compensation the intermediary has to pay decreases by at most $\Delta c = \Delta \Pi'(m^*)$ where $\Delta = m^* - m$; in contrast, the revenue of the intermediary in the downstream market decreases by $\Pi(m^*) - \Pi(m^* - \Delta)$. Again, such a deviation cannot be strictly profitable. Thus, the proposed strategy profile is an equilibrium.

By the same logic, we can show that if the intermediary buys data from $m' > m^*$ mass of consumers under some strategy profile, it can strictly increase its payoff by not purchasing data from consumers in $(m' - \Delta, m')$ (say, by offering a negative compensation) because $\Pi'(m') < c$ implies that $\Pi'(m') - \Pi'(m' - \Delta) < \Delta c$ for a small $\Delta > 0$. Similarly, if the intermediary buys data only from $m' < m^*$ consumers under some strategy profile, it can strictly increase revenue by purchasing data from consumers in $(m', m' + \Delta]$. Therefore, there is no equilibrium in which the intermediary buys data from mass $m' \neq m^*$ of consumers, which also uniquely pins down equilibrium compensations. $\square$

B Proof of Lemma 1

Proof. Take any allocation of data $(D_1, \ldots, D_K)$. Consider a strategy profile in which each intermediary $k \in K$ sets a price of Π_k in (1) and the firm buys data from all intermediaries. If an intermediary deviates and sets a higher price, then the firm buys data from all but the deviating intermediary. I show that this is an equilibrium (of the subgame). First, if no intermediary deviates, it is optimal for the firm to buy all data: Assumption 1 implies that $\Pi(\cup_{j\in K'\cup\{k\}} D_j) - \Pi(\cup_{j\in K'} D_j) - \Pi_k \geq 0$ for any $K' \subset K$. Thus, the firm is willing to buy D_k at price Π_k regardless of the prices posted by other intermediaries. Second, if intermediary k unilaterally deviates and sets a price of $p_k > \Pi_k$, then the firm prefers to buy data from intermediaries in $K \setminus \{k\}$. Thus, k cannot benefit by raising a price. Finally, any price $p_k < \Pi_k$ strictly lowers the payoff of intermediary k. $\square$

C Proof of Lemma 2

Proof. Take any equilibrium, and let $D \subset [0, 1]$ denote the set of all consumers who share their data with at least one intermediary. First, I show Point 1. Lemma 1 implies that the firm acquires the data of all consumers in D. For each $(k, j) \in K^2$ with $k \neq j$, let $\hat{D}_{kj}$ denote the set of all consumers in D to whom both intermediaries k and j offer positive compensations. Suppose to the contrary that there is $(k, j) \in K^2$ such that $|\hat{D}_{kj}| > 0$. Then, each consumer $i \in \hat{D}_{kj}$ accepts the offers from both k and j, because conditional on accepting one offer, accepting other positive compensations does not affect consumer i's loss from the firm's use of data, but it strictly increases her payoff from total compensations. However, intermediary k can then profitably deviate by offering consumers in $\hat{D}_{kj}$ zero compensation. This is because the deviation does not change the revenue in the downstream market (Corollary 1) but strictly decreases the total compensation k has to pay. This implies that $|\hat{D}| \leq \cup_{(k,j)\in K^2} |\hat{D}_{kj}| = 0$.

To show Point 2, take any intermediary k, and let $D_k \subset D$ denote the set of all consumers in D to whom intermediary k offers positive compensations. By Point 1, for (almost) all consumers in D_k, intermediary k is the only intermediary that offers positive compensations. Now, if a positive mass of consumers in D_k also share their data with other intermediaries that offer non-positive compensations, then intermediary k can profitably deviate by offering zero compensations to those consumers. This is because such a deviation strictly reduces the total compensation without affecting k's revenue in the downstream market. This is a contradiction. Thus, almost all consumers in D_k share their data exclusively with intermediary k. □

D Proof of Theorem 1

Proof. First, I prove that Point 2 implies Point 1. Take any allocation and compensations described in Point 2. Consider the following strategy profile. Intermediary k offers τ_i to consumer $i \in D_k$ and offers zero compensation to consumers in $D_{-k} := [0,1] \setminus D_k$. On the equilibrium path, consumers $i \in [0,m]$ accept the offer of only intermediary k with $i \in D_k$. The equilibrium prices in the downstream market are given by Lemma 1.

The off-path behaviors of consumers are as follows. Suppose that a consumer detects a deviation by any intermediary in the upstream market. 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. Note that this is different from the on-path behavior, where each consumer rejects the offer of zero compensation by intermediary j such that $i \notin D_j$, even though she is indifferent between rejecting and accepting it given that she accepts the offer of k with $i \in D_k$.

I show that this strategy profile is an equilibrium. First of all, it is optimal for (almost) every consumer to accept an offer because $\tau_i \geq c$. Second, suppose that intermediary k unilaterally deviates in the first stage and offers $\hat{\tau}_i^k$ to each consumer i. Consider a deviation such that $\hat{\tau}_i^k > 0$ for a positive mass of consumers $i \in D' \subset [0,1] \setminus D_k$. There are two cases to consider. First, if consumer $i \in D'$ rejects $\hat{\tau}_i^k$, intermediary k can obtain the same payoff by offering $\hat{\tau}_i^k = 0$. Second, if a positive mass of consumers accept $\hat{\tau}_i^k$, intermediary k has to pay a positive compensation. However, because these consumers accept the offers of other intermediaries as well (by construction of consumers' off-path behavior), the deviation does not increase k's revenue in the downstream market (Corollary 1). Thus, setting $\hat{\tau}_i^k = 0$ for all $i \in D'$ weakly increases intermediary k's payoff relative to the original deviation.

Thus, it is sufficient to consider deviations by intermediary k that only affect consumers $i \in D_k$, i.e., $\hat{\tau}_j^k = 0$ for any $j \notin D_k$. Consider any such deviation, and let $\hat{D} := \{i \in D_k : \hat{\tau}_i^k \neq \tau_i\}$. $\hat{D}$ is the set of consumers in D_k who receive different offers from τ_i, where τ_i is the on-path level of compensation to i. Let D^0 denote the set of consumers in $\hat{D}$ who accept no offer as a result of the deviation, and let D^1 denote the set of consumers in $\hat{D}$ who accept at least one offer. Because all intermediaries other than k offer zero compensations to consumers in $\hat{D}$, consumers in D^1 accept the offer of intermediary k; moreover, by the way I define the off-path behavior, these consumers also accept offers of other intermediaries. Note that $|\hat{D}| = |D^0| + |D^1|$, and the mass of consumers whose data are bought only by intermediary k (i.e., consumers who are not affected by the deviation of k) is $|\hat{D}| - |D^0| - |D^1|$. I show that the revenue of intermediary k in the downstream market decreases by more than the total compensation that k can save. First, the total compensation that k pays in the upstream market decreases by at most $\hat{D} \cdot \Pi'(m)$, because $\tau_i \leq \Pi'(m)$ on the equilibrium path. Next, consider the equilibrium price of data held by k. After the deviation, the firm's revenue from (aggregate) data becomes $\Pi(D^* - D^0)$, where $D^* := [0, m]$. Without intermediary k's data, it would be $\Pi(D^* - D^0 - (D_k - D^0 - D^1))$, because now only $D_k - D^0 - D^1$ is exclusive to k's data. By Lemma 1, the deviation decreases k's revenue in the downstream market by

$$
\begin{aligned}
&\underbrace{\Pi(D^*) - \Pi(D^* - D_k)}_{\text{revenue without deviation}} - \underbrace{\left[\Pi(D^* - D^0) - \Pi(D^* - D^0 - (D_k - D^0 - D^1)\right]}_{\text{revenue with deviation}} \\
=&\Pi(D^*) - \Pi(D^* - D^0) + \Pi(D^* - D_k + D^1) - \Pi(D^* - D_k) \\
\geq&\Pi(D^*) - \Pi(D^* - D^0) + \Pi(D^*) - \Pi(D^* - D^1) \\
\geq&(|D^0| + |D^1|) \cdot \Pi'(m) \\
=&|\hat{D}| \cdot \Pi'(m)
\end{aligned}
$$

Here, the first and the second inequalities are by the concavity of $\Pi(\cdot)$. Therefore, the deviation cannot be strictly profitable for k. We can also verify that a unilateral deviation by a consumer, an intermediary in the downstream market, and the firm cannot be profitable.

Next, I show that Point 1 implies Point 2. First, suppose to the contrary that there is an equilibrium in which consumers in $[0, m]$ with $m > m^*$ share their data. Lemma 2 implies that there is a measure zero set D such that for each $i \in [0, m] \setminus D$, exactly one intermediary pays a positive compensation to i. Without loss of generality, suppose that intermediary 1 buys data from consumers in $[0, \Delta]$ with $\Delta > 0$. Then, intermediary 1 can strictly increase its payoff by giving zero offers to consumers in $[0, \delta]$ with a small $\delta \in (0, \Delta)$. Indeed, this deviation increases its payoff by $\delta c - [\Pi(m) - \Pi(m - \delta)] > 0$. This establishes that there is no equilibrium in which the mass of consumers sharing their data is strictly more than m^*. With the equilibrium restriction, it holds that $\cup_k D_k = [0, m]$ for $m \leq m^*$. Also, Lemma 2 implies that the allocation of data is disjoint.

Next, I show that the compensation that each consumer i earns satisfies the inequalities (described in the theorem) for almost every $i \in [0, 1]$. Note that for almost every $i \in [0, m]$, Lemma 2 implies that only one intermediary, say k, offers i a positive compensation $\tau_i^k > 0$. Suppose that intermediary k offers $\tau_i^k > \Pi'(m)$ to a positive mass of consumers. Then, intermediary k can profitably deviate by offering the compensation to zero to a small but positive mass (say ε) of those consumers. Indeed, this increases the payoff of intermediary k by $\varepsilon \tau_i^k - [\Pi(m) - \Pi(m - \varepsilon)] > 0$. Also, if $\tau_i^k < c$, then consumer i would reject the offer. Thus, $\tau_i^k \in [c, \Pi'(m)]$ for almost every $i \in [0, m]$ in equilibrium. Almost every consumer in $(m, 1]$ earns no compensation because she reject any offer. $\square$

E Proof of Proposition 2

Proof. Let $N := [0, 1]$. For $X, Y \subset N$, $X - Y$ stands for $X \setminus Y$. Let $(D'_k)_{k \in K}$ and $(D_k)_{k \in K}$ denote two disjoint allocations of data such that the former is more concentrated than the latter. Without loss of generality, assume that $\cup_k D'_k = \cup_k D_k = [0, 1]$. Note that in general, for any $N_0 \subset N$ and a partition $(M_1, \ldots, M_K)$ of N_0, we have

$$
\begin{aligned}
& \Pi(N) - \Pi(N - N_0) \\
= & \Pi(N) - \Pi(N - M_1) + \Pi(N - M_1) - \Pi(N - M_1 - M_2) + \cdots \\
& + \Pi(N - M_1 - M_2 - \cdots - M_{K-1}) - \Pi(N - M_1 - M_2 - \cdots - M_K) \\
\geq & \sum_{k \in K} [\Pi(N) - \Pi(N - M_k)],
\end{aligned}
$$

where the last inequality follows from the concavity of $\Pi(\cdot)$. Now, for any $\ell \in K$, let $K(\ell) \subset K$ satisfy $D'_\ell = \sum_{k \in K(\ell)} D_k$. The above inequality implies

$$\Pi(N) - \Pi(N - D'_\ell) \geq \sum_{k \in K(\ell)} \left[\Pi(N) - \Pi(N - D_k)\right], \forall \ell \in K$$

$$\Rightarrow \sum_{\ell \in K} \left[\Pi(N) - \Pi(N - D'_\ell)\right] \geq \sum_{\ell \in K} \sum_{k \in K(\ell)} \left[\Pi(N) - \Pi(N - D_k)\right].$$

In the last inequality, the left and the right hand sides are the total revenue for intermediaries in the downstream market under (D'_k) and (D_k), respectively. □

F Proof of Claim 1

In this appendix, I prove Claim 1 and the existence of an equilibrium in which consumers extract full surplus.

Proof of Claim 1. 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. Let D_k denote the set of consumers from whom intermediary k buys data, and define $D^* := \cup_k D_k$. Suppose that intermediary 2 deviates and offers each consumer $i \in D_1$ a compensation of $\tau_i^1 + \varepsilon$, where τ_i^1 is the compensation by intermediary 1. Then, all consumers in D_1 accept the offer of *only* intermediary 2. In the downstream market, the revenue of intermediary 2 increases from $\Pi(D^*) - \Pi(D^* - D_2)$ to $\Pi(D^*) - \Pi(D^* - D_1 - D_2)$, which yields a net gain of $\Pi(D^* - D_2) - \Pi(D^* - D_1 - D_2)$. By Assumption 1, $\Pi(D^* - D_2) - \Pi(D^* - D_1 - D_2) \geq \Pi(D^*) - \Pi(D^* - D_1)$. As intermediary 1 obtains a positive payoff if intermediary 2 does not deviate, it holds that $\Pi(D^*) - \Pi(D^* - D_1) - \int_{D_1} \tau_i^1 di > 0$, which implies $\Pi(D^* - D_2) - \Pi(D^* - D_1 - D_2) - \int_{D_1} (\tau_i^1 + \varepsilon) di > 0$ for a small $\varepsilon > 0$. Thus, intermediary 2 has a profitable deviation, which is a contradiction. □

Existence of a Consumer-Optimal Equilibrium

Consider the following strategy profile: All intermediaries offer compensation $\Pi'(i)$ to consumer $i \in [0, m^*]$ and zero compensation to consumer $i \in (m^*, 1]$. Each consumer $i \leq m^*$ accepts the offer of intermediary 1. Each player's action in other information sets are naturally defined. In particular, the equilibrium pricing in the downstream market is given by Lemma 1.

No consumer has a profitable deviation because $\Pi'(i) \geq c$ for all $i \leq m^*$. Next, I show that no intermediary has a profitable deviation. Suppose that intermediary $k \geq 2$ deviates and obtains data $D \subset [0, m^*]$ and $D' \subset (m^*, 1]$. To simplify notation, below I write $m^* + D$ instead of $m^* + |D|$. The deviation increases the revenue by $\Pi(m^* + D') - \Pi(m^* - D)$ in the downstream market, but k has to pay the compensation of weakly greater than $\int_{m^*-D}^{m^*+D'} \Pi'(i)di = \Pi(m^* + D') - \Pi(m^* - D)$. (Note that for consumers in D', the intermediary has to pay at least $c \geq \Pi'(m^*)$.) Thus, $k \geq 2$ has no profitable deviation.

Finally, suppose that intermediary 1 deviates, so that it does not obtain data $D \subset [0, m^*]$ but obtains data $D' \subset (m^*, 1]$. Following this deviation, $i \in D$ provides her data exclusively to another intermediary. Thus, the deviation reduces the total compensation 1 has to pay by at most $\int_D \Pi'(i)di - |D'|c$ but decreases the revenue in the downstream market by $\Pi(m^*) - [\Pi(m^* + D') - \Pi(D)]$. Thus, to show that the deviation is not profitable, it is enough to show that

$$\Pi(m^*) - [\Pi(m^* + D') - \Pi(D)] \geq \int_D \Pi'(i)di - |D'|c. \tag{4}$$

First, the concavity of Π implies that $\Pi(D) = \int_0^{|D|} \Pi'(i)di \geq \int_D \Pi'(i)di$. Second,

$$|D'|c = |D'| \cdot \Pi'(m^*) \geq \Pi(m^* + D') - \Pi(m^*).$$

These inequalities imply inequality (4). Thus, the deviation cannot be profitable.

G Proof of Proposition 3

Proof. Let $\tau_i^k = \emptyset$ denote intermediary k's choice of making no offer to consumer i. To see that an equilibrium exists, consider the following strategy profile: For each $i \leq m(\gamma)$, $\tau_i^1 = c$; for each $i > m(\gamma)$, $\tau_i^1 = \emptyset$; for each $k \neq 1$ and $i \in [0, 1]$, $\tau_i^k = \emptyset$. On the equilibrium path, consumers $i \in [0, m(\gamma)]$ accept offers from intermediary 1. The equilibrium price in the downstream market is given by Lemma 1. If an intermediary deviates in the first stage, each consumer accepts a set of offers to maximize her payoff. By the same argument as Theorem 1 where I show that Point 2 implies Point 1, we can confirm that this consists of an equilibrium.

Now, take any equilibrium where consumers $[0, m]$ share their data. I show that consumers $i \leq m$ and $i > m$ earn compensations c and zero, respectively. First, any consumer $i > m$ does not receive compensation because she does not share her data. Second, suppose to the contrary that there is an equilibrium in which a positive mass of consumers in $D \subset [0, m]$ receive compensation $\tau > c$ from, say, intermediary 1. Suppose that intermediary 1 unilaterally deviates and offers all consumers $i \in D$ a compensation of $\tau' \in (c, \tau)$. Suppose that for a positive mass of consumers in D, τ' is the only offer they receive. Let D' denote the set of those consumers. Then they continue to accept the offer. However, this means that intermediary 1 can profitably deviate by lowering offers to consumers in D', which is a contradiction. Thus, $|D'| = 0$. In other words, (almost) every consumer in D provides her data to another intermediary $k \neq 1$ following this deviation. By the same logic as Lemma 2, we can show that such intermediary k must be offering a non-positive compensation that is reject on the equilibrium path. However, this is a contradiction, because intermediary k can strictly increase its payoff by not sending an offer to consumers in D. Thus, intermediary 1 can profitably deviate by lowering the compensation, which is a contradiction.

Next, I show that there is no equilibrium where mass $m \neq m(\gamma)$ of consumers share their data. First, take any equilibrium, and suppose that mass $m < m(\gamma)$ of consumers share their data. Suppose that intermediary 1 deviates and offers consumers $(m, m + \Delta]$ a compensation strictly greater than but close to c. Consumer $i \in (m, m + \Delta]$ accepts this offer, and importantly, intermediary 1 is the only one that acquires the data of consumer i. Indeed, given the transaction costs, there cannot be other intermediaries making offers, which are rejected for sure in the proposed equilibrium. Thus, intermediary 1's deviation increases its payoff by (arbitrarily close to) $\Pi(m + \Delta) - \Pi(m) - \Delta(c + \gamma) > 0$, which is a contradiction. We can also show that there is no equilibrium in which mass $m > m^*$ of consumers provide their data, in the same way as the proof of Theorem 1 (with c there being replaced by $c + \gamma$). □

H Proof of Proposition 4

Proof. Take any equilibrium, and suppose to the contrary that the mass of consumers sharing their data is $m < 1$. This means that any offer that consumers in $(m, 1]$ face contains a fee of weakly greater than b. Then, intermediary (say) 1 can weakly increase its payoff by offering a compensation of 0 to consumers in $(m, 1]$. Note that following this deviation, consumers in $(m, 1]$ accept offers of only intermediary 1. This increases the net payoff of intermediary 1 by $\Pi(1) - \Pi(m) > 0$. This shows that in any equilibrium, almost every consumer shares her data.

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

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

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

I Proof of Proposition 6

Proof. Point 3 follows from Lemma 1. Also, the convexity of $c(\cdot)$ implies that each consumer i weakly prefers to provide Q_i^k to intermediary k in exchange for τ_i^k. Thus, it remains to show that there is no profitable deviation for intermediaries in the upstream market. Take any $k \in K$ and any deviation of intermediary k, $(\hat{Q}_i^k, \hat{\tau}_i^k)_{i \in N}$. Without loss of generality, assume that $\hat{Q}_i^k \cap Q_i^j = \emptyset$, $\forall j \neq k$, and that each consumer accepts the deviating offer. (Otherwise, I could set $\hat{Q}_i^k = \emptyset$.) Define $Q^{-k} := \cup_{j \in K \setminus \{k\}} Q^j$ and $Q_i^{-k} := \cup_{j \in K \setminus \{k\}} Q_i^j$, $\hat{q}_i^{-k} := |Q_i^{-k}|$, and $\hat{q}_i := |Q_i^{-k} \cup \hat{Q}_i^k|$. Note that k has to pay at least $c\left(\hat{q}_i\right) - c\left(\hat{q}_i^{-k}\right)$ to obtain $\hat{Q}_i^k$, because of the convexity of $c(\cdot)$. Now, suppose that there are consumers, say 1 and 2, such that $\hat{q}_1 < \hat{q}_2$. Then, intermediary k can find another more profitable deviation. Namely, it can reduce by $\varepsilon > 0$ the amount of consumer 2's data and increase by ε the amount of consumer 1's data. This does not change the price of k's data in the downstream market, because the firm's revenue only depends on the total amount of data. However, the new deviation decreases the total compensation k has to pay (relative to the original deviation), because of the convexity of $c(\cdot)$. Thus, I can focus on deviations such that $\hat{q}_i$ is independent of $i \in N$. Now, if $\hat{q}_i < q^*$ for all i, then intermediary k can strictly increase its payoff by slightly increasing the amount of data to obtain from each consumer, because $\Pi'(Nq) > c'(q)$. Similarly, any deviation with $\hat{q}_i > q^*$ is inferior to obtaining q^* from each consumer. Therefore, intermediary k has no profitable deviation in the upstream market. $\square$