Competing Data Intermediaries

Shota Ichihashi*

September 27, 2020

Abstract

I study a model of competition between data intermediaries, which collect personal data from consumers and sell them to downstream firms. Competition has a limited impact on benefiting consumers: If intermediaries offer high compensation for data, consumers share data with multiple intermediaries, which lowers the downstream price of data and hurts intermediaries. Anticipating this, intermediaries offer low compensation for data. Although consumers are exclusive suppliers of data, the nonrivalry of data leads to concentration and low consumer welfare in data markets. In particular, competing intermediaries earn a monopoly profit when downstream firms' data acquisition unambiguously harms consumers.

JEL Classification: D42, D43, L12, D80

*Bank of Canada, 234 Wellington Street West, Ottawa, ON K1A 0G9, Canada. Email: shotaichihashi@gmail.com.
I thank Jason Allen, Jonathan Chiu, Antoine Dubus (discussant), Itay Fainmesser, Matthew Gentzkow, Byung-Cheol Kim, Sitian Liu, Paul Milgrom, Shunya Noda, Makoto Watanabe, and seminar and conference participants at the Bank of Canada, CEA Conference 2019, Decentralization Conference 2019, Yokohama National University, the 30th Stony Brook Game Theory Conference, EARIE 2019, Keio University, NUS, HKU, HKUST, Western University, University of Montreal, ShanghaiTech University, SUFE, and Digital Economics Conference at Toulouse. The opinions expressed in this article are the author's own and do not reflect the views of Bank of Canada.

1 Introduction

Online platforms, such as Google and Facebook, collect user data and share them indirectly through targeted advertising. Data brokers, such as Acxiom and Nielsen, collect consumer data and sell them to retailers and advertisers (Federal Trade Commission, 2014).[1] This paper studies competition between data intermediaries, which collect and distribute personal data between consumers and downstream firms.

As an example, consider online platforms that collect consumer data and share them with third parties. The use of data by third parties may harm consumers through price discrimination and intrusive advertising, or benefit them through improved products and personalized offerings. Depending on these effects, platforms may offer compensation or charge fees for collecting data from consumers. Compensation may be in non-monetary benefits, such as online services (e.g., web-mapping services).

The main question of this paper is whether competition between data intermediaries improves consumer welfare. This is an important question in recent policy debates on competition in digital markets (Crémer et al., 2019; Furman et al., 2019; Morton et al., 2019).

The baseline model consists of a consumer (she), data intermediaries, and a downstream firm. The consumer has a finite set of data (or data labels), such as her email address, location, and purchase history. First, each intermediary chooses the set of data to collect and how much compensation to offer. Second, the consumer decides whether to accept each offer. Then each intermediary observes what data other intermediaries have collected.[2] Finally, the intermediaries post prices and sell the data to the downstream firm.

The model captures two features of personal data. First, data are nonrivalrous—i.e., the consumer can provide the same data to multiple intermediaries. Second, the consumer's payoff may depend non-monotonically on what data the downstream firm obtains. For example, a consumer may be comfortable with sharing either their place of birth or date of birth. However, the consumer may demand compensation to share both, because companies may use them to infer their Social Security number (Acquisti and Gross, 2009). The model is rich enough to capture such a situation.

Section 5 considers a consumer who holds one unit of data. Competing intermediaries sustain a

[1] Section 3.2 discusses these applications in detail.
[2] Section 3.1 motivates this assumption and discusses several ways to relax it.

monopoly outcome if and only if the downstream firm's data acquisition lowers consumer welfare. Even when the firm's data acquisition increases consumer welfare, the benefit of competition is lower than in markets for rivalrous goods. The result shows that the nonrivalry of data relaxes competition between data intermediaries and lowers consumer welfare, and that the effect is strong when the downstream firm's data usage harms consumers.

Section 6 considers a consumer with any finite set of data. The consumer may benefit or lose, depending on the set of data the downstream firm obtains. Under general preferences, I characterize an equilibrium that maximizes intermediary surplus and minimizes consumer surplus across all equilibria. Competition occurs only for the data the downstream firm uses to benefit the consumer. As a result, in this equilibrium, consumer surplus and intermediary surplus fall between those in the monopoly market and those in markets for rivalrous goods.

With an additional assumption that the consumer incurs an increasing marginal loss of sharing her data, I characterize the equilibria that have the following properties. First, the intermediaries collect disjoint sets of data. Second, each intermediary acts as a local monopsony—i.e., it pays the consumer the minimum compensation to cover her losses of sharing the data. I compare these equilibria in terms of their degrees of data concentration, then derive welfare implications. I connect this result with the welfare impact of "breaking up platforms."[3]

To clarify the intuition, the baseline model assumes a single consumer and homogeneous intermediaries. However, the main insight is robust to various extensions, such as multiple consumers and differentiated intermediaries (Section 7).

The contribution of this paper is to clarify when competition between data intermediaries benefits consumers. I show that (i) competition may not benefit consumers when they incur losses from third parties' data usage, but (ii) competition works when consumers benefit from the data usage, because competition eliminates positive fees a monopolist would charge. Points (i) and (ii) lead to the main insight: Competition for personal data may benefit consumers, but not as much as in markets for physical goods. The results help us understand why consumers do not seem to be compensated for their data provision (Arrieta-Ibarra et al., 2018).

The rest of the paper is as follows. Section 2 discusses related literature, and Section 3 describes

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

the model. Section 4 studies two benchmarks: a model of a monopoly intermediary, and that of competition for rivalrous goods. Section 5 assumes the consumer has one unit of data. Section 6 allows general preferences, and shows an equilibrium that maximizes intermediary profit and minimizes consumer surplus. Under an additional assumption, I characterize equilibria that have different degrees of data concentration. Section 7 provides extensions, and Section 8 concludes.

2 Related Literature

This paper relates to three strands of literature: markets for data, two-sided markets, and contracting with externalities.

Markets for data. Recent work, such as Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019), Bergemann, Bonatti, and Gan (2019), and Choi, Jeon, and Kim (2019), study models in which platforms collect data from consumers. These papers focus on a monopolist, and assume that data usage by firms or platforms harms consumers. In contrast, I study competition, and the consumer in my model may benefit or lose depending on what pieces of data are collected. These points lead to new insights: The effect of competition for data depends on how downstream firms use consumer data, and competition may not occur even if homogeneous intermediaries compete for the same data.

The papers cited above show that data externalities (or information externalities) between consumers may create inefficiency. In contrast, my focus is not on efficiency or market failure, but on the division of surplus created by data—i.e., how the nonrivalry of data relaxes competition and reduces consumer welfare, and how this effect depends on downstream firms' data usage.[4] Even if the outcome is efficient, these questions are relevant as long as we care about whether consumers receive surplus generated by their data.

The downstream market of my model relates to Gu, Madio, and Reggiani (2020). They study how data brokers' incentives to merge data depend on the downstream firm's revenue function. I abstract away from contracting between intermediaries, but consider data collection in the upstream market. Jones and Tonetti (2018) study a semi-endogenous growth model that incorporates

[4] I abstract away from other important issues regarding privacy, such as the privacy paradox, behavioral biases, and the ethical aspects of privacy (see, e.g., Acquisti et al. 2016 and Zuboff 2019).

data intermediaries.

Recent work studies rich models of how online platforms monetize data. De Corniere and De Nijs (2016) study the design of an online advertising auction, in which a platform uses data to improve the quality of the matches between users and advertisers. Fainmesser, Galeotti, and Momot (2019) study the optimal design of data-storage and data-protection policies by a monopoly platform. Kim (2018) incorporates privacy concerns and competition in a model of an advertising platform. Bonatti and Cisternas (2020) study the aggregation of consumers' purchasing histories and how data aggregation and transparency affect a strategic consumer's incentives. De Cornière and Taylor (2020) employ the competition-in-utilities approach to study the issue of data and competition.

Finally, this paper relates to the literature on other information goods, such as patents and digital goods (e.g., Shapiro and Varian 1998; Lerner and Tirole 2004; Sartori 2018). Relative to this literature, the novelty of my paper is to consider the upstream market in which consumers provide data to intermediaries. To keep the model simple, I abstract away from important issues relevant to information goods, such as network effects and versioning.

Two-sided markets. This paper relates to the literature on two-sided markets (e.g., Caillaud and Jullien 2003; Rochet and Tirole 2003; Armstrong 2006; Galeotti and Moraga-González 2009; Hagiu and Wright 2014; Carrillo and Tan 2015; Rhodes, Watanabe, and Zhou 2018). I show that the non-rivalry of data relaxes competition, which echoes the finding of the literature that multi-homing by one side relaxes the platform competition for that side (e.g., Caillaud and Jullien 2003 and Tan and Zhou 2020). The main difference is that in my model, the consumer's benefit or loss of sharing data and the resulting impact of competition depend on which pieces of data are collected (see, e.g., Proposition 6). To my knowledge, such a setting does not have a counterpart in the literature, in which consumers usually enjoy benefits on platforms.[5]

Contracting with externalities. We can interpret my model as that of contracting with externalities (McAfee and Schwartz, 1994; Segal, 1999; Rey and Tirole, 2007). Namely, a supplier (a consumer) provides goods (data) to retailers (intermediaries) who later compete in the downstream market. Compared to a typical model of vertical contracting, my model differs in the supplier's

[5] Anderson and Coate (2005) and Reisinger (2012) consider a model of platform competition in which advertisers impose negative externalities on consumers.

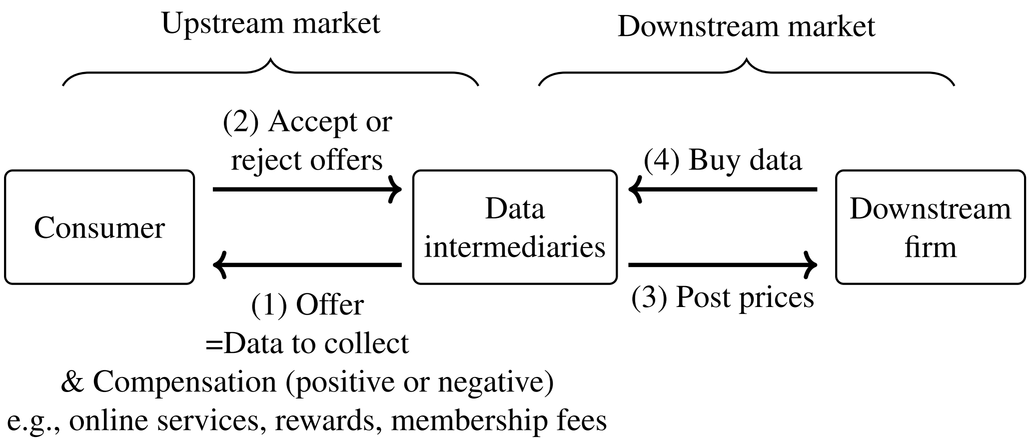

Figure 1: Timing of moves.

"cost" of producing goods: The cost (i.e., the consumer's utility from sharing data) can be non-monotone in what set of goods to produce. Also, the marginal cost of producing the same good is decreasing, because the consumer's payoff from the firm's data usage does not depend on how many intermediaries resell their data.

3 Model

There are $K \in \mathbb{N}$ data intermediaries, one consumer (she), and one downstream firm. We use K for the number and the set of the intermediaries. Figure 1 depicts the game: Intermediaries buy data in the upstream market and sell them in the downstream market. The detail is as follows.

Upstream Market

The consumer has a finite set $\mathcal{D}$ of data. Elements of $\mathcal{D}$ represent data labels, such as location and health data. They can also be different versions of the same data, such as health data of different qualities. Each element of $\mathcal{D}$ is an indivisible and nonrivalrous good (Section 3.1 discusses this assumption).

At the beginning of the game, each intermediary $k \in K$ simultaneously makes an *offer* (D_k, τ_k), where $\tau_k \in \mathbb{R}$ is compensation intermediary k is willing to pay for data $D_k \subset \mathcal{D}$. Compensation can be monetary rewards a consumer can enjoy by sharing data; it could also be the quality of a service that has a monetary value to consumers equal to the cost of provision for an intermediary. A negative compensation corresponds to a fee.

The consumer observes the offers, then chooses a set $K_C \subset K$ of offers to accept. Here, $k \in K_C$ means the consumer receives τ_k and provides the requested data D_k to intermediary k. The consumer can accept any set of offers, which reflects the nonrivalry of data. All intermediaries and the firm observe the data $\hat{D}_k \in \{D_k, \emptyset\}$ that each intermediary k has collected. I call $(\hat{D}_k)_{k\in K}$ the *allocation of data*.

Downstream Market

Each intermediary k simultaneously posts a price $p_k \in \mathbb{R}$ for $\hat{D}_k$. The firm then chooses a set $K_F \subset K$ of intermediaries, from which it buys data $\cup_{k\in K_F} \hat{D}_k$ at total price $\sum_{k\in K_F} p_k$.

Preferences

All players maximize their expected payoffs, and their ex post payoffs are as follows. The payoff of each intermediary is revenue from the downstream firm minus compensation to the consumer.

Suppose the consumer earns compensation τ_k from each intermediary $k \in K_C$, and the firm obtains data $D \subset \mathcal{D}$ from intermediaries. Then, the consumer receives a payoff of $U(D) + \sum_{k\in K_C} \tau_k$, where $U(D)$ is her gross payoff when the firm acquires D. I normalize $U(\emptyset) = 0$, so the firm's acquisition of D harms the consumer if $U(D) < 0$.

Suppose the firm obtains data $D \subset \mathcal{D}$ and pays a total price of p to intermediaries. Then the firm obtains a payoff of $\Pi(D) - p$, where $\Pi(D)$ is the firm's revenue from data D. The revenue function $\Pi(\cdot)$ is strictly increasing and satisfies $\Pi(\emptyset) = 0$.[6]

Throughout the paper, we assume total surplus is maximized when the downstream firm obtains all data, $\mathcal{D}$:

Assumption 1. The set functions $U(\cdot)$ and $\Pi(\cdot)$ satisfy $\mathcal{D} \in \arg\max_{D\subset\mathcal{D}} U(D) + \Pi(D)$.

When the consumer holds one unit of data (i.e., $|\mathcal{D}| = 1$) as in Section 5, the assumption is necessary for nontrivial results. For a general $\mathcal{D}$, the assumption holds, for example, if the firm sells products and can use data $\mathcal{D}$ to efficiently price discriminate consumers (Appendix N microfounds $U(\cdot)$ and $\Pi(\cdot)$ with this interpretation). In terms of primitives, the assumption holds if the firm's marginal revenue from data is high relative to the consumer's marginal loss of sharing

[6] $\Pi(\cdot)$ is strictly increasing if and only if for any $X, Y \subset \mathcal{D}$ such that $X \subsetneq Y$, $\Pi(X) < \Pi(Y)$.

data. Appendix L studies an extension with multiple consumers, and argues that the counterpart of Assumption 1 is likely to hold if there are negative data externalities between consumers.

Timing

The timing of the game is as follows (see Figure 1). First, intermediaries simultaneously make offers to the consumer, who then chooses the set of offers to accept. After observing the allocation of data, intermediaries simultaneously posts prices to the firm. The firm then chooses the set of intermediaries from which it buys data.

Solution

The solution concept is pure-strategy subgame perfect equilibrium (SPE) that is Pareto undominated from the perspectives of the intermediaries. Unless otherwise noted, "equilibrium" refers to SPE that satisfies this restriction.

3.1 Modeling Assumptions

I discuss important modeling assumptions.

Data as indivisible and nonrivalrous goods. I do not model the "realization" of data. For example, a consumer's location—the realization of her location data—is initially her private information. Depending on her location, she may prefer to disclose or conceal the information. However, I abstract away from such uncertain realizations of data, by assuming that the consumer has no private information. This assumption follows recent work on data markets, which focuses on consumers' ex ante incentive to provide data (see Section 2).

A single consumer. I assume a single consumer. The results do not depend on this assumption, but a model with multiple consumers is useful in two ways. First, it enables us to distinguish between the types of data (e.g., location or health data) and data subjects (e.g., consumer i's data or j's data). Second, a multiple-consumer model enables us to introduce data externalities, under which consumers may receive negative payoffs in equilibrium. Section 7 presents these extensions.

The allocation of data is publicly observable. The baseline model assumes that intermediaries and the firm observe what data each intermediary collects. In practice, some data intermediaries disclose what kinds of data they collect. For example, a data broker CoreLogic states it holds property data that cover more than 99.9% of U.S. property records.[7] Also, when an intermediary collects data from consumers, it needs to reveal what data it collects—e.g., Nielsen Homescan states it will collect purchase records.

One concern is that an intermediary may not observe *whose* data other intermediaries hold. For example, we may know Google holds search queries, but we may not know whose search queries it holds. A model with a single consumer cannot address this concern. However, Section 7.1 shows that the main result continues to hold when (i) there is a continuum of consumers, and (ii) each intermediary does not observe the identities of consumers in other intermediaries' data. As I discuss next, this extension also justifies the assumption on the space of possible contracts.

The restriction on the contract space. I assume that data intermediaries cannot offer compensation that depends on outcomes such as the allocation of data. In the baseline model, this restriction is crucial. For example, suppose intermediary k can commit to pay positive compensation if and only if the consumer provides her data *only* to intermediary k (i.e., an exclusive offer as in Bernheim and Whinston 1998). Exclusive offers render data rivalrous, and competition provides full surplus to the consumer. The baseline model excludes such offers by assuming that compensation from intermediary k depends only on whether the consumer provides her data to intermediary k.

There are two motivations for this restriction. First, the restriction arises if each intermediary cannot observe or verify the identities of consumers who interact with other intermediaries. The baseline model with a single consumer cannot capture this idea. Thus, in Section 7.1, I study a setting in which (i) there are many consumers and (ii) intermediaries do not observe the identities of consumers in their rivals' datasets. There, even though intermediaries can commit to compensation that depends on observable outcomes, we obtain the same equilibrium as in the baseline model.

Second, when compensation is the provision of a service, consumers may receive it simultaneously as they provide data. For example, a consumer provides location data and benefits from the web-mapping service at the same time. In such a case, intermediaries would not be able to change compensation based on subsequent outcomes, such as whether the consumer uses other services.

Pure-strategy equilibrium. I study pure-strategy equilibrium (PSE) for two reasons. First, PSE clearly captures the intuition that the nonrivalry of data relaxes competition—e.g., there is a PSE in which one intermediary acts as a monopolist. Second, PSE facilitates the analysis. For instance,

[7] https://www.corelogic.com/about-us/our-company.aspx (accessed July 4, 2020)

we can derive a PSE even if the consumer's payoff is non-monotone in the set of data acquired. For another instance, we can compare pure-strategy equilibria in terms of their degrees of data concentration. Such a comparison is difficult in mixed strategy equilibrium (MSE), because the allocation of data is ex ante uncertain. However, Section 7.3 shows the main insight can hold even in an MSE.

Timing. In practice, consumers first decide which platforms to join, after which they use the services and generate data. However, I assume that participation, service usage, and data collection occur simultaneously. Such a timing assumption would be a reasonable approximation so long as a platform does not change the value of its service after consumers join it. Also, data collection may affect consumers long after they provide data (e.g., data breach). We can incorporate such a situation by interpreting $U(\cdot)$ as the consumer's discounted utility.

In the downstream market, intermediaries first observe the allocation of data, then choose prices. The assumption is similar to that of endogenous product differentiation, in which sellers choose prices after observing their chosen product design (e.g., D'Aspremont et al. 1979). What data an intermediary collects (i.e., offer) can be a part of platform design or a company's policy. For example, a web-mapping service, such as Google Maps, could be an offer (D_k, τ_k) such that D_k consists of location data, and τ_k is the service quality. After collecting data, platforms and data brokers typically share the data in exchange for money. Then, it is reasonable to assume that intermediaries can adjust downstream prices more quickly than adjusting what data to collect.

3.2 Applications

I provide several applications and discuss the validity of assumptions in light of them.

Online platforms. The model can capture competition for data between online platforms, such as Google and Facebook. Given an offer (D_k, τ_k), a consumer provides data D_k to use platform k, whose service quality is τ_k. Platforms share data with advertisers and retailers, which may benefit or hurt consumers (e.g., beneficial targeting or harmful price discrimination). The utility $U(D)$ captures the net effect of these data usages.

Several remarks are in order. First, in practice, advertising platforms use consumer data to match users with advertisers, instead of reselling data.[8] The downstream market of my model abstracts away from such details. However, the model captures a general idea—that platforms have a higher willingness to pay for the data that others do not hold (Lemma 1 in Appendix F shows that the price of data is high when other intermediaries do not hold them). For example, a platform will have a higher willingness to acquire health data when its rivals do not have them, because the platform will be the only one that can display ads based on users' health profiles. When this economic force is present in the downstream interactions, the insights in this paper would be relevant (see also the discussion in Section 5.1).

Second, $U(\cdot)$ is exogenous—i.e., intermediaries cannot affect how the firm uses data. This reflects the difficulty of writing a contract over how third parties use data. A similar assumption appears in recent papers, such as Huck and Weizsacker (2016) and Jones and Tonetti (2018).

Third, compensation is one-to-one transfer. If we interpret compensation as the value of a service, this assumption implies that a consumer's benefit from a service does not depend on what other services she uses. Section 7.2 relaxes this assumption and shows that the main insight holds if services offered by intermediaries are not very substitutable.

Finally, some data are more likely to satisfy the nonrivalry assumption than other data. To see this, compare location data with browsing history: Consumers may easily share their location with multiple intermediaries—e.g., they sign up for multiple online services that track location (potentially in the background). In contrast, consumers generate browsing data only when they use a browser, and it is unclear whether they can share the data with multiple services. Thus, even though data are in principle nonrivalrous, some data are easier for consumers to share than other data. The current application is suitable when we consider data that consumers can easily share with multiple platforms.

Data brokers. We can interpret intermediaries as data brokers, such as LiveRamp, Nielsen, and Oracle. Data brokers collect personal data from online and offline sources, and resell or share that data with others, such as retailers and advertisers (Federal Trade Commission, 2014).

Some data brokers obtain data from consumers in exchange for monetary compensation (e.g., Nielsen Home Scan). At the same time, data brokers commonly obtain personal data without in-

[8]For example, the privacy policies of Google and Facebook state that they do not resell personal information. See `https://policies.google.com/privacy?hl=en-US` and `https://www.facebook.com/policy.php` (accessed on September 21, 2020).

teracting with consumers. The model could also fit such a situation. For example, suppose data brokers obtain individual purchase records from retailers. Consider the following chain of transactions: Retailers compensate customers and record their purchases—e.g., they offer discounts to customers who sign up for loyalty cards. Retailers then sell these records to data brokers, which resell the data to third parties. We can regard retailers in this example as consumers in the model.

The model can also be useful for understanding how the incentives of data brokers would look like if they had to source data directly from consumers. This question is important, since awareness of data sharing practices increases, and policymakers try to ensure consumers have control over their data (e.g., the EU's GDPR and California Consumer Privacy Act).

Mobile application industry. Kummer and Schulte (2019) empirically show that mobile application developers trade greater access to personal information for lower app prices, and consumers trade off lower prices and greater privacy. Also, 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.

4 Two Benchmarks

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

4.1 Monopoly Intermediary ($K = 1$)

In the upstream market, a monopoly intermediary collects data D by paying a compensation of $-U(D)$. In the downstream market, it sets a price of $\Pi(D)$ to extract full surplus from the firm. Assumption 1 leads to the following result.

Claim 1. *In any equilibrium, a monopoly intermediary extracts full surplus* $\Pi(\mathcal{D}) + U(\mathcal{D})$*, and the consumer and the firm obtain zero payoffs.*

4.2 Competition for Rivalrous Goods

Suppose data are rivalrous—i.e., the consumer can provide each piece of data to at most one intermediary.[9] This model captures competition between intermediaries for physical goods (e.g., Stahl 1988). See Appendix A for the proof of the following result.

Claim 2. *Suppose that data are rivalrous and there are multiple intermediaries. In any equilibrium, the consumer extracts full surplus,* $\Pi(\mathcal{D}) + U(\mathcal{D})$*, and all intermediaries and the firm obtain zero payoffs.*

The result follows from Bertrand competition in the upstream market: If one intermediary earned a positive profit by obtaining some data, another intermediary could profitably deviate by offering the consumer slightly higher compensation to exclusively obtain the data.

5 Single Unit Data

We now assume that the consumer holds one unit of data (i.e., $\mathcal{D} = \{d\}$), and there are multiple intermediaries (i.e., $K \geq 2$). We write $U := U(\{d\})$ and $\Pi := \Pi(\{d\})$. The following result characterizes the equilibrium (see Appendix B for the proof).

Proposition 1. *In any equilibrium, one intermediary obtains data at compensation* $\max(0, -U)$*. The consumer obtains* $\max(0, U)$*, the intermediary obtains* $\Pi - \max(0, -U)$*, and other intermediaries and the firm obtain zero payoffs. In particular, one intermediary earns a monopoly profit* $\Pi + U$ *in any equilibrium if and only if data collection is harmful, i.e.,* $U \leq 0$*.*

If (and only if) $U > 0$, the consumer receives a strictly greater payoff under competition than under monopoly in Claim 1. If $U \leq 0$, the equilibrium coincides with monopoly. In either case, consumer surplus is lower than in the rivalrous-goods benchmark, in which the consumer receives $\Pi + U$ (Claim 2).

The intuition is that competition incentivizes intermediaries to decrease positive fees to zero, but does not incentivize them to increase positive compensation beyond a monopoly level. To see this, suppose intermediary 1 collects data at a positive fee. Then, intermediary 2 can undercut the

[9] "Rivalrous data" refer to the model in which the consumer can accept a collection of offers $(D_k, \tau_k)_{k \in K_C}$ if and only if $D_k \cap D_j = \emptyset$ for any distinct $j, k \in K_C$.

fee to exclusively obtain the data: Indeed, when the consumer faces the two offers with positive fees, she shares her data with *only* intermediary 2, because she receives a gross utility of U so long as the firm obtains data from at least one intermediary. As a result, competing intermediaries cannot charge a positive fee. Thus if $U > 0$, the consumer receives a payoff of at least U.

In markets for rivalrous goods, this Bertrand competition in the upstream market raises the equilibrium compensation to $\Pi + U > 0$. However, in the market for data, competition of increasing compensation (above zero) may fail. For example, suppose intermediary 1 collects data at monopoly compensation $-U$ when $U < 0$. If intermediary 2 also offers positive compensation, the consumer provides her data to *both* intermediaries. Then the downstream price of data will be zero. Anticipating this, intermediary 2 makes no competing offer.

The effect of competition depends on how downstream firms use data. To see this, we fix total surplus $TS = \Pi + U$. First, the joint profit of intermediaries, $TS - \max(0, U)$, is maximized when $U < 0$: The intermediation of data is more profitable when the downstream firm's data usage harms consumers. As U increases from negative to $TS > 0$, the equilibrium outcome changes from the monopoly outcome to the consumer-best one. This observation contrasts with the case of rivalrous goods, in which the outcome depends only on TS.

Proposition 1 may also justify the frequently used assumption in the literature that the market consists of a monopoly data seller.[10] We may view this assumption as a subgame of the extended game in which data sellers first acquire information at cost, then sell collected data.

5.1 Discussion on Exclusive Data Acquisition

Proposition 1 implies we do not observe multiple intermediaries that hold the same data. However, in practice there seem to be counterexamples—e.g., online advertising intermediaries sell the same targeting data to advertisers via their ad networks.

In this respect, we should interpret Proposition 1 (and other results in this paper) as an approximation: As in models of product differentiation, intermediaries will have an incentive to differentiate themselves in terms of what data they hold (Gartner, 2016; Gu et al., 2020; Madio

[10] See, e.g., Babaioff et al. (2012), Bergemann et al. (2018), Bergemann and Bonatti (2019), Bimpikis et al. (2019), and references therein. Also, Sarvary and Parker (1997) study competition between information sellers.

et al., 2019).[11] This incentive to differentiate and the nonrivlary of data discourage intermediaries from paying for data their rivals already hold. As a result, the upstream market for data becomes less competitive than that for rivalrous goods. The model captures this intuition in an extreme way: Overlapping data will have a downstream price of zero, and thus intermediaries have no incentive to collect data their rivals already hold. Section 7.4 relaxes this assumption, and studies an extension in which intermediaries may earn positive revenue even if they hold the same data in the downstream market.

5.2 How to Improve Consumer Surplus?

I discuss two institutional changes that could improve consumer surplus.

Stronger bargaining power of consumers. One is to give consumers more bargaining power. For example, if the consumer can make a take it or leave it offer to any intermediary, she will offer $(\{d\}, \Pi)$ to extract full surplus. This solution relates to the idea of "data labour union" in Arrieta-Ibarra et al. (2018).

Rich contract space. Another way is to enable intermediaries to offer richer contracts. For example, suppose each intermediary can set compensation based on whether the consumer shares her data with other intermediaries. In this case, there is an equilibrium with exclusive contracts: Each intermediary k commits to pay all of its downstream revenue (i.e., Π) if and only if the consumer provides data to only intermediary k. In this equilibrium, the consumer extracts full surplus.

However, enforcing richer contracts require transparency. For example, to use exclusive contracts, an intermediary needs to monitor how consumers interact with its rivals. Thus, each intermediary needs to track the identities of consumers across its competitors, which may be difficult if consumers could change their identities to transact with different intermediaries. Section 7 shows that in the absence of intermediaries' ability to track consumers, we continue to have the same equilibrium as in Proposition 1, even if intermediaries can use intricate compensation mechanisms.

[11] Although empirical evidence on data markets is sparse, an industry expert Gartner (2016) describes that "data brokers tend to specialize in certain industries in order to gain a competitive advantage." Gu et al. (2020) describe that "different data brokers may be particularly strong in different areas. For example, Acxiom and Datalogix may profile more consumers for targeting purposes than other data brokers, collecting information such as demographics, sociographics, lifestyle, and purchasing behaviours. Data brokers like Corelogic and eBureau mostly sell in-depth financial and property data analytics."

6 General Preferences

I now consider the consumer with any finite set of data. I generalize Proposition 1, then point out the multiplicity of equilibria due to the nonrivalry of data.

6.1 Partially Monopolistic Equilibrium

In this section, we allow any gross utility function $U(\cdot)$ and any strictly increasing revenue function $\Pi(\cdot)$ that satisfy Assumption 1. The following result generalizes Proposition 1 (see Appendix C for the proof).

Proposition 2 (Partially Monopolistic Equilibrium (PME)). *There is a subgame perfect equilibrium in which one intermediary obtains all data at compensation $\max_{D\subset\mathcal{D}} U(D) - U(\mathcal{D})$. The consumer receives an equilibrium payoff of $\max_{D\subset\mathcal{D}} U(D)$. This equilibrium coincides with the monopoly outcome if and only if $U(D) \leq 0$ for all $D \subset \mathcal{D}$.*

If the consumer holds a single piece of data d, then $\max_{D\subset\mathcal{D}} U(D) = \max(0, U(\{d\}))$. As a result, the PME equals the unique equilibrium in Proposition 1. Proposition 2 states that the intuition for Proposition 1 applies to a broad class of preferences.

To see the intuition, consider Figure 2, which depicts $U(\cdot)$ and $\Pi(\cdot)$ as functions of the amount of data acquired by the firm. The gross utility function $U(\cdot)$ is non-monotone. First, a monopoly intermediary will obtain all data at compensation $-U(\mathcal{D})$ (i.e., short red dotted arrow). We can decompose this compensation $-U(\mathcal{D})$ as follows: The monopolist extracts surplus created by $D^* \in \arg\max_{D\subset\mathcal{D}} U(D)$ from the consumer by charging a fee of $U(D^*) > 0$, and it additionally obtains data $\mathcal{D} \setminus D^*$ at the minimum compensation of $U(D^*) - U(\mathcal{D})$ (i.e., long blue dotted arrow). In contrast, when there are multiple intermediaries, competition prevents intermediaries from extracting surplus $U(D^*)$ from the consumer. However, competition does not increase compensation for data $\mathcal{D} \setminus D^*$, the sharing of which harms the consumer. As a result, in the PME, a single intermediary obtains all data and compensates the consumer according to her loss $U(D^*) - U(\mathcal{D})$ of sharing $\mathcal{D} \setminus D^*$. Finally, the compensation in the PME is lower than $\Pi(\mathcal{D})$, which is the compensation she would receive in markets for rivalrous goods (i.e., black dashed arrow).

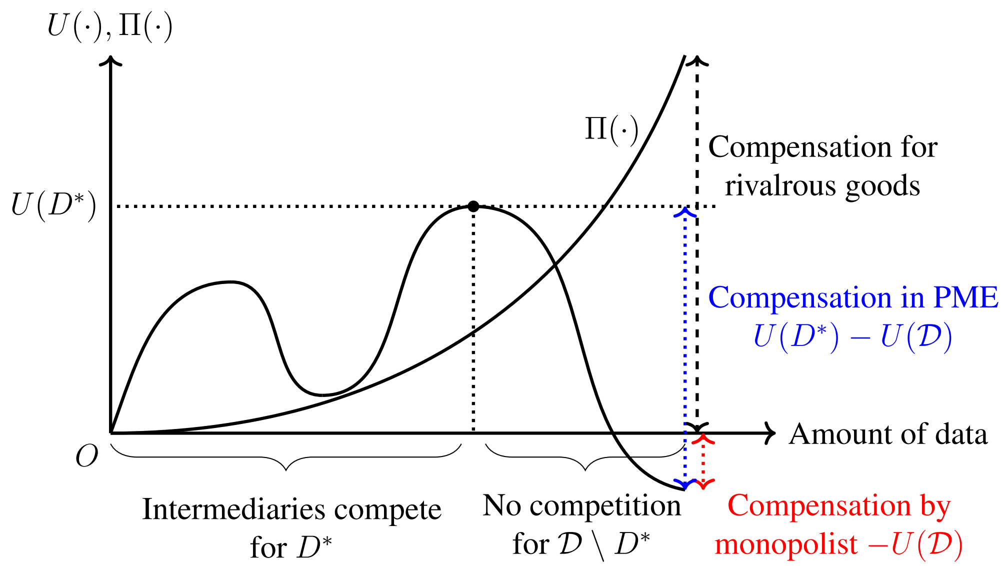

Figure 2: Partially monopolistic equilibrium.

The next result shows that if there are many intermediaries, the PME minimizes consumer surplus and maximizes intermediary surplus across all equilibria (see Appendix D for the proof). Thus, the PME is a natural extension of the monopoly equilibrium. Let $CS(K) \subset \mathbb{R}_+$ denote the set of all pure-strategy subgame perfect equilibrium (SPE) payoffs of the consumer when there are K intermediaries.

Proposition 3. *There is a $K^* \in \mathbb{N}$ such that for any $K \geq K^*$, the following holds.*

1. *The PME minimizes consumer surplus:* $\min CS(K) = \max_{D \subset \mathcal{D}} U(D)$.

2. *The PME maximizes the intermediaries' joint profit across all pure-strategy SPE.*

The intuition is as follows. Suppose there are K intermediaries, and in some equilibrium the consumer obtains a payoff of $U(D^*) - \delta_K$ with $\delta_K > 0$. If an intermediary offers (D^*, ε) with $\varepsilon < \delta_K$, the consumer will accept it. Because any intermediary can always deviate and offer (D^*, ε), each intermediary obtains a payoff of at least δ_K. Thus, intermediary surplus is at least $K \cdot \delta_K$. However, intermediary surplus is bounded from above by $\Pi(\mathcal{D}) + U(\mathcal{D}) < \infty$. As a result, δ_K goes to 0 as K grows large—i.e., as the number of intermediaries grows large, the worst consumer surplus converges to $U(D^*)$, which is the consumer's payoff in the PME. In the proof, I show that δ_K hits zero for a finite K, when $\Pi(\cdot)$ is strictly increasing. In the PME, total surplus is maximized and consumer surplus is $U(D^*)$. Thus, the PME is intermediary-optimal for a large K.

The main takeaway of the above propositions is that the impact of competition for data depends on how downstream firms use data. In a frictionless market for rivalrous goods, for any $U(\cdot)$, competition gives full surplus to agents (e.g., consumers) in the upstream market. In markets for data, the shape of $U(\cdot)$ affects the division of surplus. If data usage benefits consumers, competition eliminates fees that consumers would have to pay under monopoly. However, if data usage harms consumers, competition may not increase compensation. As a result, when data usage may benefit or harm consumers depending on the set of data to be used, competition weakly increases consumer welfare and decreases intermediary profit, but not as much as in markets for rivalrous goods.

6.2 Partitional Equilibria

With an additional assumption, I characterize a class of equilibria to examine the impact of market concentration. Specifically, I assume the increasing convex cost of sharing data for the consumer, and the decreasing marginal revenue for the downstream firm.

Assumption 2. As set functions, $U(\cdot)$ is decreasing and submodular, and $\Pi(\cdot)$ is strictly increasing and submodular.[12]

Definition 1. A *partitional equilibrium* is an equilibrium in which the allocation of data $(\hat{D}_k)_{k\in K}$ is a partition of $\mathcal{D}$, i.e., $\hat{D}_k \cap \hat{D}_j = \emptyset$ for any distinct $j, k \in K$, and $\cup_{k\in K} D_k = \mathcal{D}$.

In the case of rivalrous goods (i.e., Claim 2), the allocation is typically a trivial partition (see Appendix E for the proof).

Claim 3. *Suppose data are rivalrous, and $\Pi(\cdot)$ is strictly submodular. In any equilibrium, at most one intermediary collects a non-empty set of data.*

In contrast, any partition can be an equilibrium allocation of data (see Appendix F).

Proposition 4. *Under Assumption 2, if an allocation of data $(D^*_k)_{k\in K}$, compensation $(\tau^*_k)_{k\in K}$, and downstream prices $(p^*_k)_{k\in K}$ consist of a partitional equilibrium, then they satisfy the following conditions:*

[12] $U(\cdot)$ is submodular if for any $X, Y \subset \mathcal{D}$ with $X \subsetneq Y$ and $d \in \mathcal{D} \setminus Y$, we have $U(Y \cup \{d\}) - U(Y) \leq U(X \cup \{d\}) - U(X)$. If the strictly inequalities hold, $U(\cdot)$ is strictly submodular.

1. $D_j^* \cap D_k^* = \emptyset$ *for any distinct* $j, k \in K$*, and* $\cup_{k \in K} D_k^* = \mathcal{D}$.

2. *Each intermediary* k *offers* $\tau_k^* = U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$ *to collect* D_k^**, whenever the right-hand side is positive.*

3. *Each intermediary* k *sets a price of* $p_k^* = \Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D_k^*)$ *for* D_k^**, whenever the right-hand side is positive.*

Also, any $(D_k^*, \tau_k^*, p_k^*)_{k \in K}$ *that satisfies the above conditions is an outcome of some subgame perfect equilibrium.*

Partitional equilibria have three features. First, although data are nonrivalrous, intermediaries never collect overlapping data, because such data will have no value in the downstream market.

Second, any partition of $\mathcal{D}$ may arise in an equilibrium. For example, if the consumer holds data x_1 and x_2, then in one equilibrium, intermediaries 1 and 2 collect x_1 and x_2, respectively. In the rivalrous-goods case, intermediary (say) 1 could profitably deviate by offering the consumer to collect $\{x_1, x_2\}$ at a higher compensation. For nonrivalrous data, intermediary 1 does not benefit from such a deviation because the consumer will share data x_2 with both intermediaries.

Third, each intermediary compensates the consumer according to her incremental loss of sharing D_k, conditional on sharing data $\cup_{j \neq k} D_j$ with other intermediaries. In contrast, in the rivalrous-goods case, the equilibrium compensation depends on the downstream firm's willingness to pay.

6.3 Data Concentration

Proposition 4 implies any partition of $\mathcal{D}$ can arise as an equilibrium allocation of data. We can interpret a coarser partition as greater concentration of data:

Definition 2. Take two partitional equilibria, $\mathcal{E}$ and $\mathcal{E}'$. Let $(D_k)_{k \in K}$ and $(D'_k)_{k \in K}$ denote the equilibrium allocations of data in $\mathcal{E}$ and $\mathcal{E}'$, respectively. We say that $\mathcal{E}$ is *more concentrated than* $\mathcal{E}'$ if for each $k \in K$, there is $\ell \in K$ such that $D'_k \subset D_\ell$.

The following result shows the welfare implications of data concentration (see Appendix G for the proof).

Proposition 5. *Take two partitional equilibria such that one is more concentrated than the other. In the more concentrated equilibrium, intermediaries' joint profit is higher, and consumer surplus and the firm's profit are lower.*

The downstream price of data D_k is the firm's marginal revenue $\Pi(\mathcal{D}) - \Pi(\cup_{j \in K \setminus \{k\}} D_j)$ from D_k. If each of many intermediaries has a small subset of $\mathcal{D}$, the contribution of each dataset is close to $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus \{d\})$. In contrast, if a few intermediaries jointly hold $\mathcal{D}$, each of them can charge a high price to extract the infra-marginal value of its data. Because $\Pi(\cdot)$ is submodular, concentration leads to a greater total revenue for intermediaries. Similarly, if $U(\cdot)$ is submodular, data concentration harms consumers, because a large intermediary compensates the consumer based on her infra-marginal cost of sharing data. The following example relates the result to the idea of breaking up platforms.

Example 1. The consumer has location and financial data. The downstream firm profits from data, but there is a risk of data leakage. The consumer incurs an expected loss of \$20 from the potential data leakage if only if the firm holds *both* location and financial data (otherwise, she incurs no loss). A monopoly intermediary will obtain both location and financial data and pay \$20 to the consumer, leading to zero consumer surplus. For example, the intermediary may operate an online service that requires consumers to provide those data. Suppose now that a regulator breaks up the monopolist into two intermediaries, 1 and 2. Proposition 4 implies that in one of the equilibria, intermediaries 1 and 2 collect location and financial data, respectively, and each intermediary pays compensation \$20. In this equilibrium, the consumer earns a net surplus of \$20. Therefore, breaking up the monopolist may change the allocation of data, increase compensation, and benefit consumers.

7 Extensions

I study several extensions. The purpose is to show that the main insight—competition for data will have a limited impact on benefiting consumers—continues to hold in more realistic settings.

7.1 Multiple Consumers with a Rich Contract Space

This extension has two features relevant in practice—i.e., the market consists of many consumers, and intermediaries may not observe the identities of consumers in other intermediaries' datasets. As discussed in Section 3.1, this extension also justifies the restriction on the contract space in the baseline model.

We now consider a unit mass of consumers. Consumer $i \in [0,1]$ has a set of data, $\mathcal{D}_i = \left\{d_i^\ell\right\}_{\ell \in \mathcal{L}}$, where $\mathcal{L}$ is a finite set of data labels, such as $\mathcal{L} = \{location, health\}$. The set of all data is $\mathcal{D} := \cup_{i \in [0,1], \ell \in \mathcal{L}} \left\{d_i^\ell\right\}$, and the allocation of data is written as $(D_k)_{k \in K}$, where $D_k \subset \mathcal{D}$ is the set of data that intermediary k holds. For any $D \subset \mathcal{D}$, we write $q^\ell(D) \in [0,1]$ for the amount of data $\ell \in \mathcal{L}$ contained in D.[13] Given an allocation of data $(D_k)_{k \in K}$, the amount of data ℓ intermediary k holds is $q^\ell(D_k)$. I call $(q^\ell(D_k))_{\ell \in \mathcal{L}, k \in K}$ a *quantity vector*. The choice of an individual consumer, who is atomless, does not affect the quantity vector. Finally, let $\mathcal{Q}_{-k} := [0,1]^{|\mathcal{L}| \times (K-1)}$ denote the set of all quantity vectors for intermediaries other than k.

Importantly, each intermediary does not observe the identities of consumers in the data collected by other intermediaries. Formally, given the realized allocation of data $(D_k)_{k \in K}$, each intermediary k observes only D_k and $(q^\ell(D_j))_{\ell \in \mathcal{L}, j \in K \setminus \{k\}}$.

Intermediaries have limited information, but can use richer contracts than in the baseline model: At the beginning of the game, each intermediary k chooses a *contract*, which is a mapping $\gamma_k : 2^{\mathcal{L}} \times \mathcal{Q}_{-k} \to \mathbb{R}$. For any $L \subset \mathcal{L}$ and $q_{-k} \in \mathcal{Q}_{-k}$, any consumer i who provides data $\left\{d_i^\ell\right\}_{\ell \in L}$ to intermediary k receives compensation $\gamma_k(L, q_{-k})$, if the quantity vector of other intermediaries is q_{-k}. Intermediaries can choose any contract such that $\gamma_k(\emptyset, q) = 0$ for all q—i.e., there is no transfer if a consumer does not share any data. In equilibrium, consumers choose what data to share with each intermediary k, taking q_{-k} as exogenous.

I consider the following payoffs: If consumer i receives compensation τ and the firm acquires i's data $\left\{d_i^\ell\right\}_{\ell \in L}$, her payoff is $U(L) + \tau$, where $U(\emptyset) = 0$. Consumers are homogeneous, in that $U(\cdot)$ is independent of i.[14] The firm's payoff from buying data $D \subset \mathcal{D}$ is $\Pi(D) - p$, where p is the total payment to intermediaries. The revenue function $\Pi(\cdot)$ is an increasing set function, and $\Pi(D)$

[13]Formally, $q^\ell(D) = \lambda\left(\left\{i \in [0,1] : d_i^\ell \in D\right\}\right)$, where $\lambda(\cdot)$ is the Lebesgue measure. In the equilibrium I consider, $q^\ell(D)$ is well-defined on-path and after any unilateral deviation.

[14]If intermediaries can make discriminatory offers, we can allow heterogeneous preferences.

depends only on $(q^\ell(D))_{\ell\in\mathcal{L}}$. The payoff of each intermediary is revenue minus compensation.

The timing of the game is the same as before. First, intermediaries simultaneously offer contracts. Then each consumer decides the set of data to share with each intermediary. Intermediaries and the firm observe the realized quantity vector. Finally, intermediaries simultaneously post prices for their datasets, after which the firm makes a purchasing decision. The solution concept is perfect Bayesian equilibrium.[15] I impose the following straightforward extension of Assumption 1.

Assumption 3. The primitives $U(\cdot)$ and $\Pi(\cdot)$ are such that a monopoly intermediary collects all data in some equilibrium.[16]

The proof of the following result is in Appendix H.

Claim 4 (PME Under Rich Contract Space). *Suppose $K \geq 2$. There is an equilibrium in which one intermediary obtains all data by paying each consumer $\max_{L\subset\mathcal{L}} U(L) - U(\mathcal{L})$. This equilibrium coincides with a monopoly equilibrium if and only if $U(L) \leq 0$ for all $L \subset \mathcal{L}$.*

In the current model, an intermediary does not observe the identities of consumers who interact with other intermediaries. This limited observability prevents intermediaries from designing a contract that punishes consumers for sharing the same data with multiple intermediaries. As a result, competition for data gets relaxed, and we obtain the same equilibrium as in the baseline model, even though intermediaries can choose from a rich space of contracts.

7.2 Non-Additive Compensation

In the baseline model, the consumer's payoff is additively separable across compensations from intermediaries. However, if compensation is the value of a service, this additive separability may not hold. Thus, I extend the model as follows. For simplicity, assume that the consumer holds one unit of data d, and data collection is harmful, i.e., $L := -U(\{d\}) > 0$. Suppose the consumer shares her data with n intermediaries, receives compensation τ_k from each intermediary $k \in K$,

[15] In PBE, we need to specify the beliefs of intermediaries about whose data other intermediaries hold. However, in the equilibrium I consider, we can assign any beliefs on-path and after unilateral deviation, because the firm's revenue is independent of consumers' identities.

[16] In terms of $U(\cdot)$ and $\Pi(\cdot)$, the condition is written as follows: For each $D \subset \mathcal{D}$ and $L \subset \mathcal{L}$, let $q^L(D)$ denote the mass of consumers who have their data $\{d_i^\ell\}_{\ell\in L}$ collected under D. Then, the assumption means $\mathcal{D} \in \arg\max_{D\subset\mathcal{D}} \Pi(D) + \sum_{L\subset\mathcal{L}} q^L(D)U(L)$, where the summation is across all the subsets of $\mathcal{L}$.

and the firm obtains $D \subset \{d\}$. Then the consumer's payoff is $U(D) + T(\tau_1, \ldots, \tau_K) - n \cdot c$. The last term $n \cdot c$ is the cost of sharing data with n intermediaries, and $c \geq 0$ is exogenous. It captures the opportunity cost of using the service provided by an intermediary. The second term $T(\tau_1, \ldots, \tau_K)$ is the effective compensation, which maps a profile $(\tau_1, \ldots, \tau_K)$ of compensations to the consumer's utility.

Assumption 4. The function $T : \mathbb{R}^K \to \mathbb{R}$ satisfies the following: For each coordinate, T is strictly increasing and continuous. Also, T is symmetric, $T(0, \ldots, 0) = 0$, and $\lim_{x\to\infty} T(x, 0, \ldots, 0) = \infty$. Finally, T is submodular.

The assumption holds if $T(\tau_1, ..., \tau_K) = \hat{T}(\sum_{k\in K} \tau_k)$ for an increasing concave function $\hat{T}(\cdot)$, which subsumes the original setting, $T(\tau_1, ..., \tau_K) = \sum_{k\in K} \tau_k$. Submodularity captures the substitutability of services provided by intermediaries. The payoffs of intermediaries and the downstream firm remain the same. In particular, $\Pi > 0$ is the firm's gross revenue from the data. The proof of the following result is in Appendix I.

Claim 5. *Assume $T(\Pi, 0, \ldots, 0) > L + c$. Let τ^* denote the lowest value that satisfies*

$$T(\tau^*, 0, \ldots, 0) \geq L + c \quad \textit{and} \tag{1}$$

$$T(\tau^*, \Pi, 0, \ldots, 0) - T(0, \Pi, 0, \ldots, 0) \geq c. \tag{2}$$

At least one of these inequalities binds at τ^*. If only inequality (1) binds at τ^*, there is a subgame perfect equilibrium (SPE) in which one intermediary earns a monopoly profit, and the consumer obtains a payoff of 0. This occurs whenever $c = 0$. If inequality (2) binds at τ^*, there is an SPE in which all intermediaries earn zero profits, and the consumer receives a positive payoff.

Claim 5 states that depending on which of (1) and (2) binds, we have a monopoly outcome or a competitive outcome. Intuitively, (2) is likely to bind when services are substitutable. Formally, suppose $T(\tau_1, \ldots, \tau_K) = (1-\sigma)\sum_{k\in K} \tau_k + \sigma \max(\tau_1, \ldots, \tau_K)$, where $\sigma \in [0, 1]$ is the degree of substitutability. The main insight of this paper holds when the services provided by intermediaries are not too substitutable:

Claim 6. *Suppose $\Pi > L + c$. There is a $\sigma^* \in (0, 1)$ such that (i) if $\sigma < \sigma^*$, there is an SPE in which one intermediary earns a monopoly profit and the consumer obtains a payoff of zero, and* *(ii) if $\sigma > \sigma^*$, there is an SPE in which intermediaries earn zero profits and the consumer receives full surplus $\Pi - L - c$.*

Proof. We can write (1) and (2) respectively as

$$\tau^* \geq L + c \quad \text{and} \tag{3}$$

$$(1-\sigma)(\tau^* + \Pi) + \sigma \max(\tau^*, \Pi) - \Pi \geq c. \tag{4}$$

The left-hand side of (4) is strictly decreasing in σ, so it is more likely to bind for a larger σ. If $\sigma = 0$, then (3) implies (4), in which case (3) binds (at $\tau^* = L + c$). If $\sigma = 1$, then $\tau^* = \Pi + c$, and only (4) binds. As a result, we obtain a cutoff σ^* with the desired properties. □

7.3 Mixed Strategy Equilibrium

So far, I have focused on pure strategy equilibrium. This subsection considers mixed strategy equilibrium (MSE), assuming that the consumer holds one unit of data, and data collection is harmful, i.e., $L := -U(\{d\}) > 0$. I focus on a symmetric MSE in which (i) each intermediary draws compensation from the same distribution, (ii) data collection occurs with a positive probability, and (iii) the consumer rejects a compensation of zero.[17] The main insight extends to this MSE (see Appendix J for the proof).

Claim 7. *Take any $K \geq 2$ and $\Pi > L > 0$. In any equilibrium that satisfies (i) - (iii) above, consumer surplus is at most half of total surplus. This bound is tight: The ratio of consumer surplus to total surplus converges to $\frac{1}{2}$ as $\frac{L}{\Pi} \to 1$.*

In the case of rivalrous goods, the consumer extracts full surplus of the efficient outcome. Therefore, Claim 7 implies that across all parameters, consumer surplus in this mixed strategy equilibrium is less than half of that in the case of rivalrous goods.

[17] The pure strategy equilibrium of the baseline analysis satisfies (ii) and (iii) when $L > 0$. Without these restrictions, there will be other equilibria in which the consumer breaks ties arbitrarily when she faces offers with zero compensation on-path.

7.4 Differentiated Intermediaries

So far, I have assumed homogeneous intermediaries. This subsection considers intermediaries that are ex ante differentiated in the upstream and downstream markets. I continue to assume that the consumer holds one unit of data, and data collection is harmful, i.e., $L := -U(\{d\}) > 0$.

Suppose that the consumer shares her data with intermediaries in K_C and receives total compensation τ, and the firm obtains $D \subset \{d\}$. Then the consumer's payoff is $U(D) + \tau - \sum_{k \in K_C} c_k$, where $c_k > 0$ is the exogenous cost of sharing data with intermediary k. An intermediary with a low c_k can collect data at low compensation.

Second, if the downstream firm does not buy the data, it obtains a payoff of zero. If the firm buys data d from intermediaries in $K_F \subset K$ at a total price of p, it receives a payoff of

$$\Pi + \sigma \max_{k \in K_F} \Delta_k + (1 - \sigma) \sum_{k \in K_F} \Delta_k - p,$$

where $\Pi > 0$, $\sigma \in [0, 1]$, and $\Delta_k \geq 0$ for each $k \in K$. The first term Π is the base value of the data. The parameters σ and $(\Delta_k)_{k \in K}$ respectively capture the substitutability and qualities of intermediaries in the downstream market. To see this, assume $K = 2$. A higher $\Delta_1 - \Delta_2$ implies that the offering of intermediary 1 has a higher quality than that of intermediary 2, and a lower σ implies that the offerings of two intermediaries are less substitutable. For example, suppose two intermediaries have the same underlying data and offer a similar targeting compaign, but intermediary 1's campaign has a higher accuracy. Then $\Delta_1 > \Delta_2$ and σ is high. As another example, suppose intermediaries 1 and 2 use the same data to offer targeting and fraud detection, respectively, and their services have high quality. Then Δ_1 and Δ_2 are high, and σ is low.

The following result provides conditions under which one intermediary, possibly an inefficient one, acts as a monopolist in equilibrium. To obtain a nontrivial result, we assume $\Pi + \Delta_k > L + c_k$ for some $k \in K$ (see Appendix K for the proof).

Claim 8. *There is a $\sigma_1 < 1$ such that for any $\sigma \in [\sigma_1, 1]$, the following holds: There is an efficient equilibrium in which intermediary $k^* \in \arg\max_{k \in K} \Delta_k - c_k$ acts as a monopolist. If there is a $\hat{k} \notin \arg\max_{k \in K} \Delta_k - c_k$ such that $\hat{k} \in \arg\max_{k \in K} \Delta_k$ and $\Pi + \Delta_{\hat{k}} > L + c_{\hat{k}}$, there is also an inefficient equilibrium in which intermediary $\hat{k}$ acts as a monopolist.*

The result implies that if the degree of downstream differentiation is sufficiently is small, there is a monopoly equilibrium. The result also shows an inefficient equilibrium: An intermediary with a high c_k can monopolize the market, if it offers the highest value to the firm. This inefficiency stems from the nonrivalry of data: In the rivalrous-goods counterpart, the intermediary that generates the highest total surplus obtains the goods in any equilibrium.

The inefficient equilibrium points to a challenge for emerging "personal data marketplaces," such as Killi and Hu-manity.co.[18] These companies claim to provide greater transparency and privacy protection to consumers. We may interpret such companies as intermediaries that have a lower c_k than existing platforms or data brokers, which may collect data at the cost of a potential data breach and misuse (e.g., Cambridge Analytica scandal). The existence of an inefficient equilibrium suggests that if the efficiency advantage of those new companies comes from a low c_k and not from a high Δ_k, they may fail to replace less efficient incumbents.

Claim 8 focuses on a high σ; if σ is low, there may not be a monopolistic equilibrium. However, for a sufficiently high Δ_k's, the main insight—that the nonrivalry of data reduces consumer welfare—continues to hold.

Claim 9. *Fix any $\sigma \in [0, 1)$. There is a $\Delta^* > 0$ such that if $\min_{k \in K} \Delta_k \geq \Delta^*$, there is an equilibrium in which all intermediaries collect data and the consumer obtains a payoff of zero.*

Proof. Consider the following strategy profile: Each intermediary k offers $c_k + \frac{L}{K}$ to the consumer, who provides data to all intermediaries. If an intermediary deviates and increases compensation, the consumer continues to share her data with all intermediaries. If an intermediary deviates and decreases compensation, the consumer does not share her data with any intermediary. For any $(\Delta)_{k \in K}$, the consumer's strategy is optimal. For a large Δ^*, each intermediary earns a payoff of at least $(1 - \sigma)\Delta^* - c_k - \frac{L}{K} > 0$, and thus each intermediary's strategy is optimal. □

7.5 Multiple Consumers with Data Externalities

The baseline model assumes a single consumer. Appendix L studies an extension in which there are multiple consumers, and the payoff of each consumer may depend on what data other consumers

[18] See https://techcrunch.com/2018/07/18/hu-manity-wants-to-create-a-health-data-marketplac (accessed on March 20, 2020) and https://www.forbes.com/sites/curtissilver/2020/04/28/killis-fair-trade-data-program-enables-you-to-profit-off-your-data/ (accessed July 28, 2020).

provide to the firm. The main insight continues to hold in this setting: There is an equilibrium in which a single intermediary collects all data. Also, if the downstream firm's data usage harms consumers, there is a monopolistic equilibrium.

In this general setting, an equilibrium may be inefficient, because consumers do not consider how providing data affects the welfare of other consumers. As a result, consumer surplus can be lower than in the absence of intermediaries. These findings are in line with recent work on data externalities (Acemoglu et al., 2019; Bergemann et al., 2019; Choi et al., 2019).

7.6 Multiple Downstream Firms

We can incorporate multiple downstream firms, provided they do not interact with each other. Then we can define $\Pi(\cdot)$ as the sum of the revenues of all the firms, and $U(\cdot)$ as the aggregate effect of data acquisition. Appendix M formalizes this idea.

8 Conclusion

This paper studies competition between data intermediaries, which obtain data from consumers and sell them to downstream firms. The model incorporates two properties of personal data: Data are nonrivalrous, and the use of data by third parties can increase or decrease consumer welfare. The nonrivalry of data relaxes competition between intermediaries: If a downstream firm's data usage harms consumers, the equilibrium may coincide with the monopoly outcome. For general preferences, competition may benefit consumers but less than in the case of physical goods. These insights are robust to a number of extensions.

References

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar (2019), "Too much data: Prices and inefficiencies in data markets." Technical report, National Bureau of Economic Research.

Acquisti, Alessandro and Ralph Gross (2009), "Predicting social security numbers from public data." *Proceedings of the National academy of sciences*, 106, 10975–10980.

Acquisti, Alessandro, Curtis R Taylor, and Liad Wagman (2016), "The economics of privacy." *Available at SSRN 2580411*.

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

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2019), "The economics of social data." *Cowles Foundation Discussion Paper*.

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.

Bernheim, B Douglas and Michael D Whinston (1998), "Exclusive dealing." *Journal of political Economy*, 106, 64–103.

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

Bonatti, Alessandro and Gonzalo Cisternas (2020), "Consumer scores and price discrimination." *The Review of Economic Studies*, 87, 750–791.

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

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

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim (2019), "Privacy and personal data collection with information externalities." *Journal of Public Economics*, 173, 113–124.

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

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

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

De Cornière, Alexandre and Greg Taylor (2020), "Data and competition: a general framework with applications to mergers, market structure, and privacy policy."

Fainmesser, Itay P, Andrea Galeotti, and Ruslan Momot (2019), "Digital privacy." *Available at SSRN*.

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

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

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

Gartner (2016), "How to choose a data broker." URL https://www.gartner.com/smarterwithgartner/how-to-choose-a-data-broker/.

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 (2020), "Data brokers co-opetition." *Available at SSRN 3308384*.

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

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

Ichihashi, Shota (2020), "Online privacy and information disclosure by consumers." *American Economic Review*, 110, 569–95.

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

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

Kummer, Michael and Patrick Schulte (2019), "When private information settles the bill: Money and privacy in googles market for smartphone applications." *Management Science*, 65, 3470–3494.

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

Madio, Leonardo, Yiquan Gu, and Carlo Reggiani (2019), "Exclusive data, price manipulation and market leadership."

McAfee, R Preston and Marius Schwartz (1994), "Opportunism in multilateral vertical contracting: Nondiscrimination, exclusivity, and uniformity." *The American Economic Review*, 210–230.

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

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

Rey, Patrick and Jean Tirole (2007), "A primer on foreclosure." *Handbook of industrial organization*, 3, 2145–2220.

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

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

Sartori, Elia (2018), "Competitive provision of digital goods."

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

Segal, Ilya (1999), "Contracting with externalities." *The Quarterly Journal of Economics*, 114, 337–388.

Shapiro, Carl and Hal Varian (1998), *Information rules: a strategic guide to the network economy*. Harvard Business Press.

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

Tan, Guofu and Junjie Zhou (2020), "The effects of competition and entry in multi-sided markets." *The Review of Economic Studies*.

Zuboff, Shoshana (2019), *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Appendix

A Competition for Rivalrous Goods: Proof of Claim 2

Take any $K \geq 2$, and let $TS^* := \Pi(\mathcal{D}) + U(\mathcal{D})$ denote the efficient total surplus (because of Assumption 1). Because $\Pi(\cdot)$ is strictly increasing, in any equilibrium, the firm buys all (rivalrous) data collected by the intermediaries. Take any equilibrium, in which the consumer's equilibrium payoff is u^*. First, I show that all intermediaries and the firm earn a payoff of zero. Suppose to the contrary that some intermediary k^* or the firm obtains a positive payoff of $y^* > 0$. Suppose that intermediary $j \neq k^*$ offers $(\mathcal{D}, u^*+\varepsilon-U(\mathcal{D}))$ with $\varepsilon \in (0, y^*)$. The consumer accepts this offer and rejects other non-empty offers, and obtains a payoff of $u^* + \varepsilon$, because the goods are rivalrous (i.e., non-empty offers are the offers that ask for non-empty sets of data). The deviation of intermediary j increases the consumer's payoff by ε, reduces the sum of payoffs of intermediaries $k \neq j$ and the firm by at least y^*. Because the deviation weakly increases total surplus, intermediary j's payoff increases by at least $y^* - \varepsilon > 0$. This is a contradiction.

A similar argument implies that in any equilibrium, the firm buys a set of data that maximizes total surplus (otherwise, an intermediary can deviate in the upstream market). Thus, in any equilibrium, the consumer receives a payoff of TS^*. Finally, such an equilibrium exists: We can consider a strategy profile such that all intermediaries offer $(\mathcal{D}, \Pi(\mathcal{D}))$, and the consumer accepts one of them. $\square$

B Equilibrium for Single Unit Data: Proof of Proposition 1

Throughout the proof, I consider the following strategies in the downstream market: An intermediary sets a price of Π if it is the only one that holds d. An intermediary sets a price of zero if multiple intermediaries hold d. In either case, the firm buys data from all intermediaries. Any equilibrium of the downstream-market subgame is payoff-equivalent to this equilibrium.

First, I show that for any equilibrium in which the consumer sells data to at least one intermediary, the total compensation τ^* that she earns is weakly greater than $\max(0, -U)$. First, consider $U \geq 0$ and suppose to the contrary that $\tau^* < \max(0, -U) = 0$. This implies that all intermediaries that collect d charge positive fees (negative compensation), and the consumer provides data only to intermediary (say) k^* that charges the lowest fee $-\tau^* > 0$. However, intermediary $j \neq k^*$ can then offer $(\{d\}, \tau)$ with $\tau \in (\tau^*, 0)$, exclusively obtain d, and earn a positive profit. This is a contradiction. If $U < 0$, then $\tau^* \geq \max(0, -U) = -U$ holds; otherwise, the consumer would obtain a negative payoff, so she would not sell her data to any intermediary.

Second, I show that there is an equilibrium in which one intermediary collects data at compensation $\max(0, -U)$ and sets a downstream price of Π. Consider the following strategy profile: Intermediary (say) 1 offers $(\{d\}, \max(0, -U))$, and all other intermediaries offer $(\{d\}, 0)$. On the path of play, the consumer accepts the offer of intermediary 1 and rejects others. If intermediary k unilaterally deviates to $(\{d\}, \tau)$, then the consumer accepts a set K_C of offers such that (A) K_C maximizes her payoff and (B) if $k \in K_C$, then there is some $j \neq k$ with $j \in K_C$.

The proposed strategy profile is an SPE. First, no intermediary has a profitable deviation: Suppose intermediary k offers $(\{d\}, \tau)$. If $\tau < 0$, then the consumer rejects it, because another intermediary offers non-negative compensation. If $\tau \geq 0$, then the consumer may accept it, but she also accepts the offer of another intermediary. Then, the downstream price of data is zero. Thus, the deviation is not profitable. Second, the consumer's strategy is optimal. In particular, suppose that intermediary k deviates to a non-empty offer (i.e., $D_k = \{d\}$). Suppose also that K_C that satisfies Point (A) contains intermediary k. Then, the consumer can add any $j \neq k$ that offers non-negative compensation to K_C in order to satisfy Point (B). Adding j to K_C weakly increases the consumer's payoff because it weakly increases total compensation without affecting her gross payoff from data usage.

The above SPE maximizes the joint profit of intermediaries among all SPEs, because the consumer receives the minimum possible compensation, the firm obtains zero profit, and the outcome (i.e., the firm acquiring d) maximizes total surplus. Also, in this equilibrium one intermediary extracts this maximized joint profit. This implies that if there is another equilibrium that is Pareto-undominated from the perspectives of intermediaries, then in such an equilibrium, multiple intermediaries must be earning positive profits. However, there is no such equilibrium because an intermediary earns positive profit only by selling d to the firm at a positive price, which occurs only if one intermediary collects d.

The above arguments imply that in any equilibrium (i.e., any pure-strategy SPE that is Pareto-undominated for intermediaries), one intermediary collects d at compensation $\max(0 - U)$ and sets a price of Π to the firm. As a result, the consumer obtains a payoff of $\max(0, U)$ and the firm obtains a payoff of zero. If $U < 0$, this is a monopoly outcome in which all players except one intermediary receive zero payoffs. $\square$

C Partially Monopolistic Equilibrium: Proof of Proposition 2

Take any $D^* \in \arg\max_{D \subset \mathcal{D}} U(D)$. Consider the following strategy profile: In the upstream market, intermediary 1 offers $(\mathcal{D}, U(D^*) - U(\mathcal{D}))$. Other intermediaries offer $(D^*, 0)$. The consumer accepts only the offer of intermediary 1. If an intermediary deviates, then the consumer optimally decides which intermediaries to share data with, breaking ties in favor of sharing data. In the downstream market, if intermediary 1 obtains $\mathcal{D}$ 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 we assume that the players follow any equilibrium of the corresponding subgame. In the downstream market, this strategy profile consists of an equilibrium.

The suggested strategy profile is an equilibrium. First, intermediary 1 has no incentive to deviate. To see this, suppose intermediary 1 deviates and obtains data $D_1 \subset \mathcal{D}$. Let $\hat{D}$ denote the set of all data that the consumer shares as a result of intermediary 1's deviation ($D_1 \subsetneq \hat{D}$ if she also shares data with some intermediary $j \neq 1$). The revenue of intermediary 1 in the downstream market is at most $\Pi(\hat{D})$. The compensation τ to the consumer has to satisfy $\tau \geq U(D^*) - U(\hat{D})$. To see this, suppose $U(D^*) > U(\hat{D}) + \tau$. The left-hand side is the payoff that the consumer 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, $U(D^*) > U(\hat{D}) + \tau$ implies the consumer strictly prefers to reject the offer from intermediary 1. These bounds on revenue and cost imply intermediary 1's payoff after the deviation is at most $\Pi(\hat{D}) - [U(D^*) - U(\hat{D})] = \Pi(\hat{D}) + U(\hat{D}) - U(D^*)$. Assumption 1 implies this expression is at most $\Pi(\mathcal{D}) + U(\mathcal{D}) - U(D^*) = \Pi(\mathcal{D}) - [U(D^*) - U(\mathcal{D})]$, which is intermediary 1's payoff without deviation. Thus no deviation is profitable for intermediary 1.

Second, suppose intermediary 2 deviates and offers (D_2, τ_2). Without loss of generality, assume the consumer accepts the offer. Let D^{-1} denote the set of data the consumer provides to intermediaries in $K \setminus \{1\}$ after the deviation. If the consumer accepts the offer of intermediary 1 in addition to sharing D^{-1}, her payoff increases by $U(\mathcal{D}) - U(D^{-1}) + U(D^*) - U(\mathcal{D}) \geq 0$. The inequality follows from $U(D^*) \geq U(D^{-1})$. Thus, the consumer prefers to accept the offer of intermediary 1. If $\tau_2 \geq 0$, this implies that intermediary 2 could be better off (relative to the deviation) by not collecting D_2, because it can save compensation without losing revenue in the downstream market. Indeed, intermediary 2's revenue in the downstream market is zero. If $\tau_2 < 0$, the consumer 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. The optimality of each player's strategy on other nodes holds by construction. $\square$

D Welfare Properties of PME: Proof of Proposition 3

We prepare several notations. Define $U^* := \max_{D \subset \mathcal{D}} U(D)$, and $TS^* := \Pi(\mathcal{D}) + U(\mathcal{D}) \geq 0$. Assumption 1 implies TS^* is the maximum total surplus. Define $m := \min_{d \in \mathcal{D}, D \subset \mathcal{D}} \Pi(D) - \Pi(D \setminus \{d\}) > 0$. Let K^* satisfy $K^* > TS^*/m$. Suppose there are $K \geq K^*$ intermediaries, and take any equilibrium. Suppose to the contrary that the consumer's payoff is $U(D^*) - \delta$ with $\delta > 0$. I derive a contradiction by assuming that any intermediary obtains a payoff of at least m. If intermediary k deviates and offers (D^*, ε) with $\varepsilon \in (0, \delta)$, the consumer accepts this offer. Let D_{-k} denote the data the consumer shares with intermediaries in $K \setminus \{k\}$ as a result of k's deviation. Then, $D^* \setminus D_{-k} \neq \emptyset$ holds. To see this, suppose to the contrary that $D^* \subset D_{-k}$. Then, the consumer could be strictly better off by rejecting intermediary k's offer (D^*, ε) because $\varepsilon > 0$. However, conditional on rejecting k's deviating offer, the set of offers the consumer faces shrinks relative to the original equilibrium. Thus, the maximum payoff the consumer can achieve by rejecting k's deviating offer is at most $U(D^*) - \delta < U(D^*) - \varepsilon$, which is a contradiction. Because the consumer accepts the offer of intermediary k and $D^* \setminus D_{-k} \neq \emptyset$, intermediary k can earn a profit arbitrarily close to m. This implies that in the equilibrium, any intermediary earns a payoff of at least m. However, if each intermediary earns at least m, the sum of payoffs of all intermediaries is at least $Km > TS^*$. This implies that the consumer or the firm obtains a negative payoff, which is contradiction. Thus in any equilibrium, the consumer obtains a payoff of at least $U(D^*)$. The PME then minimizes the consumer's payoff across all pure-strategy SPE. Because the PME maximizes total surplus while giving the lowest payoffs to the consumer and the firm, it maximizes the intermediaries' joint profit for any $K \geq K^*$. $\square$

E Proof of Claim 3

Take any equilibrium, and let $(D_k)_{k\in K}$ denote the allocation of data (i.e., rivalrous goods). Without loss of generality, suppose $D_1 \neq \emptyset$. Suppose to the contrary that $D_k \neq \emptyset$ for some $k \neq 1$. Let τ_j denote compensation from each intermediary j. Suppose intermediary 1 offers $(\cup_{k\in K} D_k, \sum_{k\in K} \tau_k + \varepsilon)$ with $\varepsilon > 0$. Then, the consumer only accepts this offer. Thus, intermediary 1 earns a downstream revenue of $\Pi(\cup_{k\in K} D_k)$. Without intermediary 1's deviation, the joint downstream revenue is $\sum_{k\in K} [\Pi(\cup_{j\in K} D_j) - \Pi(D_{-k})]$, where $D_{-k} = \cup_{j\neq k} D_j$ (this follows from Lemma 1 below). By the same logic as the proof of Proposition 5 below, $\Pi(\cup_{k\in K} D_k) > \sum_{k\in K} [\Pi(\cup_{j\in K} D_j) - \Pi(D_{-k})]$ holds. Thus, intermediary 1 strictly benefits from the deviation with a sufficiently small $\varepsilon > 0$, which is a contradiction. $\square$

F Partitional Equilibria: Proof of Proposition 4

I first present the unique equilibrium outcome of the downstream market.[19]

Lemma 1. *Suppose $\Pi(\cdot)$ is submodular. Suppose each intermediary k has collected data D_k. In any pure-strategy subgame perfect equilibrium of the downstream market, intermediary k obtains a revenue of*

$$\Pi_k := \Pi\left(\bigcup_{j\in K} D_j\right) - \Pi\left(\bigcup_{j\in K\setminus\{k\}} D_j\right). \tag{5}$$

If $\Pi_k > 0$, then intermediary k sets a price of Π_k and the firm buys D_k with probability 1. The downstream firm obtains a payoff of $\Pi\left(\bigcup_{j\in K} D_j\right) - \sum_{k\in K} \Pi_k$.

Proof. Take any allocation of data $(D_1, \ldots, D_K)$. I show that there is an equilibrium (of the downstream market) in which each intermediary k posts a price of Π_k and the firm buys all data. First, the submodularity of Π implies that $\Pi(\cup_{k\in K'\cup\{j\}} D_k) - \Pi(\cup_{k\in K'} D_k) \geq \Pi_j$ for all $K' \subset$

[19] Lemma 1 generalizes Proposition 18 of Bergemann et al. (2019) in that the equilibrium payoff profile in the downstream market is shown to be unique even if $D_k \subset D_j$ for some k and $j \neq k$. Gu et al. (2020) assume $K = 2$ and consider not only submodularity but also supermodularity.

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.

To prove the uniqueness of equilibrium payoffs, I first 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 $K^* := \{\ell \in K : \ell \notin K', p_\ell = 0\} \cup K'$. 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 $j \notin K^*$. Then, it holds that

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

To see this, suppose that one side is greater than the other. If the left-hand side is strictly greater, then intermediary 1 can profitably deviate by slightly increasing its price. If the right hand side is strictly greater, then the firm would not buy D_1. In either case, we obtain a contradiction.

Let J^* denote a solution of the right hand side of (6). I consider two cases. First, suppose that there exists some $j \in J^* \setminus K^*$. By the construction of K^*, $p_j > 0$. Then, intermediary j can profitably deviate by slightly lowering p_j. To see this, note that

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

where $\hat{p}_k = p_k$ for all $k \neq j$ and $\hat{p}_j = p_j - \varepsilon > 0$ for a small $\varepsilon > 0$. This implies that after the deviation by intermediary j, the firm buys data D_j. This is because the left-hand side of (7) is the maximum revenue that the firm can obtain if it cannot buy data D_j, and the right hand side is the lower bound of the revenue that the firm can achieve by buying D_j. Thus, the firm always buy data

D_j, which is a contradiction.

Second, suppose that $J^* \backslash K^* = \emptyset$, i.e., $J^* \subset K^*$. This implies that the right hand side of (6) can be maximized by $J^* = K^* \setminus \{1\}$, because Π is submodular and $\Pi(\cup_{k \in K^*} D_k) - \Pi(\cup_{k \in K^* \setminus \{\ell\}} D_k) \geq p_\ell$ for all $\ell \in K^*$. Plugging $J^* = K^* \setminus \{1\}$, we obtain

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

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

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

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

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

By submodularity, this implies that

$$\Pi(\cup_{k \in K \setminus \{1\}} D_k) = \Pi(\cup_{k \in K \setminus \{1\}} D_k).$$

Then, we can write (8) as

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

which implies $\Pi_1 = p_1$. This is a contradiction. Thus, there must be $j \notin K^*$ such that (9) holds. Such intermediary j can again profitably deviate by lowering its price, which is a contradiction. Therefore, intermediary k's revenue is at most Π_k.

Next, I show that in equilibrium, each intermediary k gets a revenue of at least Π_k. This follows from the submodularity of 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.

Finally, the payoff of the downstream firm is $\Pi(\cup_{k\in K} D_k) - \sum_{k\in K} \Pi_k$, because the firms' gross revenue from data is $\Pi(\cup_{k\in K} D_k)$ whereas it pays Π_k to each intermediary k. □

I now prove Proposition 4.

Proof of Proposition 4. We begin with proving the second part. Take any $(D_k^*, \tau_k^*, p_k^*)_{k\in K}$ that satisfies Points 1 - 3 of Proposition 4. Consider the following strategy profile: Each intermediary k offers (D_k^*, τ_k^*) and sets the price of data following Lemma 1 (if $\Pi_k = 0$, then k sets a price of zero). On the path of play, the consumer accepts all offers. After a unilateral deviation of an intermediary, the consumer accepts all offers from non-deviating intermediaries and decides whether to accept the deviating offer, breaking a tie in favor of acceptance.

I show that this strategy profile is an equilibrium. First, the strategy of the consumer is optimal because $U(\cdot)$ is decreasing and submodular. Second, Lemma 1 implies that there is no profitable deviation in the downstream market. Third, suppose that intermediary k deviates and offers $(\tilde{D}_k, \tilde{\tau}_k)$. Without loss of generality, we can assume that $\tilde{D}_k \subset D_k^*$ for the following reason. If the consumer rejects $(\tilde{D}_k, \tilde{\tau}_k)$, then intermediary k can replace such an offer with $(\emptyset, 0)$. If the consumer accepts $(\tilde{D}_k, \tilde{\tau}_k)$ but $\tilde{D}_k \subsetneq D_k^*$, it means that k obtains some data $d \in \tilde{D}_k \setminus D_k^*$. Because $\cup_k D_k^* = \mathcal{D}$, there is another intermediary that obtains data d. By Lemma 1, intermediary k is indifferent between offering $(\tilde{D}_k \setminus \{d\}, \tilde{\tau}_k)$ and $(\tilde{D}_k, \tilde{\tau}_k)$. Now, let $D^- := D_k^* \setminus \tilde{D}_k$ denote the set of data that are not acquired by the firm as a result of k's deviation. If intermediary k deviates in this way, its revenue in the downstream market decreases by $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D_k^*) - [\Pi(\mathcal{D} \setminus D^-) - \Pi(\mathcal{D} \setminus D_k^*)] = \Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D^-)$. In the upstream market, if the consumer provides data $\tilde{D}_k$ to k, then it is optimal for the consumer to accept other offers from non-deviating intermediaries, because $U(\cdot)$ is submodular. This implies that the minimum compensation that k has to pay is $U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D} \setminus D^-)$. Thus, k's compensation in the upstream market decreases by $U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}) - [U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D} \setminus D^-)] = U(\mathcal{D} \setminus D^-) - U(\mathcal{D})$. Because collecting $\mathcal{D}$ is an optimal choice of the monopolist, it holds that $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D^-) - [U(\mathcal{D} \setminus D^-) - U(\mathcal{D})] \geq 0$. Therefore, the deviation does not strictly increase intermediary k's payoff.

Next, we prove the first part. Points 1 and 3 follow from the definition of partitional equilibrium and Lemma 1, respectively. Let τ_k^* denote the compensation k pays for collecting D_k^*. To show Point 2, suppose to the contrary that $\tau_k^* \neq U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}) > 0$. Suppose that $\tau_k^* < U(\mathcal{D} \setminus$

$D_k^*) - U(\mathcal{D})$. Then, the consumer rejects the offer from intermediary k, which is a contradiction. Next, suppose $\tau_k^* > U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. Then, by the second part of the proposition proved above, we can find an equilibrium that has the same outcome except intermediary k offers a lower compensation $\tau_k' \in (U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}), \tau_k^*)$ for collecting D_k^*. This equilibrium Pareto-dominates the original equilibrium from the perspectives of intermediaries, which is a contradiction. Thus, we obtain $\tau_k^* = U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. □

G Data Concentration: Proof of Proposition 5

Proof. Let $(\hat{D}_k)_{k \in K}$ and $(D_k)_{k \in K}$ denote two partitions of $\mathcal{D}$ such that the former is more concentrated than the latter. 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 $\Pi(\cdot)$. 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. By replacing Π with $-U$, we can show that the consumer receives a lower total compensation in a more concentrated equilibrium. This completes the proof. □

H Multiple Consumers with a Rich Contract Space: Proof of Claim 4

Proof. We prepare notations. For $L \subset \mathcal{L}$ and $\tau \in \mathbb{R}$, let $\gamma^*(L, \tau)$ denote the contract such that (i) it pays compensation τ regardless of the realized quantity vector, whenever any consumer i shares data $\left\{d_i^\ell\right\}_{\ell \in L}$, and (ii) it pays zero compensation if consumer i shares data other than $\left\{d_i^\ell\right\}_{\ell \in L}$. For any $L \subset \mathcal{L}$, define $D_L = \left\{d_i^\ell\right\}_{\ell \in L, i \in [0,1]}$, which is the dataset consisting of data labels in L on all consumers. When each consumer i shares her data $\left\{d_i^\ell\right\}_{\ell \in L}$ to the same intermediary (say) k, we simply say "consumers share data D_L with intermediary k" or "consumers share data L with intermediary k." Finally, let L^{**} denote a maximizer of $U(\cdot)$.

Consider the following strategy profile. Intermediary (say) 1 offers $\gamma^*(\mathcal{L}, U(L^{**}) - U(\mathcal{L}))$, and any other intermediary offers $\gamma^*(L^{**}, 0)$ to all consumers. On the path of play, each consumer i provides all of her data $\mathcal{L}$ to only intermediary 1, and rejects other offers. Off the path of play, consumers act as follows. First, suppose intermediary $k \neq 1$ unilaterally deviates and offers a contract $\hat{\gamma}$. Let $q_{-k}^1 \in \mathcal{Q}_{-k}$ denote the quantity vector such that intermediary 1 holds all data, and any intermediary $j \neq 1$ has no data. Take any $L^* \in \arg\max_{L \subset \mathcal{L}} \hat{\gamma}(L, q_{-k}^1)$. Following the deviation of intermediary $k \neq 1$, consumers continue to share all data $\mathcal{L}$ with intermediary 1, and share data L^* with intermediary k. Second, after intermediary 1's deviation in the upstream market, I assign any equilibrium in which all consumers act in the same way (e.g., all consumers first provide data L^{**} to intermediary $j \neq 1$, then provide the set of data L to intermediary 1 in order to maximize the payoff, given the contract chosen by intermediary 1). The interaction in the downstream market remains the same; in particular, if intermediary k has all data $\mathcal{D}$ and j has data D, then j posts a price of zero, k posts a price of $\Pi(\mathcal{D}) - \Pi(D)$, and the firm buys data from k. Finally, we need to specify intermediaries' beliefs regarding whose data other intermediaries hold. However, we can choose arbitrary beliefs on any information sets that can reach after no or unilateral deviation. This strategy profile satisfies the conditions described in the proposition.

I show that the suggested strategy profile is an equilibrium. On the path of play, if consumer i shares $\mathcal{L}$ with intermediary 1, she obtains a net payoff of $U(L^{**}) \geq 0$. If she deviates, her payoff is either zero or $U(L^{**})$. Thus, it is optimal for i to share all data $\mathcal{L}$ with intermediary 1 and share nothing with others.

Second, consider the incentive of consumers after intermediary $k \neq 1$ deviates and offers a contract $\hat{\gamma}$. Because consumers are atomless, the action of each consumer does not affect the quantity vector. Thus, each consumer chooses what data to share, taking a compensation schedule (as a function of the set of data she provides) as exogenous. As a result, it is optimal for each consumer to share all data $\mathcal{L}$ with intermediary 1 and data L^* with intermediary k (recall that $L^* \in \arg\max_{L \subset \mathcal{L}} \hat{\gamma}(L, q^1_{-k})$).

Third, I show that intermediary 1 has no incentive to deviate. Suppose intermediary 1 deviates in the upstream market. Let $\hat{L}$ denote the set of data labels that consumers share as a result of 1's deviation. The set $\hat{L}$ is well-defined, because we construct the strategy profile so that all consumers act in the same way after intermediary 1's deviation. The revenue of intermediary 1 in the downstream market is at most $\Pi(D_{\hat{L}})$. Compensation τ from intermediary 1 to each consumer has to satisfy $\tau \geq U(L^{**}) - U(\hat{L})$. To see this, suppose $U(L^{**}) > U(\hat{L}) + \tau$. The left-hand side is the payoff a consumer can attain by sharing data exclusively with intermediary $k \neq 1$. The right-hand side is her maximum payoff conditional on sharing data with intermediary 1. Then, this inequality implies that the consumer would strictly prefer to reject the offer from intermediary 1. Now, these bounds on revenue and cost imply that intermediary 1's payoff after the deviation is at most $\Pi(D_{\hat{L}}) + U(\hat{L}) - U(L^{**})$. Because of Assumption 3, this is at most $\Pi(D_{\mathcal{L}}) + U(\mathcal{L}) - U(L^{**})$, i.e., intermediary 1's payoff without deviation. Thus there is no profitable deviation for intermediary 1 in the upstream market.

Finally, suppose intermediary $k \neq 1$ deviates in the upstream market. Because consumers continue to provide all data to intermediary 1, the set of data that intermediary k can acquire will be a subset of intermediary 1's data. Thus, intermediary k earns zero revenue in the downstream market. As a result, the only way intermediary k earns a positive profit from the deviation is that the consumers pay to intermediary k. However, given any deviation by intermediary k, consumers share data $L^* \in \arg\max_{L \subset \mathcal{L}} \hat{\gamma}(L, q^1_{-k})$. Because $\hat{\gamma}(\emptyset, q^1_{-k}) = 0$, a transfer from k to consumers (i.e., $\hat{\gamma}(L^*, q^1_{-k})$) is non-negative. Thus, intermediary k has no profitable deviation in the upstream market. $\square$

I Non-Additive Compensation: Proof of Claim 5

Proof. Because $\lim_{x \to \infty} T(x, 0, \ldots, 0) = \infty$ and T is increasing, there is some τ that satisfies (1) and (2). The function T is continuous, so τ^* is well-defined. Because $T(0, \ldots, 0) = 0$, at least one of the inequalities binds at τ^* for any $c \geq 0$. We consider three cases.

(Case 1: Only inequality (1) binds) In this case, $\tau^* \leq \Pi$. Consider the following strategy profile: Intermediary (say) 1 offers $(\{d\}, \tau^*)$, and all other intermediaries offer $(\{d\}, 0)$. On the path of play, the consumer accepts the offer of intermediary 1 and rejects others. If intermediary ℓ unilaterally deviates to $(\{d\}, \tau)$, then the consumer accepts a set K_C of offers to maximize her payoff.

The proposed strategy profile is an SPE. The consumer prefers to provide data d to only intermediary 1, because $T(\tau^*, 0, \ldots, 0) = L + c$ and other intermediaries offer zero compensation. If intermediary 1 increases or decreases compensation, then either it pays more compensation for the same data, or the consumer rejects the offer. Next, suppose intermediary $k \neq 1$ deviates and offers $(\{d\}, \tau)$. Assume $\tau \leq \Pi$; otherwise, the deviation cannot be profitable for k. Because $T(\tau^*, \Pi, 0, \ldots, 0) - T(0, \Pi, 0, \ldots, 0) > c$ and T is submodular, we have $T(\tau^*, \tau, 0, \ldots, 0) - T(0, \tau, 0, \ldots, 0) > c$. As a result, whenever the consumer accepts the offer of intermediary k, she also accepts the offer of intermediary 1, in which case the downstream price of data is zero. Thus, k's deviation cannot be profitable. As a result, there is an SPE in which intermediary 1 earns a monopoly profit.

(Case 2: Inequality (2) binds and $\tau^* > \Pi$) Consider the following strategy profile: Intermediaries 1 and 2 offer $(\{d\}, \Pi)$, and other intermediaries offer $(\{d\}, 0)$. On the path of play, the consumer accepts the offer of intermediary 1 and rejects others. If intermediary ℓ unilaterally deviates to $(\{d\}, \tau)$, then the consumer accepts a set K_C of offers to maximize her payoff, breaking ties in favor of acceptance.

The proposed strategy profile is an SPE. If intermediary 1 increases compensation, it will pay more compensation for the same data. If it decreases compensation to $\tau < \Pi$, then the consumer will accept the offer of intermediary 2; thus, the deviation cannot be profitable for intermediary 1. Similarly, other intermediaries cannot increase its payoff by increasing or decreasing compensation. On the equilibrium path, the consumer only accepts the offer of intermediary 1, because

$T(\tau^*, \Pi, 0, \ldots, 0) - T(0, \Pi, 0, \ldots, 0) = c$ implies $T(\Pi, \Pi, 0, \ldots, 0) - T(0, \Pi, 0, \ldots, 0) < c$. In this equilibrium, all intermediaries earn zero profits, and the consumer obtains a payoff of $T(\Pi, 0, \ldots, 0) - L - c > 0$.

(Case 3: Inequality (2) binds and $\tau^* \leq \Pi$) Consider the following strategy profile: Intermediary 1 offers $(\{d\}, \Pi)$, and intermediary 2 offers $(\{d\}, \tau^*)$. All other intermediaries offer $(\{d\}, 0)$. On the path of play, the consumer accepts the offer of intermediary 1 and rejects others. If intermediary ℓ unilaterally deviates to $(\{d\}, \tau)$, then the consumer accepts a set K_C of offers to maximize her payoff, breaking ties in favor of acceptance.

The proposed strategy profile is an SPE. If intermediary 1 increases compensation, it will pay more compensation for the same data. Suppose it decreases compensation to $\tau < \Pi$. Then, the consumer will accept the offer of intermediary 2 whenever she accepts the offer of intermediary 1, because we have $T(\tau^*, \tau, 0, \ldots, 0) - T(0, \tau, 0, \ldots, 0) = T(\tau, \tau^*, 0, \ldots, 0) - T(\tau, 0, \ldots, 0) \geq c$, because of submodularity. Thus, the deviation cannot be profitable for intermediary 1. Similarly, other intermediaries cannot increase its payoff by increasing or decreasing compensation. Indeed, the only case in which intermediary $k \neq 1$ exclusively obtains data is when it offers compensation strictly greater Π (i.e., intermediary 1's offer), but then it will earn a negative payoff. In this equilibrium, all intermediaries earn zero profits, and the consumer obtains a positive payoff. $\square$

J Mixed Strategy Equilibrium: Proof of Claim 7

Proof. In this proof, "equilibrium" refers to subgame perfect equilibrium that satisfies (i), (ii), and (iii) in Section 7.3. Let F^* denote the equilibrium distribution of compensation, and let S^* denote its support. First, I show that we can without loss of generality assume $S^* = \{0, L\}$. The reason is as follows. Suppose, to the contrary, that there is $x \in S^*$ such that $x > L$. The only event in which intermediary k can earn a non-negative profit by paying x is when all other intermediaries offer compensation equal to or less than zero. On such an event, intermediary k can still obtain data by offering compensation $x - \varepsilon$ for a small $\varepsilon > 0$, which is a contradiction. Also, because the consumer will reject any negative compensation on and off the paths of play, we can replace any negative compensation with zero compensation, and the resulting strategy profile continues to be an equilibrium. Finally, suppose there is $x \in S^*$ such that $x \in (0, L)$. x does not fully cover the consumer's loss of sharing data. Thus, the only event in which intermediary k obtains data by paying x is when another intermediary pays positive compensation. But the data will then have a downstream price of zero, so intermediary k will earn negative profits in such an event. Thus, if $S^* \cap (0, L) \neq \emptyset$, the consumer must reject any offer in equilibrium. However, this contradicts the condition (ii).

The above argument enables us to focus on an equilibrium in which each intermediary offers compensation L and 0 with probabilities p and $1 - p$, respectively, and the consumer accepts all offers with compensation L. The only event in which an intermediary earns a positive payoff is when all other intermediaries offer zero compensation, which occurs with probability $(1 - p)^{K-1}$. If so, the intermediary earns a downstream revenue of Π. Thus, intermediary k's indifference condition is $(1 - p)^{K-1}\Pi - L = 0$. Solving this equation, we obtain $p^* = 1 - r^{\frac{1}{K-1}}$, where $r = \frac{L}{\Pi} \in (0, 1)$.

We show consumer surplus is at most half of total surplus. Because intermediaries earn zero profits in equilibrium, it suffices to prove that the downstream firm receives at least half of total surplus. The ratio S_F of the downstream firm's revenue to total surplus is

$$S_F = \frac{\left(1 - r^{\frac{K}{K-1}} - K(1 - r^{\frac{1}{K-1}})r\right)\Pi}{(\Pi - L)(1 - r^{\frac{K}{K-1}})} = \frac{1 - r^{\frac{K}{K-1}} - K(1 - r^{\frac{1}{K-1}})r}{(1 - r)(1 - r^{\frac{K}{K-1}})} = \frac{1 - rK + (K-1)r^{\frac{K}{K-1}}}{(1 - r)(1 - r^{\frac{K}{K-1}})}.$$

I show $S_F \geq 1/2$ for all $(r, K) \in [0, 1] \times (\mathbb{N} \setminus \{1\})$. We have

$$\begin{aligned}
&\frac{1 - rK + (K-1)r^{\frac{K}{K-1}}}{(1 - r)(1 - r^{\frac{K}{K-1}})} \geq \frac{1}{2} \\
\iff& 2 - 2rK + 2(K-1)r^{\frac{K}{K-1}} \geq (1 - r)(1 - r^{\frac{K}{K-1}}) \\
\iff& 2 - 2rK + 2(K-1)r^{\frac{K}{K-1}} \geq 1 - r^{\frac{K}{K-1}} - r + r^{\frac{2K-1}{K-1}} \\
\iff& 1 + r(1 - 2K) + r^{\frac{K}{K-1}}(2K - 1) - r^{\frac{2K-1}{K-1}} \geq 0. \qquad (11)
\end{aligned}$$

To show inequality (11), define $f(r, K) = 1 + r(1 - 2K) + r^{\frac{K}{K-1}}(2K - 1) - r^{\frac{2K-1}{K-1}}$. We have

$$\frac{\partial f}{\partial r} = 1 - 2K + (2K - 1) \cdot \frac{K}{K-1} \cdot r^{\frac{1}{K-1}} - \frac{2K-1}{K-1} r^{\frac{K}{K-1}} = 1 - 2K + \frac{2K-1}{K-1} \cdot r^{\frac{1}{K-1}}(K - r).$$

$\frac{\partial f}{\partial r} = 0$ at $r = 1$. To show $\frac{\partial f}{\partial r}$ is non-positive, we prove $g(r) = (K - r) \cdot r^{\frac{1}{K-1}}$ is increasing in r:

$$\frac{\partial g}{\partial r} = -r^{\frac{1}{K-1}} + \frac{K-r}{K-1} \cdot r^{\frac{2-K}{K-1}} = r^{\frac{1}{K-1}} \cdot \left(\frac{K-r}{K-1} \cdot r^{\frac{1-K}{K-1}} - 1 \right) \geq r^{\frac{1}{K-1}} \cdot \left(\frac{K-1}{K-1} \cdot \frac{1}{r} - 1 \right) \geq 0.$$

Thus, $\frac{\partial f}{\partial r} \leq 0$ for any (r, K). Because $f(r, K)$ is non-increasing in r, (11) holds if and only if it holds for $r = 1$. The inequality (11) holds for $r = 1$ because $f(1, K) = 0$. To show $\lim_{r \to 1} S_F = 1/2$, we apply L'Hospital's rule twice, then obtain

$$\lim_{r \to 1} \frac{1 - rK + (K-1)r^{\frac{K}{K-1}}}{(1-r)(1 - r^{\frac{K}{K-1}})} = \lim_{r \to 1} \frac{-K + Kr^{\frac{1}{K-1}}}{-1 - \frac{K}{K-1} r^{\frac{1}{K-1}} + \frac{2K-1}{K-1} r^{\frac{K}{K-1}}} = \lim_{r \to 1} \frac{\frac{K}{K-1} r^{\frac{2-K}{K-1}}}{-\frac{K}{(K-1)^2} r^{\frac{2-K}{K-1}} + \frac{(2K-1)K}{(K-1)^2} r^{\frac{1}{K-1}}} = \frac{1}{2}.$$

□

K Differentiated Intermediaries: Proof of Claim 8

Proof. First, we construct an inefficient equilibrium. Without loss of generality, suppose $1 \in \arg\max_{k \in K} \Delta_k$ and $\Pi + \Delta_1 > L + c_1$, but $1 \notin \arg\max_{k \in K} \Delta_k - c_k$. Consider the following strategy profile: Intermediary 1 offers to collect d at compensation $L + c_1$. Other intermediaries offer zero compensation. On the path of play, the consumer provides her data only to intermediary 1. If any intermediary unilaterally deviates in the upstream market, the consumer optimally decides which intermediaries to share her data with, breaking a tie in favor of sharing data with intermediary 1. The equilibrium in the downstream market remains the same as Lemma 1, because the firm's gross revenue is a submodular function of datasets held by intermediaries.

We show this strategy profile is an equilibrium for a σ close to 1. The consumer's strategy is optimal by construction, and so are the strategies of intermediaries and the firm in the downstream market. If intermediary 1 deviates and lowers compensation, the consumer does not share data. Because $\Pi + \Delta_1 > L + c_1$, intermediary 1 is worse off. If intermediary 1 deviates and increases compensation, then the consumer shares the same data and obtains a higher compensation. Thus, intermediary 1 does not benefit from the deviation: If intermediary $k > 1$ deviates in the upstream market, then its profit is at most $(1-\sigma)\Delta_k - c_k$, which is negative for $\sigma > \max_k \left(1 - \frac{c_k}{\Delta_k}\right)$. Finally, this strategy profile is inefficient. Indeed, if intermediary $k^* \in \arg\max_{k \in K} \Delta_k - c_k$ exclusively collects data, then compared to the above outcome, total surplus increases by $\Delta_{k^*} - c_{k^*} - (\Delta_1 -$

$c_1) > 0$.

By the same argument as above, we can construct an equilibrium in which intermediary 1 acts as a monopolist when $1 \in \arg\max_{k \in K} \Delta_k - c_k$. In particular, if intermediary $k \neq 1$ with $\Delta_k \geq \Delta_1$ deviates and collects data, then its profit increases by at most $\sigma(\Delta_k - \Delta_1) + (1-\sigma)\Delta_k - c_k < \sigma[\Delta_k - c_k - (\Delta_1 - c_1)] + (1-\sigma)(\Delta_k - c_k)$. The right-hand side of this inequality is non-positive at $\sigma = 1$. As a result, for a σ close to 1, $\sigma(\Delta_k - \Delta_1) + (1-\sigma)\Delta_k - c_k < 0$, i.e., an intermediary $k \neq 1$ has no profitable deviation in the upstream market. Also, by the same argument as in the previous paragraph, intermediary $k \neq 1$ with $\Delta_k < \Delta_1$ has no profitable deviation for a σ close to 1. This equilibrium is efficient: If another intermediary k also collects data, total surplus increases by at most $\Delta_k - \Delta_1 - c_k < \Delta_k - c_k - (\Delta_1 - c_1) \leq 0$. Thus, for a σ close to 1, it is efficient that only intermediary 1 collects data. By taking the maximum of all the σ's found above, we obtain the desired cutoff σ_1. $\square$

L Multiple Consumers with Data Externalities: Appendix for Section 7.5

We extend the model as follows. Let $I \in \mathbb{N}$ denote the number and the set of consumers. Each consumer $i \in I$ has a finite set $\mathcal{D}_i$ of data. Define $\mathcal{D} := \cup_{i \in I}\mathcal{D}_i$ and $\mathcal{D}_{-i} := \cup_{j \in I \setminus \{i\}}\mathcal{D}_j$. If the firm acquires data $D \subset \mathcal{D}$, then consumer i obtains a gross payoff of $U_i(D_i, D_{-i})$, where $D_i = D \cap \mathcal{D}_i$ and $D_{-i} = D \cap \mathcal{D}_{-i}$. Intermediaries know $(U_i(\cdot,\cdot))_{i \in I}$ and can make different offers to different consumers. As before, the firm's revenue from data $D \subset \mathcal{D}$ is $\Pi(D)$, where $\Pi(\cdot)$ is an increasing set function such that $\Pi(\emptyset) = 0$.

The timing of the game is as follows. First, each intermediary $k \in K$ makes an offer (D_i^k, τ_i^k) to each consumer $i \in I$, where $D_i^k \subset \mathcal{D}_i$. Then, each consumer i privately observes $\{(D_i^k, \tau_i^k)\}_{k \in K}$, and chooses a set $K_i \subset K$ of offers to accept. The consumers' decisions determine the allocation of data, in which intermediary k holds $D^k := \cup_{i: k \in K_i} D_i^k$. After observing the allocation of data, each intermediary simultaneously posts a price for D^k. Finally, the firm decides from which intermediaries to buy data.

To accommodate the case in which $U_i(D_i, D_{-i})$ depends on D_{-i}, I modify the informational assumption and the solution concept. First, I assume each consumer i does not observe offers made to other consumers. Second, the solution concept is perfect Bayesian equilibrium in which consumers have passive beliefs—i.e., after consumer i detects deviations of intermediaries, she does not change her beliefs regarding what offers other consumers are receiving. The following result extends Proposition 1 (we do not require Assumption 1).

Claim 10. *Suppose $\mathcal{D}_i = \{d_i\}$ for each $i \in I$, and $\Pi(\cdot)$ is submodular. Take any set $D^M \subset \{d_1, \ldots, d_I\}$ of data a monopoly intermediary collects in some equilibrium. For any $K \geq 2$ and a partition $(D_1^M, \ldots, D_K^M)$ of D^M, there is an equilibrium in which each intermediary k collects each piece of data $d_i \in D_k^M$ at compensation $\max\left\{0, -U_i\left(\{d_i\}, D_{-i}^M\right)\right\}$, where $D_{-i}^M = D^M \cap \mathcal{D}_{-i}$.*

Proof. Take any set $D^M \subset \{d_1, \ldots, d_I\}$ of data a monopoly intermediary collects in some equilibrium. Take any partition $(D_1^M, \ldots, D_K^M)$ of D^M. I show that the following strategy profile is an equilibrium: For each $i \in I$, if $d_i \in D_k^M$, then intermediary k offers $\left(\{d_i\}, \max\{0, -U_i(d_i, D_{-i}^M)\}\right)$ to consumer i. If $d_i \notin D_k^M$, intermediary k offers $(\{d_i\}, 0)$ to i. On the path of play, each consumer i accepts the offer of intermediary k if and only if $d_i \in D_k^M$. After a deviation of any intermediary, a consumer chooses the set of offers to accept, breaking ties in favor of acceptance. The equilibrium in the downstream market follows Lemma 1.

The optimality of each consumer's strategy follows the proof of Proposition 1, with $U_i(\cdot)$ replaced by $U_i(\cdot, D_{-i}^M)$. Because consumers hold passive beliefs, after any deviation of an intermediary, consumer i maximizes her utility given $U_i(\cdot, D_{-i}^M)$.

Next, I show the optimality of each intermediary's strategy. First, it is not optimal for intermediary k to collect data d_i such that $d_i \notin D_k^M$, because consumer i will then share the same data with other intermediaries. Second, suppose intermediary k chooses to not collect d_i such that $d_i \in D_k^M$. This weakly decreases k's payoff if $U_i(d_i, D_{-i}^M) \geq 0$, because k collects d_i for free. Suppose $U_i(d_i, D_{-i}^M) < 0$. Not collecting d_i will reduce k's downstream revenue by at least $\Pi(D^M) - \Pi(D^M \setminus \{d_i\})$ and decreases compensation by $-U_i(d_i, D_{-i}^M) > 0$ (the downstream revenue may decrease by more than $\Pi(D^M) - \Pi(D^M \setminus \{d_i\})$ if intermediary k does not collect more than one d's such that $d \in D_k^M$). Because a monopoly intermediary finds it optimal to collect d_i under the passive beliefs, $\Pi(D^M) - \Pi(D^M \setminus \{d_i\}) \geq -U_i(d_i, D_{-i}^M)$. Thus, the strategy of each intermediary is optimal. $\square$

To extend other results, I modify Assumptions 1 and 2. I replace Assumption 1 with the assumption that $\Pi(\cdot)$ and $(U_i(\cdot))_{i\in I}$ are such that a monopoly intermediary collects and sells all data $\cup_{i\in I}\mathcal{D}_i$ in some equilibrium.

Claim 11 (Extending Proposition 2). *Under the modified version of Assumption 1, there is an equilibrium in which a single intermediary collects all data at compensation* $\max_{D_i\subset\mathcal{D}_i} U_i(D_i,\mathcal{D}_{-i}) - U_i(\mathcal{D}_i,\mathcal{D}_{-i})$ *for each* $i\in I$.

Proof. Take any $D_i^* \in \arg\max_{D\subset\mathcal{D}_i} U_i(D,\mathcal{D}_{-i})$. Consider the following strategy profile. Intermediary 1 offers $(\mathcal{D}_i, U_i(D_i^*,\mathcal{D}_{-i}) - U_i(\mathcal{D}_i,\mathcal{D}_{-i}))$, and intermediary $k\neq 1$ offers $(D_i^*, 0)$ to each consumer $i\in I$. On the path of play, each consumer i accepts the offer of intermediary 1. After a deviation of an intermediary, a consumer choose the set of offers to accept, breaking ties in favor of acceptance. Assign any equilibrium to each subgame of the downstream market. We can apply the proof of Proposition 2 by replacing $U_i(\cdot)$ with $U_i(\cdot,\mathcal{D}_{-i})$. □

The modified version of Assumption 1 is likely to hold if there are negative data externalities between many consumers. As Bergemann et al. (2019) show, the externality creates a gap between an intermediary's revenue from selling data and necessary compensation. Thus the externality renders it more likely that a monopoly intermediary collects all data (see Section 2).

We can also extend Proposition 4. To do so, I modify Assumption 2 so that for each $i\in I$ and $D_{-i}\subset\mathcal{D}$, $U_i(\cdot,D_{-i})$ is a decreasing submodular set function (and $\Pi(\cdot)$ is submodular). Then, the set of partitional equilibria is characterized by the allocation of data, compensation, and prices such that each intermediary compensates consumer i according to her marginal loss of sharing data calculated by $U_i(\cdot,\mathcal{D}_{-i})$. As in the original setting, consumers are worse off and intermediaries are better off in an equilibrium with a more concentrated allocation of data.

M Multiple Downstream Firms: Appendix for Section 7.6

Suppose that there are L firms, where firm $\ell\in L$ has revenue function Π^ℓ that depends only on data available to ℓ. The consumer gross payoff of sharing data is $\sum_{\ell\in L} U^\ell$, where each U^ℓ depends on the set of data firm ℓ obtains.

This setting reduces to the one with a single firm. For example, suppose intermediary 1 has all data $\mathcal{D}$ and intermediary 2 had $D\subset\mathcal{D}$. Then, intermediary 1 posts $\Pi_\ell(\mathcal{D})-\Pi_\ell(D)$ and intermediary 2 posts 0 to each firm ℓ. If each $\Pi_\ell(\cdot)$ is submodular, Lemma 1 implies that for any allocation of data, each intermediary k posts a price of $\Pi_\ell(\cup_k D_k) - \Pi_\ell(\cup_{j\neq k} D_j)$ to firm ℓ. In either case, this is as if the downstream market consists of one firm with revenue function $\sum_{\ell \in L} \Pi_\ell$.

Second, once the consumer shares her data with one intermediary, the data is sold to all firms. This means that in equilibrium, the consumer decides which offers to accept to maximize the sum of total compensation and $\sum_{\ell \in L} U^\ell(D)$. Therefore, we can apply the same analysis as before by defining $U := \sum_{\ell \in L} U^\ell$.

N Information Design by Data Intermediaries

I apply results in Section 6 to a setting in which the firm uses data to learn about a consumer's willingness to pay, and tailors pricing and product recommendation. The downstream firm is now a seller that provides $M \in \mathbb{N}$ products, $1, \ldots, M$. The consumer has a unit demand, and her values for products, $\mathbf{u} := (u_1, \ldots, u_M)$, are independently and identically distributed according to a cumulative distribution function F with a finite support $V \subset (0, +\infty)$.[20] Each $d \in \mathcal{D}$ is a signal (Blackwell experiment) from which the seller can learn about $\mathbf{u}$. The set $\mathcal{D}$ consists of all signals with finite realization spaces. Intermediaries can request the consumer any set of signals.[21] The consumer decides which offers to accept before observing $\mathbf{u}$.

After buying a set of data $D \subset \mathcal{D}$ from intermediaries, the seller learns about $\mathbf{u}$ from signals in D. Then, the seller sets a price and recommends one of M products to the consumer. Finally, the consumer observes the value and the price of the recommended product, and she decides whether or not to buy it.[22] A recommendation could be an advertiser displaying a targeted advertisement or an online retailer showing a product as a recommendation. If the consumer buys product m at price p, her payoff from this transaction is $u_m - p$. Otherwise, her payoff is zero. The seller's payoff is its revenue. I consider pure-strategy perfect Bayesian equilibrium such that, both on and

[20] 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.

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

[22] 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 (2020) for a detailed discussion of the motivation behind this formulation.

off the equilibrium paths, all players calculate its posterior beliefs based on the prior F, signals in D, and Bayes' rule.[23]

Importantly, Assumption 1 holds: If the seller has all data, it can access a fully informative signal and perfectly learn $\mathbf{u}$. The seller can then recommend the highest value product and perfectly price discriminate, which maximizes total surplus.

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. $\Pi(D)$ is an increasing set function because a larger D corresponds to a more informative signal. Define $p(F) := \min(\arg\max_{p \in V} p[1 - F(p)])$. $p(F)$ is the lowest monopoly price given a prior distribution F.

A monopoly intermediary can collect a fully informative signal (or any signal that achieves an efficient outcome) and extract full surplus from the consumer and the seller. In equilibrium, consumer surplus is $U(\emptyset)$, which is the payoff that the consumer would earn if the seller recommended a product randomly at a price of $p(F)$.

If there are multiple intermediaries, consumer surplus in the partially monopolistic equilibrium, $\max_{d \in \mathcal{D}} U(\{d\})$, is equal to the one in a hypothetical scenario where the consumer directly discloses information to the seller.

Proposition 6. *Suppose 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* $\max_{d \in \mathcal{D}} U(\{d\})$*. This equilibrium satisfies the following.*

1. *If the seller provides a single product (*$M = 1$*), then all intermediaries earn zero payoffs and the consumer receives full surplus created by data.*

2. *If the seller provides multiple products (*$M \geq 2$*), then 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.*[24]

[23] Thus, I omit the description of players' beliefs in the following results. I also assume that the seller breaks ties in favor of the consumer when the seller sets a price and recommends a product. The existence of an equilibrium is shown in Ichihashi (2020).

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

Proof. Note that Theorem 2 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 (2020) shows that if $M = 2$, then for a generic F satisfying $p(F) > \min V$, any signal $d^{**} \in \arg\max_{d \in \mathcal{D}} U(d)$ leads to an inefficient outcome. This implies $\Pi(d_{FULL}) + U(d_{FULL}) > \Pi(d^{**}) + U(d^{**}) \geq \Pi(\emptyset) + U(d^{**})$. Then, $\Pi(d_{FULL}) - \Pi(\emptyset) - [U(d^{**}) - U(d_{FULL})] > 0$. Thus, intermediary 1 earns a positive profit. $\square$