Competing Data Intermediaries

Shota Ichihashi*

March 23, 2020

Abstract

I study a model of competition among data intermediaries (e.g. online platforms and data brokers), 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, then consumers may share data with multiple intermediaries, which lowers the downstream price of data and hurts intermediaries. Anticipating this, intermediaries offer low compensation for data. Competing intermediaries can earn a monopoly profit if and only if firms' data acquisition unambiguously hurts consumers. The result explains, for example, why the market for personal data seems to be missing.

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] I study competition among such "data intermediaries," which collect and distribute personal data between consumers and downstream firms.

Concretely, consider online platforms that collect consumer data and share them with third parties. The use of data by third parties may hurt consumers via price discrimination and intrusive advertising. Alternatively, it may benefit consumers via improved products and personalized offerings. Depending on the sign of this effect, platforms offer compensation or charge a fee for collecting data from consumers. Compensation might be monetary transfers or non-monetary benefits such as online services (e.g. web mapping services).

The main question is whether competition among data intermediaries improves consumer and total surplus. Specifically, does competition incentivize intermediaries to offer consumers better compensation and lower fees? Does competition drive cost-inefficient intermediaries out of the market? This is a key question in recent policy debates on competition in digital markets (Crémer et al., 2019; Furman et al., 2019; Morton et al., 2019).

The baseline model consists of a consumer, data intermediaries, and a downstream firm. The consumer has a finite set of data (or data labels), say, 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, intermediaries post prices and sell collected data to the firm.

The model captures two features of personal data. First, data are non-rivalrous, that is, a consumer can provide the same data to multiple intermediaries. Second, the payoff of a consumer can depend non-monotonically on what data the downstream firm obtains. For example, the consumer may be comfortable with sharing only one of the state of birth and the date of birth. However, she may require compensation to share both, which allow the inference of her Social Security number

[1] Section 3 discusses these applications in detail.
[2] Subsection 3.1 motivates this assumption.

(Acquisti and Gross, 2009). The model is rich enough to capture such a situation.

Section 5 considers a consumer with one unit of data. Competing intermediaries sustain a monopoly outcome if and only if the downstream firm's data acquisition lowers consumer welfare. Moreover, even if competition benefits the consumer relative to monopoly, the magnitude of the benefit is smaller than in markets for rivalrous goods. This is because competition does not incentivize intermediaries to raise compensation: If multiple intermediaries offer high compensation, then the consumer shares her data with all of them, which lowers the downstream price of data. Also, the non-rivalry of data can create inefficiency in which a cost-inefficient intermediary excludes a more efficient one from the market.

Section 6 considers the consumer with any finite set of data. The consumer may benefit or lose depending on the set of data the downstream firm acquires. For general preferences, I characterize an equilibrium that maximizes intermediary profit and minimizes consumer surplus among all equilibria under a weak condition. The analysis shows that competition occurs only for a set of data that the firm uses to benefit the consumer. As a result, in this equilibrium, consumer surplus and intermediary profit 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 cost of sharing data, I characterize a class of equilibria with the following two properties. First, intermediaries collect mutually exclusive sets of data. Second, each intermediary, as a local monopsony, pays the consumer just enough compensation to cover her losses from sharing the data. I compare these equilibria in terms of the degrees of *data concentration*, and show that intermediaries are better off and the consumer is worse off in a more concentrated equilibrium. I connect this result with the welfare impact of "breaking up platforms."[3]

Finally, I use the results to study information design by data intermediaries. A downstream firm uses data for price discrimination and product recommendation. Intermediaries can potentially obtain any Blackwell experiments about the consumer's willingness to pay. In the intermediary-optimal but consumer-worst equilibrium described above, the resulting consumer surplus is equal to the one under hypothetical Bayesian persuasion (Kamenica and Gentzkow, 2011) in which the

[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

consumer directly discloses information to the firm.

The contribution of the paper is to clarify when competition among data intermediaries benefits consumers. I show that (i) competition does not work when consumers require positive compensation for sharing data, but (ii) it partially works when consumers benefit from data collection, in which case a monopoly intermediary would charge a fee. (i) and (ii) lead to the main insight that competition for data benefits consumers but not as much as in traditional markets. The results help us understand why consumers do not seem to be compensated properly for their data provision (Arrieta-Ibarra et al., 2018). The paper also highlights the limit of competition on improving efficiency: Proposition 2 potentially explains why incumbent data brokers may not be replaced by emerging "data marketplaces," even if the latter can collect data more efficiently.[4]

The rest of the paper is organized as follows. Section 2 discusses related works and Section 3 describes the model. Section 4 considers two benchmarks: a model of a monopoly intermediary, and a model of multiple intermediaries with rivalrous goods. Section 5 assumes that the consumer has one unit of data. I characterize equilibrium, and also show that a cost-inefficient intermediary may earn a monopoly profit even if there is a more efficient competitor. Section 6 allows general consumer preferences. I present the intermediary-best and consumer-worst equilibrium, which generalizes the case of a single unit data. With an additional assumption, I characterize a class of equilibria which tells us the impact of data concentration. Section 7 considers information design by data intermediaries. Section 8 provides extensions, and Section 9 concludes.

2 Literature Review

This paper relates to three strands of literature.

Markets for Data: Recent works such as Acemoglu et al. (2019) and Bergemann et al. (2019) consider models of data collection by platforms. In particular, Bergemann et al. (2019) study a model of data intermediaries, including the one with competing intermediaries. Their baseline model assumes that (i) data collection by intermediaries unambiguously hurts consumers, and (ii) under competition, different intermediaries exclusively access different pieces of data. Relative to

[4] See, for example, https://www.wired.com/story/i-sold-my-data-for-crypto/ (accessed on March 10, 2020).

(i), the consumer in my model may benefit or lose depending on what data are collected. Relative to (ii), I consider intermediaries that can collect the same set of data. These two points lead to the new insights: The magnitude of the benefit of competition on consumers depends on how downstream firms use data, and competition may not occur even if homogeneous intermediaries compete for the same data. The economic mechanism of my paper is amenable to but independent of data externality, which is the key idea of Acemoglu et al. (2019) and Bergemann et al. (2019).

The downstream market of my model relates to Gu et al. (2018). They study how data brokers' incentives to merge data depends on the downstream firm's revenue function. I abstract from contracting among intermediaries, but consider endogenous data collection in the upstream market. By modeling the upstream market, we can conduct consumer welfare analysis. Jones et al. (2018) consider a semi-endogenous growth model incorporating data intermediaries.

My paper considers pure data intermediaries, which only buy and sell data. Several works consider richer formulations of how online platforms monetize data. De Corniere and De Nijs (2016) study the design of an online advertising auction where a platform can use consumer data to improve the quality of match between consumers and advertisements. Fainmesser et al. (2019) study the optimal design of data storage and data protection policies by a monopoly platform. Choi et al. (2019) consider consumers' privacy choices in the presence of an information externality. Kim (2018) considers a model of a monopoly advertising platform and studies consumers' privacy concerns, market competition, and vertical integration between the platform and sellers. Bonatti and Cisternas (2020) study the aggregation of consumers' purchase histories and study how data aggregation and transparency impact a strategic consumer's incentives. De Cornière and Taylor (2020) employ competition-in-utilities approach to study the issue of data and competition.

Finally, the paper relates to a broader literature on information goods other than personal data, such as patent and digital goods (e.g., Shapiro et al. 1998; Lerner and Tirole 2004; Sartori 2018). Relative to this literature, the main novelty is to consider the upstream market in which consumers provide data to intermediaries. To keep the model simple, I abstract from many important issues relevant to information goods, such as versioning and network effects.

Two-sided Markets: The 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 et al. 2018). That the nonrivalry of data relaxes competition among intermediaries echoes the finding of the literature that multi-homing by one side relaxes competition for that side (e.g., Caillaud and Jullien 2003 and Tan and Zhou 2019). Nonetheless, there are three differences. First, in my model, depending on how downstream firms use data, intermediaries may earn a monopoly profit, or the consumer may extract full surplus (e.g. Proposition 7). This is more nuanced than the non-rivalry of data relaxing competition. Second, in terms of the modeling, the consumer may have multiple pieces of data, and she may benefit or hurt depending on what set of data to share. Such a situation is important for data markets but has no counterpart in the literature, where consumers typically choose whether or not to join a platform to earn some benefit.[5] Moreover, many of my results—such as the analysis of data concentration and information design—have no counterpart in the literature. Third, in my model, the consumer shares the same data with multiple intermediaries only *off* the equilibrium path. This is in contrast to the literature where consumers multi-home on the equilibrium path. The difference arises partly because compensation is endogenous.

Vertical Contracting and Contracting with Externalities: We can interpret the model as contracting with externalities (McAfee and Schwartz, 1994; Segal, 1999; Rey and Tirole, 2007). Namely, a supplier (consumer) provides goods to retailers (intermediaries), which later compete in the downstream market. The model departs from a typical model of vertical contracting in terms of the supplier's "cost" of producing goods: The cost can be positive or negative and depend non-monotonically on what combination of goods to produce. Also, the marginal cost of producing the second unit of the same good is zero, because the consumer's payoff does not depend on how many intermediaries resell her data.

3 Model

There are a consumer, $K \in \mathbb{N}$ data intermediaries, and a single downstream firm (Section 8 considers multiple consumers and firms). Abusing notation, I use K for both the number and the set of intermediaries. Figure 1 depicts the game: Intermediaries obtain data in the upstream market and sell them in the downstream market. The detail is as follows.

[5] Anderson and Coate (2005) and Reisinger (2012) consider platform competition with single-homing such that the presence of advertisers imposes negative externalities on viewers.

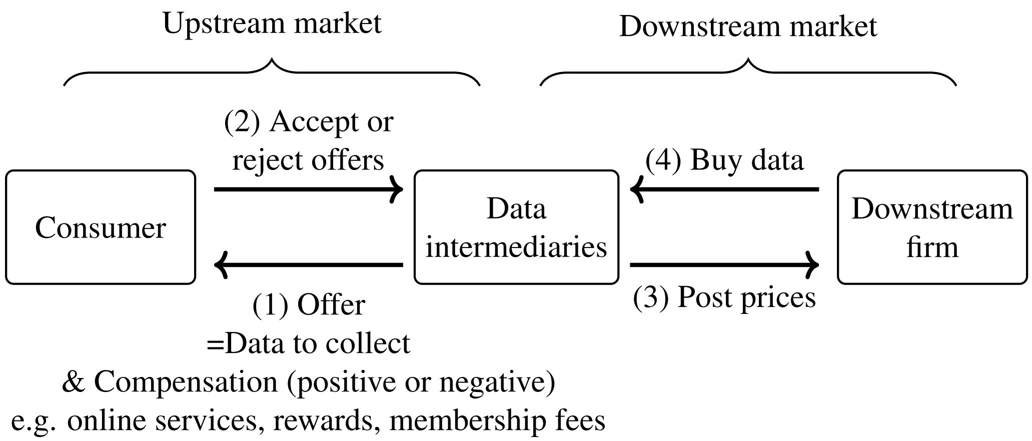

Figure 1: Timing of Moves

Upstream Market

The consumer (she) has a finite set $\mathcal{D}$ of data. Elements of $\mathcal{D}$ represent data labels such as her email address, location, and browsing history. They may also be different versions of the same data (e.g. browsing histories of different lengths). Each element of $\mathcal{D}$ is an indivisible and non-rivalrous good. See the next subsection for the discussion of this modeling approach.

At the beginning of the game, each intermediary $k \in K$ simultaneously makes an *offer* (D_k, τ_k). $\tau_k \in \mathbb{R}$ is the amount of compensation that intermediary k is willing to pay for $D_k \subset \mathcal{D}$. Compensation represents the quality of online services or the amount of monetary rewards that a consumer can enjoy by sharing data. $\tau_k < 0$ represents a fee. If $D_k \neq \emptyset$, I call (D_k, τ_k) a *non-empty offer*.

The consumer then chooses a set $K_C \subset K$ of offers to accept. $k \in K_C$ means that the consumer provides the requested data D_k to intermediary k and receives τ_k. The consumer can accept any set of offers, which reflects the non-rivalry of data.

All intermediaries and the firm observe the set $\hat{D}_k \in \{D_k, \emptyset\}$ of data 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 expected payoffs, and their ex post payoffs are as follows. The payoff of each intermediary is revenue from the firm minus compensation to the consumer.

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

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

Timing

The timing of the game, depicted in Figure 1, is as follows. First, intermediaries simultaneously make offers to the consumer. She 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 with this restriction.

3.1 Discussion of Assumptions

I comment on several important modeling assumptions.

Data as indivisible and non-rivalrous goods

In this paper, I do not model the "realization" of data. For example, before sharing location data, the consumer's exact location (i.e. "realization") is her private information. Moreover, depending on her location, she may have different preferences over sharing and not sharing the data. This may lead to a situation in which the consumer is privately informed of $U(\cdot)$. However, I assume

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

that contracting takes place ex ante, and do not model the uncertainty regarding the realization of data. As a result, the consumer has personal data but no private information. This is in line with recent works on data markets, such as Acemoglu et al. (2019), Bergemann et al. (2019), and Choi et al. (2019).

Observable allocation of data

It is crucial that intermediaries observe what data other intermediaries collect before setting downstream prices. There are several motivations for this assumption. First, in practice, some data intermediaries disclose what kind of data they collect. For example, a data broker CoreLogic states that it holds property data covering more than 99.9% of U.S. property records.[7] Also, if an intermediary collects data directly from consumers, it needs to communicate what data it collects (e.g. Nielsen Homescan). Moreover, it is practically necessary for downstream firms to know what data intermediaries hold in order to make purchase decisions.

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

Timing

I assume that intermediaries set prices after observing the allocation of data. The idea is similar to models of endogenous product differentiation such as d'Aspremont et al. (1979), where sellers set prices after observing their choices of product design. What data an intermediary collects (i.e., offer) is often a part of platform design or a company's policy. For example, a web mapping service

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

such as Google maps could correspond to an offer (D_k, τ_k) such that D_k consists of location data and τ_k reflects the value of service, which can depend on costly investment. In contrast, after collecting data, online platforms and data brokers typically share the data in exchange for money. Then, it is reasonable to assume that intermediaries can adjust downstream prices of data more quickly than adjusting what data they collect.

3.2 Applications

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

Online platforms

The model can capture competition for data among online platforms such as Google and Facebook. Given an offer (D_k, τ_k), D_k represents the set of data that a consumer needs to provide to use platform k, and τ_k represents the quality of k's service. Platforms may share data with advertisers, retailers, and political consulting firms, which benefits or hurts data subjects (e.g. beneficial targeting or harmful price discrimination). The net effect is summarized by $U(D)$.

Several remarks are in order. First, $U(\cdot)$ is exogenous, that is, intermediaries cannot influence how the firm's use of data affects the consumer. This reflects the difficulty of writing a fully contingent contract over how and which third parties can use personal information. The lack of commitment over the sharing and use of data plays an important role in other models of markets for data such as Huck and Weizsacker (2016) and Jones et al. (2018).

Second, if we interpret compensation as the value of a service, then modeling it as a one-to-one transfer means that a consumer's gross benefit from one service does not depend on what other services she uses. This requires that the consumer does not perceive services offered by different platforms as substitutes. Thus, the model is not appropriate, for example, if two platforms offer search engines. The assumption of costly compensation is natural if an intermediary needs to invest to improve the quality of its service.

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

Data brokers

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

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

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

Mobile application industry

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

4 Two Benchmarks

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

4.1 Monopoly Intermediary ($K = 1$)

In the upstream market, a monopoly intermediary can collect data D by paying a compensation of $-U(D)$. In the downstream market, it can set a price of $\Pi(D)$ to extract full surplus from the firm. Thus, I obtain

Claim 1. *In any equilibrium, a monopoly intermediary obtains and sells data $D^M \subset \mathcal{D}$ that satisfies $D^M \in \arg\max_{D \subset \mathcal{D}} \Pi(D) + U(D)$. The consumer and the firm obtain zero payoffs.*

4.2 Competition for Rivalrous Goods

Suppose that data are rivalrous—the consumer can provide each piece of data to *at most one* intermediary.[8] This model captures competition among intermediaries for physical goods (cf. 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, all intermediaries and the firm obtain zero payoffs. There is an equilibrium in which the consumer extracts full surplus* $\max_{D \subset \mathcal{D}} \Pi(D) + U(D)$.

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

5 Single Unit Data

This section considers a consumer with one unit of data. I first characterize equilibrium, and then point to an inefficiency coming from the non-rivalry of data. Then, I discuss two implications: why the market for personal data is missing, and how to improve consumer welfare and total welfare.

Formally, assume that there are multiple intermediaries ($K \geq 2$), and $\mathcal{D} = \{d\}$. Define $U := U(\{d\})$ and $\Pi := \Pi(\{d\})$. To obtain non-trivial results, assume $\Pi + U > 0$, which means that data collection is efficient. The following result characterizes the equilibrium (see Appendix B for the proof).

[8]Formally, I assume that 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 $j, k \in K_C$.

Proposition 1. *In any equilibrium, one intermediary obtains data at compensation* $\max(0, -U)$*. The consumer obtains* $\max(0, U)$*, one 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.,* $U < 0$*.*

If (and only if) $U > 0$, the consumer obtains a positive equilibrium payoff, which is strictly greater than under the monopoly outcome in Claim 1. If $U < 0$, the equilibrium coincides with monopoly. In either case, consumer surplus is lower than $\Pi + U$, which is the one in the case of rivalrous goods (Claim 2).

The intuition is that competition forces intermediaries to reduce positive fees to zero, but does not incentivize them to raise non-negative compensation beyond a monopoly level. To see this, suppose that intermediary 1 collects data at a positive fee. Then, intermediary 2 can undercut it. Importantly, facing the competing offer from intermediary 2, the consumer shares her data with *only* intermediary 2, because she earns a gross benefit of U as long as the downstream firm buys data from at least one intermediary. This implies that competing intermediaries have to offer non-negative compensation. If the firm's data usage is beneficial to the consumer ($U > 0$), then this logic implies that the consumer enjoys 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 this market for data, competition of raising compensation does not work. For example, suppose $U < 0$ and that intermediary 1 collects data at monopoly compensation $-U > 0$. If intermediary 2 offers positive compensation, then the consumer shares her data with *both* intermediaries. This intensifies price competition in the downstream market and reduces the price of data to zero. Anticipating this, intermediary 2 makes no competing offer.

Consider how the equilibrium depends on U given a fixed total surplus $TS = \Pi + U$. First, the joint profit of intermediaries, $TS - \max(0, U)$, is maximized if $U < 0$. Thus, Proposition 1 suggests that the intermediation of data is more profitable when the downstream firm uses it to hurt consumers. As U increases from a negative value to $TS > 0$, the equilibrium outcome changes from the monopoly outcome to the one in which the consumer extracts full surplus. This underscores the importance of considering a firm's data usage when we examine competition among data intermediaries. We may think $(\Pi, U) = (0, TS)$ as a knife-edge case, however, Section 7 shows that $(\Pi, U) = (0, TS)$ can be relevant when the firm uses data for price discrimination.

Finally, the result gives a rationale to the frequently used assumption in the literature that the market consists of a monopoly data seller.[9] We can justify the assumption as a subgame of the extended game in which data sellers acquire information at cost and then sell collected data.

5.1 Competition and Inefficiency

Because of the non-rivalry of data, competition may fail to drive out a "less efficient" intermediary out of the market. To see this, consider two intermediaries. Modify the consumer's payoffs so that, if the consumer shares d with intermediary k, then she incurs a cost of c_k. Thus, the consumer's payoff from sharing d with the set K_C of intermediaries equals her original payoff minus $\sum_{k \in K_C} c_k$. Assume $c_1 = 0$ but $0 < c_2 < \Pi + U$. For example, intermediary 2 shares collected data with malicious third parties, which lowers consumer welfare. The efficient outcome is that intermediary 1 collects and sells data. However, the non-rivalry of data can create an inefficiency in which intermediary 2 acts as a monopolist (see Appendix C for the proof).

Proposition 2. *The following holds.*

1. *In the case of rivalrous goods, in any equilibrium, intermediary* 1 *collects data at compensation* $\Pi - c_2$.

2. *In the case of non-rivalrous data, if* $U < 0$*, then there is an equilibrium in which intermediary* 2 *earns a monopoly profit* $\Pi + U - c_2$.

The intuition is as follows. In the case of rivalrous goods, if intermediary 2 earned a nonnegative profit by collecting data, then intermediary 1 could offer the same compensation, exclusively obtain data, and earn a positive profit. Thus, the less efficient intermediary is never active in equilibrium. In the case of non-rivalrous data, if intermediary 2 collects data and intermediary 1 makes a competing offer, then the consumer shares data with both intermediaries. As before, this will reduce the downstream price of data. Anticipating this, intermediary 1 does not make a competing offer.

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

For rivalrous goods (Point 1), the presence of an inefficient intermediary increases consumer surplus by $\Pi + U - c_2 > 0$ without changing total surplus. For non-rivalrous data (Point 2), the presence of an inefficient intermediary may lower total surplus without changing consumer surplus. In particular, if c_2 is close to $\Pi + U$, then the inefficient intermediary may destroy most of the surplus without benefiting the consumer.

Intuitively, the negative welfare implication is likely to materialize when c_k reflects the quality of privacy enhancing technology, and an entrant has a better technology (a lower c_k) than an incumbent. To see this, consider a variant of the game in which (i) intermediary 2 (incumbent) makes an offer, and (ii) after observing (i), intermediary 1 (entrant) makes an offer. This game selects the equilibrium in which intermediary 2 earns a monopoly profit.

5.2 Why is "Market for Personal Data" Missing?

Right now, any consumer can technically download all their data from Facebook and Google and sell it.[10] However, the market price for such data seems to be zero.[11] The results provide an economic explanation for why markets for personal data is missing: Nobody will pay you much for your data, because the data is already available from existing platforms in the downstream market.

The results also point to a challenge faced by small but emerging "personal data marketplaces" such as Killi or Hu-manity.co.[12] Their claimed advantages seem to be greater transparency and privacy protection. This contrasts with existing platforms or data brokers, whose data collection may impose the cost of data breach or mismanagement on consumers (e.g. Cambridge Analytica scandal). The result suggests that even if those new companies can create a higher total surplus, they may not give competitive pressure to existing players to better compensate consumers for providing data.

[10] For Facebook, see https://www.facebook.com/help/212802592074644 (accessed on March 23, 2020). For Google, see https://support.google.com/accounts/answer/3024190?hl=en (accessed on March 23, 2020).

[11] See https://www.wired.com/story/i-sold-my-data-for-crypto/ for anecdotal evidence (accessed on March 23, 2020).

[12] https://techcrunch.com/2018/07/18/hu-manity-wants-to-create-a-health-data-marketplace- (accessed on March 20, 2020).

5.3 How to Improve Consumer Surplus and Total Surplus?

This subsection asks how we can change the rule of the game to increase consumer surplus and total surplus. I propose two potential solutions.

Giving Bargaining Power to the Consumer One straightforward solution is to give bargaining power to the consumer. Namely, if the consumer can make a take-it-or-leave-it offer to any intermediary, then she offers $(\{d\}, \Pi)$ and extracts full surplus. This modification also eliminates an inefficient equilibrium discussed in the previous subsection. This solution relates to the recent idea of "data labour union" discussed in Arrieta-Ibarra et al. (2018).

Richer Contract Space A more subtle way is to enable intermediaries to offer richer contracts. Here, I discuss two possibilities. First, suppose that each intermediary k can offer a "revenue-sharing contract" of the form (D_k, α_k), where $\alpha_k \in [0, 1]$ is the fraction of k's downstream revenue that the consumer earns. In this case, there is an equilibrium in which each intermediary offers $(\{d\}, 1)$, and the consumer shares data with the most efficient intermediary and extracts full surplus. Facing revenue-sharing contracts, the consumer never provides the same data to multiple intermediaries, because it will reduce the downstream price of data. This restores competition for data and incentivizes intermediaries to make the most attractive offer $\alpha_k = 1$.

Second, suppose that each intermediary can offer compensation that depends on the consumer's data sharing decision with respect to other intermediaries. Then, there is an equilibrium with exclusive contracts: Each intermediary k commits to pay Π if and only if the consumer provides data to only k. In this equilibrium, the consumer shares data to the most efficient intermediary and extracts full surplus. Finally, note that implementing these richer contracts implicitly requires the greater commitment power of intermediaries and the greater transparency of market outcomes such as downstream transaction.

6 General Preferences

I now consider the consumer with any finite set of data. The purpose is to generalize some of the previous results, and also to point out the multiplicity of equilibria due to the non-rivalry of data.

Intermediaries are homogeneous as in the baseline model. I allow any $U(\cdot)$ and any increasing $\Pi(\cdot)$ such that

Assumption 1. $\mathcal{D} \in \arg\max_{D \subset \mathcal{D}} \Pi(D) + U(D)$.

I maintain Assumption 1 throughout this section. The assumption implies that total surplus is maximized when the firm acquires all data. It holds, for example, if the firm is a seller and can use all data $\mathcal{D}$ to efficiently price discriminate the consumer. Section 7 microfounds U and Π 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 the data. In Section 8, where I consider the extension with multiple consumers, I argue that (a version of) Assumption 1 is likely to hold when there is an information externality among many consumers.

6.1 Partially Monopolistic Equilibrium

The following result generalizes Proposition 1 (see Appendix D for the proof).

Proposition 3 (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 obtains an equilibrium payoff of* $\max_{D \subset \mathcal{D}} U(D)$.

If the consumer has one piece of data d, then $\max_{D \subset \mathcal{D}} U(D) = \max(0, U(\{d\}))$, and thus the PME coincides with the unique equilibrium characterized by Proposition 1. Proposition 3 states that the intuition for Proposition 1 applies to arbitrary 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. $U(\cdot)$ is non-monotone, and $\Pi(\cdot)$ exhibits increasing returns to scale. First, a monopoly intermediary obtains all data at compensation $-U(\mathcal{D})$ (short red dotted arrow). We can decompose $-U(\mathcal{D})$ into two parts: The monopolist extracts surplus created by $D^* \in \arg\max_{D \subset \mathcal{D}} U(D)$ from the consumer by charging $U(D^*) > 0$, and it obtains additional data $\mathcal{D} \setminus D^*$ at the minimum compensation $U(D^*) - U(\mathcal{D})$ (long blue dotted arrow). In contrast, when there are multiple intermediaries, competition prevents intermediaries from extracting surplus $U(D^*) > 0$. This guarantees that the consumer obtains a payoff of at least $U(D^*)$. However, competition does not increase compensation for data $\mathcal{D} \setminus D^*$, the sharing of which hurts the consumer. Thus, in the PME, a single intermediary acquires all data but compensates the consumer according to the loss $U(D^*) - U(\mathcal{D})$ of sharing $\mathcal{D} \setminus D^*$. Finally, the compensation in the PME is still lower than $\Pi(\mathcal{D})$, which is the compensation that the consumer would receive in markets for rivalrous goods (black dashed arrow).

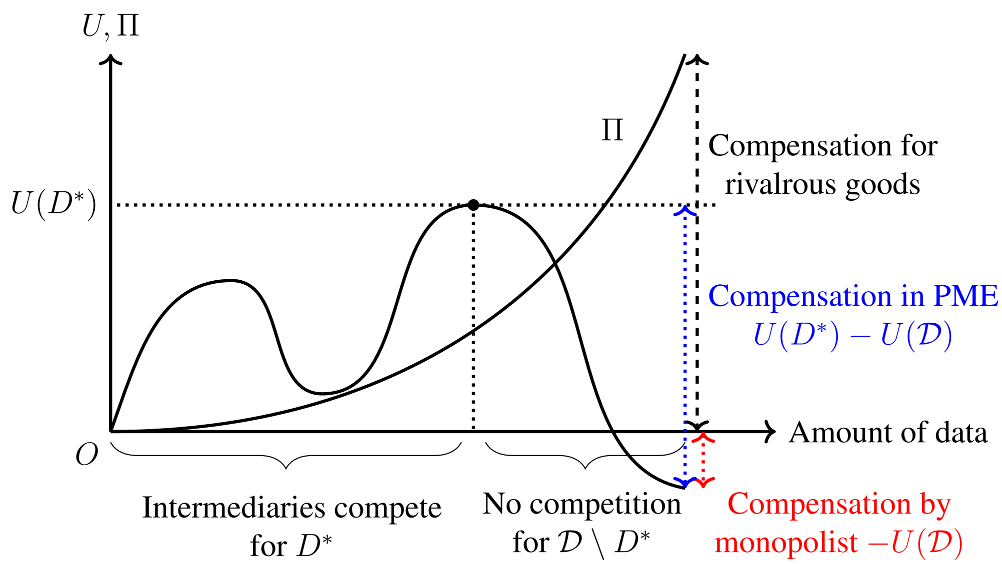

Figure 2: Partially monopolistic equilibrium

The next result shows that if there are many intermediaries, then the PME minimizes consumer surplus and maximizes intermediary surplus across all equilibria (see Appendix E for the proof). In this sense, the PME is a natural extension of the monopoly equilibrium. To state the result, let $CS(K)$ denote the set of all possible subgame perfect equilibrium payoffs of the consumer when there are K intermediaries.

Proposition 4. *As the number K of intermediaries grows large, the worst consumer surplus converges to the one in the PME:*

1. $\lim_{K \to \infty} (\inf CS(K)) = \max_{D \subset \mathcal{D}} U(D)$. *$\mathcal{D}$ can be an infinite set if the right-hand side is well-defined.*

2. *If $\mathcal{D}$ is finite and $\Pi(\cdot)$ is strictly increasing, then there is a $K^* \in \mathbb{N}$ such that, for any $K \geq K^*$,* $\min CS(K) = \max_{D \subset \mathcal{D}} U(D)$.

Thus, the best intermediary surplus (among all subgame perfect equilibria) also converges to the one in the PME.

The intuition is as follows. Suppose that 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 $\varepsilon < \delta_K$, then the consumer prefers to accept it. Because any intermediary can always deviate and offer (D^*, ε), each intermediary obtains a payoff of at least δ_K. This implies that intermediary surplus is at least $K \cdot \delta_K$. However, intermediary surplus is bounded from above by $\Pi(\mathcal{D}) < \infty$. Thus, $\delta_K \to 0$ as K grows large, i.e., the worst consumer surplus converges to $U(D^*)$ as the number of intermediaries grows large. Point 2 shows that under a stronger assumption, $U(D^*)$ is exactly the lowest equilibrium payoff of the consumer for a sufficiently large but finite K. Finally, in the PME, total surplus is maximized, and consumer surplus is $U(D^*)$. Thus, the PME is (approximately) an intermediary-optimal outcome 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 among intermediaries gives the full surplus to those in the upstream market. In markets for data, the non-rivalry of data makes $U(\cdot)$ relevant. If the use of data benefits consumers, then competition eliminates fees that consumers would have to pay in a monopoly market. However, if the use of data hurts consumers, then competition may have no impact on increasing compensation. In a general setting, both effects are relevant. As a result, competition may increase consumer welfare and decrease intermediary profit, but not as much as in markets for rivalrous goods.

6.2 Partitional Equilibria

It is beyond the scope of the paper to characterize all equilibria for any $(U(\cdot), \Pi(\cdot))$. In this subsection, I assume the increasing and convex cost of sharing data for the consumer, and the decreasing marginal revenue for the downstream firm.

Assumption 2. $U(\cdot)$ is decreasing and submodular, and $\Pi(\cdot)$ is increasing and submodular.[13]

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

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

In the "rivalrous goods" model (i.e. Claim 2), any equilibrium has a trivial partition under a mild additional restriction (see Appendix F for the proof).

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

In contrast, any partition can arise as an equilibrium allocation of data.

Proposition 5. *The allocation of data $(D_k^*)_{k\in K}$, compensation $(\tau_k^*)_{k\in K}$, and prices $(p_k^*)_{k\in K}$ consist of a partitional equilibrium if and only if*

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

Partitional equilibria have three features. First, although data are non-rivalrous, no two intermediaries obtain the same piece of data. This is because such a piece of data will have no value in the downstream market.

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

Third, each intermediary compensates the consumer according to the marginal (or precisely, incremental) loss that she incurs by sharing $\hat{D}_k$ conditional on sharing data with other intermediaries. This contrasts with the rivalrous goods case in which the equilibrium compensation depends on the downstream firm's willingness to pay.

6.3 Data Concentration

Proposition 5 implies that any partition of $\mathcal{D}$ can arise as an allocation of data in some equilibrium. We can interpret an equilibrium associated with a coarser partition as an equilibrium with a greater concentration of data among intermediaries:

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

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

Proposition 6. *Take two partitional equilibria such that one is more concentrated than the other. Intermediaries' joint profit is higher, and consumer surplus and the firm's profit are lower in the more concentrated equilibrium.*

The intuition is as follows. 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 there are many intermediaries each of which has a small subset of $\mathcal{D}$, then the contribution of each piece of data 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. Since $\Pi(\cdot)$ is submodular, the latter leads to a greater total revenue for intermediaries. Symmetrically, if $U(\cdot)$ is submodular, data concentration hurts consumers. This is because a large intermediary compensates the consumer based on the infra-marginal cost of sharing data. The following example relates the result to the idea of breaking up big platforms.

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

If the market consists of a monopoly intermediary, then it obtains both location and financial data and pays \$20 to each consumer. This leads to a consumer surplus of zero. For example, the intermediary may operate an online service that requires consumers to provide these data.

Suppose now that a regulator breaks up the monopolist into two intermediaries, 1 and 2. Proposition 5 implies that in one of the equilibria, intermediaries 1 and 2 collect location and financial data, respectively, and each intermediary pays a compensation of \$20. For example, two intermediaries may operate mobile applications that collect different data, and each application delivers the value of \$20 to consumers. In this equilibrium, each consumer obtains a net surplus of \$20. Thus, breaking up a monopolist may change the equilibrium allocation of data, increase compensation, and benefit consumers.

7 Application: Information Design by Data Intermediaries

So far, I have treated data as indivisible and non-rivalrous goods. However, in practice, firms eventually use consumer data to learn about their private information. To illustrate this point, this section applies the model to a setting in which the firm uses data to learn about a consumer's willingness to pay. The firm then tailors pricing and product recommendation.

The formal description is as follows. 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 $V \subset (0, +\infty)$.[14]

Each $d \in \mathcal{D}$ is a signal (Blackwell experiment) from which the seller can learn about $\mathbf{u}$. $\mathcal{D}$ consists of all signals with finite realization spaces. Intermediaries can request the consumer any set of signals.[15] 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.[16] A recommendation could be an advertiser displaying a targeted advertisement

[14] 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.

[15] 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.

[16] The model assumes that the seller only recommends one product, and thus the consumer cannot buy non-

or an online retailer showing a product as a personalized recommendation. If the consumer buys product m at price p, then 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 off the equilibrium paths, all players calculate its posterior beliefs based on the prior F, signals in D, and Bayes' rule.[17]

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

The following notations are useful: 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, then 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. In other words, consumer surplus is equal to the one in Bayesian persuasion (see Appendix I for the proof).

Proposition 7. *Suppose that there are multiple intermediaries. In the partially monopolistic equilibrium, one intermediary (say* 1*) obtains a fully informative signal, and the consumer obtains a payoff of* $\max_{d \in \mathcal{D}} U(\{d\})$*. Moreover, this equilibrium satisfies the following.*

1. *If the seller provides a single product (*$M = 1$*), then all intermediaries earn zero payoffs,*

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.

[17] 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).

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.*[18]

The intuition is as follows. First, consider Point 1. Bergemann et al. (2015) show that there is a signal d^* such that (i) d^* maximizes the consumer's payoff, i.e., $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$, (ii) the seller is indifferent between obtaining d^* and nothing, i.e., $\Pi(d^*) = \Pi(\emptyset)$, and (iii) d^* maximizes total surplus $U(d) + \Pi(d)$. (i) implies that competing intermediaries cannot charge the consumer a positive fee for d^*. (ii) implies that they cannot charge the firm a positive price for d^*. Moreover, (iii) implies that intermediaries cannot make a profit by obtaining and selling additional information. Thus, in the PME, the consumer obtains a payoff of $U(d^*)$, and no intermediaries can make a positive profit. In this case, competition among intermediaries yields the consumer all welfare gain from her information. Proposition 4 implies that, if K is large, then this equilibrium (PME) is worst for the consumer. This implies when $M = 1$ and K is large, the equilibrium outcome is (almost) unique.

Second, consider Point 2. Ichihashi (2020) shows that if the prior F satisfies the condition in Point 2, then any consumer-optimal signal $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$ leads to inefficiency. Intuitively, d^* conceals some information about which product is most valuable to the consumer. This benefits the consumer by inducing the seller to lower prices, but it leads to inefficiency due to product mismatch. This inefficiency (under the hypothetical Bayesian persuasion) creates a room for competing intermediaries to earn a positive profit: An intermediary can additionally obtain information that enables the seller to perfectly learn the consumer's values. The consumer requires a positive compensation to share such information. This, in turn, implies that a single intermediary can act as a monopoly of that information. Thus, competition benefits the consumer relative to monopoly but it does not completely dissipate intermediaries' profits.

[18] 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 $\Delta(V)$.

8 Extensions

8.1 Other Market Structures

The baseline model assumes that intermediaries simultaneously move when they make offers or set prices. The main insight is robust to other timing assumptions. For example, suppose that each intermediary sequentially makes an offer to the consumer, and after observing all offers, the consumer chooses which offers to accept. For the cases of Propositions 1, 3 (for a large K), and 5, this game selects the equilibrium such that the first-mover collects all data.

We can also think of various games for the downstream market. For example, we could assume that the downstream firm can buy data from only one intermediary. Alternatively, we could think of a game in which the downstream firm is randomly matched with one intermediary, which can make take-it-or-leave-it offer to the firm. In either case, there is an equilibrium in which intermediaries jointly earn a monopoly profit for the case of Propositions 1 (for $U < 0$) and 5.

8.2 Multiple Consumers with Information Externality

The baseline model assumes a single consumer. However, the results extend to multiple consumers (see Appendix J for the detailed description and the proofs of the following claims). Formally, let $I \in \mathbb{N}$ denote the number and the set of consumers. Each consumer $i \in I$ has a 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.

To accommodate a general case in which $U_i(D_i, D_{-i})$ depends on D_{-i} ("information exteranlities"), I need several modifications. First, I assume private offers, that is, each consumer i does not observe offers made to other consumers. Second, a solution concept is perfect Bayesian equilibrium such that consumers have passive beliefs. In other words, after consumer i detects deviations of intermediaries, she does not change her beliefs regarding what offers other consumers are receiving.

Claim 4 (Extending Proposition 1). *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 that a monopoly intermediary collects in some* *equilibrium. Then, 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 data $d_i \in D_k^M$ at compensation $\max(0, -U_i(\{d_i\}, D_{-i}^M))$ where $D_{-i}^M = D^M \cap \mathcal{D}_{-i}$.*

To extend other results, I modify Assumptions 1 and 2. I replace Assumption 1 with the assumption that Π and $(U_i)_{i \in I}$ are such that a monopoly intermediary collects and sells all data $\cup_{i \in I} \mathcal{D}_i$ in some equilibrium. Without information externalities, this modified assumption is equivalent to the original Assumption 1.

Claim 5 (Extending Proposition 3). *Under the modified Assumption 1, there is an equilibrium such that a single intermediary collects all data at compensation of $\max_{D_i \subset \mathcal{D}_i} U_i(D_i, \mathcal{D}_{-i}) - U_i(\mathcal{D}_i, \mathcal{D}_{-i})$ for each $i \in I$.*

The modified Assumption 1 is likely to hold if there are informational externalities among many consumers. As Bergemann et al. (2019) show, the externality creates a gap between an intermediary's revenue from selling data and compensation consumers demand. This makes it more likely that a monopoly intermediary transfers all data.

Finally, To extend Proposition 5, 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. Given this modified assumption, 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 $U_i(\cdot, \mathcal{D}_{-i})$. The welfare implication of data concentration naturally extends.

8.3 Multiple Downstream Firms

On top of multiple consumers, the model can take into account multiple downstream firms if they do not interact with each other: Suppose that there are L firms, where firm $\ell \in L$ has revenue function Π^ℓ that depends only on data available to ℓ. Each consumer i's gross payoff of sharing data is $\sum_{\ell \in L} U_i^\ell$, where each U_i^ℓ depends on the set of i's data that firm ℓ obtains.

This setting is equivalent to the one with a single firm. For example, suppose that intermediary 1 has all data $\mathcal{D}$ and intermediary 2 had some data 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, then 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, intermediaries cannot commit to not sell data to downstream firms. Thus, once a consumer shares her data with one intermediary, the data is sold to all firms. This means that in equilibrium, each consumer i decides which offers to accept in order to maximize the sum of total compensation and $\sum_{\ell\in L} U_i^\ell(D_i)$. Therefore, we can apply the same analysis as before by defining $U_i := \sum_{\ell\in L} U_i^\ell$.

9 Conclusion

This paper studies competition among data intermediaries, which obtain data from consumers and sell them to downstream firms. The model incorporates two key features of personal data: Data are non-rivalrous, and the use of data by third parties can increase or decrease consumer welfare. I show that the non-rivalry of data relaxes competition among intermediaries. If a downstream firm's data usage hurts consumers, then the equilibrium may coincide with the monopoly outcome. Unlike markets for rivalrous goods, the entry of a more efficient intermediary may not kick out a less efficient incumbent. Under a certain condition, an equilibrium with greater data concentration is associated with higher profits of intermediaries and lower consumer welfare.

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.

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.”

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

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

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

Bonatti, Alessandro and Gonzalo Cisternas (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.

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

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

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

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

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

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

Kamenica, Emir and Matthew Gentzkow (2011), “Bayesian persuasion.” *American Economic Review*, 101, 2590–2615.

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

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

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

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.”

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, Shapiro Carl, Hal R Varian, et al. (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 (2019), “Price competition in multi-sided markets.” *Available at SSRN 3029134*.

Appendix

A Competition for Rivalrouds Goods: Proof of Claim 2

Take any $K \geq 2$. Suppose to the contrary that there is an equilibrium in which intermediary k^* obtains a positive payoff of $\pi^* > 0$. For each intermediary $k \in K$, let (D_k^*, τ_k^*) denote its equilibrium offer. Let K_C denote the set of intermediaries whose offers are accepted. Take any $j \neq k^*$, and suppose that j offers $(\cup_{k \in K_C} D_k^*, \sum_{k \in K_C} \tau_k^* + \epsilon)$ with $\varepsilon \in (0, \pi^*)$. Then, the consumer accepts this offer and rejects other offers. The deviation of intermediary j increases the consumer's payoff by ε, reduces the sum of payoffs of other intermediaries $k \neq j$ by at least π^*, and weakly reduces the firm's payoff. Because the deviation does not change total surplus, this means that j's payoff increases by at least $\pi^* - \varepsilon > 0$. This is a contradiction.

To show the second part, take any $D^* \in \arg\max_{D \subset \mathcal{D}} \Pi(D) + U(D)$. Consider the following strategy profile: All intermediaries offer $(D^*, \Pi(D^*))$, and the consumer accepts one of them. In the downstream market, an intermediary that has D^* sets a price of $\Pi(D^*)$. Assign arbitrary equilibrium strategy in any subgame following deviations. This strategy profile is an equilibrium.

Indeed, if an intermediary could deviate and earn a positive payoff, this weakly increases the consumer's payoff (as she can at least accept $(D^*, \Pi(D^*))$), weakly increases the firm's payoff, and strictly increases the intermediaries' joint profits. This contradicts D^* maximizing total surplus.

B Equilibrium for Single Unit Data: Proof of Proposition 1

Throughout the proof, I consider the following strategies of intermediaries and the firm in the downstream market: An intermediary sets a price of Π if it is the only one holding 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 subgame corresponding to the downstream market is equivalent to this strategy profile in terms of the payoff of each player.

First, I show that, in any equilibrium such that 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 make non-empty offers 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 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 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 a 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. Suppose also that K_C satisfying (A) contains k. Then, the consumer can add any $j \neq k$ that offers non-negative compensation to K_C in order to satisfy (B). Adding j to K_C weakly increases the consumer's payoff because it weakly increases total compensation without affecting her gross payoff.

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, 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, 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.

C Competition and Inefficiency: Proof of Proposition 2

First, consider the case of rivalrous goods. Take any SPE. If intermediary 2 collects data, then intermediary 1 can make the same offer and exclusively collects d, which is a contradiction. Thus, intermediary 1 collects d. Let τ^* denote the equilibrium compensation of 1. If $\tau^* > \Pi - c_2$, then intermediary 1 can instead offer $\tau \in (\Pi - c_2, \tau^*)$. The consumer continues to accept this, because the maximum payoff that intermediary 2 can give is $U + \Pi - c_2$. This is a contradiction, and thus $\tau^* \leq \Pi - c_2$. If $\tau^* < \Pi - c_2$, then intermediary 2 can collect data by offering compensation $\tau \in (\tau^*, \Pi - c_2)$, which is a contradiction. Thus, $\tau^* = \Pi - c_2$. Moreover, there exists an equilibrium in which intermediary 1 offers $(\{d\}, \Pi - c_2)$, intermediary 2 offers $(\{d\}, \Pi)$, and the consumer accepts the offer from 1. This completes the proof of Point 1.

Second, consider the case of non-rivalrous data. Consider the following strategy profile: Intermediary 1 offers $(\{d\}, 0)$, intermediary 2 offers $(\{d\}, -U + c_2,)$, and the consumer accepts only

2's offer. Assign arbtirary equilibrium strategy after deviations. 1 has no profitable deviation: If it raises compensation, the consumer shares d with both intermediaries, following which the downstream price of data becomes zero. If 1 offers negative compensation, the consumer continues to reject it. 2 has no profitable deviation: In particular, the consumer's payoff from accepting 2's offer is 0. Thus, if 2 lowers compensation, the consumer rejects 2's offer. However, as intermediary 2 obtains a payoff of $\Pi + U - c_2 > 0$ on the path of play, 2 does not benefit from lowering compensation. Finally, that this equilibrium is Pareto-undominated follows from the same argument as the proof of Proposition 1.

D Partially Monopolistic Equilibrium: Proof of Proposition 3

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 assume that players play any equilibrium of the corresponding subgame. That this strategy profile consists of an equilibrium in the downstream market is straightforward.

I show that the suggested strategy profile is an equilibrium. First, I show that intermediary 1 has no incentive to deviate. Suppose that intermediary 1 deviates and obtains data D_1. Let $\hat{D}$ denote the set of all data that the consumer shares as a result of 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 that the consumer would strictly prefer to reject the offer from intermediary $k \neq 1$. Now, these bounds on revenue and cost imply that intermediary

1's payoff after the deviation is at most $\Pi(\hat{D}) - [U(D^*) - U(\hat{D})] = \Pi(\hat{D}) + U(\hat{D}) - U(D^*)$. Since the efficient outcome involves full data sharing, this 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, there is no profitable deviation for intermediary 1.

Second, suppose that intermediary 2 deviates and offers (D_2, τ_2). Without loss of generality, assume that each consumer accepts the offer. Let D_{-1} denote the set of data that 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 $U(\mathcal{D}) - U(D_{-1}) + U(D^*) - U(\mathcal{D}) \geq U(\mathcal{D}) - U(D^*) + U(D^*) - U(\mathcal{D}) = 0$. 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's 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 for any increasing Π. 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.

E Welfare Properties of PME: Proof of Proposition 4

(Point 1) I prepare several notations. Define $U^* := \max_{D \subset \mathcal{D}} U(D)$, and $TS^* := \Pi(\mathcal{D}) + U(\mathcal{D}) > 0$. Assumption 1 implies that TS^* is the maximum total surplus. As U^* is an equilibrium payoff in the PME, $\inf CS(K) \leq U^*$ holds for all $K \in \mathbb{N}$. Thus, we obtain $\limsup_{K \to \infty} (\inf CS(K)) \leq U^*$. Thus, it suffices to show that

$$\liminf_{K \to \infty} (\inf CS(K)) \geq U^*.$$

Suppose to the contrary that $\liminf_{K \to \infty} (\inf CS_i(K)) < U^* - 3\delta$ for some $\delta > 0$. This implies that there exists a strictly increasing subsequence $\{K_n\} \subset \mathbb{N}$ such that $\inf CS(K_n) < \liminf_{K \to \infty} (\inf CS(K_n)) + \delta < U^* - 2\delta$. This implies that for each K_n, there exists an equilibrium E_n in which the payoff of the consumer, denoted by CS^n, satisfies $CS^n < U^* - \delta$.

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

(Point 2) Define $m := \min_{d\in\mathcal{D}, D\subset\mathcal{D}} \Pi(D) - \Pi(D \setminus \{d\}) > 0$. Let K^* satisfy $K^* > TS^*/m$. Suppose that there are $K \geq K^*$ intermediaries, and take any equilibrium. Suppose (to the contrary) that the 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. Suppose to the contrary that intermediary k earns a strictly lower payoff than m. If intermediary k deviates and offers (D^*, ε) with $\varepsilon \in (0, \delta)$, then she accepts this offer. Let D_{-k} denote the data that 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 that 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. Since 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 one of consumers and the firm obtains a negative payoff, which is contradiction. Therefore, in any equilibrium, any consumer obtains a payoff of at least $U(D^*)$. Because the PME gives the consumer a payoff of $U(D^*)$, we obtain the result.

F 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_k denote compensation that intermediary k pays to the consumer. 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 1's deviation, the joint downstream revenue is $\sum_{k\in K} [\Pi(\cup_{j\in K} D_j) - \Pi(D_{-k})]$ (this follows from Lemma 1 below). By the same logic as the proof of Proposition 6 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 can strictly benefits from the deviation with a sufficiently small $\varepsilon > 0$. This concludes $\cup_{k\in K} D_k = D_1$.

G Partitional Equilibria: Proof of Proposition 5

To characterize partitional equilibria, I first show that the downstream market has a unique equilibrium outcome if the firm's revenue function is submodular.[19]

Lemma 1 (Unique Equilibrium Payoffs in the Downstream Market). *Suppose $\Pi(\cdot)$ is submodular. Suppose that 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*

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

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 $(D_1, \dots, D_K)$. 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 $K' \subset K$. 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.

[19] Lemma 1 is more general than 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. (2018) assume $K = 2$ and consider not only submodularity but also supermodularity. Relative to Gu et al. (2018), the uniqueness of the equilibrium revenue for any K is a new result.

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

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} p_k = \max_{J\subset K\setminus\{1\}} \left(\Pi(\cup_{k\in J} D_k) - \sum_{k\in J} p_k \right). \tag{2}$$

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 (2). 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

$$\Pi(\cup_{k\in K^*} D_k) - \sum_{k\in K^*} \hat{p}_k < \Pi(\cup_{k\in J^*} D_k) - \sum_{k\in J^*} \hat{p}_k, \tag{3}$$

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 (3) is the maximum revenue that the firm can obtain if it cannot buy data D_j, and the right hand side is the lower bound of the revenue that the firm can achieve by buying D_j. Thus, the firm always buy data D_j, which is a contradiction.

Second, suppose that $J^* \setminus K^* = \emptyset$, i.e., $J^* \subset K^*$. This implies that the right hand side of (2) 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_ℓ for all $\ell \in K^*$. Plugging $J^* = K^* \setminus \{1\}$, we obtain

$$\Pi(\cup_{k\in K^*} D^k) - \sum_{k\in K^*} p_k = \Pi(\cup_{k\in K^*\setminus\{1\}} D_k) - \sum_{k\in K^*\setminus\{1\}} p_k. \tag{4}$$

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

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D_k) < \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D_k). \tag{5}$$

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

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D_k) = \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D_k). \tag{6}$$

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 (4) 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 (5) 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. $\square$

A direct corollary of this lemma is that the revenue of an intermediary is determined by the part of its data that other intermediaries do not hold.

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

I now prove Proposition 5.

Proof of Proposition 5. First, I prove "if" part. Take any $(D_k^*)_{k \in K}$, $(\tau_k^*)_{k \in K}$, and $(p_k^*)_{k \in K}$ that satisfy Points 1 - 3 of Proposition 5. Suppose that each intermediary k offers (D_k^*, τ_k^*) and sets a price of data following Lemma 1 (if $\Pi_k = 0$, then k sets a price of zero). On the equilibrium path, 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 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 Corollary 1, 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 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.

Second, I prove "only if" 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 $\tau_k^* \neq U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. Suppose that $\tau_k^* < U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$ although the right-hand side is positive. Then, the consumer rejects at least one non-empty offer on the equilibrium path. This contradicts a condition for partitional equilibrium that the intermediaries jointly collect $\mathcal{D}$. Next, suppose $\tau_k^* > U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. Then, by the "if" part, we can find an equilibrium that has the same outcome except intermediary k offers $\tau_k' \in (\tau_k^*, U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}))$ for collecting D_k^*. This equilibrium Pareto-dominates the original equilibrium, which is a contradiction. Thus, we obtain $\tau_k^* \leq U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. □

H Data Concentration: Proof of Proposition 6

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. □

I Proof of Proposition 7

Proof. Note that Theorem 3 holds even when $\mathcal{D}$ is not finite. Let d_{FULL} denote a fully informative signal. I show Point 1. Assuming that there is a single product ($M = 1$), Bergemann et al. (2015) show that there is a signal d^* that satisfies the following conditions: $d^* \in \arg\max_{d \in \mathcal{D}} U(d)$; $\Pi(d^*) = \Pi(\emptyset)$; d^* maximizes total surplus, i.e., $U(d^*) + \Pi(d^*) = U(d_{FULL}) + \Pi(d_{FULL})$. Namely, d^* simultaneously maximizes consumer surplus and total surplus without increasing the seller's revenue. These properties imply that intermediary 1's revenue in the downstream market is equal to the compensation it pays in the upstream market: $\Pi(d_{FULL}) - \Pi(\emptyset) = \Pi(d_{FULL}) - \Pi(d^*) = U(d^*) - U(d_{FULL})$. Thus, all intermediaries earn zero payoffs.

I show Point 2. Ichihashi (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. $\square$

J Multiple Consumers with Information Externalities: Appendix for Subsection 8.2

First, I describe the timing of the game when there are multiple consumers. 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 $\left\{(D_i^k, \tau_i^k)\right\}_{k \in K}$, and chooses a set K_i of offers to accept. This leads to the allocation of data such that 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. As discussed in the main text, the gross payoff of consumer i is given by $U_i(D_i, D_{-i})$. The solution concept is pure-strategy perfect Bayesian equilibrium with passive beliefs.

Proof of Claim 4. Take any set $D^M \subset \{d_1, \ldots, d_I\}$ of data that a monopoly intermediary collects in some equilibrium. Take any partition $(D_1^M, \ldots, D_K^M)$ of D^M. Consider the following strategy profile. Take any $i \in I$. 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 $d_i \notin D_k^M$, then intermediary k offers $(\{d_i\}, 0)$ to consumer i. On the path of play, each consumer i accepts the offer of k if and only if $d_i \in D_k^M$. After a deviation of an intermediary, a consumer choose 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 $U_i(\cdot, D^M_{-i})$. The passive belief implies that after any deviation, consumer i's (perceived) gross payoff is given by $U_i(\cdot, D^M_{-i})$.

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 $d_i \not\in D^M_k$, because consumer i will then share the same data with other intermediaries. Second, suppose that intermediary k chooses to not collect d_i such that $d_i \in D^M_k$. This weakly decreases k's payoff if $U_i(d_i, D^M_{-i}) \geq 0$, because k collects d_i for free. Suppose $U_i(d_i, D^M_{-i}) < 0$. 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 $-U_i(d_i, D^M_{-i}) > 0$. Since a monopoly intermediary finds it optimal to collect d_i, $\Pi(D^M) - \Pi(D^M \setminus \{d_i\}) \geq -U_i(d_i, D^M_{-i})$. This completes the proof. □

Proof of Claim 5. 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 $(\mathcal{D}_i, U_i(\mathcal{D}_i, \mathcal{D}_{-i}) - U_i(D_i^*, \mathcal{D}_{-i}))$, 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 3 by replacing $U_i(\cdot)$ with $U_i(\cdot, \mathcal{D}_{-i})$. □