Competing Data Intermediaries

Forthcoming at Rand Journal of Economics

Shota Ichihashi*

Abstract

I study a model of competition between data intermediaries, which collect personal data from consumers and sell them to downstream firms. Competition has a limited impact on benefiting consumers: If intermediaries offer high compensation for data, consumers share data with multiple intermediaries, which lowers the downstream price of data and hurts intermediaries. Anticipating this, intermediaries offer low compensation for data. Although consumers are exclusive suppliers of data, the nonrivalry of data can lead to concentration and high intermediary profits in data markets. In particular, if downstream firms use data to extract surplus from consumers, competing intermediaries sustain a monopoly outcome.

*Bank of Canada; shotaichihashi@gmail.com. The previous version of the paper was circulated under the title "Non-Competing Data Intermediaries." I thank the editor, Gary Biglaiser, and two referees for comments and suggestions that greatly improved the paper. For valuable suggestions, 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, Digital Economics Conference at Toulouse, and DERN Workshop on Data and Competition. The opinions expressed in this paper 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 through targeted advertising. Data brokers, such as Acxiom and Nielsen, collect consumer data and sell them to retailers and advertisers (**?**). I model these companies as data intermediaries, which distribute personal data between consumers and downstream firms.

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

I ask whether competition between data intermediaries benefits consumers, and how competition for data differs from competition for physical goods in traditional markets. The question relates to recent policy debates on competition in digital markets (**???**).

To answer the question, I study a model that consists of a consumer (she), data intermediaries, and a downstream firm. The consumer has a finite set of data (or data labels), such as her email address, location, and purchase history. First, each intermediary chooses the set of data to collect and how much compensation to offer. Second, the consumer decides whether to accept the offer of each intermediary; if she accepts, she provides the requested data and receives compensation. Finally, after observing what data other intermediaries have collected, the intermediaries post prices and sell the data to the downstream firm.[1]

The first key feature of the model is that data are nonrivalrous—i.e., the consumer can provide the same data to multiple intermediaries and earn compensation from all of them. The second key feature is that the consumer's payoff depends on the set of data the downstream firm acquires. For example, the firm may use acquired data for price discrimination, which can benefit or harm the consumer, depending on what data the firm uses to infer their willingness to pay.

[1] Section **??** motivates the assumption that each intermediary observes what data other intermediaries collect. The section also discusses several ways to relax the assumption.

To highlight the difference between competition for nonrivalrous data and competition for rivalrous goods, I begin with the benchmark case of rivalry, in which the consumer can provide her data to at most one intermediary. In this case, the intermediary that offers the highest compensation exclusively obtains the data and becomes a downstream monopolist. In equilibrium, the consumer receives full surplus and intermediaries make no profits, because they spend the downstream monopoly profit to acquire data.

The nonrivalry of data changes the nature of competition. Section **??** assumes that the consumer holds one unit of data. If intermediaries offer high compensation for nonrivalrous data, the consumer will share her data with all of them, which reduces the price of data and the revenues of intermediaries in the downstream market. Anticipating such an outcome, intermediaries choose to not increase compensation for data. As a result, unlike the case of rivalry, intermediaries avoid competition for data and earn a positive profit.

The extent to which the nonrivalry of data relaxes competition depends on the consumer's preferences on the downstream firm's data acquisition. When the preferences are such that the consumer incurs a loss from the firm's data acquisition (e.g., harmful price discrimination), a single intermediary earns a monopoly profit: In equilibrium the intermediary collects data at positive compensation that just covers the consumer's loss from the downstream data acquisition. In contrast, when the consumer benefits from data acquisition, intermediaries cannot sustain a monopoly outcome. A monopoly intermediary would offer negative compensation (or equivalently, charge a positive fee) to extract the consumer's gain that stems from data acquisition. Competition eliminates such negative compensation, because the consumer shares her data with only one of the intermediaries that offer negative compensation. In such a case, competition shifts surplus from intermediaries to the consumer.

To provide general insights, Section **??** assumes that the consumer has any set of data, and her payoff from the downstream firm's data acquisition depends arbitrarily on the subset of data it acquires. I characterize an equilibrium that maximizes intermediary surplus and minimizes consumer surplus. Competition occurs only for the subset of data that the downstream firm uses to

benefit the consumer. As a result, consumer surplus and intermediary surplus fall between those in the monopoly market and those in markets for rivalrous goods.

With an additional assumption that the consumer incurs an increasing marginal loss of sharing her data, I characterize the equilibria with the following properties. First, intermediaries collect disjoint sets of data. Second, each intermediary acts as a local monopsony—i.e., it pays the consumer the minimum compensation to cover her losses of sharing the data. I compare these equilibria in terms of their degrees of data concentration. In a more concentrated equilibrium, a few intermediaries collect large sets of data at low compensation, leading to high intermediary surplus and low consumer welfare.

The key takeaway of the paper is as follows. The nonrivalry of data enables consumers to share the same data with multiple intermediaries and earn rewards from all of them. Anticipating such opportunistic behavior, intermediaries compete for data less aggressively, possibly resulting in high intermediary surplus and low consumer welfare. The magnitude of this effect depends on how the downstream firm uses data: Competition is less effective and the intermediation of data is more profitable when the gain from the final data usage is skewed toward the downstream firm. The result provides insights into the policy discussion on data: First, emerging services that pay for data may not promote competition. Second, weak competition can cause data concentration, and a policy that mitigates concentration can benefit consumers. The paper also points to key variables that affect competition, such as the degree of differentiation between intermediaries and the kind of contracts they can use (see extensions in Section **??**).

The rest of the paper is as follows. Section **??** discusses related literature, and Section **??** describes the model. Section **??** studies two benchmarks: the case of a monopoly intermediary and the case of competition for rivalrous goods. Section **??** and Section **??** present the main analysis. Section **??** provides extensions and Section **??** concludes.

2 Related Literature

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

Vertical contracting. The paper relates to the literature of vertical contracting, common agency, and contracting with externalities (**???????**). In my model the consumer's opportunism—i.e., her inability to commit to not share data with multiple intermediaries—may decrease her welfare. Similarly, in the secret bilateral contracting, the upstream manufacturer's opportunism lowers its profit (e.g., **?**).

However, my paper and the literature differ in two important aspects. First, in this literature the upstream manufacturer's profit depends on the production cost and payments from retailers, but it does not depend on how downstream buyers use the final goods. In contrast, consumers in the upstream data market care about how firms use their data—e.g., price discrimination, targeting, or fraud detection. As a result, in my model the consumer's payoff depends on the set of data the downstream firm acquires. This feature enables me to study the relation between (i) the impact of the consumer's opportunism and (ii) externalities that the downstream firm's data usage imposes on the consumer. The relation between (i) and (ii) can differ from what the literature predicts. For example, the consumer may not suffer from her opportunism when the externality in (ii) is positive.

Second, in secret bilateral contracting, opportunism lowers the profits of intermediaries. In contrast, in my model the consumer's opportunistic behavior never arises on the equilibrium path, and whenever it occurs off the equilibrium path, it will increase the equilibrium profits of intermediaries. As Online Appendix B shows, the difference arises because intermediaries holding the same data will face more severe downstream price competition than retailers buying and selling the same physical goods in the secret contracting model. The difference is important for understanding when competition can fail to curb the market power of data intermediaries.

Markets for data. Recent work, such as **?**, **?**, and **?**, studies platforms that collect data. They focus on a monopolist and assume that data usage harms consumers. I study competition, and the

consumer in my model may benefit or lose depending on the data collected. The papers cited above show the inefficiency of data markets. In contrast I study how competition affects the division of surplus created by data, and how the effect of competition depends on downstream firms' data usage.[2]

Recent work also studies how firms monetize data. **?** study data brokers' incentives to merge data. I abstract away from contracting between intermediaries but consider data collection in the upstream market. **?** study an online advertising auction, in which data improve the quality of the matches between users and advertisers. **?** study the optimal design of a platform's data-protection policies. **?** studies a semi-endogenous growth model that incorporates data intermediaries. **?** incorporates privacy concerns and competition on an advertising platform. **?** study the aggregation of consumers' purchase histories in a dynamic model. **?** employ the competition-in-utilities approach to study competition and data.

Finally, this paper relates to the literature on information goods, such as patents and digital goods (e.g., **???**). The novelty of my paper is to consider the upstream market in which consumers provide data. My paper abstracts from important issues relevant to information goods, such as network effects and versioning.

Two-sided markets. This paper also relates to the literature on two-sided markets (e.g., **???????**). I show that the nonrivalry of data relaxes competition, which echoes the finding of the literature that multi-homing can relax platform competition (e.g., **?** and **?**). The main difference is that in my model the consumer's benefit or loss of sharing data depends on what data are collected. To my knowledge, such a setting does not have a counterpart in the literature, in which consumers usually enjoy benefits on platforms.[3]

[2] I abstract away from other important issues on privacy, such as the privacy paradox, behavioral biases, and the ethical aspects of privacy (e.g., **?** and **?**).

[3] **?** and **?** consider a model of platform competition in which advertisers impose negative externalities on consumers.

3 Model

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

Upstream Market

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

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

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

Downstream Market

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

Preferences

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

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

Suppose the firm obtains data $D \subset \mathcal{D}$ and pays a total price of p to intermediaries. The firm obtains a payoff of $\Pi(D) - p$, where $\Pi(D)$ is the firm's revenue from data D. The revenue function $\Pi(\cdot)$ is strictly increasing and satisfies $\Pi(\emptyset) = 0$.[4]

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

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

When the consumer holds one unit of data (i.e., $|\mathcal{D}| = 1$) as in Section **??**, the assumption is necessary for nontrivial results. For a general $\mathcal{D}$, the assumption holds, for example, if the firm sells products and can use data $\mathcal{D}$ to efficiently price discriminate consumers (Online Appendix C microfounds $U(\cdot)$ and $\Pi(\cdot)$ with this interpretation). In terms of primitives, the assumption holds if the firm's marginal revenue from data is high relative to the consumer's marginal loss of sharing data. Online Appendix A considers multiple consumers, and argues that the counterpart of Assumption **??** is likely to hold if there are negative data externalities between consumers.

Timing

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

Solution

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

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

Modeling Assumptions

I discuss important modeling assumptions.

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

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

The allocation of data is publicly observable. The model assumes that intermediaries and the firm observe what data each intermediary collects. In practice, some data intermediaries disclose what kinds of data they collect. For example, a data broker CoreLogic states it holds property data that cover more than 99.9% of U.S. property records.[5] Also, when an intermediary collects data from consumers or sell data to firms, the intermediary needs to reveal what data it deals with—e.g., Nielsen Homescan states it collects purchase records. Finally, the assumption is natural when a regulation requires that companies disclose what kinds of data they collect.[6]

One concern is that an intermediary may not observe *whose* data other intermediaries hold. For example, we may know Google holds search queries but may not know whose search queries it holds. To capture such a situation, Section **??** shows that the main result continues to hold when (i) there is a continuum of consumers, and (ii) each intermediary does not observe the identities of

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

[6] For example, the General Data Protection Regulation requires that a company inform users the categories of personal data it collects. See, e.g., https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/principles-gdpr/.

consumers in other intermediaries' data. The extension also justifies the assumption on the space of possible contracts, as I discuss next.

The restriction on the contract space. Intermediaries cannot offer compensation that depends on outcomes such as the allocation of data. The assumption is crucial in the baseline model. For example, suppose intermediary k could commit to pay positive compensation if and only if the consumer provides her data *only* to intermediary k. Exclusive offers render data rivalrous, and competition yields the consumer full surplus. The baseline model excludes such offers.

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

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

Pure-strategy equilibrium. I study pure-strategy equilibrium (PSE) for two reasons. First, PSE captures the intuition that the nonrivalry of data relaxes competition—e.g., there is a PSE in which one intermediary acts as a monopolist. Second, PSE facilitates the analysis. For instance, we can derive a PSE even if the consumer's payoff is non-monotone in the set of data acquired. Also, we can compare pure-strategy equilibria in terms of their degrees of data concentration. Such a comparison is difficult in mixed strategy equilibrium (MSE), because the allocation of data is ex ante uncertain. However, Section **??** shows the main insight can hold even in MSE.

Timing. In practice, consumers first decide which platforms to join, after which they use the services and generate data. I assume these events occur simultaneously. Such a timing assumption would be reasonable if a platform does not change the value of its service after consumers join

it. Data collection may affect consumers long after they provide data (e.g., data breach). We can model such a situation by interpreting $U(\cdot)$ as a discounted utility.

In the downstream market, intermediaries first observe the allocation of data, then choose prices. The assumption is similar to that of endogenous product differentiation, in which sellers choose prices after observing their product design (e.g., **?**). What data an intermediary collects can be a part of platform design or a company's policy. After collecting data, platforms and data brokers typically share the data in exchange for money. If so, intermediaries can adjust downstream prices more quickly than adjusting what data to collect.

Applications

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

Several remarks are in order. First, in practice, advertising platforms use data to match users with advertisers, instead of reselling data.[7] The downstream market of my model abstracts away from such details. However, the model captures a general idea—that platforms have a higher willingness to pay for the data that others do not hold (Lemma **??** in the Appendix shows that the price of data is high when other intermediaries do not hold them). For example, a platform will have a higher willingness to acquire health data when its rivals do not have them, because the platform will be the only one that can display ads based on users' health profiles. When this economic force is present in the downstream interactions, the insights in this paper would be relevant (see also the discussion in Section **??**).

Second, $U(\cdot)$ is exogenous—i.e., intermediaries cannot affect how the firm uses data. This reflects the difficulty of writing a contract over how third parties use data. A similar assumption

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

appears in recent papers, such as **?** and **?**.

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

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

Data brokers. Data brokers collect personal data from online and offline sources, and resell or share that data with others, such as retailers and advertisers (**?**). Some data brokers obtain data from consumers in exchange for monetary compensation (e.g., Nielsen Home Scan). At the same time, data brokers commonly obtain data without interacting with consumers. The model could also fit such a situation. For example, suppose data brokers obtain individual purchase records from retailers. Consider the following chain of transactions: Retailers compensate customers and record their purchases—e.g., they offer discounts to customers who sign up for loyalty cards. Retailers then sell these records to data brokers, which resell the data to third parties. We can regard retailers in this example as consumers in the model.

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

Personal data marketplaces. Startups, such as Datum, Killi, and Universal Basic Data Income, attempt to provide marketplaces in which consumers can monetize their data. For example, Killi

pays users according to the types of data they provide.[8] Even an incumbent provides such a service—e.g., Amazon has launched Amazon Shopper Panel, which pays consumers for data on non-Amazon purchases.[9]

Mobile application industry. **?** empirically show that mobile application developers trade greater access to personal information for lower app prices, and consumers trade off lower prices and greater privacy. Also, app developers share collected data with third parties for direct monetary benefit (see **?** and references therein).

4 Two Benchmarks

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

Monopoly Intermediary ($K = 1$)

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

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

Competition for Rivalrous Goods

Suppose data are rivalrous—i.e., the consumer can provide each piece of data to at most one intermediary.[10] The model captures competition between intermediaries for physical goods (all proofs are in the Appendix).

[8] For example, a user who shares their individual profile alone will earn $1 per month, whereas a user who additionally shares shopping and browsing data (collected through a browser extension) will earn $3 per month with their home currency. See https://killi.io/ (accessed on December 25, 2020).

[9] See https://techcrunch.com/2020/10/20/amazon-launches-a-program-to-pay-consumers-for-the (accessed on January 8, 2021).

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

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

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

5 Single Unit Data

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

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

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

The intuition is that competition incentivizes intermediaries to decrease positive fees to zero, but does not incentivize them to increase positive compensation beyond a monopoly level. To see this, suppose intermediary 1 collects data at a positive fee. Then intermediary 2 can undercut the fee to exclusively obtain the data: When the consumer faces the two offers with positive fees, she shares her data with *only* intermediary 2, because she receives a gross utility of U so long as the firm obtains data from at least one intermediary. As a result, competing intermediaries cannot charge a positive fee. If $U > 0$ the consumer receives a payoff of at least U.

In markets for rivalrous goods, the Bertrand competition in the upstream raises compensation to $\Pi + U > 0$. In contrast, such competition may fail in data markets. For example, suppose intermediary 1 collects data at monopoly compensation $-U$ with $U < 0$. If intermediary 2 also offers positive compensation, the consumer provides her data to *both* intermediaries, and the downstream price of data will become zero. Anticipating such an outcome, intermediary 2 makes no competing offer.

The effect of competition depends on how the downstream firm uses data. To see this, we fix any $TS > 0$ and study how the equilibrium changes along the iso-total surplus line—i.e., the set of all $(\Pi, U) \in \mathbb{R}^2$ such that $\Pi + U = TS$. Along the line, the equilibrium surplus of intermediaries, $TS - \max(0, U)$, is decreasing in U and maximized at any (Π, U) such that $U < 0$. That is, the intermediation of data is more profitable when the gain from data usage is skewed toward the downstream firm. Correspondingly, as U increases from a negative value to TS (and Π decreases to maintain $\Pi + U = TS$), consumer surplus increases from zero to the maximum value, TS. In contrast, for rivalrous goods the equilibrium is independent of the composition of Π and U with a fixed $\Pi + U$.

Discussion on Exclusive Data Acquisition

Proposition **??** implies intermediaries do not hold the same data. In practice there seem to be counterexamples—e.g., online advertising intermediaries sell the same targeting data to advertisers via their ad networks. In this respect, we should interpret Proposition **??** as an approximation: As in models of product differentiation, intermediaries will have an incentive to differentiate themselves in terms of what data they hold (**???**).[11] The incentive to differentiate and the nonrivlary of data discourage intermediaries from paying for data their rivals already hold. As a result, the upstream market for data becomes less competitive than the market for rivalrous goods. The model captures this intuition in an extreme way: Overlapping data will have a downstream price of zero,

[11] Although empirical evidence on data markets is sparse, an industry expert **?** describes that "data brokers tend to specialize in certain industries in order to gain a competitive advantage." **?** describe that different data brokers may be particularly strong in collecting different kinds of consumer information.

so intermediaries have no incentive to collect data their rivals already hold.

How to Improve Consumer Surplus?

We now relate the results to a policy discussion on consumer data. The idea that consumers may not be properly compensated for their data provision is not new (**?**). Policy reports, such as **?** and **?**, suggest such a situation may result from the lack of competition. For example, **?** states "it might have been that with more competition consumers would have given up less in terms of privacy or might even have been paid for their data." At the same time, an increasing number of startups and tech companies pay for consumer data (e.g., Killi and Amazon Shopper Panel; see Section **??**). An important question is whether these services intensify competition for data and improve consumer surplus. The paper provides an intuition about when such a competition may fail. If consumers can easily share the same data with multiple intermediaries, an intermediary that offers the highest compensation may not become a monopolist for that data (e.g., consumers may have provided similar data to incumbents). If so, competition for consumer data does not work in a way we expect in traditional markets.

To provide further insights, Section **??** illustrates two factors that contribute to low consumer welfare. First, I assume that the consumer may view compensation from intermediaries as substitutes. Second, I study differentiated intermediaries that may earn positive revenues in the downstream market even if they hold the same data. The extensions show that the monopolistic equilibrium continues to exist if the degree of upstream differentiation is high and the degree of downstream differentiation is low. The result that the consumer obtains zero payoffs is more general: It holds, for example, if intermediaries are so differentiated that they collect the same data in equilibrium. The extensions clarify when a regulator can rely on competition to curb the market power of data intermediaries, in terms of their degrees of differentiation.

Finally, the paper points to two possible ways to increase consumer surplus.

1. Stronger bargaining power of consumers. Consumer surplus increases if they have more bargaining power. For example, if the consumer can make a take it or leave it offer to intermediaries,

she will offer $(\{d\}, \Pi)$ to extract full surplus. Shifting bargaining power can be challenging, but **?** suggest that wireless carriers could bargain with platforms on behalf of its subscribers for payments for their data. Competition between carriers would transfer the surplus to consumers, provided they do not subscribe multiple wireless carriers.[12]

2. Rich contract space. Another way is to enable intermediaries to use richer contracts. For example, suppose each intermediary can base compensation on whether the consumer shares her data with other intermediaries. In equilibrium each intermediary k commits to pay all of its downstream revenue (i.e., Π) if and only if the consumer provides data to only intermediary k. Enforcing richer contracts require transparency. For example, an intermediary may need to monitor how consumers interact with its rivals. Tracking consumers across its competitors is difficult if consumers could change their identities to transact with different intermediaries. In the absence of intermediaries' ability to track consumers, we continue to have the same equilibrium as in Proposition **??**, even if intermediaries can use intricate compensation mechanisms. Section **??** shows such a result.

6 General Preferences

I generalize Proposition **??** to any consumer preferences, then show the multiplicity of equilibria that stems from the nonrivalry of data. The consumer now has any finite set of data and any gross utility function $U(\cdot)$. The downstream firm has any strictly increasing revenue function $\Pi(\cdot)$. The functions $U(\cdot)$ and $\Pi(\cdot)$ satisfy Assumption **??**. The following result generalizes Proposition **??**.

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

If the consumer holds a single piece of data d, then $\max_{D \subset \mathcal{D}} U(D) = \max(0, U(\{d\}))$. As

[12] If the consumer chooses one of the "carriers" that will make take it or leave it offers to sell data to intermediaries, competing carriers offer to pay Π to the consumer in equilibrium. The consumer's single-homing is natural if carriers apply compensation as discounts to positive net prices of their services.

a result, the PME equals the unique equilibrium in Proposition **??**. Proposition **??** states that the intuition for Proposition **??** applies to general preferences.

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

The PME extends the monopoly equilibrium: If there are many intermediaries, it minimizes consumer surplus and maximizes intermediary surplus across all equilibria. Let $CS(K) \subset \mathbb{R}_+$ denote the set of all pure-strategy subgame perfect equilibrium (SPE) payoffs of the consumer when there are K intermediaries.

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

1. *The PME minimizes consumer surplus:* $\min CS(K) = \max_{D \subset \mathcal{D}} U(D)$.
2. *The PME maximizes the intermediaries' joint profit across all pure-strategy SPE.*

The intuition is as follows. Suppose there are K intermediaries, and in some equilibrium the consumer obtains a payoff of $U(D^*) - \delta_K$ with $\delta_K > 0$. If an intermediary offers (D^*, ε) with $\varepsilon < \delta_K$, the consumer will accept it. Because any intermediary can always deviate and offer (D^*, ε), each intermediary obtains a payoff of at least δ_K. Thus intermediary surplus is at least $K \cdot \delta_K$. However, intermediary surplus is bounded from above by $\Pi(\mathcal{D}) + U(\mathcal{D}) < \infty$. As a

result, δ_K goes to 0 as K grows large—i.e., as the number of intermediaries grows large, the worst consumer surplus converges to $U(D^*)$, which is the consumer's payoff in the PME. In the proof, I show that δ_K hits zero for a finite K, when $\Pi(\cdot)$ is strictly increasing. In the PME total surplus is maximized and consumer surplus is $U(D^*)$. Thus the PME is intermediary-optimal for a large K.

The results imply that the impact of competition for data depends on how downstream firms use data. In a frictionless market for rivalrous goods, for any $U(\cdot)$, competition gives full surplus to players in the upstream market. In markets for data, the shape of $U(\cdot)$ affects the division of surplus. If data usage benefits consumers, competition eliminates fees that consumers would have to pay under monopoly. If data usage harms consumers, competition may not increase compensation. As a result, under general preferences, competition weakly increases consumer welfare and decreases intermediary profit, but not as much as in markets for rivalrous goods.

With an additional assumption I examine the impact of data concentration. The consumer now incurs an increasing convex cost of sharing data, and the downstream firm faces a decreasing marginal revenue from data.

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

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

In the case of rivalrous goods (i.e., Claim **??**), the equilibrium allocation is typically a trivial partition.

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

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

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

Proposition 4. *Under Assumption* **??***, if an allocation of data* $(D_k^*)_{k\in K}$*, compensation* $(\tau_k^*)_{k\in K}$*, and downstream prices* $(p_k^*)_{k\in K}$ *consist of a partitional equilibrium, they satisfy the following conditions:*

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

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

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

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

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

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

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

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

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

The following result shows the welfare implications of data concentration.

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

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

The recent discussion on data suggests that the concentration of data within a small number of companies may harm consumers by relaxing subsequent competition, such as the one in an online advertising market (e.g., **?**). Proposition **??** shows that data concentration can also harm consumers by reducing compensation for data. The following example shows that a policy which limits data concentration can benefit the consumer.

Example 1. The consumer has location and financial data. The downstream firm profits from data at the risk of data leakage. The consumer incurs an expected loss of \$20 from the potential data leakage if only if the firm holds both location and financial data. Otherwise she incurs no loss. Two intermediaries collect and sell data. In one equilibrium, intermediary 1 obtains both location and financial data and pays \$20 to the consumer, leading to zero consumer surplus. In another equilibrium, intermediaries 1 and 2 collect location and financial data, respectively, and each intermediary pays compensation \$20. The consumer earns a net surplus of \$20.

A policy that prevents data concentration can benefit consumers. In this example, a regulator can achieve the goal by requiring that each intermediary collect at most one type of data. Alternatively, if intermediaries try to attain a monopoly outcome through a merger and otherwise they would play the less concentrated equilibrium, blocking the merger prevents data concentration.

7 Extensions

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

Multiple Consumers with a Rich Contract Space

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

We now consider a unit mass of consumers. Consumer $i \in [0,1]$ has a set of data, $\mathcal{D}_i = \left\{d_i^\ell\right\}_{\ell\in\mathcal{L}}$, where $\mathcal{L}$ is a finite set of data labels, such as $\mathcal{L} = \{location, health\}$. The set of all data is $\mathcal{D} := \cup_{i\in[0,1],\ell\in\mathcal{L}} \left\{d_i^\ell\right\}$, and the allocation of data is written as $(D_k)_{k\in K}$, where $D_k \subset \mathcal{D}$ is the set of data that intermediary k holds. For any $D \subset \mathcal{D}$, we write $q^\ell(D) \in [0,1]$ for the amount of data $\ell \in \mathcal{L}$ contained in D.[14] Given an allocation of data $(D_k)_{k\in K}$, the amount of data ℓ intermediary k holds is $q^\ell(D_k)$. I call $(q^\ell(D_k))_{\ell\in\mathcal{L},k\in K}$ a *quantity vector*. The choice of an individual consumer, who is atomless, does not affect the quantity vector. Finally, let $\mathcal{Q}_{-k} := [0,1]^{|\mathcal{L}|\times(K-1)}$ denote the set of all quantity vectors for intermediaries other than k.

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

Intermediaries have limited information, but can use richer contracts than in the baseline model: At the beginning of the game, each intermediary k chooses a *contract*, which is a mapping $\gamma_k : 2^{\mathcal{L}} \times \mathcal{Q}_{-k} \to \mathbb{R}$. For any $L \subset \mathcal{L}$ and $q_{-k} \in \mathcal{Q}_{-k}$, any consumer i who provides data $\left\{d_i^\ell\right\}_{\ell\in L}$ to intermediary k receives compensation $\gamma_k(L, q_{-k})$, if the quantity vector of other intermediaries is q_{-k}. Intermediaries can choose any contract such that $\gamma_k(\emptyset, q) = 0$ for all q—i.e., a consumer does

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

not pay if she does not share any data. In equilibrium, consumers choose what data to share with each intermediary k, taking q_{-k} as exogenous.

I consider the following payoffs: If consumer i receives compensation τ and the firm acquires i's data $\left\{d_i^\ell\right\}_{\ell \in L}$, her payoff is $U(L) + \tau$, where $U(\emptyset) = 0$. Consumers are homogeneous, in that $U(\cdot)$ is independent of i.[15] The firm's payoff from buying data $D \subset \mathcal{D}$ is $\Pi(D) - p$, where p is the total payment to intermediaries. The revenue function $\Pi(\cdot)$ is an increasing set function, and $\Pi(D)$ depends only on $(q^\ell(D))_{\ell \in \mathcal{L}}$. The payoff of each intermediary is revenue minus compensation.

The timing of the game is the same as before. First, intermediaries simultaneously offer contracts. Then each consumer decides the set of data to share with each intermediary. Intermediaries and the firm observe the realized quantity vector. Finally, intermediaries simultaneously post prices for their datasets, after which the firm makes a purchasing decision. The solution concept is perfect Bayesian equilibrium.[16] I impose the following straightforward extension of Assumption **??**.

Assumption 3. The primitives $U(\cdot)$ and $\Pi(\cdot)$ are such that a monopoly intermediary collects all data in some equilibrium.[17]

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

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

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

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

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

Non-Additive Compensation

In the baseline model the consumer's payoff is additively separable across compensations from intermediaries. However, if compensation is the value of a service, this additive separability may not hold. Thus, I extend the model as follows. For simplicity, assume that the consumer holds one unit of data d, and data collection is harmful, i.e., $L := -U(\{d\}) > 0$. Suppose the consumer shares her data with n intermediaries, receives compensation τ_k from each intermediary $k \in K$, and the firm obtains $D \subset \{d\}$. The consumer's payoff is now $U(D) + T(\tau_1, \ldots, \tau_K) - n \cdot c$. The last term $n \cdot c$ is the cost of sharing data with n intermediaries, and $c \geq 0$ is exogenous. It captures the opportunity cost of using the service provided by an intermediary. The second term $T(\tau_1, \ldots, \tau_K)$ is the effective compensation, which maps a profile $(\tau_1, \ldots, \tau_K)$ of compensations to the consumer's utility.

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

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

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

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

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

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

in which all intermediaries earn zero profits, and the consumer receives a positive payoff.

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

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

Differentiated Intermediaries

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

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

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

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

where $\Pi > 0$, $\sigma \in [0,1]$, and $\Delta_k \geq 0$ for each $k \in K$. The first term Π is the base value of the data. The parameters σ and $(\Delta_k)_{k\in K}$ respectively capture the substitutability and qualities of intermediaries in the downstream market. To see this, assume $K = 2$. A higher $\Delta_1 - \Delta_2$

implies that the offering of intermediary 1 has a higher quality than that of intermediary 2, and a lower σ implies that the offerings of two intermediaries are less substitutable. For example, suppose two intermediaries have the same underlying data and offer a similar targeting compaign, but intermediary 1's campaign has a higher accuracy. Then $\Delta_1 > \Delta_2$ and σ is high. As another example, suppose intermediaries 1 and 2 use the same data to offer targeting and fraud detection, respectively, and their services have high quality. Then Δ_1 and Δ_2 are high, and σ is low.

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

Claim 7. *There is a $\sigma_1 < 1$ such that for any $\sigma \in [\sigma_1, 1]$, the following holds: There is an efficient equilibrium in which intermediary $k^* \in \arg\max_{k \in K} \Delta_k - c_k$ acts as a monopolist. If there is a $\hat{k} \notin \arg\max_{k \in K} \Delta_k - c_k$ such that $\hat{k} \in \arg\max_{k \in K} \Delta_k$ and $\Pi + \Delta_{\hat{k}} > L + c_{\hat{k}}$, there is also an inefficient equilibrium in which intermediary $\hat{k}$ acts as a monopolist.*

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

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

[18] See, e.g., https://www.forbes.com/sites/curtissilver/2020/04/28/killis-fair-trade-data-program-enables-you-to-profit-off-your-data/ (accessed Feb 12, 2021).

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

Claim 8. *Fix any $\sigma \in [0,1)$. There is a $\Delta^* > 0$ such that if $\min_{k\in K} \Delta_k \geq \Delta^*$, there is an equilibrium in which all intermediaries collect data and the consumer obtains a payoff of zero.*

Mixed Strategy Equilibrium

This subsection considers mixed strategy equilibrium (MSE), assuming that the consumer holds one unit of data, and data collection is harmful, i.e., $L := -U(\{d\}) > 0$. I focus on a symmetric MSE in which (i) each intermediary draws compensation from the same distribution, (ii) data collection occurs with a positive probability, and (iii) the consumer rejects a compensation of zero.[19] The main insight extends to this MSE.

Claim 9. *Take any $K \geq 2$ and $\Pi > L > 0$. In any equilibrium that satisfies (i) - (iii) above, consumer surplus is at most half of total surplus. This bound is tight: The ratio of consumer surplus to total surplus converges to $\frac{1}{2}$ as $\frac{L}{\Pi} \to 1$.*

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

Multiple Consumers with Data Externalities

Online Appendix A studies an extension in which there are multiple consumers, and the payoff of each consumer may depend on what data other consumers provide to the firm. The main insight remains the same: There is an equilibrium in which one intermediary collects all data. If the

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

downstream firm's data usage harms consumers, there is a monopolistic equilibrium. In this general setting, an equilibrium may be inefficient, because consumers do not consider how providing data affects the welfare of other consumers. As a result, consumer surplus can be lower than in the absence of intermediaries. The findings are in line with recent work on data externalities (**???**).

Multiple Downstream Firms

I have assumed that the downstream market consists of one firm. However, Proposition **??** holds even if the downstream market consists of multiple firms that have heterogeneous values for the data, and intermediaries only know the distribution of their values. Under a stronger assumption that intermediaries know downstream firms' willingness to pay and can price discriminate them, all of the results in this paper hold. Indeed, provided downstream firms do not interact with each other, we can define $\Pi(\cdot)$ as the sum of the revenues of all the firms, and $U(\cdot)$ as the aggregate effect of data acquisition. Online Appendix A formalizes this idea.

8 Conclusion

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

Appendix

Competition for Rivalrous Goods: Proof of Claim ??

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

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

Equilibrium for Single Unit Data: Proof of Proposition ??

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

First, I show that for any equilibrium in which the consumer sells data to at least one intermediary, the total compensation τ^* that she earns is weakly greater than $\max(0, -U)$. First, consider

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

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

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

The above SPE maximizes the joint profit of intermediaries among all SPEs, because the consumer receives the minimum possible compensation, the firm obtains zero profit, and the outcome (i.e., the firm acquiring d) maximizes total surplus. Also, in this equilibrium one intermediary extracts this maximized joint profit. This implies that if there is another equilibrium that is Pareto-undominated from the perspectives of intermediaries, then in such an equilibrium, multiple in-

termediaries must be earning positive profits. However, there is no such equilibrium because an intermediary earns positive profit only by selling d to the firm at a positive price, which occurs only if one intermediary collects d.

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

Partially Monopolistic Equilibrium: Proof of Proposition ??

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

The suggested strategy profile is an equilibrium. First, intermediary 1 has no incentive to deviate. To see this, suppose intermediary 1 deviates and obtains data $D_1 \subset \mathcal{D}$. Let $\hat{D}$ denote the set of all data that the consumer shares as a result of intermediary 1's deviation ($D_1 \subsetneq \hat{D}$ if she also shares data with some intermediary $j \neq 1$). The revenue of intermediary 1 in the downstream market is at most $\Pi(\hat{D})$. The compensation τ to the consumer has to satisfy $\tau \geq U(D^*) - U(\hat{D})$. To see this, suppose $U(D^*) > U(\hat{D}) + \tau$. The left-hand side is the payoff that the consumer can attain by sharing data exclusively with intermediary $k > 1$. The right hand side is her maximum payoff conditional on sharing data with intermediary 1. Note that all intermediaries other than 1

offer zero compensation. Then, $U(D^*) > U(\hat{D}) + \tau$ implies the consumer strictly prefers to reject the offer from intermediary 1. These bounds on revenue and cost imply intermediary 1's payoff after the deviation is at most $\Pi(\hat{D}) - [U(D^*) - U(\hat{D})] = \Pi(\hat{D}) + U(\hat{D}) - U(D^*)$. Assumption **??** implies this expression is at most $\Pi(\mathcal{D}) + U(\mathcal{D}) - U(D^*) = \Pi(\mathcal{D}) - [U(D^*) - U(\mathcal{D})]$, which is intermediary 1's payoff without deviation. Thus no deviation is profitable for intermediary 1.

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

Welfare Properties of PME: Proof of Proposition ??

We prepare several notations. Define $U^* := \max_{D \subset \mathcal{D}} U(D)$, and $TS^* := \Pi(\mathcal{D}) + U(\mathcal{D}) \geq 0$. Assumption **??** implies TS^* is the maximum total surplus. Define $m := \min_{d \in \mathcal{D}, D \subset \mathcal{D}} \Pi(D) - \Pi(D \setminus \{d\}) > 0$. Let K^* satisfy $K^* > TS^*/m$. Suppose there are $K \geq K^*$ intermediaries, and take any equilibrium. Suppose to the contrary that the consumer's payoff is $U(D^*) - \delta$ with $\delta > 0$. I derive a contradiction by assuming that any intermediary obtains a payoff of at least m. If intermediary k deviates and offers (D^*, ε) with $\varepsilon \in (0, \delta)$, the consumer accepts this offer. Let D_{-k} denote the data the consumer shares with intermediaries in $K \setminus \{k\}$ as a result of k's deviation. Then, $D^* \setminus D_{-k} \neq \emptyset$ holds. To see this, suppose to the contrary that $D^* \subset D_{-k}$. Then, the consumer could be strictly better off by rejecting intermediary k's offer (D^*, ε) because

$\varepsilon > 0$. However, conditional on rejecting k's deviating offer, the set of offers the consumer faces shrinks relative to the original equilibrium. Thus, the maximum payoff the consumer can achieve by rejecting k's deviating offer is at most $U(D^*) - \delta < U(D^*) - \varepsilon$, which is a contradiction. Because the consumer accepts the offer of intermediary k and $D^* \setminus D_{-k} \neq \emptyset$, intermediary k can earn a profit arbitrarily close to m. This implies that in the equilibrium, any intermediary earns a payoff of at least m. However, if each intermediary earns at least m, the sum of payoffs of all intermediaries is at least $Km > TS^*$. This implies that the consumer or the firm obtains a negative payoff, which is contradiction. Thus in any equilibrium, the consumer obtains a payoff of at least $U(D^*)$. The PME then minimizes the consumer's payoff across all pure-strategy SPE. Because the PME maximizes total surplus while giving the lowest payoffs to the consumer and the firm, it maximizes the intermediaries' joint profit for any $K \geq K^*$. □

Proof of Claim ??

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

Partitional Equilibria: Proof of Proposition ??

I first present the unique equilibrium outcome of the downstream market.[20]

[20] Lemma **??** generalizes Proposition 18 of **?** 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 $j \neq k$. **?** assume $K = 2$ and consider not only submodularity but also supermodularity.

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

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

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

Proof. Take any allocation of data $(D_1, \ldots, D_K)$. I show that there is an equilibrium (of the downstream market) in which each intermediary k posts a price of Π_k and the firm buys all data. First, the submodularity of Π implies that $\Pi(\cup_{k\in K'\cup\{j\}} D_k) - \Pi(\cup_{k\in K'} D_k) \geq \Pi_j$ for all $K' \subset K$. Thus, if each intermediary k sets a price of Π_k, the firm prefers to buy all data. Second, if intermediary k increases its price, the firm strictly prefers buying data from intermediaries in $K \setminus \{k\}$ to buying data from a set of intermediaries containing k. Finally, if an intermediary lowers the price, it earns a lower revenue. Thus, no intermediary has a profitable deviation.

To prove the uniqueness of equilibrium payoffs, I first show that the equilibrium revenue of each intermediary k is at most Π_k. Suppose to the contrary that (without loss of generality) intermediary 1 obtains a strictly greater revenue than Π_1. Let $K' \ni 1$ denote the set of intermediaries from which the firm buys data.

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

Second, define $K^* := \{\ell \in K : \ell \notin K', p_\ell = 0\} \cup K'$. Note that K^* satisfies $\Pi(\cup_{k\in K'} D_k) = \Pi(\cup_{k\in K} D_k) = \Pi(\cup_{k\in K^*} D_k)$, $\sum_{k\in K'} p_k = \sum_{k\in K^*} p_k$, and $p_j > 0$ for all $j \notin K^*$. Then, it holds that

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

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 (**??**). 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

$$\Pi(\cup_{k\in K^*} D_k) - \sum_{k\in K^*} \hat{p}_k < \Pi(\cup_{k\in J^*} D_k) - \sum_{k\in J^*} \hat{p}_k, \tag{5}$$

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 (**??**) 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 (**??**) can be maximized by $J^* = K^* \setminus \{1\}$, because Π is submodular and $\Pi(\cup_{k\in K^*} D_k) - \Pi(\cup_{k\in K^*\setminus\{\ell\}} D_k) \geq p_\ell$ for all $\ell \in K^*$. Plugging $J^* = K^* \setminus \{1\}$, we obtain

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

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

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D_k) < \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D_k). \tag{7}$$

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

$$\Pi(\cup_{k\in K^*\setminus\{1\}} D_k) = \Pi(\cup_{k\in (K^*\setminus\{1\})\cup\{j\}} D_k). \tag{8}$$

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 (**??**) 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 (**??**) holds. Such intermediary j can again profitably deviate by lowering its price, which is a contradiction. Therefore, intermediary k's revenue is at most Π_k.

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

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

I now prove Proposition **??**.

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

I show that this strategy profile is an equilibrium. First, the strategy of the consumer is optimal because $U(\cdot)$ is decreasing and submodular. Second, Lemma **??** implies that there is no

profitable deviation in the downstream market. Third, suppose that intermediary k deviates and offers $(\tilde{D}_k, \tilde{\tau}_k)$. Without loss of generality, we can assume that $\tilde{D}_k \subset D_k^*$ for the following reason. If the consumer rejects $(\tilde{D}_k, \tilde{\tau}_k)$, then intermediary k can replace such an offer with $(\emptyset, 0)$. If the consumer accepts $(\tilde{D}_k, \tilde{\tau}_k)$ but $\tilde{D}_k \subsetneq D_k^*$, it means that k obtains some data $d \in \tilde{D}_k \setminus D_k^*$. Because $\cup_k D_k^* = \mathcal{D}$, there is another intermediary that obtains data d. By Lemma **??**, intermediary k is indifferent between offering $(\tilde{D}_k \setminus \{d\}, \tilde{\tau}_k)$ and $(\tilde{D}_k, \tilde{\tau}_k)$. Now, let $D^- := D_k^* \setminus \tilde{D}_k$ denote the set of data that are not acquired by the firm as a result of k's deviation. If intermediary k deviates in this way, its revenue in the downstream market decreases by $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D_k^*) - [\Pi(\mathcal{D} \setminus D^-) - \Pi(\mathcal{D} \setminus D_k^*)] = \Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D^-)$. In the upstream market, if the consumer provides data $\tilde{D}_k$ to k, then it is optimal for the consumer to accept other offers from non-deviating intermediaries, because $U(\cdot)$ is submodular. This implies that the minimum compensation that k has to pay is $U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D} \setminus D^-)$. Thus, k's compensation in the upstream market decreases by $U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}) - [U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D} \setminus D^-)] = U(\mathcal{D} \setminus D^-) - U(\mathcal{D})$. Because collecting $\mathcal{D}$ is an optimal choice of the monopolist, it holds that $\Pi(\mathcal{D}) - \Pi(\mathcal{D} \setminus D^-) - [U(\mathcal{D} \setminus D^-) - U(\mathcal{D})] \geq 0$. Therefore, the deviation does not strictly increase intermediary k's payoff.

Next, we prove the first part. Points 1 and 3 follow from the definition of partitional equilibrium and Lemma **??**, respectively. Let τ_k^* denote the compensation k pays for collecting D_k^*. To show Point 2, suppose to the contrary that $\tau_k^* \neq U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}) > 0$. Suppose that $\tau_k^* < U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. Then, the consumer rejects the offer from intermediary k, which is a contradiction. Next, suppose $\tau_k^* > U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. Then, by the second part of the proposition proved above, we can find an equilibrium that has the same outcome except intermediary k offers a lower compensation $\tau_k' \in (U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D}), \tau_k^*)$ for collecting D_k^*. This equilibrium Pareto-dominates the original equilibrium from the perspectives of intermediaries, which is a contradiction. Thus, we obtain $\tau_k^* = U(\mathcal{D} \setminus D_k^*) - U(\mathcal{D})$. □

Data Concentration: Proof of Proposition ??

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. □

References

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. E. (2019), "Too much data: Prices and inefficiencies in data markets." *NBER Working Paper*.

Acquisti, A., Taylor, C. R., and Wagman, L. (2016), "The economics of privacy." *Available at SSRN 2580411*.

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

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

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., and Weyl, E. G. (2018), "Should we treat data as labor? Moving beyond 'Free'." 108, 38–42.

Bergemann, D., Bonatti, A., and Gan, T. (2019), "The economics of social data." *Cowles Foundation Discussion Paper*.

Bernheim, B. D. and Whinston, M. D. (1986), "Common agency." *Econometrica*, 923–942.

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

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

Carrillo, J. and Tan, G. (2015), "Platform competition with complementary products." *Working paper*.

Choi, J. P., Jeon, D.-S., and Kim, B.-C. (2019), "Privacy and personal data collection with information externalities." *Journal of Public Economics*, 173, 113–124.

Crémer, J., de Montjoye, Y.-A., and Schweitzer, H. (2019), "Competition policy for the digital era." *Report for the European Commission*.

D'Aspremont, C., Gabszewicz, J. J., and Thisse, J.-F. (1979), "On hotelling's 'stability in competition'." *Econometrica*, 1145–1150.

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

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

Fainmesser, I. P., Galeotti, A., and Momot, R. (2019), "Digital privacy." *Available at SSRN*.

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

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

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

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

Gu, Y., Madio, L., and Reggiani, C. (2020), "Data brokers co-opetition." *Available at SSRN 3308384*.

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

Hart, O. and Tirole, J. (1990), "Vertical integration and market foreclosure." *Brookings papers on economic activity. Microeconomics*, 1990, 205–286.

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

Jones, C. I. and Tonetti, C. (2020), "Nonrivalry and the economics of data." *American Economic Review*, 110, 2819–58.

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

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

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

Madio, L., Gu, Y., and Reggiani, C. (2019), "Exclusive data, price manipulation and market leadership." *CESifo Working Paper*.

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

Morton, F. S., Nierenberg, T., Bouvier, P., Ezrachi, A., Jullien, B., Katz, R., Kimmelman, G., Melamed, A. D., and Morgenstern, J. (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*.

O'Brien, D. P. and Shaffer, G. (1992), "Vertical control with bilateral contracts." *The RAND Journal of Economics*, 299–308.

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

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

Rey, P. and Vergé, T. (2004), "Bilateral control with vertical contracts." *RAND Journal of Economics*, 728–746.

Rhodes, A., Watanabe, M., and Zhou, J. (2018), "Multiproduct intermediaries." *Working Paper*.

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

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

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

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

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

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

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

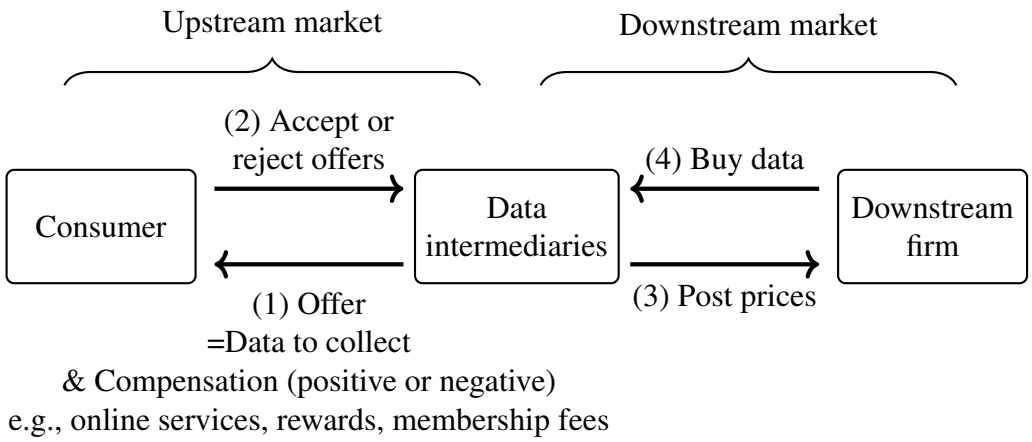

Figure 1: Timing of moves.

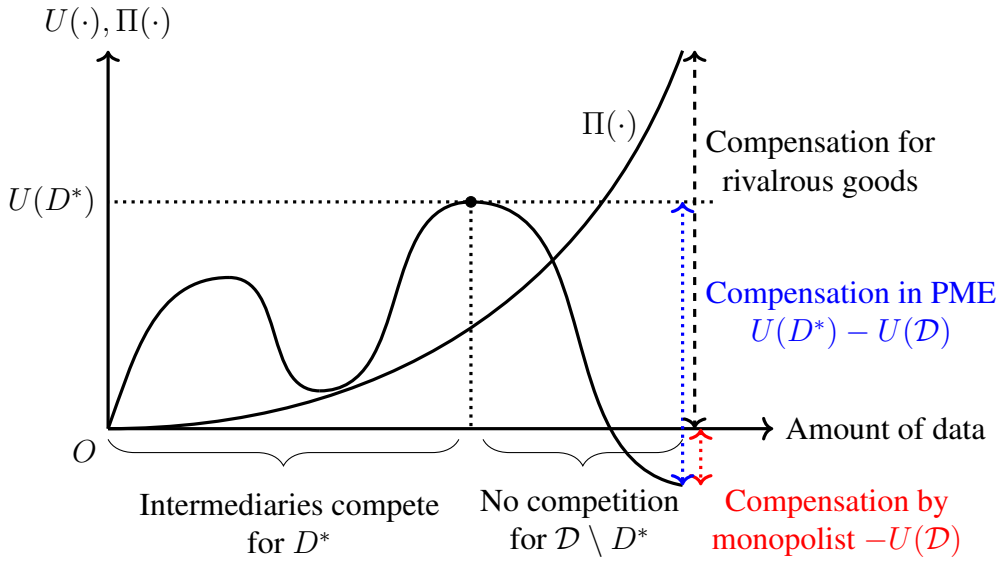

Figure 2: Partially monopolistic equilibrium.