The Economics of Social Data*

Dirk Bergemann† Alessandro Bonatti‡ Tan Gan§

January 21, 2021

Abstract

We propose a model of data intermediation to analyze the incentives for sharing individual data in the presence of informational externalities. A data intermediary acquires signals from individual consumers regarding their preferences. The intermediary resells the information in a product market wherein firms and consumers can tailor their choices to the demand data. The social dimension of the individual data—whereby an individual's data are predictive of the behavior of others—generates a *data externality* that can reduce the intermediary's cost of acquiring the information. We derive the intermediary's optimal data policy and establish that it preserves the privacy of consumer identities while providing precise information about market demand to the firms. This policy enables the intermediary to capture the total value of the information as the number of consumers becomes large.

KEYWORDS: social data; personal information; consumer privacy; privacy paradox; data intermediaries; data externality; data policy; data rights; collaborative filtering.

JEL CLASSIFICATION: D44, D82, D83.

*Bergemann and Bonatti acknowledge financial support through NSF Grant SES-1948692. We thank Joseph Abadi, Daron Acemoğlu, Susan Athey, Steve Berry, Nima Haghpanah, Nicole Immorlica, Al Klevorick, Scott Kominers, Annie Liang, Roger McNamee, Jeanine Miklós-Thal, Enrico Moretti, Stephen Morris, Denis Nekipelov, Asu Özdağlar, Fiona Scott-Morton, Shoshana Vasserman, Glen Weyl, and Kai-Hao Yang for helpful discussions. We also thank Michelle Fang and Miho Hong for valuable research assistance and the audiences at numerous seminars and conferences for their productive comments.

†Department of Economics, Yale University, New Haven, CT 06511, dirk.bergemann@yale.edu.

‡MIT Sloan School of Management, Cambridge, MA 02142, bonatti@mit.edu.

§Department of Economics, Yale University, New Haven, CT 06511, tan.gan@yale.edu.

1 Introduction

Individual Data and Data Intermediaries The rise of large digital platforms—such as Facebook, Google, and Amazon in the US and JD, Tencent and Alibaba in China—has led to the unprecedented collection and commercial use of individual data. The steadily increasing user bases of these platforms generate massive amounts of data about individual consumers, including their preferences, locations, friends, political views, and nearly every facet of their lives. In turn, many of the services provided by large Internet platforms rely critically on these data. The availability of individual-level data allows these companies to offer refined search results, personalized product recommendations, informative ratings, timely traffic data, and targeted advertisements.

A central feature of the data collected from individuals is their social dimension—data captured from an individual user are informative not only about that individual but also about other individuals with similar characteristics or behaviors. In the context of shopping data, an individual's purchases can convey information to a third party about the willingness to pay for a given product among consumers with similar purchase histories. More importantly, data from other individuals can also be informative to a specific individual. For instance, in the context of geolocation data, an individual conveys information about traffic conditions for nearby drivers who can use this information to improve their decisions. Thus, these *individual data* are actually *social data*. The social nature of the data generates a *data externality*, the sign and magnitude of which are not clear a priori. Instead, the sign and magnitude of the data externality depend on the structure of the data and on the use of the gained information.

In this paper, we analyze three critical aspects of the economics of social data. First, we consider how the collection and transmission of individual data change the terms of trade among consumers, firms (advertisers), and data intermediaries (e.g., large Internet platforms that sell targeted advertising space). Second, we examine how the social dimension of the data magnifies the value of individual data for platforms and facilitates the acquisition of large datasets. Third, we analyze how data intermediaries with market power manipulate the trade-offs induced by social data through the aggregation and the precision of the information that they provide about consumers.

A Model of Data Intermediation We develop a framework to evaluate the flow and allocation of individual data in the presence of data externalities. Our model focuses on three types of economic agents: consumers, firms, and data intermediaries. These agents interact in two distinct but linked markets: a *data market* and a *product market*.

In the product market, each consumer (she) determines the quantity that she wishes to purchase, and a single producer (he) sets the unit price at which he offers a product to the consumers. Initially, each consumer has private information about her willingness to pay for the firm's product. This information consists of a signal with two additive components: a *fundamental* component and a *noise* component. The fundamental component represents her willingness to pay, and the noise component reflects that her initial information might be imperfect. Both components can be correlated across consumers: in practice, different consumers' preferences can exhibit common traits, and consumers might undergo similar experiences that induce correlation in their errors. The social dimension of the data—whereby a consumer's data are also predictive of the behavior of others—is central to understanding the consumer's incentives to participate in the data market.

In the data market, a monopolist intermediary acquires demand information from the individual consumers in exchange for a monetary payment. The intermediary then chooses how much information to share with the other consumers and how much information to sell to the producer. In particular, sharing data with each consumer is analogous to providing a personalized purchase recommendation on the basis of other consumers' signals. Selling data to the producer enables him to choose more precise, potentially personalized prices. Thus, the data intermediary has control over the volume and the structure of the information flows across all of the product market participants. However, we maintain the assumption that large data platforms monetize their data by selling the data only to producers.

Direct vs. Indirect Sale of Information In our model, each consumer is compensated directly with a monetary transfer for her individual data. While there exist concrete examples of such transactions (e.g., Nielsen offers monetary rewards to consumers for access to their browsing and purchasing data), most data intermediaries (e.g., Facebook and Google) compensate their users via the quality of the free services they offer (e.g., social networks, search, mail, video). Likewise, these intermediaries do not transfer the consumers' data to merchants for a fee, but they sell targeted advertising space. This enables the merchants to reap the value of information, by conditioning their messages and their prices on the consumers' preferences, without directly observing their data. All these transactions amount to indirect sales of information, as discussed in Bergemann and Bonatti (2019). An augmented model along these lines would add complexity to the interaction between the consumer and data intermediary, but would not affect the fundamental nature of the data externality, which is the focus of our paper.

The Value of Social Data Collecting data from multiple consumers helps any market participant to predict the fundamental component of each individual signal. This process can occur through two channels. First, in a market wherein the noise terms are largely idiosyncratic, a large sample size helps filter out errors and identify common fundamentals. Second, in a market with largely idiosyncratic fundamentals, many observations help filter out demand shocks and identify common noise terms, thereby estimating individual fundamentals by differencing.

The welfare consequences of such informational gains are complex. On the one hand, when a consumer shares her data with the intermediary, the other consumers benefit from learning her signal—better information allows them to tailor their demand to their true preferences. On the other hand, when the intermediary sells the additional data to the producer, selling the additional data enables more accurate price discrimination, which reduces all consumers' welfare.[1]

However, the choice by each consumer to share her information with the intermediary is guided only by her private benefits and costs, not by the data externality she generates with her actions. Thus, the intermediary must compensate each individual consumer only to the extent that the disclosed information affects her own welfare. Conversely, the platform does not have to compensate the individual consumer for any changes she causes in the welfare of others or any changes in her welfare caused by the information revealed by others.

Therefore, social data drive a wedge between the socially efficient and profitable uses of information. First, the cost of acquiring individual data can be substantially less than the value of the information to the platform. Second, although many uses of consumer information exhibit positive externalities, very little prevents the platform from trading data for profitable uses that are in fact harmful to consumers. We thus seek to identify under which conditions there might be too much or too little trade in data.

Recent empirical work on the effects of privacy regulation such as the European Union's General Data Protection Regulation (e.g., Aridor, Che, and Salz (2020) and Johnson, Shriver, and Goldberg (2020)), indicates that data externalities are relevant for consumers' and businesses' decisions to share their data. In the United States, legislators are also increasingly aware of the consequences of data externalities. In particular, the US House Committee on the Judiciary (2020) reports that "[...] the social data gathered through [a platform's] services may exceed their economic value to consumers."

[1] As we argue in Section 2, data externalities are not limited to settings where the consumers' information enables price discrimination. Instead, our insights apply to any product market where (a) data sharing teaches consumers about their preferences, and (b) the consumers' data is sold to a firm that seeks to extract their surplus.

Equilibrium Data-Sharing Policies We begin the equilibrium analysis by restricting the data intermediary to complete data sharing—collecting the consumers' signals and revealing them perfectly to all market participants. We first identify the factors that support complete data sharing as an equilibrium outcome and then emphasize the gap between profitable and socially efficient data sharing. In particular, when facing many consumers whose true preferences are strongly correlated, the intermediary can profitably trade the consumers' information: the producers' willingness to pay is substantial, and thanks to a strong (negative) data externality, the intermediary can acquire the consumers' data in exchange for minimal compensation. However, if the consumers' signals are also sufficiently precise, data sharing is detrimental to consumer welfare: consumers have very little to learn from others' signals, while the producer learns very precisely the willingness to pay of all consumers.[2]

We then ask whether the data market imposes any limitations at all on equilibrium information sharing. To do so, we remove the restriction of complete data sharing and allow the data intermediary to determine several dimensions of its information policy. We first consider the choice of whether to reveal the consumers' identities to the producer or to collect anonymous data. When consumers are homogeneous ex ante, we show that the intermediary prefers to collect anonymous data; collecting such data amounts to selling aggregate, market-level information to the producer. With this choice, the intermediary does not enable the producer to set personalized prices: the data are transmitted but disconnected from the users' personal profiles. In other words, the role of social data provides a more nuanced ability to determine the modality of information acquisition and use.[3]

The gap between the social value of the data and the price of the data widens when the number of consumers increases. In particular, under aggregate data intermediation, as the sources of data are multiplying, the contribution of each individual consumer to the aggregate information is shrinking. The presence of a data externality thus provides an explanation for the *digital privacy paradox* (e.g., Athey, Catalini, and Tucker (2017)), whereby small monetary incentives significantly affect the willingness of the subjects to relinquish their private data. In practice, this force also likely drives the extraordinary appetite of Internet platforms to gather information.[4]

[2]Conversely, there are data structures (e.g., ones with independent fundamentals and strongly correlated error terms) for which data sharing is beneficial to consumers but unprofitable for the data intermediary.

[3]The importance of social data is also manifest in the optimal information design. In particular, the intermediary might find it profitable to introduce correlated noise terms into the information elicited from each consumer. Noise reduces the value for the producer but exacerbates the data externality by rendering the consumers' reports more correlated. Thus, noise severely reduces the cost of procuring the data.

[4]The recent Furman reports identifies "the central importance of data as a driver of concentration and a barrier to competition in digital markets" (Digital Competition Expert Panel (2019))—a theme echoed in the reports by Cremèr, de Montjoye, and Schweitzer (2019) and by the Stigler Committee on Digital Platforms (2019). The social dimension of data we highlight also helps explain these forces.

We explore the limitations of our anonymization result by extending the model in several directions. In particular, we introduce consumer heterogeneity by considering multiple market segments, i.e., heterogeneous groups of consumers. Indeed, we find that data are aggregated at least to the level of the coarsest partition of homogeneous consumers, although further aggregation is profitable for the intermediary when the number of consumers is small. The resulting group pricing (which can be interpreted as discriminatory based on observable characteristics, such as location) has welfare consequences between those of complete privacy and those of price personalization.

We then consider a model in which the producer can choose prices and product characteristics to match an additional horizontal (taste) dimension of the consumers' preferences. The resulting data policy then aggregates the vertical dimension but not the horizontal dimension, thereby enabling the producer to offer personalized product recommendations but not personalized prices. More generally, with homogeneous consumers, we show that the intermediary collects anonymous data if and only if the transmission of these data reduces the total surplus. Therefore, even if the data *transmission* is socially detrimental, as in the case of price discrimination, the equilibrium level of data *aggregation* is socially efficient.

Related Literature This paper contributes to the growing data market literature recently surveyed in Bergemann and Bonatti (2019). In particular, the role of data externalities in the socially excessive diffusion of personal data has been a central concern in Choi, Jeon, and Kim (2019); and Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019).

Choi, Jeon, and Kim (2019) introduce information externalities into a model of monopoly pricing with unit demand. Each consumer is described by two *independent* random variables: her willingness to pay for the monopolist's service and her sensitivity to a loss of privacy. The purchase of the service by the consumer requires the transmission of personal data. From the collected data, the seller gains additional revenue, depending on the proportion of units sold and the volume of data collected. The total nuisance cost paid by each consumer depends on the total number of consumers sharing their personal data. Thus, the optimal pricing policy of the monopolist yields excessive loss of privacy, relative to the social welfare maximizing policy. In contrast, we consider the interaction between distinct data and product markets. Importantly, our data policy and data flow are determined explicitly as part of the equilibrium analysis, rather than as being represented by a reduced-form loss of privacy.

In contemporaneous work, Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019) also analyze data acquisition in the presence of information externalities. As in Choi, Jeon, and Kim (2019), they consider a model with many consumers and a single data-acquiring firm. Like the current analysis, Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019) propose

an explicit statistical model for their data; the model allows the authors to assess the loss of privacy for the consumer and the gains in prediction accuracy for the firm. Their analysis then pursues a different, and largely complementary, direction from ours. In particular, they analyze how consumers with heterogeneous privacy concerns trade information with a data platform. The authors derive conditions under which the equilibrium allocation of information is (in)efficient. In contrast, we endogenize privacy concerns to quantify the downstream welfare impact of data intermediation. We further investigate when and how privacy can be partially or fully preserved through aggregation, anonymization, and noise.

An early and influential paper on consumer privacy is Taylor (2004), who analyzes the sales of consumer purchase histories without data externalities.[5] More recently, Cummings, Ligett, Pai, and Roth (2016) investigate how privacy policies affect user and advertiser behavior in a simple model of targeted advertising. The low level of compensation that users command for their personal data is discussed in Arrieta-Ibarra, Goff, Jimenez-Hernandez, Lanier, and Weyl (2018), who propose sources of countervailing market power.

Fainmesser, Galeotti, and Momot (2020) provide a digital privacy model in which data collection improves the service provided to consumers. However, as the collected data can also leak to third parties and thus impose privacy costs, an optimal digital privacy policy must be established. Similarly, Jullien, Lefouili, and Riordan (2020) analyze the equilibrium privacy policy of websites that monetize information collected from users by charging third parties for targeted access. Gradwohl (2017) considers a network game in which the level of beneficial information sharing among the players is limited by the possibility of leakage and a decrease in informational interdependence. Ali, Lewis, and Vasserman (2019) study a model of personalized pricing with disclosure by an informed consumer, and they analyze how different disclosure policies affect consumer surplus. Ichihashi (2020b) studies both personalized pricing and product recommendations, and shows that a seller benefits from committing not to use the consumer's information to set prices. Our result on optimal anonymization and market-level pricing has similar implications, but is entirely driven by the data externality that appears when multiple consumers are present.

Finally, Liang and Madsen (2020) investigate how data policies can provide incentives in principal-agent relationships. They emphasize the structure of individual data and how the substitutes or complements nature of individual signals determines the impact of data on incentives. Ichihashi (2020a) considers a single data intermediary and asks how the complements or substitutes nature of the consumer signals affects the equilibrium price of the individual data.

[5] Acquisti, Taylor, and Wagman (2016) provide a recent literature survey of the economics of privacy.

2 Model

We consider an idealized trading environment with many consumers, a single intermediary in the data market, and a single producer in the product market.

2.1 Product Market

There are finitely many consumers, labeled $i = 1, ..., N$. In the product market, each consumer (she) chooses a quantity level q_i to maximize her net utility given a unit price p_i offered by the producer (he):

$$u_i(w_i, q_i, p_i) \triangleq w_i q_i - p_i q_i - \frac{1}{2} q_i^2. \tag{1}$$

Each consumer i has a baseline willingness to pay for the product $w_i \in \mathbb{R}$.

The producer sets the unit price p_i at which he offers his product to each consumer i. The producer has a linear production cost

$$c(q) \triangleq c \cdot q, \text{ for some } c \geq 0.$$

The producer's operating profits are given by

$$\pi(p_i, q_i) \triangleq \sum_i (p_i - c) q_i. \tag{2}$$

2.2 Data Environment

The consumers' willingness to pay is distributed according to

$$w \sim F_w.$$

Initially, each consumer may have only imperfect information about her willingness to pay. In particular, consumer i observes a signal

$$s_i \triangleq w_i + \sigma \cdot e_i,$$

where $\sigma > 0$ and e_i is consumer i's error term. The error terms e are independent of the willingness to pay w, and they are distributed according to

$$e \sim F_e.$$

We denote by S the information structure generated by the complete vector of consumer signals $s \in \mathbb{R}^N$.

Throughout the paper, we allow for arbitrary correlation structures across consumers, under a symmetry restriction namely that the distributions (F_w, F_e) are symmetric across individuals. Without loss of generality we assume that (i) each individual willingness to pay w_i has mean μ and variance 1; (ii) individual errors e_i have mean 0 and variance 1 (which is scaled by the parameter σ).

This data environment has two important features. First, any demand information beyond the common prior comes from the signals of the individual consumers. Second, with any amount of noise in the signals (i.e., if $\sigma > 0$), each consumer can learn more about her own demand from the signals of the other consumers.

The producer knows the structure of demand and thus the common prior distribution of consumers' willingness to pay. However, absent any additional information, the producer does not know the realized willingness to pay w_i of any consumer (or her signal s_i) prior to setting prices.

Additive Data Structure In some cases, we will specialize our model to a more tractable environment, which we refer to as the *additive data structure.* Specifically, we shall assume the willingness to pay of consumer i is the sum of two components:

$$w_i = \theta + \theta_i. \tag{3}$$

The term θ is *common* to all consumers in the market, while the term θ_i is *idiosyncratic* to consumer i. Similarly, the error term of consumer i is given by

$$e_i \triangleq \varepsilon + \varepsilon_i, \tag{4}$$

where the terms ε and ε_i refer to a common and an idiosyncratic error, respectively. We also refer to the willingness to pay w_i as the fundamental as opposed to the error term e_i.

As we vary the number of consumers N, the additive data structure described by (3) and (4) allows us to hold the pairwise correlation between any two consumers' fundamentals and noise terms constant. In particular, let α denote the correlation coefficient between any two (w_i, w_j), and let β denote the correlation coefficient between (e_i, e_j).

2.3 Data Market

The data market is run by a single data intermediary (it). As a monopolist market maker, the data intermediary decides how to collect the available information (s_i) from each consumer and how to share it with the other consumers and the producer. Thus, the data intermediary faces both an information design problem and a pricing problem.

We consider bilateral contracts between the individual consumers and the intermediary and between the producer and the intermediary. The data intermediary offers these bilateral contracts *ex ante*, that is, before the realization of any demand shocks. Each bilateral contract defines a *data policy* and a *data price*.

The data contract with consumer i specifies a *data inflow* policy X_i and a fee $m_i \in \mathbb{R}$ paid to the consumer. The data inflow policy describes how each signal s_i enters the database of the intermediary. We restrict attention to the following two policies: (i) the *complete (identity-revealing)* data policy $X = S$, where the intermediary collects each consumer's signal s_i; and (ii) the *anonymized* data policy $X = A$, where the intermediary collects individual signals without identifying information. We model the anonymized data policy as

$$A : S \to \delta(S), \tag{5}$$

for a random permutation of the consumers' indices $i \to \delta(i)$.

In our product market model, where the consumer's demand is linear in her signal, the anonymized data policy A is equivalent to an *aggregate* data policy that conveys information about the average willingness to pay. Intuitively, the anonymized data policy prevents the producer from matching signals to consumers, i.e., from profitably charging personalized prices. In Section 5.3, we enrich the intermediary's strategy space by allowing for data policies that collect partial information about the consumers' signals.

A data contract with the producer specifies a *data outflow* policy Y and a fee $m_0 \in \mathbb{R}$ paid by the producer. The data outflow policy determines how each consumer's collected signal is transmitted to the producer and to other consumers. In particular, letting X denote the intermediary's *realized* data inflow, a data outflow policy $Y = (Y_0, Y_1, \ldots Y_N)$ describes how the collected data are released to the seller,

$$Y_0 : X \to \Delta(\mathbb{R}^N),$$

and to each consumer,

$$Y_i : X \to \Delta(\mathbb{R}^N).$$

Sharing data with other consumers is a critical design element because doing so allows each

consumer to adjust her quantity demanded at any price. Therefore, the information received by consumers also impacts the *producer's* willingness to pay for the intermediary's data.

The data intermediary maximizes the net revenue

$$R \triangleq m_0 - \sum\nolimits_{i=1}^{N} m_i. \tag{6}$$

2.4 Equilibrium and Timing

The game proceeds sequentially. First, the terms of trade on the data market and then the terms of trade on the product market are established. The timing of the game is as follows:

1. The data intermediary offers a data inflow policy (m_i, X_i) to each consumer i. Consumers simultaneously accept or reject the intermediary's offer.

2. The data intermediary offers a data outflow policy (m_0, Y) to the producer. The producer accepts or rejects the offer.

3. Consumers observe their signals s, and the information flows (x, y) are transmitted according to the terms of the data policies.

4. The producer sets a unit price p_i for each consumer i who makes a purchase decision q_i, given her available information about w_i.

We analyze the Perfect Bayesian Equilibria of the game. Under the timing described above, the information is imperfect but symmetric at the contracting stage. Furthermore, when the consumer receives the intermediary's offer, she must anticipate the intermediary's subsequent choice of data outflow policy, which determines what data is shared with her, as well as with the producer. We denote by a_0 and a_i the participation decisions by the producer and by consumer i, respectively (where $a = 0$ indicates rejection and $a = 1$ indicates acceptance). A Perfect Bayesian Equilibrium is then a tuple of inflow and outflow data policies, data and product pricing policies, and participation decisions:

$$\left\{(X^*, Y^*, m^*)\,;\; p^*\left(X, Y\right);\; a^*\right\}, \tag{7}$$

where

$$a_0^* : X \times Y \times \mathbb{R} \to \{0, 1\}, \qquad a_i^* : X_i \times \mathbb{R} \to \{0, 1\}, \tag{8}$$

such that (i) the producer maximizes his expected profits, (ii) the intermediary maximizes its expected revenue, and (iii) each consumer maximizes her net utility. In our baseline analysis, we focus on the best equilibrium for the data intermediary; in the best equilibrium,

every consumer accepts the offer from the data intermediary. We then discuss a unique implementation in Section 4.4.

Figure 1 summarizes the information and value flow in the data and product markets.

Figure 1: Data and Value Flows

2.5 Discussion of Model Features

Monetary Payments In our model, both the intermediary and the consumers are compensated with a monetary transfer for the data they transfer. In practice, many data intermediaries offer consumers free services rather than money in exchange for personal data. An augmented model in which services are rendered in exchange for data adds complexity to the interaction between the consumer and data intermediary but does not affect the nature of the data externality, which is the focus of our analysis.

Participation Constraints The participation constraints of every consumer and of the producer are required to hold at the *ex ante* level. Thus, the consumers agree to the data policy before the realization of their signals. The choice of *ex ante* participation constraints captures the prevailing conditions under which users interact with large digital platforms. For instance, users of services on Amazon, Facebook, or Google typically establish an account and accept the "terms of service" before making any specific query or post. Through the lens of our model, the consumer requires a level of compensation that allows her to profitably share the information in expectation. Upon agreeing to participate, there are no further incentive compatibility constraints on the transmission of her information.

Lack of Commitment As we mentioned above, targeted advertising is the primary source of revenue for digital platforms. Consequently, the data intermediary in our sequential

game sells the consumers' data only to the producer, and cannot commit to withhold any information from him. Similarly, the intermediary's choice of data outflow policy occurs after consumers are enlisted but before their data are realized. This assumption captures the limited ability of a platform to write advertising contracts contingent on, say, the volume of activity taking place or the number of registered users. Under these assumptions, Theorem 1 shows that selling data *only* to the producer entails no loss of revenue for the intermediary. In Section 5.4, we discuss additional data-sharing policies (such as charging consumers for product recommendations) that may emerge in equilibrium if the intermediary can commit to withholding information privacy.

Linear Pricing The producer in our model uses the consumers' data to set (possibly discriminatory) prices. While this canonical example facilitates the interpretation of our anonymization result (Theorem 2), any model where the data buyer can use the consumers' information against them would yield similar intuitions. Our producer is further restricted to charging a (possibly personalized) unit price to each consumer. The gains from data sharing would arguably change if richer pricing instruments enabled the producer to extract more of the surplus generated through better information. In turn, this would affect the value of the social data and the price of individual data. However, the presence of the data externality would continue to drive a gap between equilibrium and socially efficient allocation.

3 Complete Data Sharing

The costs and benefits of data intermediation display similar features across all data-sharing policies. We therefore begin the analysis of the data market by considering the complete (identity-revealing) data-sharing policy. We first identify the welfare impact of (exogenously imposed) complete data sharing for all market participants. We then exhibit conditions under which complete data sharing generates positive profits for the intermediary in equilibrium. In Section 4, we allow the data intermediary to choose between complete and anonymized data inflows, and to restrict the data outflow to producers and consumers.

3.1 Data Sharing and Product Markets

Under the complete data-sharing policy, all the consumers and the producer observe the entire vector of signals s. In other words, every signal s_i enters the database without any modification: $X_i(s_i) = s_i$ for all i and s_i; and the outflow policy simply returns the inflow: $Y(s) = s$, for all s. Given this data policy, the optimal pricing policy for the producer consists of a vector of personalized prices $p^*(s) \in \mathbb{R}^N$, thus resulting in a vector of individual

quantities purchased $q_i^*(s)$. In particular, let

$$\widehat{w}_i(s) \triangleq \mathbb{E}[w_i \mid s] \tag{9}$$

denote the predicted value of consumer i's willingness to pay, given the signal profile s. The realized demand function of consumer i is

$$q_i(s,p) = \widehat{w}_i(s) - p. \tag{10}$$

Therefore, the producer charges consumer i the optimal personalized price

$$p_i^*(s) = \frac{\widehat{w}_i(s) + c}{2}, \tag{11}$$

which results into the equilibrium quantity

$$q_i^*(s) = \frac{\widehat{w}_i(s) - c}{2}. \tag{12}$$

We now quantify the value of information for consumers and producers. The shared data help each consumer estimate her own willingness to pay. For the producer, the shared data enable a more informed pricing policy. The net revenue of the producer from interacting with consumer i is given by

$$\Pi_i(S,S) \triangleq \mathbb{E}[\pi(p_i^*(s), q_i^*(s)) | S] = \frac{1}{4}\mathbb{E}\left[(\widehat{w}_i(s) - c)^2 | S\right].$$

The first argument in $\Pi_i(\cdot,\cdot)$ refers to consumer i's information structure and the second argument refers to the producer's information structure. With complete data sharing, the two structures coincide, and thus we write $\Pi_i(S,S)$. Similarly, we denote the gross expected utility of consumer i from complete data sharing by

$$U_i(S,S) \triangleq \mathbb{E}[u_i(w_i, q_i^*(s), p_i^*(s)) | S] = \frac{1}{8}\mathbb{E}\left[(\widehat{w}_i(s) - c)^2 | S\right].$$

The model with quadratic payoffs yields explicit expressions for the value of information for all product market participants. In particular, since prices and quantities are linear functions of the posterior mean $\widehat{w}_i$, the ex ante average prices and quantities $\mathbb{E}[p^*]$ and $\mathbb{E}[q^*]$ are constant across all information structures. Consequently, all surplus levels under complete data sharing depend only on the ex ante variance of the posterior mean $\widehat{w}_i(s)$.

We therefore define the *quantity of payoff-relevant information* under any information

structure S as the *gain* function:

$$G(S) \triangleq \text{var}\left[\widehat{w}_i(s)\right]. \tag{13}$$

Because we normalized the variance of the fundamental w_i to 1, the gain function $G(S)$ represents the fraction of the variance of w_i explained by the vector s. In particular, if the posterior expectation is linear in the signal realization, then G is just the R^2 of a regression of w_i on s.

We now turn to the consequences of complete data sharing relative to no information sharing. Without information, the producer charges a constant price $\bar{p}$ for all consumers based on the prior mean. In contrast, the consumer already has an initial signal s_i, according to which she can adjust her quantity. The producer's net revenue and the consumer's expected utility are given by

$$\begin{aligned}
\Pi_i(S_i, \varnothing) &\triangleq \mathbb{E}\left[\pi\left(\bar{p}, q_i^*(s_i)\right)\right], \\
U_i(S_i, \varnothing) &\triangleq \mathbb{E}\left[u_i\left(w_i, q_i^*(s_i), \bar{p}\right) | S_i\right].
\end{aligned}$$

The welfare of an individual consumer i then depends on the quantity of information carried by her own signal, as measured by

$$G(S_i) \triangleq \text{var}\left[\widehat{w}_i(s_i)\right]. \tag{14}$$

We can now express the value of complete data sharing for the consumers and the producer in terms of the respective information gains.

Proposition 1 (Value of Complete Data Sharing)

1. *The value of complete data sharing for the producer is*

$$\Pi_i(S, S) - \Pi_i(S_i, \varnothing) = \frac{1}{4} G(S). \tag{15}$$

2. *The value of complete data sharing for consumer i is*

$$U_i(S, S) - U_i(S_i, \varnothing) = \frac{1}{2}\left(G(S) - G(S_i)\right) - \frac{3}{8} G(S). \tag{16}$$

3. *The social value of complete data sharing is*

$$W_i(S, S) - W_i(S_i, \varnothing) = \frac{1}{2}\left(G(S) - G(S_i)\right) - \frac{1}{8} G(S). \tag{17}$$

The welfare consequences of complete data sharing operate through two channels. First, with more information about her own preferences, the demand of each consumer is more responsive to her willingness to pay; this responsiveness is beneficial for the consumers and (weakly) for the producer. Second, with access to the complete data, the producer pursues a *personalized* pricing policy for each individual consumer. As the producer adapts his pricing policy to the estimate of each consumer's willingness to pay, or $\widehat{w}_i$, some of the quantity responsiveness is dampened by the price responsiveness. While beneficial for the producer, this second channel *reduces* the consumer surplus and total welfare. In specific data environments, one of these two channels may dominate.

Corollary 1 (Inefficient Sharing)

1. *If consumers observe the fundamentals w_i perfectly ($\sigma = 0$), complete data sharing is socially inefficient.*

2. *If consumers' fundamentals (w_i, w_j) and errors (e_i, e_j) are independent, complete data sharing is socially inefficient.*

Under the conditions of Corollary 1, each consumer learns nothing about her own willingness to pay w_i from the other consumers' signals, but the producer does. The first term in (16) and (17) is then nil. As data sharing enables only price discrimination, data sharing affects producer surplus positively, but affects consumer and social surplus negatively.[6]

Corollary 2 (Efficient Sharing)
If individual signals s_i are sufficiently uninformative but the entire vector s remains sufficiently informative, complete data sharing improves consumer (and social) surplus.

This result is obtained by letting $G(S_i)$ go to zero in Proposition 1. Corollary 2 shows that, when information remains *symmetric*, better data yields Pareto improvements in the product market—intuitively, the producer and the consumer share the additional gains from trade associated with better informed consumption and pricing decisions. The following leading examples illustrate these results.

[6]The special case in which each consumer knows her willingness to pay (i.e., signals are noiseless in our model's language) is closely related to the model of third-degree price discrimination in Robinson (1933) and Schmalensee (1981). In our setting, data sharing enables the producer to offer personalized prices; thus, price discrimination occurs across different *realizations* of the willingness to pay. In contrast, in Robinson (1933) and Schmalensee (1981), price discrimination occurs across different market segments. In both settings, the central result is that average demand does not change (with all markets served), but social welfare is lower under finer market segmentation.

Example 1 (Common Preferences) *Fundamentals w_i are perfectly correlated and errors e_i are independent: $s_i = w + \sigma \cdot e_i$. This structure represents, for example, a new technology that consumers are imperfectly informed about. In this case, the average signal $\bar{s}$ identifies the common willingness to pay as N becomes large. Data sharing is then socially beneficial if and only if σ is also large, i.e., if consumers have much to learn from others' signals.*

Example 2 (Common Experience) *Errors e_i are perfectly correlated and fundamentals w_i are independent: $s_i = w_i + \sigma \cdot e$. An example with this structure is health data, where individuals have independent needs for a given therapy, but are all exposed to common factors, or share the same risk perception. Under this structure, the average signal $\bar{s}$ identifies the common error component e as N becomes large. All market participants can then precisely estimate each w_i from the difference $s_i - \bar{s}$. In this case too, data sharing is socially beneficial if and only if the consumers' initial signals are sufficiently noisy.*

These examples illustrate two ways in which data sharing can help one learn individual willingness to pay: by filtering out idiosyncratic error *or* fundamental terms. While information sharing enables learning in both examples, the actions of consumers $-i$ impact the surplus of consumer i quite differently in the two cases. To formalize this idea, we introduce our notion of data externality.

3.2 Data Externality

Our notion of *data externality* isolates the effect on consumer i's surplus of the decision of the other consumers to share their data with all market participants.

Definition 1 (Data Externality)
The data externality imposed by consumers $-i$ on consumer i is given by

$$DE_i(S) \triangleq U_i(S, S_{-i}) - U_i(S_i, \varnothing). \tag{18}$$

To quantify the data externality, we consider what would happen if consumer i held back her signal, given that the remaining $N-1$ consumers share theirs with the producer and with consumer i. In this case, the consumer's ex ante utility is given by

$$U_i(S, S_{-i}) \triangleq \mathbb{E}\left[u_i\left(w_i, q_i^*(s), p_i^*(s_{-i})\right) | S\right]. \tag{19}$$

The data externality can be characterized in terms of the gain function as follows.

Proposition 2 (Data Externality)
The data externality is given by

$$DE_i(S) = \frac{1}{2}\left(G(S) - G(S_i)\right) - \frac{3}{8}G(S_{-i}).$$

We compare the data externality to the overall impact of data sharing on consumer surplus in Propositions 1 and 2. Because $G(S_{-i}) \leq G(S)$, it is immediately clear that $\Delta U < DE$. Intuitively, the only difference between the overall effect of data sharing and the data externality is whether the consumer's own signal is revealed to the firm. Nevertheless, the two quantities retain significant similarities. In particular, if consumers do not learn from others' signals ($G(S) = G(S_i)$), then the only effect of sharing signals s_{-i} is to help the producer learn w_i; i.e., the data externality is negative.

1. "Common preferences" ($s_i = w + \sigma e_i$): as $N \to \infty$, the information content of N and $N-1$ signals converge; i.e., $G(S) \approx G(S_{-i})$. If, in addition, σ is small enough, then consumer i is worse off when others share their signals.

2. "Common experience" ($s_i = w_i + \sigma e$): because all w_i are independent, the producer cannot learn anything about w_i from signals s_{-i} only. However, consumer i can use signals s_{-i} to filter out the common error in her own signal s_i. Therefore, the other consumers' signals help consumer i—the data externality is unambiguously positive.

Thus, while the overall effect of data sharing on consumers depends largely on the informativeness of individual signals s_i, the impact of other consumers' sharing decisions varies significantly with the data structure, particularly the correlation structure of fundamentals and noise. We next explain how the distinction between welfare effects and data externalities influences the profitability of data sharing for the intermediary.

3.3 Equilibrium with Complete Data Sharing

We now investigate whether socially efficient data sharing can occur in a market with a data intermediary. Conversely, we ask whether the data intermediary can profitably induce complete data sharing even when it is socially harmful.

To compute the profitability of complete sharing for the intermediary, we compute the compensation owed to each consumer m_i and the total payment charged to the producer m_0. For the producer, the gains from data acquisition have to at least offset the price of the data. From Proposition 1, we can write the participation constraint for the producer as

$$m_0 \leq \Pi(S, S) - \Pi(S_i, \varnothing) = \frac{NG(S)}{4}. \tag{20}$$

In an equilibrium where all consumers accept the intermediary's contract, the payment m_i to each consumer depends on what would happen if consumer i held back her signal, given that the other $N-1$ consumers share theirs. The complete data-sharing policy prescribes that a nonparticipating consumer i nonetheless receives the signals of all other consumers. Thus, by rejecting her contract, consumer i does not transmit her data but does not reduce the amount of information available to her.[7]

Therefore, the data intermediary must set payments to consumers m_i that satisfy

$$m_i \geq U_i\left(S, S_{-i}\right) - U_i\left(S, S\right), \text{ for all } i. \tag{21}$$

When (21) binds, we can conveniently rewrite consumer i's compensation as

$$\begin{aligned} m_i^* &= U_i\left(S, S_{-i}\right) - U_i\left(S, S\right) \\ &= -\underbrace{\left(U_i\left(S, S\right) - U_i\left(S_i, \varnothing\right)\right)}_{\Delta U_i(S)} + \underbrace{U_i\left(S, S_{-i}\right) - U_i\left(S_i, \varnothing\right)}_{DE_i(S)}. \end{aligned} \tag{22}$$

The first term in the data compensation m_i is the total change in consumer i's surplus, denoted by $\Delta U_i\left(S\right)$, associated with complete data sharing. The second term is the data externality $DE_i\left(S\right)$ imposed on i by consumers $j \neq i$ when they sell their data to the intermediary, who then shares the data with the producer and with all consumers. In particular, if consumers impose negative data externalities on each other, this imposition directly reduces the compensation owed to each one.

With the characterization of ΔU_i and DE_i in Propositions 1 and 2, we can write the payment owed to each consumer in (22) as

$$m_i^* = \frac{3}{8}\left(G\left(S\right) - G\left(S_{-i}\right)\right) \geq 0. \tag{23}$$

Finally, with the producer's binding participation constraint (20), we can write the intermediary's profit (6) as the sum of two terms:

$$R\left(S\right) = \sum\nolimits_{i=1}^{N}\left(\Delta W_i\left(S\right) - DE_i\left(S\right)\right). \tag{24}$$

The intermediary's profits are equal to the effect of data sharing on social surplus, net of the data externalities across all consumers. This formulation clarifies how the intermediary's objective differs from the social planner's. If the data externality is negative, intermediation can be profitable but welfare reducing. Conversely, if the data externality is positive, welfare-

[7] In Section 4.1, we show that this feature of the data policy is indeed part of the equilibrium when the data inflow and outflow are chosen optimally by the intermediary.

enhancing intermediation might not be profitable. Combining the terms in (20) and (23), we obtain a necessary and sufficient condition for profitable intermediation.

Proposition 3 (Intermediation Profits)
Complete data intermediation is profitable if and only if $3G(S_{-i}) \geq G(S)$.

Intermediation is profitable if the amount of information gained by the producer about consumer i's willingness to pay on the basis of signals s_{-i} is sufficiently large. Intuitively, it is cheaper to acquire each signal s_i if the other consumers' signals are close substitutes. Proposition 3 shows that the signals s_{-i} must explain at least $1/3$ of the variance of w_i explained by the entire vector s for complete data sharing to be profitable.

We now establish a sufficient condition for the profitability of complete data sharing as the number of consumers becomes large. For this result, we adopt the *additive data structure* described in (3) and (4), where α denotes the correlation between any pair (w_i, w_j).

Proposition 4 (Profitable Intermediation of Complete Data)
There exist $\alpha^* \in (0,1)$ *and* $N^*(\alpha)$ *such that for each* $\alpha > \alpha^*$ *and* $N > N^*(\alpha)$*, complete data intermediation is profitable.*

Intuitively, the demand of each individual consumer comes from two sources: the idiosyncratic shock and the common shock. While each consumer has an informational monopoly over the idiosyncratic shock, the producer can learn about the common shock not only from consumer i but also from all of the other consumers. The more strongly correlated the underlying fundamentals w_i and w_{-i} are, the easier it is to learn from other consumers' signals. Proposition 4 shows that intermediation is profitable if w_i and w_{-i} are sufficiently correlated, regardless of the correlation structure in the noise. Conversely, for independent fundamentals, $G(S_{-i}) = 0$, and intermediation is not profitable by Proposition 3.

We now draw the implications for the intermediary's profits in our two leading examples, again under the additive data structure in (3) and (4).

Corollary 3 (Inefficient but Profitable)
Suppose fundamentals w_i *are perfectly correlated* $(\alpha = 1)$ *and errors* e_i *are independent* $(\beta = 0)$*. As* $N \to \infty$*, the information gained from* $N-1$ *signals approaches that of* N*, and intermediation is profitable. However, if* σ *is sufficiently small, the data externality is negative, and data sharing reduces consumer surplus.*

These results echo the findings of Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019), who considered signals with diminishing marginal informativeness, and found socially excessive data intermediation. The information structure in Corollary 3 satisfies this submodularity property. In our model, however, socially insufficient intermediation can also occur. In

particular, the intermediary may be unable to generate positive profits from socially efficient information with complementary signals, such as those in Corollary 4.

Corollary 4 (Efficient but Unprofitable)
Suppose fundamentals w_i are independent ($\alpha = 0$) and errors e_i are perfectly correlated ($\beta = 1$). For sufficiently large σ, consumers benefit from data sharing as $N \to \infty$. However, because fundamentals are independent, however, no intermediation is profitable.

4 Optimal Data Intermediation

In the preceding analysis, the data intermediary collected and distributed all of the available demand data to all of the product market participants. We now allow the intermediary to design an optimal data intermediation policy along two key dimensions. First, we allow the intermediary *not* to release all of the data that it has collected, i.e., to introduce incomplete, possibly asymmetric information in the product market by limiting the data outflow relative to the data inflow. Second, we allow the intermediary to choose between collecting complete, identity-revealing information and anonymized information about each individual consumer. We now analyze the intermediary's optimal data policy along all dimensions, beginning with the choice of data outflow for any realized data inflow.[8]

4.1 Data Outflow

In the extensive form game, given the realized data inflow, the intermediary offers a data outflow policy to the producer. The outflow policy specifies both the fee m_0 and the flow of information to all market participants, including the consumers. The data outflow policy thus determines both how well-informed the producer is and how well-informed his customers are. Thus, a critical driver of the consumer's decision to share data is her ability to anticipate the intermediary's use of the information thus gained. In this case, the consumer knows that the intermediary will choose the data outflow policy that maximizes the producer's profits, which the intermediary then extracts through the fee m_0.

We now show that the complete data outflow policy maximizes the producer's gross surplus and hence the intermediary's profits. Under this policy, all collected signals are reported to the producer and to all consumers, including those who did not accept the intermediary's offer.

[8]In Section 5.3, we allow the intermediary to introduce further (possibly correlated) noise terms in any (revealing or anonymized) signals that it collects.

Theorem 1 (Data Outflow Policy)
Given any realized data inflow X, the complete data outflow policy $Y^*(X) = X$ maximizes the gross revenue of the producer among all feasible outflow data policies.

The intuition for this result is twofold. First, as we showed in Proposition 1, the producer surplus is increasing in the amount of information that is symmetrically gained by all of the market participants: when a consumer's demand responds to the intermediary's information, a producer endowed with the same information can better tailor his price. Therefore, the intermediary should not withhold any information from the producer. Second, the producer does not benefit from holding superior information relative to the consumers. If he were better informed, the prices charged would convey information to the consumers about their own willingness to pay. The ensuing signaling incentives impose a cost on the producer, because he will need to deviate from the prices that maximize his profits, holding fixed the consumers' beliefs. The intermediary can then increase the producer's profits by revealing any information contained in the equilibrium prices directly to the consumers.

Therefore, in every subgame following the consumers' participation decisions, all of the consumers and the producer receive the same information from the intermediary. For any data inflow policy $X \in \{A, S\}$, the compensation owed to consumer i in equilibrium can then be written as in (22), and the intermediary's profits can be written as

$$R(X) = \sum\nolimits_{i=1}^{N} \left(\Delta W_i(X) - DE_i(X)\right). \tag{25}$$

The results in this section obtain under the extensive form of our model without commitment. As alluded to previously, we maintain the assumption that the intermediary cannot refrain from selling information to the producer and cannot sell any acquired data inflow back to the consumers. The latter assumption, however, entails no loss. Because consumers know that the intermediary must sell the data to the producer, consumers also expect to receive all available information regardless because receiving this information maximizes the producer's fee. In Section 5.4, we comment on how endowing the intermediary with varying degrees of commitment power modifies the result in Theorem 1.

4.2 Data Anonymization

We now explore the intermediary's decision to anonymize the individual consumers' demand data. We focus on two maximally different policies along this dimension. At one extreme, the intermediary can collect and transmit identifying information about individual consumers, thereby enabling the producer to charge personalized prices, as in the previous section. At

the other extreme, the intermediary can collect anonymized data only. Under anonymized information intermediation, the producer charges the same price to all consumers who participate in the intermediary's data policy. In other words, from the point of view of the producer, anonymized data is equivalent to aggregate demand data. These data still allow the producer to perform third-degree price discrimination across realizations of the total market demand but limit his ability to extract surplus from individual consumers.[9]

Certainly, for the producer, the value of market demand data is lower than the value of individual demand data. However, the cost of acquiring such fine-grained data from consumers is also correspondingly higher. Anonymizing the consumers' information profitably reduces the intermediary's data acquisition costs.

Theorem 2 (Optimality of Data Anonymization)
The intermediary obtains strictly greater profits by collecting anonymized consumer data.

Within the confines of our policies, but independent of the distributions of fundamental and noise terms, the data intermediary finds it advantageous to not elicit the identity of the consumer. Therefore, the producer will not offer personalized prices but variable prices that adjust to the realized information about market demand. In other words, a monopolist intermediary might cause socially inefficient information transmission, but the equilibrium contractual outcome preserves privacy over the personal identity of the consumer.[10]

This finding suggests why we might see personalized prices in fewer settings than initially anticipated. In the context of direct sales of information, for example, Nielsen does not sell individual households' data to merchants. Instead, Nielsen aggregates its panel data at the local market level. Similarly, in the context of indirect sales of information, merchants on the retail platform Amazon very rarely engage in personalized pricing. However, the price of every single good or service is subject to substantial variation across both geographic markets and over time. In light of the above result, we might interpret the restraint on the use of personalized pricing in the presence of aggregate demand volatility as the optimal resolution of the intermediary's trade-off in acquiring sensitive consumer information.

To gain intuition for why the intermediary chooses data anonymization when consumers are homogeneous, consider the two components of the intermediary's profits (25): the contribution to social surplus and the data externality, beginning with the latter. Suppose consumers $-i$ reveal their signals, and consumer i does not. With access to identity information, the producer optimally aggregates the available data to form the best predictor of

[9]More formally, under the anonymized data policy A, the producer has access to the vector $\delta(s)$, i.e., to a uniformly random permutation of the consumers' signals. Because the producer faces a prediction problem for each w_i with a quadratic loss, he chooses to charge a uniform price that is optimal on average.

[10]In Section 5, we explore the boundaries of the anonymization result, under both heterogeneous consumers and alternative product-market specifications.

the missing data point. In this case, the producer charges a personalized price $p_i^*(X_{-i})$ to each nonparticipating consumer i. With anonymous data, the producer charges two prices: a single price for all participating consumers and another price for the deviating, nonparticipating consumers. Because the distribution of consumer willingness to pay and signals is symmetric, however, the producer's inference on w_i is invariant to permutations of the other consumers' signals. Therefore, a consumer who does not participate faces the same price under both data inflow policies:[11]

$$p_i^*(S_{-i}) = p_i^*(A_{-i}).$$

Furthermore, by Theorem 1, each consumer i knows that she will receive the intermediary's data outflow regardless of her participation decision. Because of symmetry, consumer i's inference about her own willingness to pay w_i does not make use of consumers $-i$'s identities either. Therefore, the consumer's expected utility off the path of play satisfies

$$U_i(S, S_{-i}) = U_i(A, A_{-i}),$$

and hence $DE_i(S) = DE_i(A)$. Thus, the data externality term DE_i in the intermediary's profits (25) is not impacted by the anonymization decision.

Along the path of play, however, the two data inflow policies yield different outcomes. In particular, the anonymized data inflow policy reduces the amount of information conveyed to the producer in equilibrium. Crucially, this reduction does not occur at the expense of the consumers' own learning, and hence

$$U_i(A, A) = U_i(S, A) > U_i(S, S), \text{ and } W_i(A, A) = W_i(S, A) > W_i(S, S).$$

At this point, it is clear that the intermediary benefits by anonymizing the consumers' individual information. Anonymization leaves the DE_i terms in (25) unchanged but reduces the information transmitted to the producer while holding fixed the consumer's information. Relative to identity-revealing data sharing, this shift increases the total surplus terms ΔW_i in (25) and hence also the intermediary's profits.

[11] The result in Proposition 2 holds even if we force the producer to charge a single price to all consumers on and off the equilibrium. With this interpretation, we intend to capture the idea that the producer offers one price "on the platform" to the participating consumers while interacting with the deviating consumer "offline." The producer then uses the available market data to tailor the offline price.

4.3 Equilibrium with Optimal Data Sharing

The anonymized data-sharing policy allows the data intermediary to operate in environments where complete data sharing is not profitable. To formalize this result, we write the optimal fees under anonymized data sharing as

$$\begin{aligned} m_0^* &= NG\left(A\right)/4, \\ m_i^* &= \frac{3}{8}(G(A) - G\left(A_{-i}\right)). \end{aligned}$$

The condition for the profitability of anonymized data intermediation is then given by

$$3G\left(A_{-i}\right) \geq G(A), \tag{26}$$

which is more permissive than the one in Proposition 3. Recall that the intermediary's revenues depend on the producer's ability to estimate the consumers' willingness to pay from subsets of the data. In particular, the producer's inference problem about w_i based on signals s_{-i} is identical under complete and anonymized learning, so $G\left(A_{-i}\right) = G\left(S_{-i}\right)$, while $G\left(A\right) < G\left(S\right)$. Condition (26) is satisfied whenever the returns from larger sample sizes are decreasing, i.e., when $N-1$ signals explain more than $1/3$ of the variation in $\bar{w}$ explained by N signals.[12]

Under the additive data structure described in (3) and (4), where α denotes the correlation coefficient between w_i and w_j, we obtain a strengthening of Proposition 4.

Proposition 5 (Profitable Intermediation of Anonymized Data)
For any $\alpha > 0$, there exists N^* such that anonymized data sharing is profitable if $N > N^*$.

We already know that a high degree of correlation in the consumers' willingness to pay allows the intermediary to profit from complete data sharing with a sufficiently large number of consumers. Under the optimal data-sharing policy, *any* degree of correlation in the consumers' willingness to pay makes the anonymized signals sufficiently close substitutes that intermediation is profitable when N is large.

Thus far, we have considered the optimal data policy for a given finite number of consumers, each of whom transmits a single signal. Perhaps, *the* defining feature of data markets is the multitude of (potential) participants, data sources, and services. We now pursue the implications of having many participants (i.e., of many data sources) for the social efficiency of data markets and the price of data.

[12] In Section 5.3, we show how the intermediary can further relax this condition by adding correlated noise terms to the data inflow.

Each additional consumer presents an additional opportunity for trade in the product market. Thus, the feasible social surplus is linear in the number of consumers. In addition, with every additional consumer, the intermediary obtains additional information about the market demand. These two effects suggest that intermediation becomes increasingly profitable in larger markets, wherein the potential revenue increases without bound, while individual consumers make a small marginal contribution to the precision of aggregate data.

In Theorem 3, we restrict attention to the additive data structure in (3) and (4), and we assume that error terms are independent. This allows us to use the sample average in order to establish a lower bound on learning from $N-1$ signals. We suspect that similar results hold more generally under correlated errors (as is the case with Gaussian distributions).

Theorem 3 (Large Markets)
Consider the additive data structure and assume that errors are independent across consumers. As $N \to \infty$,

1. *Each consumer's compensation m_i^* converges to zero.*

2. *Total consumer compensation is bounded by a constant:*

$$Nm_i^* \leq \frac{9}{8}\left(\text{var}\left[\theta_i\right] + \text{var}\left[\varepsilon_i\right]\right), \quad \forall N.$$

3. *The intermediary's revenue and profit grow linearly in N.*

As the optimal data policy aggregates the consumers' signals, each additional consumer has a rapidly decreasing marginal value. Furthermore, each consumer is paid only for her marginal contribution; this explains why the total payments Nm_i converge to a finite number. Strikingly, this convergence can occur from above*:* when the consumers' willingness to pay is sufficiently correlated, the decrease in each i's marginal contribution can be sufficiently strong to offset the increase in N. Figure 2 illustrates such an instance (with normally distributed fundamentals and errors), in which it can be less expensive for the intermediary to acquire a larger dataset than a small one.

While total costs converge to a constant, the revenue that the data intermediary can extract from the producer is linear in the number of consumers. Our model therefore implies that, as the market size grows without bound, the per capita profit of the data intermediary converges to the per capita profit when the (anonymized) data are freely available. Conversely, the impact on consumer surplus depends on the degree of correlation in the underlying fundamentals, and on the precision of the consumers' initial signals.[13]

[13]In a recent contribution, Loertscher and Marx (2020) study large digital monopoly markets, where data has the countervailing effects of improving consumer valuations and increasing monopoly prices.

Figure 2: Total Consumer Compensation ($\sigma_w = 1, \sigma_e = 0$)

Finally, we show that data anonymization is crucial for the large N properties of the intermediary's profits. Recall that, with complete data intermediation, individual consumer payments are proportional to $G(S) - G(S_{-i})$. As long as fundamentals w_i are not perfectly correlated, payments are then bounded away from zero for any finite N. Proposition 6 shows this property also holds in the limit.

Proposition 6 (Asymptotics with Complete Sharing)
Consider the additive data structure with $\text{var}[\theta_i] > 0$. *Under complete (identity-revealing) data sharing, the asymptotic individual compensation is bounded away from* 0*:*

$$\liminf_{N\to\infty} m_i^* \geq \frac{3}{8}\frac{\text{var}^2[\theta_i]}{1+\text{var}[e_i]} > 0.$$

An immediate consequence of Proposition 6 is that, with complete data sharing, total payments to consumers grow linearly in N. Thus, anonymization is critical to achieving increasing returns to scale in data intermediation: even if complete data intermediation were profitable, the per capita profits would be bounded away from the above benchmark.

4.4 Unique Implementation

Our analysis so far has characterized the intermediary's most preferred equilibrium. An ensuing question is whether the qualitative insights and the asymptotic properties discussed above would hold across all equilibria, particularly in the intermediary's least preferred equilibrium. A seminal result in the literature on contracting with externalities (see Segal (1999))

is the "divide-and-conquer" scheme that guarantees a unique equilibrium outcome (see Segal and Whinston (2000) and Miklos-Thal and Shaffer (2016)). Under this scheme, the intermediary can sequentially approach consumers and offer compensation conditional on all earlier consumers having accepted an offer. In this scheme, the first consumer receives compensation equal to her entire surplus loss, thereby guaranteeing her acceptance regardless of the other consumers' decisions. More generally, consumer i receives the optimal compensation level in the baseline equilibrium when $N = i$.

The cost of acquiring the consumers' data is strictly higher under "divide and conquer" than in the intermediary's most preferred equilibrium. Nonetheless, the impact of the ensuring unique implementation on per-capita profits vanishes in the limit.

Proposition 7 ("Divide and Conquer")
Consider the additive data structure with independent errors. Under the "Divide and Conquer" scheme, total consumer compensation satisfies

$$Nm_i^* \leq \frac{3}{4}(1 + \log N)(\text{var}[\theta_i] + [\varepsilon_i]).$$

Under divide and conquer, the total payments to the consumers do not converge to a finite constant as N grows without bound. However, the growth rate of these payments is far smaller than the rate at which the producer's willingness to pay for data diverges. Therefore, regardless of the equilibrium-selection criterion, the intermediary's per capita profits converge to the benchmark level of when anonymized consumer data are freely available.

5 Implications for Consumer Privacy

In our baseline setting, data anonymization is optimal independent of the model parameters, such as the number of consumers or the distribution of fundamentals and error terms. In this section, we enrich our model to characterize the implications of the optimal data intermediation policy for consumer privacy. In particular, we allow for (i) richer specifications of consumer heterogeneity, (ii) larger strategy spaces for the producer, and (iii) more sophisticated information design by the data intermediary.

5.1 Market Segmentation and Data

The assumption of ex ante homogeneity among consumers has enabled us to produce some of the central implications of social data. A more complete description of consumer demand should introduce heterogeneity across groups of consumers along characteristics such

as location, demographics, income, and wealth.

We now explore how these additional characteristics influence information policy and the profits of the data intermediary. To this end, we augment the description of consumer demand by splitting the population into J homogeneous groups:

$$w_{ij} \sim F_{w,j},\ e_{i,j} \sim F_{e,j},\ i = 1, 2, ..., N_j,\ j = 1, ..., J. \tag{27}$$

The intermediary's data inflow policy must now specify whether to anonymize the consumers' signals across groups and within each group. However, arguments similar to those used in Theorem 2 establish that it is always more profitable to anonymize all signals within each group, rather than revealing the consumers' identities.

Proposition 8 (No Discrimination within Groups)
The data policy that anonymizes all signals within each group $j = 1, ..., J$ and only reveals each consumer i's group identity is more profitable than the complete data-sharing policy.

By further specifying the model, we can identify conditions under which the data intermediary will collect and transmit group characteristics. By collecting information about the group characteristics, the intermediary influences the extent of price discrimination. For example, the intermediary could anonymize all signals across groups, thus forcing the producer to offer only a single price. Alternatively, the intermediary could allow the producer to discriminate between two groups of consumers by recording and transmitting the group identities. As intuition would suggest, enabling price discrimination across groups not only allows the intermediary to charge a higher fee to the producer but also increases the compensation owed to consumers.

Proposition 9 below sheds light on the optimal resolution of this trade-off. In this result, we restrict attention to the case of symmetric groups ($N_j = N$ for all j), with the additive data structure $w_i = \theta + \theta_i$, and independent noise terms in the consumers' signals.

Proposition 9 (Segmentation)
If N is large enough, inducing group-level pricing is more profitable for the intermediary than inducing uniform pricing.

While Theorem 2 stated that the intermediary will not reveal any information about consumer identity, Proposition 9 refines that result: if the market is sufficiently large, then the intermediary will convey limited identity information, i.e., each consumer's group identity. This policy allows the producer to price discriminate across, but not within, groups.

Conversely, if the producer faces few consumers and their willingness to pay are not highly correlated, then pooling all signals reduces the cost of sourcing the data.

The limited amount of price discrimination, which operates optimally at the group level rather than the individual level, can explain the behavior of many platforms. For example, Uber and Amazon claim that they do not discriminate at the individual level, but they condition prices on location, time, and other dimensions that capture group characteristics.

The result in Proposition 9 is perhaps the sharpest manifestation of the value of big data. By enabling the producer to adopt a richer pricing model, a larger database allows the intermediary to extract more surplus. Our result also clarifies the appetite of the platforms for large datasets: since having more consumers allows the platform to profitably segment the market more precisely, the value of the marginal consumer $i = N$ to the intermediary remains large even as N grows. In other words, allowing the producer to segment the market is akin to paying a fixed cost (i.e., higher compensation to the current consumers) to access a better technology (i.e., one that scales more easily with N). Figure 3 illustrates this result for an example with normally distributed fundamentals and errors.

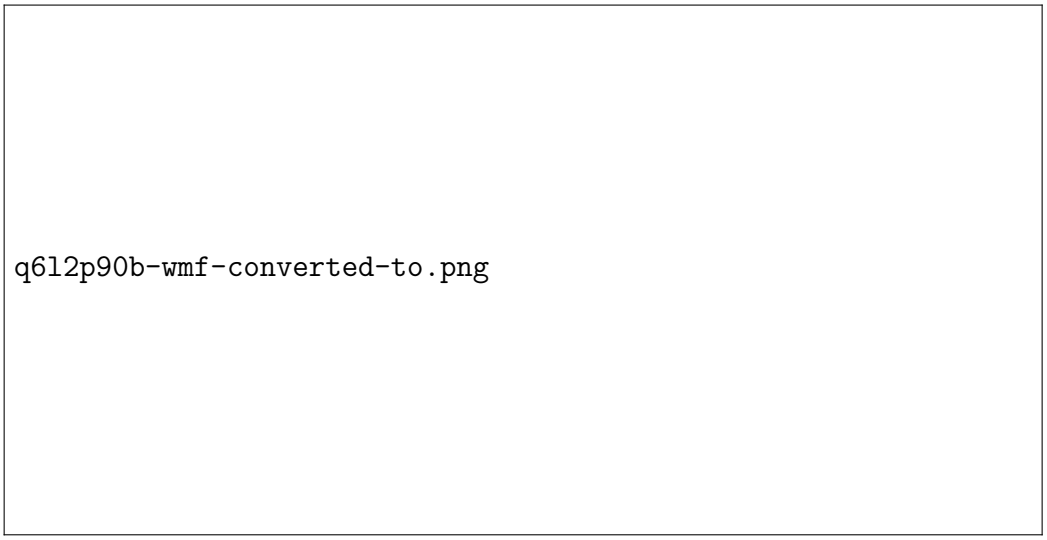

Figure 3: Marginal Value of an Additional Consumer

The optimality of using a richer pricing model when larger datasets are available is reminiscent of model selection criteria under overfitting concerns, e.g., the Akaike information criterion. In our setting, however, the optimality of inducing segmentation is not driven by econometric considerations. Instead, it is entirely driven by the intermediary's cost-benefit analysis in acquiring more precise information from consumers. As the data externality grows sufficiently strong, acquiring the data becomes cheaper as the intermediary exploits

the richer structure of consumer demand.[14]

5.2 Recommender System

In our baseline model, the data shared by the intermediary are used by the producer to set prices and by consumers to learn about their own preferences. The first assumption is, in a sense, the worst-case scenario for the intermediary: consider the case in which consumers' initial signals are very precise. As price discrimination reduces total surplus, no intermediation would be profitable without a strong, negative data externality. Consequently, data aggregation is an essential part of the optimal data intermediation policy in this case. In practice, however, consumer data can also be used by the producer in surplus-enhancing ways, for example, to facilitate targeting quality levels and other product characteristics to the consumer's tastes.

In this section, we develop a generalization of our framework; this generalization allows the producer to charge a unit price p_i and to offer a product of characteristic k_i to each consumer. Consumers differ both in their vertical willingness to pay and in their horizontal taste for the product's characteristics. Consumer i's utility function is given by

$$u_i\left(w_i, q_i, p_i, k_i, \ell_i\right) = \left(w_i - \left(k_i - \ell_i\right)^2 - p_i\right)q_i - q_i^2/2,$$

with w_i denoting the consumer's willingness to pay and ℓ_i denoting the consumer's ideal location or product characteristic. Both the willingness to pay $w \in \mathbb{R}^N$ and the locations $\ell \in \mathbb{R}^N$ of different consumers are potentially correlated. The producer has a constant marginal cost of quantity provision that we normalize to zero and can freely set the product's characteristic. Therefore, the case of a common location $\ell_i \equiv \ell$ for all consumers yields the baseline model of price discrimination.

We examine the data intermediary's optimal data inflow policy, which allows for separate aggregation policies for willingness to pay and location information. We impose the following assumptions: (i) the gains from trade under no information sharing are sufficiently large; (ii) the consumers' fundamentals have a joint Gaussian distribution; and (iii) consumer i perfectly observes (w_i, ℓ_i). The extension to noisy Gaussian signals is immediate. We then obtain another generalization of Theorem 2.

Proposition 10 (Optimal Aggregation by a Recommender System)
The intermediary's optimal policy collects aggregate data on the vertical component w_i and individual data on the horizontal component ℓ_i.

[14] Olea, Ortoleva, Pai, and Prat (2019) offered a demand-side explanation of a similar phenomenon: they showed that data buyers who employ a richer pricing model are willing to pay more for larger datasets.

Therefore, the recommender system enables the producer to offer targeted product characteristics that match k_i to ℓ_i as closely as possible. However, the system does not allow for personalized pricing. The logic is once again given by the intermediary's sources of profits, i.e., the contribution to social welfare ΔW and the data externality DE. Since the data externalities do not depend on the level of aggregation, the intermediary chooses to aggregate the vertical dimension of consumer data, thereby reducing the total surplus if transmitted to the producer. Conversely, because the distance $(k_i - \ell_i)^2$ shifts the consumer's demand function down, the intermediary allows for the personalization of product characteristics.

5.3 Information Design

We showed in Theorem 1 that the data intermediary releases all collected information to the producer. We now explore the data intermediary's ability to offer privacy guarantees in equilibrium by collecting less than perfect information about the consumers' signals.

Specifically, we consider the additive data structure in (3)-(4) and again assume that fundamental and error terms have a joint Gaussian distribution. We then specify a class of data policies that add common and idiosyncratic noise terms ξ and ξ_i to the consumers' original (noisy) signals s_i. We then have the following data inflow,

$$x_i = \underbrace{w_i + \sigma e_i}_{=s_i} + \xi + \xi_i,$$

and the intermediary chooses the variance of the additional noise terms (σ_ξ^2 and $\sigma_{\xi_i}^2$).

Theorem 4 (Optimal Data Intermediation)

1. *The optimal data inflow policy adds no idiosyncratic noise; i.e.,* $\sigma_{\xi_i}^* = 0$.
2. *The optimal data inflow policy adds (weakly) positive aggregate noise; i.e.,* $\sigma_\xi^* \geq 0$.

The optimal level of common noise σ_ξ^* is strictly positive when the number of consumers N or the correlation in their willingness to pay α is sufficiently small: if the consumers' preferences are sufficiently correlated, or if the market is sufficiently large, the intermediary does not add any noise. If the consumers' fundamentals are not sufficiently correlated, the intermediary makes their *signals* more correlated with additional common noise σ_ξ^*.

As we establish in Proposition 11, no profitable intermediation is feasible for values of α less than a threshold that decreases with N. This threshold is independent of the correlation coefficient β between the initial noise terms (e_i, e_j). Furthermore, as α approaches this threshold from above, the optimal level of common noise grows without bound. Figure 4 shows the optimal variance in the additional common noise term.

Figure 4: Optimal Additional Noise ($\sigma = 1, N = 2$)

Proposition 11 (Profitability of Data Intermediation)
Under the optimal data policy, the intermediary's profits are strictly positive if and only if

$$\alpha > \frac{N(\sqrt{3}+1)-1}{2N\left(N+1\right)-1} \in (0,1)\,. \tag{28}$$

The additional noise reduces the amount of information procured from consumers and hence the total compensation owed to them. These cost savings come at the expense of lower revenues. In this respect, aggregation and noise serve a common purpose. However, because the intermediary optimally anonymizes the consumers' signals, the correlation in the supplemental noise terms ξ renders signal s_i less valuable on the margin for estimating the average willingness to pay $\bar{w}$. In other words, the aggregate demand information without consumer i's signal s_i is a relatively better predictor of $\bar{w}$ when the intermediary uses common rather than idiosyncratic noise. Therefore, by using common noise exclusively, the intermediary can hold the information sold to the producer constant while reducing the cost of acquiring that information from consumers.

It would be misleading, however, to suggest that common noise is unambiguously more profitable for the intermediary. Indeed, these two elements of the information design— aggregation and noise— richly interact with one another. In particular, the value of common noise is deeply linked to that of aggregate data: if the intermediary is restricted to complete (identity-revealing) data sharing, one can show that supplemental idiosyncratic noise is optimal when the consumers' initial signals s_i are sufficiently precise.

5.4 Intermediary with Commitment

In our analysis, the data intermediary maintains complete control over the use of the acquired data. Given the data inflow, the data intermediary chooses the sequentially optimal data policy to be offered to the producer. This assumption reflects the substantial control that large online platforms have over the use of the data and the opacity with which the data outflow is linked to the data inflow. In other words, it is difficult to ascertain how any given data input informs an intermediary's data output. Nonetheless, it is useful to consider the implications of the data intermediary's ability to commit to a certain data policy, especially in light of the welfare properties of data sharing discussed above. To that end, suppose the data intermediary could offer the consumers contracts that specify a data inflow *and* a data outflow policy, together with a monetary transfer.

Through these richer contracts, the data intermediary can offer the consumers privacy guarantees. In particular, the intermediary can implement the socially efficient data-sharing policy, which consists of sharing all signals among all the consumers who accept the contract and not sharing any data at all with the producer. In exchange for this commitment, the data intermediary requests compensation from the consumers. In turn, the consumers are willing to pay a positive price for these data, and hence the socially optimal data sharing is always profitable.[15]

However, there are at least two reasons why the equilibrium outcome under these stronger commitment assumptions fails to capture the role of large online platforms. First, while there are examples in which consumers pay a positive price to access tailored, non-sponsored recommendations, data intermediaries choose to monetize the producers' side of their platform much more frequently. Second, the socially efficient policy is always profitable but it need not maximize the intermediary's profits. Perhaps the best example of the latter result is the case of perfectly correlated fundamentals and arbitrarily precise signals.

It is beyond the scope of this paper to characterize the optimal commitment policy for any initial data structure, but the data externality clearly remains a key driver of the equilibrium allocation of information even under stronger commitment assumptions.

[15]This environment with commitment is related to the analysis in Lizzeri (1999) but has a number of distinct features. First, in Lizzeri (1999), the private information is held by a single agent, and multiple downstream firms compete for the information and for the object offered by the agent. Second, the privately informed agent enters the contract after she has observed her private information. The shared insight is that the intermediary *with* commitment power might be able to extract a rent without any influence on the efficiency of the allocation.

6 Conclusion

We have explored the trading of information between data intermediaries with market power and multiple consumers with correlated preferences. The data externality that we have uncovered strongly suggests that levels of compensation close to zero can induce an individual consumer in a large market to relinquish precise information about her preferences. This finding holds even if the consumer's data are later sold to a firm that seeks to extract their surplus. Thus, giving consumers control rights over their data (a pillar of privacy regulation such as the EU General Data Protection Regulation or the California Privacy Rights Act) is insufficient to bring about the efficient use of information.

Of course, there are dimensions along which precise information generates surplus: for instance, ratings provide information to consumers about producers, and back-end tools render it possible to limit duplication and waste in advertising messages. There are, however, also other welfare-reducing effects, such as spillovers of consumer data to B2B markets. For example, if Amazon, Facebook, and Google use the information revealed by consumers to extract more surplus from advertisers, then consumers will pay a higher retail price, depending on the pass-through rate of the cost of advertising. In other words, the data externality that we identify is pervasive, often extending from individual consumers to other economic agents whose decisions are informed by consumer data.[16]

More constructively, our results regarding the aggregation of consumer information further suggest that privacy regulations must move away from concerns over personalized prices at the individual level. Most often, firms do not set prices in response to individual-level characteristics. Instead, segmentation of consumers occurs at the group level (e.g., as in the case of Uber) or at the temporal and spatial levels (e.g., as in the case of Staples and Amazon). Thus, our analysis points to the significant welfare effects of group-level and market-level, dynamic prices that react in real time to changes in demand.

A possible mitigator of the consequences of data externalities—echoed in Posner and Weyl (2018)—consists of facilitating the formation of consumer groups or unions to internalize the data externality when bargaining with powerful intermediaries, such as large online platforms. A different regulatory solution is based on *privacy managers*, such as internet browsers with heterogeneous privacy settings that compete for the consumers' default choice. Yet another solution—suggested by Romer (2019)—consists of making the data outflow costly for the intermediary by, for example, taxing targeted advertising. In our model, taxing the data outflow will limit efficient and inefficient intermediation alike but will affect the intermediary's choice of data policy under the assumptions of Section 5.4.

[16]This result echoes the claim in Zuboff (2019) that "privacy is a public issue."

Finally, our data intermediary collected and redistributed the consumer data but played no role in the interaction between the consumers and the producer. In contrast, a consumer can often access a given producer only through a data platform.[17] Many platforms can then be thought of as auctioning access to the consumer. The data platform provides the bidding producers with additional information that they can use to tailor their interactions with consumers. Social data platforms thus trade individual consumer information for services rather than money. The data externality then expresses itself in the quality of the services offered and in the extent of the consumers' engagement.

[17]Product data platforms, such as Amazon, Uber and Lyft, acquire individual data from the consumer through the consumers' purchase of services and products. Social data platforms, such as Google and Facebook, offer data services to individual users and sell the information to third parties, who mostly purchase the information in the form of targeted advertising space. In terms of our model, a product data platform combines the roles of data intermediation and product pricing.

7 Appendix

The Appendix collects the proofs of all the results in the paper.

Proof of Proposition 1. To prove Proposition 1, we prove a slightly more general formula for future reference. We calculate the formula when consumer i's information (σ-algebra) is Y_i and the firm's information is Y. With complete data sharing, we have $Y_i = Y = S$, while without any information sharing, we have $Y_i = S_i$ and $Y = \varnothing$. In what follows, we assume $Y_i \supset Y$ which means consumer i always knows what the firm knows. This implies prices do not signal any information. Finally, we normalize the producer's cost c to zero for notational convenience.

For any offered price p_i, consumer i demands the following quantity:

$$q_i = \mathbb{E}[w_i|Y_i] - p_i.$$

The producer finds it optimal to set the following price

$$p_i = \frac{\mathbb{E}[w_i|Y]}{2}.$$

The profit of the producer is given by

$$\begin{aligned}\Pi_i(Y_i, Y) &= \mathbb{E}\left[\frac{\mathbb{E}[w_i|Y]}{2}\left(\mathbb{E}[w_i|Y_i] - \frac{\mathbb{E}[w_i|Y]}{2}\right)\right] \\ &= \frac{\mathbb{E}\left[(\mathbb{E}[w_i|Y])^2\right]}{4} = \frac{\text{var}[\mathbb{E}[w_i|Y]] + \mathbb{E}[w_i]^2}{4},\end{aligned} \tag{29}$$

where the outside expectation represents integration over the whole probability space. The impact of complete data sharing ($Y = Y_i = S$) on producer surplus is then given by

$$\Pi_i(S, S) - \Pi_i(S_i, \varnothing) = \frac{\text{var}[\mathbb{E}[w_i|S]]}{4} =: \frac{1}{4}G(S).$$

The expected consumer surplus is given by

$$\begin{aligned}U_i(Y_i, Y) &= \mathbb{E}\left[\left(w_i - \frac{\mathbb{E}[w_i|Y]}{2}\right)\left(\mathbb{E}[w_i|Y_i] - \frac{\mathbb{E}[w_i|Y]}{2}\right) - \frac{1}{2}\left(\mathbb{E}[w_i|Y_i] - \frac{\mathbb{E}[w_i|Y]}{2}\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|Y_i])^2 - \frac{3}{4}(\mathbb{E}[w_i|Y])^2].\end{aligned} \tag{30}$$

Therefore the impact of complete data sharing on consumer surplus is given by

$$\begin{aligned} U_i(S,S) - U_i(S_i,\varnothing) &= \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|S])^2 - \frac{3}{4}(\mathbb{E}[w_i|S])^2] - \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|S_i])^2 - \frac{3}{4}(\mathbb{E}[w_i])^2] \\ &= \frac{1}{8}\operatorname{var}[\mathbb{E}[w_i|S]] - \frac{1}{2}\operatorname{var}[\mathbb{E}[w_i|S_i]] =: \frac{1}{2}\left(G\left(S\right) - G\left(S_i\right)\right) - \frac{3}{8}G\left(S\right) \end{aligned}$$

Finally, the impact on total surplus is given by the sum of the two effects. ■

Proof of Corollary 1 and 2. When consumers observe their willingness to pay perfectly, we have $G(S) = G(S_i)$. Therefore, by Proposition 1, we obtain

$$W_i(S,S) - W_i(S_i,\varnothing) = -\frac{1}{8}G(S).$$

When any two consumers' fundamentals and signals are independent, we again have $G(S) = G(S_i)$ and the same results applies.

Finally, in Corollary 2, letting $G(S_i) \to 0$ while $G(S)$ is bounded away from 0, we obtain

$$W_i(S,S) - W_i(S_i,\varnothing) = \frac{3}{8}G\left(S\right),$$

which ends the proof. ■

Proof of Proposition 2. Formula (30) in the proof of Proposition 1 implies

$$\begin{aligned} DE_i(S) &= U_i(S,S_{-i}) - U_i(S_i,\varnothing) \\ &= \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|S])^2 - \frac{3}{4}(\mathbb{E}[w_i|S_{-i}])^2] - \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|S_i])^2 - \frac{3}{4}(\mathbb{E}[w_i])^2] \\ &= \frac{1}{2}\operatorname{var}[\mathbb{E}[w_i|S]] - \frac{1}{2}\operatorname{var}[\mathbb{E}[w_i|S_i]] - \frac{3}{8}\operatorname{var}[\mathbb{E}[w_i|S_{-i}]], \end{aligned}$$

which is the expression in the statement. ■

Proof of Proposition 3. Combining the results in Propositions 1 and 2, we can write the data intermediary's revenue per capita as

$$\begin{aligned} \frac{R(S)}{N} &= \frac{1}{N}\sum_i^N \left(\Delta W_i(S) - DE_i(S)\right), \\ &= \frac{1}{2}(G(S) - G(S_i) - \frac{1}{8}G(S)) - \frac{1}{2}\left(G\left(S\right) - G\left(S_i\right)\right) + \frac{3}{8}G\left(S_{-i}\right), \\ &= \frac{3}{8}G\left(S_{-i}\right) - \frac{1}{8}G(S), \end{aligned}$$

which is the condition given in the statement. ■

In the remainder of the Appendix, we often make use of the following classical result in statistics, which we state as a lemma without proof—the result is an immediate consequence of the fact that $\mathbb{E}[X|Y]$ is the projection of X on $\mathcal{F}(Y)$ in L^2 space.

Lemma 1 *Let W and Y be two random variables. Then it holds that*

$$\operatorname{var}[\mathbb{E}[W|Y]] = \operatorname{var}[W] - \mathbb{E}[(W - \mathbb{E}[W|Y])^2] \leq \operatorname{var}[W],$$

and

$$\mathbb{E}[(W - \mathbb{E}[W|Y])^2] \leq \mathbb{E}[(W - f(Y))^2], \quad \forall f \in L^2.$$

Proof of Proposition 4. By Proposition 3, we know that intermediation of complete data sharing is profitable if and only if $3G(S_{-i}) \geq G(S)$. Using the additive data structure with $w_i = \theta + \theta_i$ and recalling that α denotes the correlation coefficient between (w_i, w_j), we now rewrite the difference as

$$\begin{aligned}
D(\alpha, N) &= 3G(S_{-i}) - G(S), \\
&= 3\operatorname{var}[\mathbb{E}[w_i|S_{-i}]] - \operatorname{var}[\mathbb{E}[w|S]], \\
&= 3\operatorname{var}[\mathbb{E}[\alpha\theta|S_{-i}]] - \operatorname{var}[\mathbb{E}[\alpha\theta + (1-\alpha)\theta_i|S]], \\
&\geq 3\alpha^2 \operatorname{var}[\mathbb{E}[\theta|S_{-i}]] - \alpha^2 \operatorname{var}[\mathbb{E}[\theta|S]] - 2\alpha(1-\alpha)\sqrt{\operatorname{var}[\theta]\operatorname{var}[\theta_i]} - (1-\alpha)^2 \operatorname{var}[\theta_i].
\end{aligned}$$

Under our symmetry assumption, the variance of the posterior expectation of the common willingness to pay $\operatorname{var}[\mathbb{E}[\theta|S_{1,..,N}]]$ can be written as a function of N. Now we argue that $\operatorname{var}[\mathbb{E}[\theta|S_{1,..,N}]]$ is increasing in N. We first define $g(S_{1,...,N-1}) \triangleq \mathbb{E}[\theta|S_{1,..,N-1}]$. Then, according to Lemma 1, we have

$$\begin{aligned}
\operatorname{var}[\mathbb{E}[\theta|S_{1,..,N}]] &= \max_{f \in L^2} \ \operatorname{var}[\theta] - \mathbb{E}[(\theta - g(S_{1,...,N}))^2], \\
&\geq \max_{f \in L^2} \ \operatorname{var}[\theta] - \mathbb{E}[(\theta - g(S_{1,...,N-1}))^2], \\
&= \operatorname{var}[\mathbb{E}[\theta|S_{1,..,N-1}]].
\end{aligned}$$

The sequence $\operatorname{var}[\mathbb{E}[\theta|S_{1,..,N}]]$ is increasing and bounded. Therefore, it converges and

$$\liminf_{N\to\infty} D(\alpha, N) \geq 2\alpha^2 \lim_{N\to\infty} \operatorname{var}[\mathbb{E}[\theta|S_{1,...,N}]] - 2\alpha(1-\alpha)\sqrt{\operatorname{var}[\theta]\operatorname{var}[\theta_i]} - (1-\alpha)^2 \operatorname{var}[\theta_i].$$

Now we let $\alpha \to 1$, and we obtain

$$\liminf_{\alpha\to 1} \liminf_{N\to\infty} D(\alpha, N) = 2 \lim_{N\to\infty} \operatorname{var}[\mathbb{E}[\theta|S_{1,...,N}]] > 0.$$

Therefore, there exists α^* such that for all $\alpha > \alpha^*$,

$$\liminf_{N\to\infty} D(\alpha, N) > \frac{3}{2} \lim_{N\to\infty} \text{var}[\mathbb{E}[\theta|S_{1,...,N}]].$$

In turn, this implies that, for any fixed $\alpha > \alpha^*$, there exists a $N^*(\alpha)$ such that for any $N > N^*(\alpha)$,

$$D(\alpha, N) > \lim_{N\to\infty} \text{var}[\mathbb{E}[\theta|S_{1,...,N}]] > 0.$$

This completes the proof. ■

Proof of Corollary 3 and 4. When fundamentals w_i are perfectly correlated,

$$\begin{aligned}
\mathbb{E}[w_i|S] &= \mathbb{E}[\theta|S] = \mathbb{E}[\theta|S_1, ..., S_N], \\
\mathbb{E}[w_i|S_{-i}] &= \mathbb{E}[\theta|S_{-i}], \\
\text{var}\,\mathbb{E}[\theta|S_{-i}] &= \text{var}\,\mathbb{E}[\theta|S_1, ..., S_{N-1}].
\end{aligned}$$

As we have argued in the proof of Proposition 4, $\text{var}\,\mathbb{E}[\theta|S_1, ..., S_{N-1}]$ is an increasing converging sequence. Therefore the informativeness of $N-1$ signals approaches that of N,

$$\lim_{N\to\infty} G(S) = \lim_{N\to\infty} G(S_{-i}),$$

and intermediation is then profitable:

$$\lim_{N\to\infty} \frac{R(S)}{N} = \frac{1}{4} \lim_{N\to\infty} G(S) > 0.$$

In the limit for $N \to \infty$, the data externality and the consumer surplus are given by

$$\begin{aligned}
\lim_{N\to\infty} U_i(S, S) - U_i(S_i, \varnothing) &= \lim_{N\to\infty} \mathbb{E}[\frac{1}{8}(\mathbb{E}[w_i|S])^2 - \frac{1}{2}(\mathbb{E}[w_i|S_i] - \mathbb{E}[w_i])^2 - \frac{1}{8}\mathbb{E}[w_i]^2] \\
&= \lim_{N\to\infty} \frac{1}{8}\,\text{var}[\mathbb{E}[w_i|S]] - \frac{1}{2}\,\text{var}[\mathbb{E}[w_i|S_i]], \\
\lim_{N\to\infty} DE_i(S) &= \lim_{N\to\infty} \frac{1}{2}\,\text{var}[\mathbb{E}[w_i|S]] - \frac{1}{2}\,\text{var}[\mathbb{E}[w_i|S_i]] - \frac{3}{8}\,\text{var}[\mathbb{E}[w_i|S_{-i}]] \\
&= \lim_{N\to\infty} \frac{1}{8}\,\text{var}[\mathbb{E}[w_i|S]] - \frac{1}{2}\,\text{var}[\mathbb{E}[w_i|S_i]].
\end{aligned}$$

Therefore, when the initial noise is sufficiently small (i.e., when $\text{var}[\mathbb{E}[w_i|S_i]]$ is close to $\text{var}[w_i]$), the data externality is negative and data sharing hurts consumers.

When $\alpha = 0$ and $\beta = 1$, since w_i is independent from the other consumers' signals, we have $\text{var}[\mathbb{E}[w_i|S_{-i}]] = 0$. Thus, intermediation is always unprofitable, and the data

externality is always positive,

$$R(S) = -\frac{1}{8}\operatorname{var}[\mathbb{E}[w_i|S]] < 0,$$
$$DE(S) = \frac{1}{2}(\operatorname{var}[\mathbb{E}[w_i|S]] - \operatorname{var}[\mathbb{E}[w_i|S_i]] \geq 0.$$

Finally, for the results on consumer surplus, we turn to Lemma 1. In particular, we know

$$\begin{aligned}
\operatorname{var}[\mathbb{E}[w_i|S]] &= \operatorname{var}[w_i] - \mathbb{E}[(w_i - \mathbb{E}[w_i|S])^2], \\
&\geq \operatorname{var}[w_i] - \mathbb{E}[(w_i - (s_i - \frac{1}{N-1}\Sigma_{j\neq i}s_j))^2], \\
&= \operatorname{var}[\theta_i] - \mathbb{E}[(\theta_i - (\theta_i + \varepsilon - \frac{1}{N-1}\Sigma_{j\neq i}\theta_j + \varepsilon))^2], \\
&= \operatorname{var}[\theta_i] - \mathbb{E}[(\frac{1}{N-1}\Sigma_{j\neq i}\theta_j)^2] = \frac{N-2}{N-1}\operatorname{var}[\theta_i] \to \operatorname{var}[w_i].
\end{aligned}$$

Thus we obtain

$$\lim_{N\to\infty} U_i(S,S) - U_i(S_i,\varnothing) = \frac{1}{8}\operatorname{var}[w_i] - \frac{1}{2}\operatorname{var}[\mathbb{E}[w_i|S_i]].$$

When σ is sufficiently large, so that $\operatorname{var}[\mathbb{E}[w_i|S_i]]$ is close to 0, intermediation increases consumer surplus. This ends the proof. ■

Proof of Theorem 1. Under an arbitrary data-inflow policy, each consumer i observes a noisy signal S_i of her own willingness to pay and sends a potentially noisier signal thereof X_i to the intermediary.[18] Consumer i knows her own S_i and X_i. Given the data inflow X, the intermediary chooses an outflow policy, namely the signal $Y = Y(X)$ sent to the producer and the signal $Y_i = Y_i(X)$ sent to each consumer i. The intermediary chooses a policy Y (and his favorite equilibrium in the ensuing game) that maximizes the producer's ex ante expected payoff, which it fully extracts through the fee m_0.

The proof of this result is organized as follows. First, we argue that it is without loss of generality to focus on outflow policy where Y is measurable with respect to Y_i, i.e., where consumer i observes all the information sent to the producer. Then, we prove it is optimal to provide full public information, $Y = Y_i = X$.

Lemma 2 *It is without loss of generality to consider information structures where $Y(X)$ is measurable with respect to $Y_i(X)$.*

[18] Under complete data sharing, for example, the consumer either reports $X_i = S_i$ or refuses to report so that X_i has infinite variance (or the corresponding $\sigma-$algebra is the empty set).

Proof. For any information structure (Y, Y_i), denote an induced signaling equilibrium as $\gamma = (\bar{q}_i, \bar{p})$, where $\bar{p} : Y \to R^+$ is the pricing strategy of the producer and $\bar{q}_i : Y_i \times S_i \times X_i \times R^+ \to R^+$ is the demand function of consumer i. We first argue that there exists an equilibrium γ^* under information structure $(\bar{p} \circ Y, (Y_i, \bar{p} \circ Y))$ that brings the producer a weakly higher ex-ante payoff. In this new information structure, instead of revealing Y to the producer, the intermediary directly recommends the price $\bar{p}(Y)$ which coincides with the equilibrium pricing strategy in $\bar{\gamma}$, and tells consumer i both Y_i and the price recommendation.

On the equilibrium path of $\bar{\gamma}$, consumer i updates her posterior $\mu(Y_i, S_i, \bar{p}_i(Y))$ using Y_i, her own private signal S_i, the report X_i, as well as the observed price p_i. We denote the consumer's demand as a function of her posterior beliefs and the price as

$$q_i(\mu(Y_i, S_i, X_i, p_i), p_i).$$

The ex-ante profit of the producer from consumer i is given by

$$\mathbb{E}\left[\bar{p}_i q_i(\mu(Y_i, S_i, X_i, \bar{p}_i), \bar{p}_i)\right].$$

Now consider the new outflow policy $(\bar{p} \circ Y, (Y_i, \bar{p} \circ Y))$. Under this policy, because consumer i knows everything that the producer knows, the price has no signaling effect. There is a natural equilibrium where consumer i forms her demand using the data outflow $(Y_i, p^* \circ Y)$ from the intermediary as well as her own signal S_i and the data inflow X_i. The price charged by the producer no longer influences the consumer's posterior, which therefore coincides with the consumer's on-path posterior in the old equilibrium $\bar{\gamma}$:

$$\mu(Y_i, S_i, X_i, \bar{p}_i(Y)).$$

Knowing this, the producer maximizes his ex ante payoff by choosing a pricing strategy $\hat{p}(\cdot)$ as a function of his signal $\bar{p} \circ Y$. Thus the producer's equilibrium profit is given by

$$\max_{\hat{p}} \hat{p}\big(\bar{p} \circ Y\big) q_i\Big(\mu(Y_i, S_i, X_i, \bar{p}_i(Y)), \hat{p}\big(\bar{p} \circ Y\big)\Big).$$

Clearly "following the intermediary's recommendation," i.e., setting $\hat{p}(p) = p$ is a feasible strategy that yields the same payoff as in the old equilibrium $\bar{\gamma}$. Consequently, the producer's equilibrium payoff is weakly higher than in $\bar{\gamma}$. ■

Now we are ready to prove Theorem 1. By Lemma 2, it is without loss of generality to assume the producer receives a signal Y and consumer receives a signal (Y_i, Y). Thus, we can focus on equilibria where prices have no signaling effect. These equilibria coincide with

those described in the proof of Proposition 1. As we have shown there, the profit of the producer is:

$$\mathbb{E}\left[\frac{\mathbb{E}[w_i|Y]}{2}\left(\mathbb{E}[w_i|Y\cup Y_i,S_i,X_i]-\frac{\mathbb{E}[w_i|Y]}{2}\right)\right]=\frac{\mathbb{E}\left[(\mathbb{E}[w_i|Y])^2\right]}{4}=\frac{\mathrm{var}[\mathbb{E}[w_i|Y]]+\mathbb{E}[w_i]^2}{4}.$$

Therefore it is optimal to maximize $\mathrm{var}[\mathbb{E}[w_i|Y]]$, which is achieved by setting $Y=X$, and hence the intermediary reveals everything both to the producer and to consumer i. ■

Proof of Theorem 2. In the main text, the data inflow from consumer i is given by $X_i=S_i$ (under complete sharing) or $X_i=\delta(S_i)$ (under anonymization). Note that X is symmetrically distributed, i.e., its joint density is unchanged under permutations.

For any fixed inflow policy X, we refer to p_{-i} as the off-path price charged to consumer i when she does not accept the intermediary's contract, and to p_i as the on-path price charged to consumer i. Now consider another inflow policy X^* identical to X up to a random permutation of the consumers' identities. Under this scheme, we refer to p^*_{-i} as the off-path price for consumer i, and to p^*_i as the on-path price for consumer i.

We first argue that $p_{-i}=p^*_{-i}$ for any realization of W,S,X. To do so, let us calculate consumer i's posterior about W_i under each inflow policy. Under the non-anonymized scheme, the posterior distribution of consumer i's willingness to pay is given by

$$\begin{aligned}&f_i(W_i=w_i|S_i=s_i,X_i=x_i,X=x)\\&\quad=\frac{\int f(W_i=w_i,W_{-i}=w'_{-i},S_i=s_i,S_{-i}=s'_{-i},X_i=x_i,X_{-i}=x_{-i})\mathrm{d}s'_{-i}\mathrm{d}w'_{-i}}{\int f(W=w',S_i=s_i,S_{-i}=s'_{-i},X_i=x_i,X_{-i}=x_{-i})\mathrm{d}s'_{-i}\mathrm{d}w'}.\end{aligned}$$

Recall from Theorem 1 that the intermediary's optimal data outflow policy consists of revealing to the consumers all the available information, even if the consumer refuses to participate. When the data is anonymized, because consumer i knows her own report X_i, the data outflow reveals to her the vector of reports X_{-i} without knowing who generated each one. We now define $\delta\in S^{n-1}$ as permutation of consumer indices. Consumer i's posterior

distribution over her willingness to pay w_i is now given by

$$\begin{aligned}
&f_i(W_i = w_i | S_i = s_i, X_i = x_i, X^*_{-i} = x_{-i}) \\
&\quad = \frac{\Pr(W_i = w_i, S_i = s_i, X_i = x_i, X^*_{-i} = x^*_{-i})}{\Pr(S_i = s_i, X_i = x_i, X^*_{-i} = x^*_{-i})} \\
&\quad = \frac{\Sigma_{\delta \in S^{n-1}} \Pr(\delta, W_i = w_i, S_i = s_i, X_i = x_i, X_{-i} = x_{\delta(-i)})}{\Sigma_{\delta \in S^{n-1}} \Pr(\delta, S_i = s_i, X_i = x_i, X_{-i} = x_{\delta(-i)})} \\
&\quad = \frac{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) \Pr(W_i = w_i, S_i = s_i, X_i = x_i, X_{-i} = x_{\delta(-i)})}{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) Pr(S_i = s_i, X_i = x_i, X_{-i} = x_{\delta(-i)})}.
\end{aligned}$$

Because of the symmetry assumption, we know that

$$\begin{aligned}
&\Pr(W_i = w_i, S_i = s_i, X_i = x_i, X_{-i} = x_{\delta(-i)}) \\
&= \int f(W_i = w_i, W_{-i} = w'_{-i}, S_i = s_i, S_{-i} = s'_{-i}, X_i = x_i, X_{-i} = x_{\delta(-i)}) \mathrm{d}s'_{-i} \mathrm{d}w'_{-i} \\
&= \int f(W_i = w_i, W_{-i} = w'_{\delta^{-1}(-i)}, S_i = s_i, S_{-i} = s'_{\delta^{-1}(-i)}, X_i = x_i, X_{-i} = x_{-i}) \mathrm{d}s'_{-i} \mathrm{d}w'_{-i} \\
&= \int f(W_i = w_i, W_{-i} = w'_{\delta^{-1}(-i)}, S_i = s_i, S_{-i} = s'_{\delta^{-1}(-i)}, X_i = x_i, X_{-i} = x_{-i}) \mathrm{d}s'_{\delta^{-1}(-i)} \mathrm{d}w'_{\delta^{-1}(-i)} \\
&= \Pr(W_i = w_i, S_i = s_i, X_i = x_i, X_{-i} = x_{-i}).
\end{aligned}$$

For the same reason, we also have

$$\Pr(S_i = s_i, X_i = x_i, X_{-i} = x_{\delta(-i)}) = \Pr(S_i = s_i, X_i = x_i, X_{-i} = x_{-i}).$$

Thus the posterior of consumer i can be simplified as:

$$\begin{aligned}
&f_i(W_i = w_i | S_i = s_i, X_i = x_i, X^*_{-i} = x_{-i}) \\
&\quad = \frac{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) \Pr(\delta, W_i = w_i, S_i = s_i, X_i = x_i, X_{-i} = x_{-i})}{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) Pr(S_i = s_i, X_i = x_i, X_{-i} = x_{-i})} \\
&\quad = \frac{\Sigma_{\delta \in S^{n-1}} \frac{1}{|S^{n-1}|} \Pr(\delta, W_i = w_i, S_i = s_i, X_i = x_i, X_{-i} = x_{-i})}{\Sigma_{\delta \in S^{n-1}} \frac{1}{|S^{n-1}|} \Pr(S_i = s_i, X_i = x_i, X_{-i} = x_{-i})} \\
&\quad = f_i(W_i = w_i | S_i = s_i, X_i = x_i, X^*_{-i} = x_{-i}).
\end{aligned}$$

We have therefore proved that consumer i has the same posterior about her willingness to pay w_i for any realization of W, S, X irrespective of whether the data is anonymized of not. Furthermore, this holds both on and off the path of play.

Next, we show the producer also has the same posterior about W_i for any realization of

W, S, X when consumer i refuses to report. Under the non-anonymized scheme, the posterior density is given by:

$$f_i(W_i = w_i | X = x) = \frac{\int f(W_i = w_i, W_{-i} = w'_{-i}, S = s', X_i = x_i, X_{-i} = x_{-i}) \mathrm{d}s' \mathrm{d}w'_{-i}}{\int f(W = w', S = s', X = x_i, X_{-i} = x_{-i}) \mathrm{d}s' \mathrm{d}w'}.$$

Under the anonymized scheme, the posterior density is given by

$$f_i(W_i = w_i | X^* = x) = \frac{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) \Pr(W_i = w_i, X = \delta(x))}{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) \Pr(X = \delta(x))}$$

By the earlier argument, we can simplify it as follows:

$$\frac{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) \Pr(W_i = w_i, X = x)}{\Sigma_{\delta \in S^{n-1}} \Pr(\delta) \Pr(X = x)} = f_i(W_i = w_i | X = x)$$

Since the posteriors for both parties are the same for any realization, so is the price, and hence the welfare impact of information

The profit of the intermediary from consumer i's data under inflow policy X is given by

$$R_i(X) = \Pi(X, X) - \Pi(S_i, \varnothing) - U_i((S_i, X_{-i}), X_{-i}) + U_i((S_i, X), X).$$

We have argued that consumer surplus off the path is the same:

$$U_i((S_i, X_{-i}), X_{-i}) = U_i((S_i, X_{-i}), X^*_{-i}).$$

We now turn to the last term–the impact on social welfare on the path of play:

$$\begin{aligned}
&\Pi((S_i, X), X) + U_i((S_i, X), X) \\
&= \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i | S_i, X_i, X] - \mathbb{E}[w_i | X])^2 + \frac{1}{4}(\mathbb{E}[w_i | X])^2] + \frac{\mathrm{var}[\mathbb{E}[w_i | X]]}{4} \\
&= \frac{1}{2}\mathrm{var}[\mathbb{E}[w_i | S_i, X_i, X]] - \frac{1}{8}\mathrm{var}[\mathbb{E}[w_i | X]].
\end{aligned}$$

Recall that consumer i has the same on path posterior under two different scheme. Therefore, the difference in the intermediary's profits under the two policies reduces to

$$\begin{aligned}
&\frac{1}{2}\mathrm{var}[\mathbb{E}[w_i | S_i, X_i, X]] - \frac{1}{8}\mathrm{var}[\mathbb{E}[w_i | X]] - \frac{1}{2}\mathrm{var}[\mathbb{E}[w_i | S_i, X_i, X^*]] + \frac{1}{8}\mathrm{var}[\mathbb{E}[w_i | X^*]] \\
&= -\frac{1}{8}\mathrm{var}[\mathbb{E}[w_i | X]] + \frac{1}{8}\mathrm{var}[\mathbb{E}[w_i | X^*]] \leq 0.
\end{aligned}$$

Therefore anonymization is more profitable than complete sharing, and strictly so whenever

anonymization makes the estimation less precise. ■

In order to prove Proposition 5, we first state a basic property of anonymized data sharing in our symmetric environment.

Lemma 3 *When the data is anonymized, the following holds:*

$$\mathbb{E}[w_i|A] = \mathbb{E}[w_j|A].$$

Proof of Lemma 3. Denote the joint distribution of W and S as $f(W = w, S = s)$ and the posterior of W_i after observing A as $f(W_i = w|A)$. Denote the permutation in S^N as ν and specially the swapping between i and j as ν_{ij}. For notational simplicity, we use Pr to denote both probability and the proper marginal density.

$$\begin{aligned} f_i(W_i = w_i|A = s) &= \frac{\Pr(W_i = w_i, A = s)}{\Pr(A = s)} = \frac{\Sigma_{\nu \in S^N} \Pr(\nu) \Pr(W_i = w_i, S_\nu = s))}{\Pr(A = s)} \\ &= \frac{\Sigma_{\nu \in S^N} \frac{1}{|S^N|} \int f(W_i = w_i, W_j = w_j, W_{-ij} = w_{-ij}, S_\nu = s) \mathrm{d}w_j \mathrm{d}w_{-ij}}{\Pr(A = s)}. \end{aligned}$$

Since f is unchanged under permutation, we can apply the following transformation:

$$\begin{aligned} f_i(W_i = w_i|A = s) &= \frac{\Sigma_{\nu \in S^N} \frac{1}{|S^N|} \int f(W_j = w_i, W_i = w_j, W_{-ij} = w_{-ij}, S_{\nu_{ij} \circ \nu} = s) \mathrm{d}w_j \mathrm{d}w_{-ij}}{\Pr(A = s)}, \\ = \frac{\Sigma_{\nu_{ij} \circ \nu \in S^N} \frac{1}{|S^N|} \int f(W_j = w_i, W_i = w_i', W_{-ij} = w_{-ij}, S_{\nu_{ij} \circ \nu} = s) \mathrm{d}w_i' \mathrm{d}w_{-ij}}{\Pr(A = s)} &= f_j(W_j = w_i|A = s). \end{aligned}$$

Because the posterior distribution is the same, so is the conditional expectation since

$$\mathbb{E}[w_i|A] = \int w_i f_i(W_i = w_i|A) \mathrm{d}w_i,$$

which ends the proof. ■

Proof of Proposition 5. Combining Lemmas 1 and 3, we obtain

$$\begin{aligned} &\mathbb{E}[w_i|A] = \mathbb{E}[\frac{1}{N}\Sigma_i w_i|A] = \mathbb{E}[\theta + \frac{1}{N}\Sigma_i \theta_i|A]; \\ &G(A) = \mathrm{var}[\mathbb{E}[\theta + \frac{1}{N}\Sigma_i \theta_i|A]] = \mathrm{var}[\theta + \frac{1}{N}\Sigma_i \theta_i] - \mathbb{E}[(\theta + \frac{1}{N}\Sigma_i \theta_i - \mathbb{E}[\theta + \frac{1}{N}\Sigma_i \theta_i|A])^2]. \end{aligned}$$

We can simplify the last term as follows:

$$\begin{aligned}
&\mathbb{E}[(\theta + \frac{1}{N}\Sigma_i\theta_i - \mathbb{E}[\theta + \frac{1}{N}\Sigma_i\theta_i|A])^2] \\
&= \mathbb{E}[(\theta - \mathbb{E}[\theta|A])^2 + \frac{1}{N^2}(\Sigma_i\theta_i - \Sigma_i\mathbb{E}[\theta_i|A])^2 - \frac{2}{N}(\theta - \mathbb{E}[\theta|A])(\Sigma_i\theta_i - \Sigma_i\mathbb{E}[\theta_i|A])] \\
&\geq \mathbb{E}[(\theta - \mathbb{E}[\theta|A])^2] - \frac{2}{N}\sqrt{\operatorname{var}[\theta - \mathbb{E}[\theta|A]]\operatorname{var}[\Sigma_i\theta_i - \Sigma_i\mathbb{E}[\theta_i|A]]} \\
&\geq \mathbb{E}[(\theta - \mathbb{E}[\theta|A])^2] - \frac{2}{N}\sqrt{N\operatorname{var}[\theta]\operatorname{var}[\theta_i]},
\end{aligned}$$

where the last inequality comes from Lemma 1. The intermediary's profit can be written as

$$\begin{aligned}
R &= 3G(A_{-i}) - G(A), \\
&= 3\operatorname{var}[\mathbb{E}[\theta|A_{-i}]] - \operatorname{var}[\theta] - \frac{1}{N}\operatorname{var}[\theta_i] + \mathbb{E}[(\theta + \frac{1}{N}\Sigma_i\theta_i - \mathbb{E}[\theta + \frac{1}{N}\Sigma_i\theta_i|A])^2], \\
&\geq 3\operatorname{var}[\mathbb{E}[\theta|A_{-i}]] - \operatorname{var}[\theta] - \frac{1}{N}\operatorname{var}[\theta_i] + \mathbb{E}[(\theta - \mathbb{E}[\theta|A])^2] - \frac{2}{N}\sqrt{N\operatorname{var}[\theta]\operatorname{var}[\theta_i]} \\
&= 3\operatorname{var}[\mathbb{E}[\theta|A_{-i}]] - \operatorname{var}[\theta] - \frac{1}{N}\operatorname{var}[\theta_i] + \mathbb{E}[(\theta + \frac{1}{N}\Sigma_i\theta_i - \mathbb{E}[\theta + \frac{1}{N}\Sigma_i\theta_i|A])^2], \\
&= 3\operatorname{var}[\mathbb{E}[\theta|A_{-i}]] - \operatorname{var}[\mathbb{E}[\theta|A]] - \frac{1}{N}\operatorname{var}[\theta_i] - \frac{2}{\sqrt{N}}\sqrt{\operatorname{var}[\theta]\operatorname{var}[\theta_i]}.
\end{aligned}$$

Therefore, in the limit we have:

$$\lim_{N\to\infty} R = 2\lim_{N\to\infty}\operatorname{var}[\mathbb{E}[\theta|A]] > 0.$$

This ends the proof. ■

Proof of Theorem 3. We first prove the total compensation is bounded from above, which immediately implies the individual compensation goes to 0 as $N \to \infty$. From Lemma 3, we know that

$$\begin{aligned}
G(A) = \operatorname{var}[\mathbb{E}[w_i|A]] &= \operatorname{var}[\mathbb{E}[\Sigma_i\frac{w_i}{N}|A]], \\
&\leq \operatorname{var}[\Sigma_i\frac{w_i}{N}] = \operatorname{var}[\theta] + \frac{\operatorname{var}[\theta_i] + \operatorname{var}[\varepsilon_i]}{N}.
\end{aligned}$$

On the other hand, we also know

$$G(A_{-i}) = \operatorname{var}[\mathbb{E}[\theta|A_{-i}]] = \operatorname{var}[\theta] - \mathbb{E}[(\theta - \mathbb{E}[\theta|A_{-i}])^2].$$

Since the conditional expectation is the best L^2 approximation, we know it leads to a smaller

error than the "sample average estimator,"

$$\mathbb{E}\left[(\theta - \mathbb{E}[\theta|A_{-i}])^2\right] \leq \mathbb{E}\left[\theta - \frac{1}{N-1}\Sigma_{j\neq i}(\theta+\theta_j+\varepsilon_j)^2\right] = \frac{1}{N-1}(\text{var}[\theta_i]+\text{var}[\varepsilon_j]).$$

Therefore, we have:

$$\begin{aligned} N(G(A)-G(A_{-i})) &\leq N\Big(\text{var}[\theta] + \frac{\text{var}[\theta_i]+\text{var}[\varepsilon_i]}{N} - \text{var}[\theta] + \frac{1}{N-1}(\text{var}[\theta_i]+\text{var}[\varepsilon_j])\Big), \\ &= N\left(\frac{\text{var}[\theta_i]+\text{var}[\varepsilon_i]}{N} + \frac{1}{N-1}(\text{var}[\theta_i]+\text{var}[\varepsilon_i])\right) \\ &\leq 3(\text{var}[\theta_i]+\text{var}[\varepsilon_i]). \end{aligned}$$

The total consumer compensation is then given by

$$\frac{3N}{8}(G(A)-G(A_{-i})) \leq \frac{9}{8}(\text{var}[\theta_i]+\text{var}[\varepsilon_i]).$$

Finally, the intermediary's profit is growing linearly in N because

$$\begin{aligned} R(S) &= \frac{N}{4}G(A) - \frac{3N}{8}(G(A)-G(A_{-i})), \\ \lim_{N\to\infty}\frac{R(S)}{N} &= \frac{1}{4}\lim_{N\to\infty}G(A), \end{aligned}$$

which ends the proof. ■

Proof of Proposition 6. When data is not anonymized we have:

$$G(S)-G(S_{-i}) = \text{var}[\mathbb{E}[\theta+\theta_i|S]] - \text{var}[\mathbb{E}[\theta|S_{-i}]].$$

Because of symmetry, we have

$$\text{cov}[\mathbb{E}[\theta|S],\mathbb{E}[\theta_i|S]] = \text{cov}[\mathbb{E}[\theta|S],\mathbb{E}[\theta_j|S]] = \text{cov}[\mathbb{E}[\theta|S],\Sigma_{j=1}^N\mathbb{E}[\theta_j/N\,|S]].$$

Because the correlation coefficient is always greater than -1, we obtain

$$\begin{aligned} \text{cov}[\mathbb{E}[\theta|S],\Sigma_{j=1}^N\mathbb{E}[\theta_j/N\,|S]] &\geq -\sqrt{\text{var}[\theta]\,\text{var}[\Sigma_{j=1}^N\mathbb{E}[\theta_j/N\,|S]]}, \\ &\geq -\sqrt{\text{var}[\theta]\,\text{var}[\Sigma_{j=1}^N\theta_j/N]}. \end{aligned}$$

Therefore, according to Lemma 1 we have:

$$\begin{aligned} G(S) - G(S_{-i}) &= \operatorname{var}[\mathbb{E}[\theta|S]] + 2\operatorname{cov}[\mathbb{E}[\theta|S], \mathbb{E}[\theta_i|S]] + \operatorname{var}[\mathbb{E}[\theta_i|S]] - \operatorname{var}[\mathbb{E}[\theta|S_{-i}]] \\ &\geq \operatorname{var}[\mathbb{E}[\theta|S]] - 2\frac{1}{\sqrt{N}}\sqrt{\operatorname{var}[\theta]\operatorname{var}[\theta_i]} + \operatorname{var}[\mathbb{E}[\theta_i|S]] - \operatorname{var}[\mathbb{E}[\theta|S_{-i}]], \end{aligned}$$

and hence

$$\liminf_{N\to\infty} G(S) - G(S_{-i}) \geq \operatorname{var}[\mathbb{E}[\theta_i|S]].$$

The last term is strictly positive because

$$\begin{aligned} \operatorname{var}[\mathbb{E}[\theta_i|S]] &= \operatorname{var}[\theta_i] - \mathbb{E}[(\theta_i - \mathbb{E}[\theta_i|S])^2] \\ &\geq \operatorname{var}[\theta_i] - \mathbb{E}[(\theta_i - \frac{\operatorname{var}[\theta_i]}{\operatorname{var}[\theta_i] + \operatorname{var}[\theta] + \operatorname{var}[e]} S_i)^2], \\ &= \operatorname{var}[\theta_i] - (\operatorname{var}[\theta_i] - \frac{\operatorname{var}^2[\theta_i]}{\operatorname{var}[\theta_i] + \operatorname{var}[\theta] + \operatorname{var}[e]}), \\ &= \frac{\operatorname{var}^2[\theta_i]}{\operatorname{var}[\theta_i] + \operatorname{var}[\theta] + \operatorname{var}[e]} > 0, \end{aligned}$$

where the first inequality again uses Lemma 1. This ends the proof. ■

Proof of Proposition 7. In the standard "divide and conquer" scheme, the compensation for the i-th consumer is the marginal loss of revealing her information given that $i-1$ consumers reveal their signals:

$$\frac{3}{8}G(S_{1,\ldots,i}) - \frac{3}{8}G(S_{1,\ldots,i-1}).$$

Since in general we do not know whether this marginal loss is decreasing in i, we consider the following revised version of divide and conquer, where consumer i receives

$$m_i = \max_{k\geq i} \frac{3}{8}G(S_{1,\ldots,k}) - \frac{3}{8}G(S_{1,\ldots,k-1}).$$

Under this payment scheme, it is dominant strategy for consumer 1 to accept the offer. Also, it is optimal for consumer i to accept the offer, given that the first $i-1$ consumers accept. Using an identical proof to Theorem 3, we obtain

$$\begin{aligned} \frac{3}{8}G(S_{1,\ldots,i}) - \frac{3}{8}G(S_{1,\ldots,i-1}) &\leq \frac{3}{8}(\frac{1}{i} + \frac{1}{i-1})(\operatorname{var}[\theta_i] + \operatorname{var}[\varepsilon_i]), \\ &\leq \frac{3}{4}\frac{1}{i-1}(\operatorname{var}[\theta_i] + \operatorname{var}[\varepsilon_i]). \end{aligned}$$

Therefore, we obtain an upper bound on the compensation paid to consumer i:

$$m_i \leq \max_{k \geq i} \frac{3}{4} \frac{1}{k-1} (\text{var}[\theta_i] + \text{var}[\varepsilon_i]) = \frac{3}{4} \frac{1}{i-1} (\text{var}[\theta_i] + \text{var}[\varepsilon_i]).$$

Finally, because we have

$$\begin{aligned} \Sigma_i \frac{3}{8} (G(S_{1,...,i}) - G(S_{1,...,i-1})) &\leq \frac{3}{4} (1 + \Sigma_{i=3}^{N} \frac{1}{i-1}) (\text{var}[\theta_i] + \text{var}[\varepsilon_i]), \\ &\leq \frac{3}{4} (1 + \log N) (\text{var}[\theta_i] + \text{var}[\varepsilon_i]), \end{aligned}$$

the total compensation grows at a speed less than $\log N$. This completes the proof. ■

Proof of Proposition 8. Consider J homogeneous groups of consumers, and denote the joint density of the fundamental and noise terms as

$$f(W_1 = w_1, ...W_j = w_j..., W_J = w_J, E_1 = e_1, ...E_j = e_j, ...E_J = e_J),$$

where W_j and E_j refers to the vector of fundamental and noise of consumers in group j. The homogeneity assumption requires:

$$\begin{aligned} &f(W_1 = w_1, ...W_j = w_j..., W_J = w_J, E_1 = e_1, ...E_j = e_j, ...E_J = e_J) \\ &\quad = f(W_1 = w_1, ...W_j = \delta(w_j)..., W_J = w_J, E_1 = e_1, ...E_j = \delta(e_j), ...E_J = e_J) \end{aligned}$$

for any j and any permutation $\delta : \{1, ..., N_j\} \to \{1, ..., N_j\}$.

The proof of this proposition is nearly identical to the one for Theorem 2, so we only sketch the argument. By Theorem 1, we know the intermediary will transmit whatever information it collected to all consumers. By homogeneity, we know the consumer's posterior about their own fundamental w_{ij} is the same whether the signals are anonymized or not, and the producer's posterior about any deviating consumer's fundamental is also the same under the two scheme. Because within-group anonymization helps reduce price discrimination on path, it increases the intermediary's revenue. ■

Proof of Proposition 9. We first consider the case where the intermediary anonymizes all data, including the group identities. Similar to the result in Lemma 3, we know the producer will offer one price to all consumers on the path of play,

$$\mathbb{E}[w_{ij}|A] = \mathbb{E}[w_{i'j'}|A].$$

Denoting $N = \Sigma_j N_j$, we have

$$\begin{aligned}\mathbb{E}[w_{i'j'}|A] &= \frac{1}{\Sigma_j N_j}\Sigma_j\Sigma_i\mathbb{E}[w_{ij}|A] = \frac{1}{\Sigma_j N_j}\Sigma_j\Sigma_i\mathbb{E}[\theta_j + \theta_{ij} + \varepsilon_{ij}|A] \\ &= \frac{1}{\Sigma_j N_j}\Sigma_j\Sigma_i\mathbb{E}[\theta_j|A] = \Sigma_j\frac{N_j}{N}\mathbb{E}[\theta_j|A].\end{aligned}$$

Therefore we obtain an upper bound on the revenue per capita

$$\begin{aligned}\frac{R(A)}{N} &= \frac{3}{8}G(A_{-ij}) - \frac{1}{8}G(A) = \frac{3}{8}\operatorname{var}[\mathbb{E}[w_{ij}]|A_{-ij}] - \frac{1}{8}\operatorname{var}[\mathbb{E}[w_{ij}]|A] \\ &\leq \frac{1}{4}\operatorname{var}[\mathbb{E}[w_{ij}]|A], = \frac{1}{4}\operatorname{var}[\Sigma_j\frac{N_j}{N}\mathbb{E}[\theta_j|A]] \\ &\leq \frac{1}{4}\operatorname{var}[\Sigma_j\frac{N_j}{N}\theta_j] = \frac{1}{4}\frac{1}{N^2}\Sigma N_j^2\operatorname{var}[\theta_j].\end{aligned}$$

Next, consider the case where the intermediary reveals the group identity. Instead of A we use A^g to denote the information that the producer receives. By an argument similar to the proof of Lemma 3, we know that (on path) the producer offers one price to all consumers in each group:

$$\begin{aligned}\mathbb{E}[w_{ij}|A^g] &= \mathbb{E}[w_{i'j}|A^g] \\ &= \frac{1}{N_j}\Sigma_{i'=1}^{N_j}\mathbb{E}[w_{i'j}|A^g] = \frac{1}{N_j}\Sigma_{i'=1}^{N_j}\mathbb{E}[w_{i'j}|A^g] = \mathbb{E}[\theta_j + \frac{1}{N_j}\Sigma_{i'=1}^{N_j}\theta_{i'j}|A^g].\end{aligned}$$

When consumer ij rejects the offer, the intermediary will know the group identity of this deviating consumer and use all the available data to estimate the demand:

$$\mathbb{E}[w_{ij}|A^g_{-ij}] = \mathbb{E}[\theta_j + \theta_{ij}|A^g_{-ij}] = \mathbb{E}[\theta_j|A^g_{-ij}].$$

The revenue that the intermediary obtains from consumer ij's data is then given by

$$\begin{aligned}&\frac{3}{8}\operatorname{var}[\mathbb{E}[w_{ij}]|A^g_{-ij}] - \frac{1}{8}\operatorname{var}[\mathbb{E}[w_{ij}]|A^g], \\ &= \frac{3}{8}\operatorname{var}[\mathbb{E}[\theta_j|A^g_{-ij}]] - \frac{1}{8}\operatorname{var}[\mathbb{E}[\theta_j + \frac{1}{N_j}\Sigma_{i'=1}^{N_j}\theta_{i'j}|A^g]], \\ &\geq \frac{3}{8}\operatorname{var}[\mathbb{E}[\theta_j|A^g_{-ij}]] - \frac{1}{8}\operatorname{var}[\theta_j] - \frac{1}{8N_j}\operatorname{var}[\theta_{ij}] - \frac{1}{4}\sqrt{\frac{1}{N_j}\operatorname{var}[\theta_j]}\sqrt{\operatorname{var}[\theta_{ij}]}.\end{aligned}$$

From Lemma 1, we know

$$\begin{aligned}\mathbb{E}[(\theta_j - \mathbb{E}[\theta_j|A^g_{-ij}])^2] &\leq \mathbb{E}[(\theta_j - \frac{1}{N_j-1}\Sigma_{i'\neq i}s_{i'j})^2],\\ &= \mathbb{E}[(-\frac{1}{N_j-1}\Sigma_{i'\neq i}s_{i'j}\theta_{ij})^2] = \frac{1}{N_j-1}\operatorname{var}[\theta_{ij}];\\ \operatorname{var}[\mathbb{E}[\theta_j|A^g_{-ij}]] &= \operatorname{var}[\theta_j] - \mathbb{E}[(\theta_j - \mathbb{E}[\theta_j|A^g_{-ij}])^2],\\ &\geq \operatorname{var}[\theta_j] - \frac{1}{N_j-1}\operatorname{var}[\theta_{ij}].\end{aligned}$$

Thus we obtain a lower bound on the revenue from consumer ij:

$$\frac{1}{4}\operatorname{var}[\theta_j] - \frac{3}{8}\frac{1}{N_j-1}\operatorname{var}[\theta_{ij}] - \frac{1}{8N_j}\operatorname{var}[\theta_{ij}] - \frac{1}{4}\sqrt{\frac{1}{N_j}\operatorname{var}[\theta_j]}\sqrt{\operatorname{var}[\theta_{ij}]}.$$

Finally we can compute the difference in the revenues

$$\begin{aligned}R(A) - R(A^g) \leq{}& \Sigma_j\frac{N_j^2}{4N}\operatorname{var}[\theta_j]\\ &- \Sigma_j\left(\frac{N_j}{4}\operatorname{var}[\theta_j] - \frac{3}{8}\frac{N_j}{N_j-1}\operatorname{var}[\theta_{ij}] - \frac{1}{8}\operatorname{var}[\theta_{ij}] - \frac{\sqrt{N_j}}{4}\sqrt{\operatorname{var}[\theta_j]}\sqrt{\operatorname{var}[\theta_{ij}]}\right).\end{aligned}$$

As long as $N_j < kN$ where $k < 1$, we know that

$$\begin{aligned}&R(A) - R(A^g)\\ &< \Sigma_j\left(-\frac{1-k}{4}N_j\operatorname{var}[\theta_j] + \frac{3}{8}\frac{N_j}{N_j-1}\operatorname{var}[\theta_{ij}] + \frac{1}{8}\operatorname{var}[\theta_{ij}] + \frac{\sqrt{N_j}}{4}\sqrt{\operatorname{var}[\theta_j]}\sqrt{\operatorname{var}[\theta_{ij}]}\right).\end{aligned}$$

The dominant linear term is decreasing in N_j, and hence we know that as $N_j \to \infty$, revealing group identities is more profitable. ■

Proof of Proposition 10. Each consumer's demand function is given by

$$q_i = w_i - (\ell_i - x_i)^2 - p_i.$$

This means the producer's profit is given by

$$\pi = \sum\nolimits_{i=1}^{N} p_i\left(w_i - (\ell_i - x_i)^2 - p_i\right).$$

Therefore, under any information structure S, the producer offers

$$\begin{aligned} p_i &= \left(\mathbb{E}\left[w_i \mid S\right] - \mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2 \mid S\right]\right)/2, \\ &= \left(\mathbb{E}\left[w_i \mid S\right] - \mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]\right)/2, \\ x_i &= \mathbb{E}\left[\ell_i \mid S\right], \end{aligned}$$

where the second line relies on the fact that the underlying random variables are normal so that $\ell_i - \mathbb{E}[\ell_i|S]$ is independent of S . The consumer's surplus is then given by

$$\begin{aligned} U_i\left(S\right) &= \frac{1}{2}\mathbb{E}\left[\left(w_i - \left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2 - \frac{\mathbb{E}\left[w_i \mid S\right] - \mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]}{2}\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(w_i - \frac{1}{2}\mathbb{E}\left[w_i \mid S\right]\right)^2\right] + \frac{1}{2}\mathbb{E}\left[\left(\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2 - \frac{1}{2}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]\right)^2\right] \\ &- \mathbb{E}\left[\left(w_i - \frac{1}{2}\mathbb{E}\left[w_i \mid S\right]\right)\right]\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2 - \frac{1}{2}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]\right] \\ &= \frac{1}{2}\mathbb{E}\left[\left(w_i - \frac{1}{2}\mathbb{E}\left[w_i \mid S\right]\right)^2\right] + \frac{1}{2}\mathbb{E}\left[\left(\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2 - \frac{1}{2}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]\right)^2\right] \\ &- \frac{1}{4}\mathbb{E}\left[w_i\right]\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[w_i^2 - \frac{3}{4}\mathbb{E}\left[w_i \mid S\right]^2\right] + \frac{1}{2}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^4 - \frac{3}{4}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]^2\right] \\ &- \frac{1}{4}\mathbb{E}\left[w_i\right]\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]. \end{aligned}$$

Therefore the difference is:

$$\begin{aligned} U_i(S) - U_i(\varnothing) &= -\frac{3}{8}\operatorname{var}\left[\mathbb{E}\left[w_i \mid S\right]\right] + \frac{1}{4}\mu\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right] + \frac{1}{2}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^4\right] \\ &\quad -\frac{3}{8}\mathbb{E}\left[\left(\ell_i - \mathbb{E}\left[\ell_i \mid S\right]\right)^2\right]^2 - \frac{1}{2}\mathbb{E}\left[\left(\ell_i - \mu_\tau\right)^4\right] + \frac{3}{8}\mathbb{E}\left[\left(\ell_i - \mu_\tau\right)^2\right]^2. \end{aligned}$$

Since every random variable is assumed to be normal, $\ell_i - \mathbb{E}[\ell_i|S]$ is also normal with zero mean. We can further simplify and obtain

$$\begin{aligned} U_i(S) - U_i(\varnothing) &= -\frac{3}{8}\operatorname{var}\left[\mathbb{E}\left[w_i \mid S\right]\right] + \frac{1}{4}\mu\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right] + \frac{3}{2}\left(\operatorname{var}\left[\ell_i\right] - \operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right]\right)^2 \\ &\quad - \frac{3}{8}\left(\operatorname{var}\left[\ell_i\right] - \operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right]\right)^2 - \frac{3}{2}\operatorname{var}[\ell_i]^2 + \frac{3}{8}\operatorname{var}[\ell_i]^2, \\ &= -\frac{3}{8}\operatorname{var}\left[\mathbb{E}\left[w_i \mid S\right]\right] + \frac{1}{4}\mu\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right] + \frac{9}{8}\left(\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right]^2 - 2\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right]\left(\sigma_\tau^2 + \sigma_{\tau_i}^2\right)\right). \end{aligned}$$

Similarly we have:

$$\begin{aligned}\Pi_i(S) &= \mathbb{E}\left[\left(w_i - (\ell_i - \mathbb{E}\left[\ell_i \mid S\right])^2 - \frac{\mathbb{E}\left[w_i \mid S\right] - \mathbb{E}\left[(\ell_i - \mathbb{E}\left[\ell_i \mid S\right])^2\right]}{2}\right)\frac{\mathbb{E}\left[w_i \mid S\right] - \mathbb{E}\left[(\ell_i - \mathbb{E}\left[\ell_i \mid S\right])^2\right]}{2}\right] \\ &= \frac{1}{4}\mathbb{E}\left[\left(\mathbb{E}\left[w_i \mid S\right] - \mathbb{E}\left[(\ell_i - \mathbb{E}\left[\ell_i \mid S\right])^2\right]\right)\right] \\ &= \frac{1}{4}\mathbb{E}\left[\mathbb{E}\left[w_i \mid S\right]^2 - 2\mathbb{E}\left[w_i \mid S\right]\mathbb{E}\left[(\ell_i - \mathbb{E}\left[\ell_i \mid S\right])^2\right] + \mathbb{E}\left[(\ell_i - \mathbb{E}\left[\ell_i \mid S\right])^2\right]^2\right],\end{aligned}$$

and hence

$$\Pi_i(S) - \Pi_i(\varnothing) = \frac{1}{4}\operatorname{var}[\mathbb{E}\left[w_i \mid S\right]] + \frac{1}{2}\mu \operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right] + \frac{1}{4}\left(\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right]^2 - 2\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right](\sigma_\tau^2 + \sigma_{\tau_i}^2)\right).$$

To summarize, the impact of data sharing on social surplus is given by

$$\begin{aligned} & W_i(S) - W_i(\varnothing) = U_i(S) - U_i(\varnothing) + \Pi_i(S) - \Pi_i(\varnothing), \\ = \; & -\frac{1}{8}\operatorname{var}\left[\mathbb{E}\left[w_i \mid S\right]\right] + \frac{3}{4}\mu \operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right] + \frac{11}{8}\left(\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right]^2 - 2\operatorname{var}\left[\mathbb{E}\left[\ell_i \mid S\right]\right](\sigma_\tau^2 + \sigma_{\tau_i}^2)\right).\end{aligned}$$

Therefore the difference $W_i(S) - W_i(\varnothing)$ is a quadratic function of the variance of the conditional expectation $x \triangleq \operatorname{var}[\mathbb{E}[\ell_i|S]]$. In particular, we let

$$g(x) \triangleq \frac{11}{8}x^2 + \left(\frac{3}{4}\mu - \frac{11}{4}(\sigma_\tau^2 + \sigma_{\tau_i}^2)\right)x.$$

As long as $3\mu > 11(\sigma_\tau^2 + \sigma_{\tau_i}^2)$, this function is positive and increasing in x, which means a higher $\operatorname{var}[\mathbb{E}[\ell_i|S]]$ increases consumer surplus.

Finally, as in the proof of Theorem 2, aggregating w_i increases $W_i(S)$ but keeps $\Pi(\varnothing)$ and $U_i(S_{-i})$ unchanged. Not aggregating ℓ_i increases $W_i(S)$ while keeping $\Pi(\varnothing)$ and $U_i(S_{-i})$ unchanged. Therefore it is optimal for the intermediary to aggregate w_i but not ℓ_i. ■

Proof of Theorem 4. Recall the formula in the proof of Proposition 1,

$$\begin{aligned}\Pi_i(Y_i, Y) &= \frac{\operatorname{var}[\mathbb{E}[w_i|Y]] + \mu^2}{4}, \\ U_i(Y_i, Y) &= \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|Y_i])^2 - \frac{3}{4}(\mathbb{E}[w_i|Y])^2].\end{aligned}$$

With a noisier report X, the consumer i will know S_i and X both on path and off path. The producer will know X on path and X_{-i} if consumer i deviates. Thus the revenue of the

intermediary is:

$$\begin{aligned}\frac{R(X)}{N} &= \Pi_i((S_i,X),X) - \Pi_i(S_i,\varnothing) + U((S_i,X),X) - U((S_i,X),X_{-i}),\\ &= \frac{\text{var}[\mathbb{E}[w_i|X]]}{4} - \frac{3\,\text{var}[\mathbb{E}[w_i|X]]}{8} + \frac{3\,\text{var}[\mathbb{E}[w_i|X_{-i}]]}{8},\\ &= -\frac{\text{var}[\mathbb{E}[w_i|X]]}{8} + \frac{3\,\text{var}[\mathbb{E}[w_i|X_{-i}]]}{8}.\end{aligned}$$

Recall that

$$X_i = w_i + \sigma e_i + \xi + \xi_i = \theta + \theta_i + (\sigma\varepsilon_i + \xi_i) + (\sigma\varepsilon + \xi).$$

For ease of exposition, we rewrite $(\sigma\varepsilon_i + \xi_i)$ as ε_i and $(\sigma\varepsilon + \xi)$ as ε. Since the intermediary can control the variance of ξ, ξ_i but not the initial precision of the consumers' signals, we effectively have a lower bound of the variance $\underline{\sigma}_i^2$ and $\underline{\sigma}^2$ on the new pair $\varepsilon_i, \varepsilon$. Denote the variance of θ as σ_θ^2 and similarly for other variables. It is straightforward to calculate that:

$$\begin{aligned}\mathbb{E}[w_i|X] &= \frac{N\sigma_\theta^2 + \sigma_{\theta_i}^2}{N^2(\sigma_\theta^2+\sigma_\varepsilon^2) + N(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\sum_{i'} x_{i'},\\ \mathbb{E}[w_i|X_{-i}] &= \frac{N\sigma_\theta^2}{(N-1)^2(\sigma_\theta^2+\sigma_\varepsilon^2) + (N-1)(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\sum_{i'\neq i} x_{i'},\\ R(X) &= \frac{3(N-1)N\sigma_\theta^4}{8((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\end{aligned}$$

Now we are ready to prove the theorem. We argue it is optimal to set $\sigma_{\varepsilon_i}^2 = \underline{\sigma}_i^2$ (i.e., to set $\sigma_{\xi_i}^2 = 0$). To show this result, suppose $\sigma_{\varepsilon_i}^2 > \underline{\sigma}_i^2$. Then there exists $\delta > 0$ such that augmenting the common noise to $\bar{\sigma}_\varepsilon^2 \triangleq \sigma_\varepsilon^2 + \delta^2$ and diminishing the idiosyncratic noise to $\bar{\sigma}_{\varepsilon_i}^2 \triangleq \sigma_{\varepsilon_i}^2 - (N-1)\delta^2 \geq \underline{\sigma}_i^2$, the profits of the intermediary will strictly increase. Too see this, consider the expression of the revenue

$$R\,(S) = \frac{3(N-1)N\sigma_\theta^4}{8((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}.$$

The first term is unchanged under new information structure, while the denominator of the second term increases, thus the total profit increases. ■

Proof of Proposition 11. Recall that α is the correlation coefficient between w_i and w_j. Because we have normalized $\text{var}\,[w_i] = 1$, under the additive structure, we have $\sigma_\theta^2 = \alpha$ and $\sigma_{\theta_i}^2 = 1-\alpha$. To establish the result in the statement, we must then show that the

intermediary obtains positive profits if and only if

$$N\left(\sqrt{3}-1\right)\sigma_\theta^2-\sigma_{\theta_i}^2>0.$$

When $N\left(\sqrt{3}-1\right)\sigma_\theta^2<\sigma_{\theta_i}^2$, we have:

$$\begin{aligned}R&=\frac{3n\sigma_\theta^4}{8(\sigma_\theta^2+\sigma_\varepsilon^2)}\left(1-\frac{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}{(n-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}\right)-\frac{(n\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8n(\sigma_\theta^2+\sigma_\varepsilon^2)}\left(1-\frac{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}{n(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}\right)\\&\leq\left(\frac{3n\sigma_\theta^4}{8(\sigma_\theta^2+\sigma_\varepsilon^2)}-\frac{(n\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8n(\sigma_\theta^2+\sigma_\varepsilon^2)}\right)\left(1-\frac{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}{n(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}\right)\leq 0.\end{aligned}$$

Conversely, when $N\left(\sqrt{3}-1\right)\sigma_\theta^2>\sigma_{\theta_i}^2$, we can rewrite R as

$$\begin{aligned}&\frac{A\sigma_\varepsilon^2+B}{8(\sigma_\varepsilon^2 n+\sigma_{\varepsilon_i}^2+n\sigma_\theta^2+\sigma_{\theta_i}^2)(\sigma_\varepsilon^2 n-\sigma_\varepsilon^2+\sigma_{\varepsilon_i}^2+n\sigma_\theta^2-\sigma_\theta^2+\sigma_{\theta_i}^2)},\\&A=(n-1)\left(2n^2\sigma_\theta^4-2n\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right)>0.\end{aligned}$$

Therefore the intermediary can get a positive profit R by setting σ_ε^2 sufficiently large through the addition of correlated noise. ■

References

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2019): "Too Much Data: Prices and Inefficiencies in Data Markets," Discussion paper, MIT.

Acquisti, A., C. Taylor, and L. Wagman (2016): "The Economics of Privacy," *Journal of Economic Literature*, 54, 442–492.

Ali, S. N., G. Lewis, and S. Vasserman (2019): "Voluntary Disclosure and Personalized Pricing," Discussion Paper 26592, National Bureau of Economic Research.

Aridor, G., Y.-K. Che, and T. Salz (2020): "The Economic Consequences of Data Privacy Regulation: Empirical Evidence from GDPR," Discussion Paper 26900, National Bureau of Economic Research.

Arrieta-Ibarra, I., L. Goff, D. Jimenez-Hernandez, J. Lanier, and G. Weyl (2018): "Should We Treat Data as Labor? Moving beyond "Free"," *American Economic Review Paper and Proceedings*, 108, 38–42.

Athey, S., C. Catalini, and C. Tucker (2017): "The Digital Privacy Paradox: Small Money, Small Costs, Small Talk," Discussion paper, National Bureau of Economic Research.

Bergemann, D., and A. Bonatti (2019): "Markets for Information: An Introduction," *Annual Review of Economics*, 11, 85–107.

Choi, J., D. Jeon, and B. Kim (2019): "Privacy and Personal Data Collection with Information Externalities," *Journal of Public Economics*, 173, 113–124.

Committee on the Judiciary (2020): "Investigation of Competition in Digital Markets," Discussion paper, United States House of Representatives.

Cremèr, J., Y.-A. de Montjoye, and H. Schweitzer (2019): "Competition policy for the digital era," Discussion paper, European Commission.

Cummings, R., K. Ligett, M. Pai, and A. Roth (2016): "The Strange Case of Privacy in Equilibrium Models," in *ACM-EC (Economics and Computation) 2016*.

Digital Competition Expert Panel (2019): "Unlocking Digital Competition," Discussion paper.

Fainmesser, I., A. Galeotti, and R. Momot (2020): "Digital Privacy," Discussion paper, Johns Hopkins University.

Gradwohl, R. (2017): "Information Sharing and Privacy in Networks," in *ACM-EC (Economics and Computation) 2017*.

Ichihashi, S. (2020a): "The Economics of Data Externalities," Discussion paper, Bank of Canada.

Ichihashi, S. (2020b): "Online Privacy and Information Disclosure by Consumers," *American Economic Review*, 110(2), 569–595.

Johnson, G., S. Shriver, and S. Goldberg (2020): "Privacy & market concentration: Intended & unintended consequences of the GDPR," Discussion paper, Boston University Questrom School of Business.

Jullien, B., Y. Lefouili, and M. Riordan (2020): "Privacy Protection, Security, and Consumer Retention," Discussion paper, Toulouse School of Economics.

Liang, A., and E. Madsen (2020): "Data and Incentives," Discussion paper, Northwestern University.

Lizzeri, A. (1999): "Information Revelation and Certification Intermediaries," *RAND Journal of Economics*, 30, 214–231.

Loertscher, S., and L. M. Marx (2020): "Digital monopolies: Privacy protection or price regulation?," *International Journal of Industrial Organization*, 71, 1–13.

Miklos-Thal, J., and G. Shaffer (2016): "Naked Exclusion with Private Offers," *American Economic Journal: Microeconomics*, 8, 174–194.

Olea, J. L. M., P. Ortoleva, M. Pai, and A. Prat (2019): "Competing Models," *arXiv preprint arXiv:1907.03809*.

Posner, E. A., and E. G. Weyl (2018): *Radical markets: Uprooting capitalism and democracy for a just society*. Princeton University Press.

Robinson, J. (1933): *The Economics of Imperfect Competition*. Macmillan, London.

Romer, P. (2019): "A Tax That Could Fix Big Tech," *The New York Times*.

Schmalensee, R. (1981): "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, 71, 242–247.

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

Segal, I., and M. Whinston (2000): "Naked Exclusion: Comment," *American Economic Review*, 90, 296–309.

Stigler Committee on Digital Platforms (2019): "Final Report," Discussion paper, Stigler Center for the Study of the Economy and the State.

Taylor, C. (2004): "Consumer Privacy and the Market for Customer Information," *RAND Journal of Economics*, 35, 631–651.

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