Yale University
EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers | Cowles Foundation

9-1-2019

The Economics of Social Data

Dirk Bergemann

Alessandro Bonatti

Tan Gan

Follow this and additional works at: https://elischolar.library.yale.edu/cowles-discussion-paper-series

 Part of the Economics Commons

THE ECONOMICS OF SOCIAL DATA

By

Dirk Bergemann, Alessandro Bonatti, and Tan Gan

August 2019

Revised July 2020

COWLES FOUNDATION DISCUSSION PAPER NO. 2203R2



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

The Economics of Social Data*

Dirk Bergemann[†] Alessandro Bonatti[‡] Tan Gan[§]

July 26, 2020

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 in which 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 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.

*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, and Glen Weyl for helpful discussions. We thank Michelle Fang and Miho Hong for valuable research assistance. We thank the audiences at the ACM-EC 2019 Plenary Lecture, ASSA 2019, ESSET 2019, and the Yale Tobin Center 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.[1]

A central feature of the data collected from individuals is their social aspect. With this term, we mean that the 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 the traffic conditions for nearby drivers which can help them to improve their decisions. Thus, these *individual data* are actually *social data*. The social dimension of the data can simultaneously lead to a loss of privacy and a gain in information. In any case, the social nature of the data generates a *data externality*, the sign and magnitude of which depend on the structure of the data and 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 of individual data changes the terms of trade among consumers, firms (advertisers), and large digital platforms. Second, we examine how the social dimension of the data magnifies the value of individual data for the platforms and facilitates data acquisition. Third, we analyze how data intermediaries with market power (e.g., large Internet platforms that sell targeted advertising space) change the level of aggregation and the precision of the information that they provide in equilibrium about individual 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*.

[1] Bergemann and Bonatti (2019) provided a recent introduction.

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. Each consumer experiences a demand shock. While the producer knows the common prior distribution of the demand shocks, he does not know the realization of the individual demand shocks.

In the data market, the data intermediary acquires demand information from the individual consumers and then sells these data to the producer in some possibly aggregated or noisy version. The data intermediary can choose how much information to buy from the consumers and how much information to share with all of the product market participants. Thus, the data intermediary has control over the volume and the structure of the information flows across consumers, as well as between the consumers and the producer. An example of an information policy is *complete data sharing* by the data intermediary. However, this example is only one of many possible information policies that the intermediary could pursue. Indeed, we find that the solution to the information design problem faced by the intermediary typically favors more limited data collection and data sharing.

The Social Nature of the Data 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 share her data with a large platform. The individual data, and hence the social data, are described by a statistical model of multivariate normal distributions. This statistical model yields a sufficiently rich, yet compact, representation of the arguably canonical factors of social data.

The personal data of the consumer consist 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. In practice, different consumers' preferences can exhibit some common traits, and consumers might also be subject to similar limitations on information that induce correlation in their errors. Thus, we model both the fundamental component and the noise as having additive *idiosyncratic* and *common* elements. The idiosyncratic elements are distributed independently across consumers, while the common elements are perfectly correlated.[2]

The Value of the Social Data The collection of the data from multiple consumers helps to predict the fundamental component of each individual. From observing many individual realizations the data intermediary can improve its estimate of the components (fundamental

[2]The twofold decomposition into fundamental and noise components and idiosyncratic and common elements is canonical in a setting with symmetric consumers. Any other multivariate normal distribution can be equivalently represented in the above format, and in this sense, this representation is canonical.

and noise) that are common across all consumers. This process improves the estimate of individual demands in two possible ways. First, in a market in which the noise terms are largely idiosyncratic, a large sample size helps to filter out the noise and identify the common fundamentals. Second, in a market with largely idiosyncratic fundamentals, many observations help to filter out demand shocks and identify the common noise term, therefore estimating individual fundamentals by differencing. As a result, data sharing can induce informational gains for third parties, such as the intermediary and the firm in the product market, as well as for individual consumers.

Data Market and Product Market Interactions in the product market are shaped by the data market in two ways. First, the consumers and the producer learn more about the realized demand. Second, the consumers and the producer respond to the additional information by changing both the demand (i.e., the choice of quantity by each consumer) and the supply (i.e., the choice of price by the producer). Thus, the linear pricing model specifies how demand information can influence market outcomes, quantities and prices. This specification of the interaction in the product market is of course only one of many models that can broadly reflect the representative interactions in the product market. In particular, the linear pricing model emphasizes that data markets can affect consumers and producers differentially and that these differences matter for the equilibrium price of data.

The Price of Individual Data Each consumer's choice to provide information is guided only by her private benefits and costs, i.e., the externality generated by the data that she provides does not enter her decision. 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. Consequently, the cost of acquiring individual data can be substantially less than the value of information to the platform.

Now we can see how social data drive a wedge between the efficient and profitable uses of information. Although many uses of consumer information exhibit positive externalities (e.g., real-time traffic information for driving directions), 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.

Data Sharing Policies We begin the analysis with complete data sharing. Initially, we restrict the data intermediary to offering only one information policy as an alternative

to no data sharing. This first step identifies the factors that support data sharing as an equilibrium outcome. It emphasizes a significant gap between the equilibrium level and the socially efficient level of data sharing.

We then remove the restriction of complete data sharing and allow the data intermediary to determine the revenue-maximizing information flow. We find that aggregation and anonymization will frequently be part of the optimal data sharing policy. This finding could contribute to a more efficient data flow, but the richer set of informational instruments also allows the intermediary to exploit the data externalities more powerfully.

In particular, we show that, when consumers are homogeneous ex ante, the intermediary prefers to collect aggregate, market-level information. 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, although from a technological standpoint, the data are freely available to the intermediary, the role of social data provides a more nuanced ability to determine the modality of information acquisition and use.[3]

More Data and More Services We then ask how the gap between the social value of the data and the price of the data behaves when the number of consumers or the number of services increases. In either of these instances, the size of the database is growing as the sources of data are multiplying, but 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 have large effects on the subjects' willingness to relinquish their private data. In practice, this force also likely drives the extraordinary appetite of Internet platforms to gather information.[4]

With a larger number of consumers in the market, we can extend the baseline model with ex-ante identical consumers to accommodate 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.

[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, it severely reduces the cost of procuring the data.

[4]The Furman report identified "the central importance of data as a driver of concentration and a barrier to competition in digital markets" (Digital Competition Expert Panel (2019)). The social dimension of data helps to explain these forces.

Different Uses of Information We eventually enrich the basic setting to consider additional features that are frequently part of data mediated transactions. We first consider a model in which the producer can choose product characteristics as well as prices and in which consumer preferences display a horizontal (taste) dimension, in addition to the vertical (willingness-to-pay) dimension. The resulting data policy then aggregates the vertical dimension but not the horizontal dimension, enabling the producer to offer personalized product recommendations but not personalized prices.

Second, we generalize our aggregation results by allowing for heterogeneous uses of information, some of which increase total surplus. We find that aggregate–market-level–information is collected if and only if it reduces total surplus. This means that even if data *transmission* is socially detrimental, as in the case of price discrimination downstream, the equilibrium level of data *aggregation* is socially efficient.

Related Literature Recently, there has been significant interest in data markets. 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) introduced information externalities into a model of monopoly pricing with unit demand. Each consumer is described by two *independent* random variables that are private information to the consumer: 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. The nuisance cost of each type is affected by the total number of consumers sharing their personal data. Finally, a proportion of consumers generate information externalities, and a proportion does not. By selling the service, the monopolist therefore offers a bundle–the service itself and the loss of privacy. The seller gains additional revenue from the collected data, depending on the proportion of units sold and the volume of data collected. The excessive loss of privacy is then established by comparing the optimal pricing policy of the monopolist with the social welfare maximizing policy. In contrast, we consider distinct data and product markets (and their interaction). Importantly, our data policy and data flow are determined explicitly as part of the equilibrium analysis, rather than represented by a reduced-form loss of privacy.

In contemporaneous work, Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019) also analyzed data acquisition in the presence of information externalities. As in Choi, Jeon, and Kim (2019), they considered a model with many consumers and a single data acquiring firm. In common with the current analysis, Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019) proposed an explicit statistical model for their data, allowing them to assess the loss

of privacy for the consumer and the gains in prediction for the data acquiring firm. Their analysis then pursued a different, and largely complementary, direction from ours. In particular, they analyzed how consumers with heterogeneous privacy concerns trade information with a data platform. They derived 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. By making the information flow endogenous to the equilibrium, we can also investigate when and how privacy can be partially or fully preserved, for example, by aggregation, anonymization, and noise.

An early and influential paper on consumer privacy and the market for customer information is Taylor (2004), who analyzed the sales of data in the absence of data externality.[5] More recently, Cummings, Ligett, Pai, and Roth (2016) investigated how privacy policies affect user and advertiser behavior in a simple economic 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 proposed sources of countervailing market power.

Fainmesser, Galeotti, and Momot (2020) provided a model of digital privacy in which the data collection improves the service provided to consumers. However, as the collected data can also be accessed by third parties imposing privacy costs, an optimal digital privacy policy must be established. Similarly, Jullien, Lefouili, and Riordan (2020) analyzed the equilibrium privacy policy of websites that monetizes information collected from users by charging third parties for targeted access. Gradwohl (2017) considered 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. In a model of personalized pricing, Ali, Lewis, and Vasserman (2019) analyzed how different choices regarding disclosure policies could affect the consumer surplus.

Liang and Madsen (2020) investigated how data policies can provide incentives in principal-agent relationships. They emphasized how the structure of individual data, as well as how the individual data form substitutes or complements the others and determines the impact of data on incentives. Ichihashi (2020) considered a single data intermediary and asked how the nature of the consumer data (under complete information for consumers), particularly complementary or substitutable data, affects the price of the individual data.

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

2 Model

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

2.1 Product Market

Consumers There are finitely many consumers, labelled $i = 1, ..., N$. 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 true willingness to pay for the product:

$$w_i \triangleq \theta + \theta_i. \tag{2}$$

The willingness to pay $w_i \in W = \mathbb{R}$ of consumer i is the sum of two components: one that is *common* to all consumers in the market, denoted by θ; and one that is *idiosyncratic* to consumer i, denoted by θ_i. Throughout, we assume that all of the random variables are normally distributed and described by a mean vector and a variance-covariance matrix,

$$\begin{pmatrix} \theta \\ \theta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_{\theta_i} \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & 0 \\ 0 & \sigma_{\theta_i}^2 \end{pmatrix} \right), \tag{3}$$

with a positive mean vector $\mu_\theta, \mu_{\theta_i} \geq 0$.

Producer The producer can choose 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

$$\pi(p_i, q_i) \triangleq \sum_i (p_i - c) q_i. \tag{4}$$

The producer knows the structure of demand and thus the common prior distribution given by (3). However, absent any additional information, the producer does not know the realized

demand shocks prior to setting prices.

2.2 Data Environment

Initially, each consumer i may only have imperfect information about her willingness to pay. The imperfect information is represented by a signal s_i that consumer i receives about her willingness to pay:

$$s_i \triangleq \theta + \theta_i + \varepsilon + \varepsilon_i. \tag{5}$$

The additive noise terms ε and ε_i refer to a common and an idiosyncratic error generated, for example, by partially overlapping sources of information about the producer's goods. Similar to the payoff components θ and θ_i above, the error terms are normally distributed,

$$\begin{pmatrix} \varepsilon \\ \varepsilon_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_{\varepsilon_i}^2 \end{pmatrix} \right). \tag{6}$$

The data environment has three 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, $\sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2 > 0$, each consumer can learn more about her own demand from the signals of the other consumers. Third, even without any noise in the signals, $\sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2 = 0$, each consumer can learn more about the composition of her own demand, i.e., the separate components θ and θ_i, from the signals of the other consumers.

The statistical model given by (3) and (6) specifies separately the idiosyncratic and common factors in the fundamental and the noise shocks. The correlation coefficients across any two consumers for payoff shocks and noise shocks are given by the respective ratios

$$\alpha \triangleq \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\theta_i}^2}, \qquad \beta \triangleq \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2}. \tag{7}$$

We refer to the pair $(\alpha, \beta) \in [0,1]^2$ as the *data structure*. The correlation coefficients summarize the commonality in the data across any pair of consumers, for fundamentals and noise separately.[6] The data structure (α, β) thus expresses in statistical terms the social dimension of the individual data.

[6] A statistically equivalent representation of the data environment could define the total payoff shock w_i and the total noise shock e_i and the correlation coefficients α, β. The underlying factors, namely (θ, θ_i) and $(\varepsilon, \varepsilon_i)$, could then be inferred from (w_i, e_i) and (α, β).

2.3 Data Market

The data market is run by a single data intermediary (it). The data intermediary can acquire demand information from the individual consumers, package it, and transmit it to the producer. We consider bilateral contracts between the individual consumers and the data intermediary, as well as between the producer and the data 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 policy determines the *inflow* of information and the *outflow* of information from the data intermediary. The data price determines the fee for the information flow. The inflow data policy describes how each signal s_i enters the database of the intermediary:

$$X_i : \mathbb{R} \to \Delta(\mathbb{R}), \forall i. \tag{8}$$

The outflow data policy determines how the processed signals are transmitted to the producer and the consumers in the product market:

$$Y : \mathbb{R}^N \to \Delta\left(\mathbb{R}^N \times \mathbb{R}^N\right). \tag{9}$$

The data contract with consumer i specifies an inflow data policy X_i and a fee $m_i \in \mathbb{R}$ paid to the consumer. Similarly, the data contract with the producer specifies an outflow data policy Y and a fee $m_0 \in \mathbb{R}$ paid by the producer. We do not restrict the sign of the payments a priori. The data intermediary maximizes the net revenue

$$R \triangleq m_0 - \sum\nolimits_{i=1}^{N} m_i. \tag{10}$$

As a monopolist market maker, the data intermediary decides how to collect the information from each consumer and how to transmit it to the other consumers and the producer. Thus, the data intermediary faces both an information design problem and a pricing problem. The data intermediary can transmit all information about each consumer i or some possibly noisy statistics of individual and market demand. Two critical dimensions of the data policy are *noise* and *aggregation*. A third critical dimension is the *sharing* of the data outflow—whether all consumers should also receive the data offered to the producer. In Section 3, we begin the equilibrium analysis under the assumption of complete data sharing. In Section 4, we allow the data intermediary to optimally determine all of the dimensions of the data policy.

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 policy (m_i, X_i) to each consumer i for data acquisition. Consumers simultaneously accept or reject the intermediary's offer.

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

3. The data s and the information flows (x, y) are realized and transmitted according to the terms of the bilateral contracts.

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. At the contracting stage, the information is imperfect but symmetric. A Perfect Bayesian Equilibrium is given by a tuple of inflow and outflow data policies, pricing policies for data and product, and participation decisions by producer and consumers

$$\left\{(X^*, Y^*, m^*)\,;\; p^*\left(X, Y\right);\; a^*\right\}, \tag{11}$$

where

$$a_0^* : X \times Y \times \mathbb{R} \to \{0, 1\}, \qquad a_i^* : X_i \times \mathbb{R} \to \{0, 1\}, \tag{12}$$

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 most of our analysis, we focus on the best equilibrium for the data intermediary and discuss unique implementation in Section 7.

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

2.5 Discussion of Model Features

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 any particular demand information s_i. The choice of *ex ante* participation constraints is meant to capture the prevailing condition under which the consumers and the producer accept the "terms of use agreement" or "terms of service" before any particular consumption or search

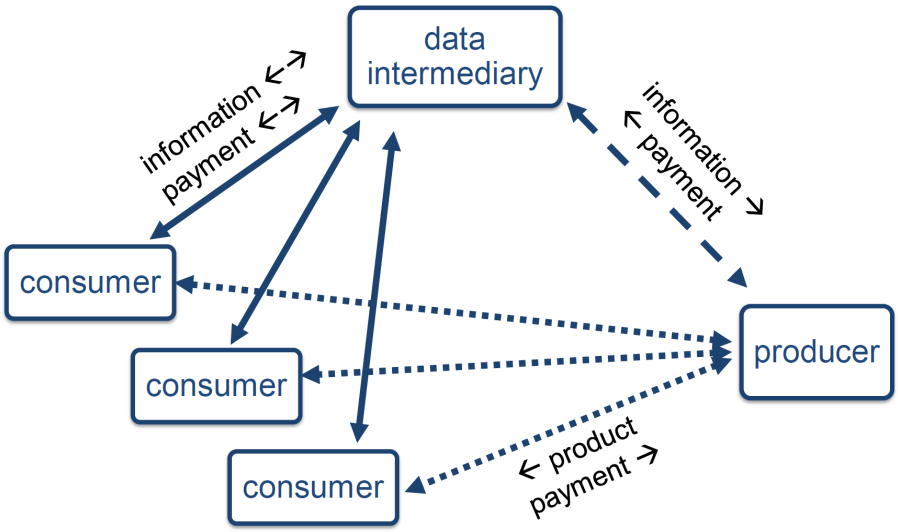


Figure 1: Data and Value Flows

event. For instance, when using Facebook, Amazon or other search engines, an account is established before making any specific choice. Through the lens of our model, the consumer evaluates the consequences of sharing her data flow ex ante. Hence, she requires a level of compensation that allows her to profitably share the information on average. Upon agreeing to share the information, there are no further incentive compatibility constraints.

Similarly, the intermediary's choice of data outflow policy occurs after consumers are enlisted but before the data are realized. This assumption is meant to capture 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.

The shared data allow each consumer to adjust demand (via quantity) and the producer to adjust supply (via price). This basic model thus contains the central elements of data mediated transactions. In this model, consumer, producer and social surplus can increase or decrease with data sharing. The rate and direction of change are largely determined by the data structure. In special case in which the shared data only improve the information of the producer and not the information of the consumer, the product market model (and its welfare implications) is closely related to that of third-degree price discrimination under linear pricing, as in Schmalensee (1981), with information sharing decreasing social welfare.

The gains from data sharing would arguably change by considering different pricing models, such as two-part tariffs or general nonlinear pricing. The instruments available to the producer would presumably 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.

Finally, in our model, each consumer is compensated with a monetary transfer for her individual data. In practice, many data intermediaries offer services rather than money in exchange for personal data. An augmented model in which services are rendered in exchange for data would add complexity to the interaction between consumer and data intermediary but would not affect the nature of the data externality, which is the focus of our analysis.

3 Complete Data Sharing

We begin the analysis of the data market by considering complete data sharing—a specific data policy whereby the intermediary collects all of the consumer information and then transmits the entirety of the data to the producer and to all consumers. We identify the costs and benefits of complete data sharing for all market participants, and exhibit conditions under which complete data sharing emerges in equilibrium. In Section 4, we allow the data intermediary to restrict the inflow and outflow of data by aggregating signals, withholding data, and introducing noise. While the optimal data policy will depart in significant ways from complete data sharing, the critical determinants of the cost and benefit of data sharing will share similar features.

3.1 Data Policy and Data Price

Under complete data sharing, 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. Furthermore, all the collected information is transmitted even to non-participating consumers, while the producer receives the data only if she participates.

Given this data policy, the optimal pricing policy for the producer consists of a vector of potentially distinct and personalized prices $p^*(s) \in \mathbb{R}^N$, resulting in a vector of individual quantities purchased $q_i^*(s)$. The net revenue of the producer is given by

$$\Pi(S) \triangleq \mathbb{E}\left[\sum\nolimits_{i=1}^{N} \pi\left(p_i^*(s), q_i^*(s)\right) | S\right].$$

If instead the producer rejects the data contract, he must offer a price using the prior information only.[7] We denote the monopoly price in the absence of any information (which is necessarily uniform across consumers) by $\bar{p}$ and the producer's net revenue by

$$\Pi(\varnothing) \triangleq \mathbb{E}\left[\sum\nolimits_{i=1}^{N} \pi\left(\bar{p}, q_i^*(s_i)\right)\right].$$

[7] If the producer rejects the intermediary's offered outflow, consumers do not receive any information either. The quantity purchased is then given by $q_i^*(s_i)$. This assumption has no consequences for payoffs.

The participation constraint for the producer to enter the data contract is thus

$$\Pi(S) - \Pi(\varnothing) \geq m_0, \tag{13}$$

i.e., the gains from data acquisition have to at least offset the price of the data.

Proceeding backward, each consumer receives an offer for a data contract (m_i, S_i). We denote the gross expected utility of consumer i from complete data sharing by

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

Holding fixed the decision of the remaining consumers, by rejecting her contract, consumer i would not transmit her data or receive compensation for her data. This reduces the outflow data policy to S_{-i} for all market participants. Thus, by withholding her information, consumer i does not change the amount of information available to her. The consumer's surplus from rejecting the intermediary's offer is given by

$$U_i(S, S_{-i}) \triangleq \mathbb{E}\left[u_i\left(w_i, q_i^*(s), p_i^*(s_{-i})\right)|S\right], \tag{14}$$

where the first argument (S) denotes the consumer's information, and the second argument (S_{-i}) refers to the producer's information.

Each pair (m_i, S_i) must satisfy the consumer's participation constraints. In particular, the data intermediary must set payments to consumers m_i that satisfy

$$m_i \geq U_i(S, S_{-i}) - U_i(S), \text{ for all } i. \tag{15}$$

The participation constraints (13) and (15) indicate how the price of the data is determined by their value in the product market.

3.2 Data Sharing and Product Markets

We now quantify the value of information for consumers and producers. For each consumer, the shared data help to improve her estimates of her willingness to pay. For the producer, the shared data enable a more informed pricing policy. The nature of the demand and of the uncertainty (i.e., the quadratic utility function and the multivariate normal distribution) yields explicit expressions for the value of information for all product market participants.

Under the complete data sharing policy, all of the consumers and the producer observe the vector of signals s. Let

$$\hat{w}_i(s) \triangleq \mathbb{E}[w_i \mid s] \tag{16}$$

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

$$q_i(s,p) = \hat{w}_i(s) - p. \tag{17}$$

Therefore, the producer charges consumer i the optimal personalized price $p_i^*(s)$, which is a linear function of the predicted type $\hat{w}_i$, i.e.,

$$p_i^*(s) = \frac{\hat{w}_i(s) + c}{2}. \tag{18}$$

From (1) and (4), the ex ante expectation of the consumer i's surplus and producer surplus in the interaction with consumer i can be written in terms of $\hat{w}_i(s)$ as

$$U_i(S) = \frac{1}{8}\mathbb{E}\left[(\hat{w}_i(s) - c)^2 \,|S\right], \tag{19}$$

$$\Pi_i(S) = \frac{1}{4}\mathbb{E}\left[(\hat{w}_i(s) - c)^2 \,|S\right]. \tag{20}$$

The social surplus generated in the interaction between consumer i and the producer is:

$$W_i(S) \triangleq U_i(S) + \Pi_i(S).$$

Since prices and quantities are linear functions of the posterior mean $\hat{w}_i$, the ex ante average prices and quantities are constant across all information policies. Consequently, all surplus levels under complete data sharing depend on the variance in the posterior mean only:

$$\text{var}\left[\hat{w}_i(s)\right] \triangleq \text{var}\left[\mathbb{E}\left[w_i \mid s\right]\right]. \tag{21}$$

For the producer, the consequences of complete data sharing relative to no information sharing are simple: without information, the producer charges a constant price based on the prior mean.[8] For an individual consumer i, the consequences of data sharing are more subtle since each consumer already has an initial signal s_i, according to which she can adjust her quantity. Her welfare without data sharing then depends on the precision her own signal, as measured by

$$\text{var}\left[\hat{w}_i(s_i)\right] \triangleq \text{var}\left[\mathbb{E}\left[w_i \mid s_i\right]\right]. \tag{22}$$

We can now express the value of complete data sharing for the consumers and the producer in terms of the variance in the posterior expectations.

[8] This price is obtained by setting $\hat{w}_i = \mu_\theta + \mu_{\theta_i}$ in (18).

Proposition 1 (Value of Complete Data Sharing)

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

$$\Pi_i(S) - \Pi_i(\varnothing) = \frac{1}{4}\operatorname{var}\left[\hat{w}_i(s)\right].$$

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

$$U_i(S) - U_i(\varnothing) = \frac{1}{8}\operatorname{var}\left[\hat{w}_i(s)\right] - \frac{1}{2}\operatorname{var}\left[\hat{w}_i(s_i)\right].$$

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

$$W_i(S) - W_i(\varnothing) = \frac{3}{8}\operatorname{var}\left[\hat{w}_i(s)\right] - \frac{1}{2}\operatorname{var}\left[\hat{w}_i(s_i)\right].$$

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, which is beneficial for both the consumers and the producer. Second, with access to the complete data, the producer pursues a *personalized* pricing policy toward each individual consumer. As the producer adapts his pricing policy to the estimate of each consumer's willingness to pay $\hat{w}_i$, some of the quantity responsiveness is dampened by the price responsiveness. While beneficial for the producer, this second channel *reduces* consumer surplus and also total welfare.

These channels have important welfare consequences in specific data environments. In particular, suppose that the consumer observes her own willingness to pay without noise or that she does not learn anything new from the data of the other consumers: data sharing simply enables price discrimination. In these special cases, data sharing has a positive impact on producer surplus but a negative impact on consumer and social surplus.[9]

3.3 Learning and Data Sharing

Proposition 1 gives us a way to describe the impact of data sharing on the creation and distribution of the surplus in the product market. In particular, the surplus levels of all

[9] This result is reminiscent of the welfare consequences of third-degree price discrimination in the linear-demand environment of Robinson (1933) and Schmalensee (1981). In these settings, each consumer knows her willingness to pay (i.e., noiseless signals, in our model's language). In the current 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 will not change (with all markets served), but social welfare is lower under finer market segmentation.

market participants can be expressed as linear combinations (with different weights) of the consumers' initial information, as measured by $\text{var}\left[\hat{w}_i\left(s_i\right)\right]$, and of the information about w_i generated by the entire social signal profile s, as measured by $\text{var}\left[\hat{w}_i\left(s\right)\right]$.

We now define the consumer's *information gain* from data sharing as

$$G \triangleq \text{var}\left[\hat{w}_i\left(s\right)\right] - \text{var}\left[\hat{w}_i\left(s_i\right)\right]. \tag{23}$$

The gain function G is a purely statistical notion that compares the variance of the posterior expectation before and after data sharing. However, we can already see from Proposition 1 that both consumer surplus and producer surplus are increasing in the information gain.

Thus, we seek to characterize the data environments that lead to smaller or larger information gains. This goal requires a more detailed analysis of the relevant statistical properties of the social data. To this end, we describe the *data structure* in terms of the correlation coefficients (α, β) introduced earlier in (7):

$$\alpha = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_{\theta_i}^2}, \qquad \beta = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2}.$$

At the same time, we hold the total variance in the fundamentals and the noise constant:

$$\sigma_w^2 \triangleq \sigma_\theta^2 + \sigma_{\theta_i}^2, \qquad \sigma_e^2 \triangleq \sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2. \tag{24}$$

For a given total variance in payoff shocks and noise shocks (σ_w^2, σ_e^2), the data structure (α, β) determines the composition of the shocks across the common and idiosyncratic components. We then define the information gain $G \triangleq G\left(\alpha, \beta\right)$ in terms of the data structure (α, β). We note from (23) that the first term $\text{var}\left[\hat{w}_i\left(s\right)\right]$ in the information gain depends on the data structure (α, β), while the second term $\text{var}\left[\hat{w}_i\left(s_i\right)\right]$ (i.e., the variance in each consumer's posterior expectation before data sharing) depends on the signal-to-noise ratio only:

$$\text{var}\left[\hat{w}_i\left(s_i\right)\right] = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_e^2}\sigma_w^2. \tag{25}$$

We now study the properties of the information gains in terms of the data structure (α, β).

Lemma 1 (Information Gains)

1. *The gain function $G\left(\alpha, \beta\right)$ satisfies $G\left(\alpha, \alpha\right) = 0$, and it is strictly increasing in $|\alpha - \beta|$ for all $\alpha, \beta \in [0, 1]$.*

2. *The gain function is maximized at $(1, 0)$ if $\sigma_w^2 > \sigma_e^2$ and at $(0, 1)$ if $\sigma_e^2 > \sigma_w^2$.*

The first part of Lemma 1 asserts that data sharing generates the most pronounced information gains when the structure of the fundamental shocks is as different as possible from the structure of the noise shocks. The intuition for this result is best grasped by considering the special data structures $(0,1)$ and $(1,0)$.

1. Suppose that payoff shocks are independent across the consumers, but noise is common to the consumers; thus $s_i = \theta_i + \varepsilon$, and the data structure is $(0,1)$. Each consumer can estimate the common noise accurately by averaging over N signals. She can then compute her true willingness to pay θ_i by differencing out the common noise term. However, if the noise terms were also independent, then the other $N-1$ signals would provide no additional information at all to consumer i and her estimation problem.

2. Conversely, suppose that there are perfectly correlated fundamentals and idiosyncratic noise; thus, $s_i = \theta + \varepsilon_i$, and the data structure $(1,0)$. Data sharing now generates additional information by filtering out idiosyncratic noise.

The second part of Lemma 1 indicates which of the special structures generates the largest information gains. In particular, the common element with the larger variance, whether it is the fundamental or noise term, can be estimated with more precision, and then the idiosyncratic element can be estimated by differencing out the common element. Thus, if the total variance of the fundamental is larger, more information can be gained by estimating the common element of the fundamental, and vice-versa, the total variance in the noise term is larger.[10] Figure 2 illustrates the level curves of the gain function when $\sigma_w = \sigma_e$, in which case G is a symmetric function of α and β.

We now relate this statistical measure of information to welfare gains and losses.

Proposition 2 (Data Structure and Welfare Gains)

1. *For every α, there exist $\underline{\beta}(\alpha) < \alpha < \overline{\beta}(\alpha)$ such that social welfare is increasing with complete data sharing if and only if $\beta \notin [\underline{\beta}(\alpha), \overline{\beta}(\alpha)]$.*

2. *For every β, there exist $\underline{\alpha}(\beta) < \beta < \overline{\alpha}(\beta)$ such that social welfare is increasing with data sharing if and only if either $\alpha < \underline{\alpha}(\beta)$ or $\alpha > \overline{\alpha}(\beta)$.*

3. *The thresholds $\underline{\beta}(\alpha)$ and $\overline{\beta}(\alpha)$ are increasing in α. The thresholds $\underline{\alpha}(\beta)$, $\overline{\alpha}(\beta)$ are increasing in β.*

[10] Indeed, when σ_e is large, for any fixed α, if the noise terms are perfectly correlated, each consumer learns her idiosyncratic type with high precision. (This occurs because the average of N signals identifies the common components $\theta + \varepsilon$.) Thus, N signals carry statistical information for any level of total noise, even as individual signals become uninformative. Conversely, with independent noise terms, as $\sigma_e \to \infty$, both the individual signals and the entire dataset lose all informativeness.

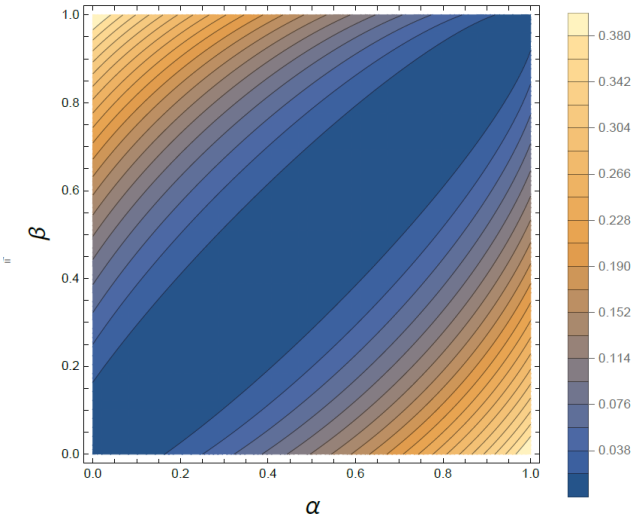


Figure 2: Information Gains $G(\alpha, \beta)$, $\sigma_w = \sigma_e = 1, N = 10$

The content of Proposition 2 mirrors the results from Lemma 1. Data sharing is socially valuable when α is sufficiently different from β. Propositions 1 and 2 show that the relationship between α and β is qualitatively symmetric, although not exactly in quantitative terms. We note that the threshold values might lie at the boundary of the unit interval; thus, only areas with welfare losses may exist.

Proposition 2 could be restated for consumer welfare rather than social welfare in similar terms. We omit a restatement here and simply note that the respective thresholds for consumer welfare would always be closer to the boundary, 0 or 1. This fact and the content of Proposition 2 are visually illustrated in Figure 3.

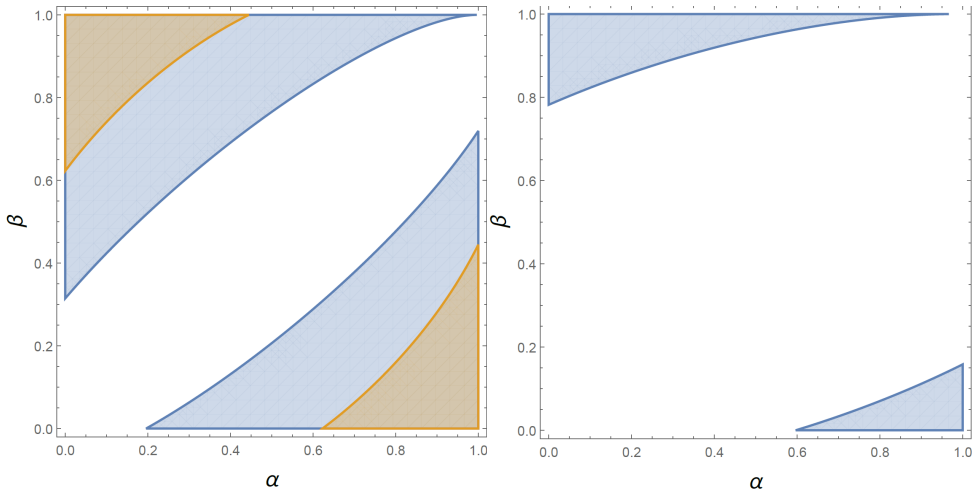


Figure 3: Social (left) and Consumer (right) Surplus Gains, $\sigma_e \in \{1, 10\}, \sigma_w = 1, N = 10$.

The blue area in each panel identifies the data structure (α, β) at which data sharing leads to an increase in social surplus (left) and consumer surplus (right), for the case of large initial noise levels ($\sigma_e = 10$). The ocher area in the left panel identifies the subset of data structures under which social surplus is increasing when consumers have precise estimates of their willingness to pay ($\sigma_e = 1$). In this case, data sharing reduces consumer surplus for any data structure—there is no ocher area on the right. In all cases, positive gains can only arise when the structures of the fundamentals and of the noise terms are sufficiently distinct, e.g., in the case of $(\alpha, \beta) = (1, 0)$ and $(\alpha, \beta) = (0, 1)$. Conversely, there are never any information gains along the diagonal. Because $G(\alpha, \beta) = 0$ when $\alpha = \beta$, consumers cannot learn anything from each other's signals, but the producer can. Thus, consumer and social welfare are diminished by data sharing.

We already know from Proposition 1 that the impact of data sharing on producer surplus is always positive. Based on Proposition 1, we can write the consumer and social surplus in terms of the gain function and the value of initial information:

$$U_i(S) - U_i(\varnothing) = \frac{1}{8} G(\alpha, \beta) - \frac{3}{8} \operatorname{var}\left[\hat{w}_i(s_i)\right],$$

and

$$W_i(S) - W_i(\varnothing) = \frac{3}{8} G(\alpha, \beta) - \frac{1}{8} \operatorname{var}\left[\hat{w}_i(s_i)\right].$$

This formulation clarifies that, for large levels of initial noise σ_e, consumers know so little about their preferences that they stand to gain from data sharing. This is only possible, however, if the data structure enables sufficiently large information gains. We now formalize how the total variance terms impact the equilibrium consumer and social welfare.

Proposition 3 (Welfare Impact of Data Sharing)

1. *For every α and β, there exist $\overline{\sigma}_e^2$ and $\overline{\overline{\sigma}}_e^2$ with $\overline{\sigma}_e^2 \leq \overline{\overline{\sigma}}_e^2 \leq \infty$ such that*

$$W(\varnothing) < W(S), \text{ if and only if } \sigma_e^2 > \overline{\sigma}_e^2, \tag{26}$$

and

$$U(\varnothing) < U(S), \text{ if and only if } \sigma_e^2 > \overline{\overline{\sigma}}_e^2. \tag{27}$$

2. *For every $\alpha < 1$, there exists β such that the thresholds $\overline{\sigma}_e^2$ and $\overline{\overline{\sigma}}_e^2$ are finite.*

Proposition 3 establishes that, if the total noise level is significant–and thus consumers stand to learn substantially from data sharing—then data sharing is going be socially beneficial. With even more noise in the individual signals, it will also be beneficial for the consumer

surplus. However, these thresholds are only finite for data structures such as those identified in Proposition 2, in which the structure α of the payoff shocks is sufficiently different from that of the noise shock β.

3.4 Equilibrium with Complete Data Sharing

We have learned that complete data sharing improves social (and consumer) welfare whenever: (a) the correlation structure of the fundamentals is sufficiently different from that of the noise shocks; and (b) the total level of noise is sufficiently large. We now investigate whether these social gains can be decentralized in a market with a data intermediary. Conversely, we are asking whether the data intermediary could 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 The payment m_i to consumers must satisfy their participation constraint (15), considering what would happen if consumer i held out her signal, *given that the other* $N-1$ *consumers share theirs.* We thus formally define a notion of *data externality.*

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

$$DE_i(S) \triangleq U_i(S, S_{-i}) - U_i(S_i, \varnothing). \tag{28}$$

In (28), we add the data S_{-i} to the information held by the producer and consumer i. Thus, the data externality for consumer i is given by the impact of the data S_{-i} provided by all other consumers. Without data externalities, the data S_{-i} would have no impact, and the difference between the two terms in (28) would be nil.

We can now represent consumer i's compensation as

$$\begin{aligned} m_i^* &= U_i(S, S_{-i}) - U_i(S, S) \\ &= -\underbrace{(U_i(S, S) - U_i(S_i, \varnothing))}_{\Delta U_i(S)} + \underbrace{U_i(S, S_{-i}) - U_i(S_i, \varnothing)}_{DE_i(S)}. \end{aligned} \tag{29}$$

The data compensation m_i can thus be written as the difference between two terms. The first term is the total change in consumer i's surplus, denoted by $\Delta U_i(S)$, associated with complete data sharing. We characterized these gains earlier in Propositions 1 and 3. The second term is the data externality $DE_i(S)$ imposed on i by consumers $j \neq i$ when they sell their data to the intermediary, who shares it with the producer and with all consumers.

Because this channel can potentially reduce consumer surplus, it creates the possibility of profitable intermediation of information, which hurts consumers overall.[11]

With the above notion of a data externality and the producer's binding participation, which pins down $m_0 = \Pi(S) - \Pi(\varnothing)$, we can write the intermediary's profit (10) as

$$R(S) = \Delta W_i(S) - \sum\nolimits_{i=1}^{N} DE_i(S), \tag{30}$$

where ΔW denotes the variation in total surplus resulting from data sharing. Thus, the intermediary's profits are given by the sum of two terms: the variation in total surplus associated with data sharing; and the data externalities across N agents. The expression in (30) clarifies the extent to which the intermediary's objective diverges from the social planner's. If consumers exert negative externalities on each other, intermediation can be profitable but welfare reducing. Conversely, if $DE_i(S) > 0$ then intermediation is profitable only if it is welfare enhancing.

The data externality can be written as

$$DE_i(S) = \frac{1}{2}\left(\text{var}\left[\hat{w}_i(s)\right] - \frac{3}{4}\text{var}\left[\hat{w}_i(s_{-i})\right] - \text{var}\left[\hat{w}_i(s_i)\right]\right). \tag{31}$$

The payments made to consumers are positive and can be written as

$$m_i^* = \frac{3}{8}\left(\text{var}\left[\hat{w}_i(s)\right] - \text{var}\left[\hat{w}_i(s_{-i})\right]\right),$$

while the payment charged to the producer m_0 comes immediately from Proposition 1,

$$m_0 = \frac{N}{4}\text{var}\left[\hat{w}_i(s)\right].$$

Finally, using Proposition 1, we can write the revenue as

$$R(S) = \frac{N}{8}\left(3\,\text{var}\left[\hat{w}_i(s_{-i})\right] - \text{var}\left[\hat{w}_i(s)\right]\right).$$

The intermediary's profits thus depend positively on $\text{var}\left[\hat{w}_i(s_{-i})\right]$, i.e., on the amount of information gained by the producer about consumer i's type on the basis of signals s_{-i} only. In Theorem 1, we characterize the data structures that allow for profitable intermediation with complete data sharing.

[11] In particular, the data externality $DE_i(S)$ is strictly negative if consumers observe their willingness to pay without noise (because in this case, $\text{var}\left[\hat{w}_i(s)\right] = \text{var}\left[\hat{w}_i(s_i)\right]$, and the only effect of sharing is to give better information to the producer).

Theorem 1 (Complete Data Sharing)
Suppose that the intermediary collects and transmits all of the consumers' signals.

1. *For every noise structure* $\beta \in [0,1]$*, there exists a threshold* $\alpha^*(\beta) > 0$ *such that the intermediary obtains positive profits if and only if* $\alpha > \alpha^*(\beta)$*.*

2. *The threshold* $\alpha^*(\beta)$ *is increasing in* β *when* $N \geq 6$*, and it has a unique fixed point* $\beta_0(N)$ *for any* N*.*

3. *When* $N \geq 6$*, the threshold* $\alpha^*(\beta)$ *is increasing in the ratio* σ_w/σ_e *for all* $\beta < \beta_0$ *and decreasing otherwise.*

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 that the underlying fundamentals w_i and w_{-i} are, the easier it is to learn from other consumers' signals. In particular, for sufficiently correlated fundamentals, intermediation is always profitable, regardless of the correlation structure in the noise, while for independent fundamentals, it is never profitable.

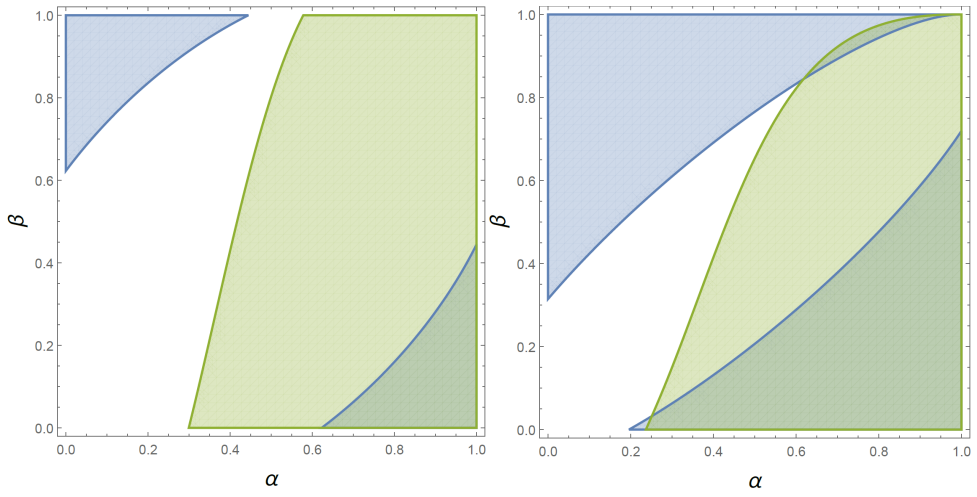


Figure 4: Profitable and Efficient Data Structures, $(\sigma_e \in \{1, 10\}, \sigma_w = 1, N = 10)$

In Figure 4 we illustrate the socially efficient data structures (blue) and profitable data structures (green) for $\sigma_e = 1$(left) and $\sigma_e = 10$(right). Proposition 1.2 shows that, when the error terms become more correlated (higher β), filtering individual willingness to pay on the basis of other consumer's signals becomes harder. Consequently, the required correlation of

fundamental terms $\alpha^*(\beta)$ increases. Proposition 1.3 shows that this problem is exacerbated when the magnitude of the error terms' variance increases, for a fixed (high) β.

Finally, we compare profitable data structures and socially efficient intermediation. When σ_e is low, as in the left panel, the only structures that yield profitable and efficient data sharing are those with high α and low β: consumers benefit overall because data sharing filters out the idiosyncratic noise and allows them to estimate their (common) willingness to pay. When σ_e is large, as in the right panel, profitable and efficient data sharing can also happen due to the filtering of the common noise, with α relatively smaller than β. We formalize this result in Corollary 1.

Corollary 1 (Efficient and Equilibrium Data Sharing)
For a given variance σ_w^2, there exists a threshold $\overline{\sigma}_e^2 > 0$ such that, for all $\sigma_e^2 < \overline{\sigma}_e^2$, efficient data sharing arises in equilibrium only if $\alpha > \beta$.

However, for any σ_e, two types of inefficiency remain possible. First, socially inefficient intermediation can be profitable, most notably when α is sufficiently large, and β is close to α. In this case, complete data sharing can lead to massive consumer and social welfare losses because the consumers learn close to nothing from each other's signals. Nonetheless, a large and negative data externality indicates that the intermediary must only compensate consumers for a small part of these losses. Second, efficient intermediation might be unprofitable: when fundamentals are largely independent, and errors are correlated (low α, high β), the presence of a positive data externality renders acquiring the consumers' signals too expensive, despite the large welfare gains associated with filtering out the common noise term across consumers.

Comparing Proposition 2 to Theorem 1 and Corollary 1, we note that the socially efficient data sharing is governed by different factors than the equilibrium data sharing. The former suggests that filtering common noise and common fundamentals are both statistically valuable operations: for equilibrium data sharing to be socially efficient, it is sufficient that any common component be precisely estimated. For the latter, the common component must be the fundamental element since it reduces the cost of information acquisition by making the data externality more severe.

When we next consider the optimal rather than the complete data sharing, we find that the bias toward inefficient data sharing driven by the common fundamental becomes even more pronounced.

3.5 Data Complements and Data Substitutes

The statistical properties of the consumers' signals can further improve our understanding of the data sharing equilibrium. In particular, we define consumers' signals as complements if the increment in the variance in the posterior expectation due to the individual signal s_i is larger in the presence of the other signals s_{-i} than in their absence. Thus, signals are complements if

$$\text{var}\left[\hat{w}_i\left(s\right)\right] - \text{var}\left[\hat{w}_i\left(s_{-i}\right)\right] > \text{var}\left[\hat{w}_i\left(s_i\right)\right] - \text{var}\left[\hat{w}_i\left(\varnothing\right)\right], \tag{32}$$

and they are substitutes otherwise. Equivalently, the data are complements if, for agent i, the gains in information from the other signals s_{-i} are larger in the presence of her signal s_i than in the absence of her signal:

$$\text{var}\left[\hat{w}_i\left(s\right)\right] - \text{var}\left[\hat{w}_i\left(s_i\right)\right] > \text{var}\left[\hat{w}_i\left(s_{-i}\right)\right] - \text{var}\left[\hat{w}_i\left(\varnothing\right)\right]. \tag{33}$$

The left-hand side of (33) is simply G. As is intuitive, the right-hand side is increasing in α—the more correlated the fundamentals, the more there is to learn from signals s_{-i} only. We now relate the data structure to informational interaction of the signals.

Proposition 4 (Data Complements vs Data Substitutes)
For any $\beta > 0$, there exists $\tilde{\alpha} \in [0, \beta]$ such that signals are complements if $\alpha \leq \tilde{\alpha}\left(\beta\right)$, and substitutes otherwise.

In particular, signals are always complements when the payoff shocks are completely idiosyncratic ($\alpha = 0$). Conversely, for any noise structure β, as the correlation in the payoff shocks across agents increases (namely, for all $\alpha \geq \beta$), the signals become substitutes.

The intuition for this result is most easily grasped by considering the special case of $s_i = \theta_i + \varepsilon$ (i.e., $\alpha = 0, \beta = 1$). In this case, observing the individual signal s_i provides partial information about w_i, but observing a large sample of signals s_{-i} allows one to estimate the common noise ε precisely; hence, adding signal s_i perfectly reveals θ_i: the signals are complements. Conversely, if $s = \theta + \varepsilon_i$, each signal increases the sample size toward estimating the common mean. With Gaussian information, this process has diminishing returns: signals are substitutes.

The profitability of intermediation is intuitively related to the substitutability between signals. With complement signals, the consumer knows they are giving up more information by participating as a consequence of the presence of others, relative to the case of a single agent. Because the intermediary must pay the marginal loss suffered by i as a consequence

of her revealing her signal, complement signals depress the profitability of intermediation.[12]

More generally, when signals are complements, the intermediary's profits are smaller than the effect of data sharing on social welfare. Indeed, comparing the left-hand side of (32) and the right-hand side of (31), we observe that the data externality is necessarily positive whenever signals are complemented (but not the converse). Thus, profitable intermediation of complement signals is possible, but only if it increases social surplus. This result is consistent with expression (30) for the intermediary's profits: when the data externality is positive, the intermediary must derive its revenue from an increase in total welfare.

Viewed in this light, these results reconcile the findings of our paper with those of Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019), who assumed *submodularity* of information and found socially excessive data intermediation. In our model, socially insufficient intermediation can also occur. In particular, the intermediary may not be able to generate positive profits from (socially efficient) information with complement signals.

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 (inflow, outflow) policy along three key dimensions. First, we allow the intermediary *not* to release all of the data that it has collected, i.e., to introduce incomplete and 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 information about each individual consumer i or some aggregate information about market demand. Third, we allow the intermediary to introduce further (possibly correlated) noise terms in any (individual or aggregate) signals that it collects.

More specifically, if the intermediary chooses to collect individual information, a data inflow policy consists of signals

$$x_i \triangleq s_i + \xi + \xi_i, \tag{34}$$

for each consumer $i = 1, \ldots, N$ who accepts the intermediary's offer. The choice of the variance terms σ_ξ^2 and $\sigma_{\xi_i}^2$ is an instrument available to the data intermediary to guarantee some degree of privacy to consumer i. Conversely, if the intermediary chooses to collect

[12] These properties are also present in Liang and Madsen (2020), who related the statistical nature of the signals to the sign of the data externality and to the strategic nature of the agents' participation decisions. In both their model and ours, negative data externalities render participation decisions strategic complements (and vice versa). However, in our game, information sharing has a direct effect on total surplus, indicating that even substitute signals (those that filter out idiosyncratic noise) can yield positive externalities.

aggregate information, a data inflow policy consists of a single signal

$$\bar{x} \triangleq \frac{1}{N} \sum\nolimits_{j=1}^{N} \left(s_j + \xi + \xi_j\right) a_j, \tag{35}$$

where $a_j \in \{0, 1\}$ represents the participation decision of consumer j. In both cases, we assume the consumer observes the realization of the noisy signal x_i that the intermediary collects about her. However, if consumer j does not participate ($a_j = 0$), she does not share her data with the data intermediary, and no signal is collected about her willingness to pay.

We now analyze the intermediary's optimal data policy along all dimensions, beginning with the choice of data outflow for any realized data inflow.

4.1 Data Outflow

Recall that, given the realized data inflow, the intermediary offers a data outflow policy to the producer specifying both the fee m_0 and the flow of information to market participants, including the consumers. The data outflow policy thus determines how well-informed the producer's customers are. A critical driver of the consumer's decision to share the data is then 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 it then extracts through the fee m_0.

Consistent with the previous section, we define a complete data outflow policy as one by which all collected signals are reported to the producer and to all consumers, including those who do not accept the intermediary's offer. We then establish that the complete data outflow maximizes the producer's gross surplus and hence the intermediary's profits.

Proposition 5 (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 3, 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. Second, the producer does not benefit from holding superior information relative to the consumers. If he did, the prices charged would convey information to the consumers. In any equilibrium, this reduces the producer's profits because of the ensuing costly signaling.

Therefore, in every subgame following the consumers' participation decisions in the data inflow policy, all of the consumers and the producer receive the same information from

the intermediary. For any data inflow policy X, the compensation owed to consumer i in equilibrium can then be written as in (29)

$$m_i^* = -\underbrace{\left(U_i\left(X, X\right) - U_i\left(X_i, \varnothing\right)\right)}_{\Delta U_i(X)} + \underbrace{\left(U_i\left(X, X_{-i}\right) - U_i\left(X_i, \varnothing\right)\right)}_{DE_i(X)},$$

where the data externality $DE_i\left(X\right)$ is defined as

$$DE_i\left(X\right) \triangleq U_i\left(X, X_{-i}\right) - U_i\left(X_i, \varnothing\right). \tag{36}$$

Finally, the intermediary's profits can be written as in (30):

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

4.2 Data Aggregation

We now explore the intermediary's decision to aggregate the individual consumers' demand data. We focus on two maximally different policies along this dimension. At one extreme, the intermediary can collect and transmit individual demand data, thereby enabling the producer to charge personalized prices, as in the previous section. At the other extreme, the intermediary can collect market demand data only. Under aggregate information intermediation, the producer charges the same price to all consumers j who participate in the intermediary's data policy. In other words, the observation of aggregate demand data still allows the producer to perform third-degree price discrimination across realizations of the total market demand but limits his ability to extract surplus from the individual consumers.[13]

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

Theorem 2 (Optimality of Data Aggregation)
For any noise level in the consumers' signals $(\sigma_\xi^2, \sigma_{\xi_i}^2)$*, the intermediary obtains strictly greater profits by collecting aggregate information.*

Within the confines of our policies, but independent of the initial and additional noise levels, 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

[13] Equivalently, the producer could have access to individual data, including identifying information about the consumer, but could not identify consumers when he offers his product.

the realized information about market demand. In other words, the presence of a monopolist intermediary might induce socially inefficient information transmission, but the equilibrium contractual outcome will preserve the personal identity of the consumer.

This finding suggests why we might see personalized prices in fewer settings than initially anticipated. For example, the retail platform Amazon and the transportation platform Uber 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 the acquisition of sensitive consumer information.

To gain intuition for why the intermediary resolves this trade-off in favor of aggregation, we turn to the data externality. According to Proposition 5, each consumer i knows that she will receive the same data outflow as all of other consumers $-i$ regardless of her participation decision. Therefore, her decision only impacts the information available to the *producer* about her (and others') willingness to pay.

Suppose that the intermediary acquires individual data; i.e., it collects signals

$$x_i = s_i + \xi + \xi_i \tag{37}$$

with some fixed noise levels $(\sigma_\xi^2, \sigma_{\xi_i}^2)$. Now consider the data externality (36) imposed on consumer i by all other consumers $-i$: if no consumer reports her signal, the producer sets a single price using the prior distribution only. If consumers $-i$ reveal their signals, and consumer i does not, the producer charges price $p_i^*(X_{-i})$ to consumer i.

When the producer observes the $N-1$ signals x_{-i} only, he is restricted in his ability to offer a personalized price to consumer i. Since the demand shock of each consumer has an idiosyncratic and a common component, the producer can use the aggregate demand data from *within* his entire sample to estimate the demand of any specific consumer *out of* sample. The optimal price is a convex combination of the prior mean and the *average* of all signals x_{-i} with some weight $\lambda \in (0,1)$:

$$p_i^*(X_{-i}) = \frac{\mathbb{E}[w_i \mid X_{-i}] + c}{2} = \frac{1}{2}\left(c + \lambda \frac{\Sigma_{j \neq i} x_j}{N-1} + (1-\lambda)\mu\right). \tag{38}$$

The extent to which the mean of signals x_{-i} is informative about w_i then depends on the variances of the common shock σ_θ^2 and the idiosyncratic shock $\sigma_{\theta_i}^2$, i.e., on the degree of correlation of the consumers' types.

Now, contrast this case with the collection of aggregate data

$$\bar{x} = \frac{1}{N} \sum\nolimits_{j=1}^{N} \left(s_j + \xi + \xi_j \right).$$

Along the path of play, the producer charges one price to all consumers. If a consumer rejects its offer, however, the producer charges two prices: a single price for the $N-1$ participating consumers about whom the producer has aggregate information; and another price to the deviating "anonymous" consumer. The latter price is again based on market data and is, in fact, equal to $p_i^*(X_{-i})$ in (38). In other words, with individual signals, the producer optimally aggregates the available data to form the best predictor of the missing data point. Therefore, aggregate data has no impact on the price faced by consumer i off the equilibrium path, hence generating the same level of the data externality as individual data.[14]

At this point, it is clear that the intermediary profits from aggregating the individual information of all consumers for any given noise level. Aggregation leaves the DE_i terms in (30) unchanged but 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 because the mean of the signals $\bar{x}_{-i}$ is a sufficient statistic for the problem of each consumer i. Thus, signal aggregation mitigates the information transmitted to the producer while holding fixed the consumer's information. This shift increases total surplus and hence also the intermediary's profits, relative to individual data sharing.

4.3 Information Design

The data intermediary can jointly choose the level of aggregation and noise in the data inflow by choosing the variance levels of the additional noise terms, σ_ξ^2 and $\sigma_{\xi_i}^2$. We begin with the information structure that maximizes the intermediary's profits. In the next section, we then consider how this policy changes with the number of consumers and how it adapts to a richer demand specification with consumer heterogeneity.

Theorem 3 (Optimal Data Intermediation)
The optimal data policy includes data aggregation and

1. *no additional idiosyncratic noise, $\sigma_{\xi_i}^* = 0$; or*
2. *(weakly) positive additional aggregate noise $\sigma_\xi^* \geq 0$.*

[14] The result in Proposition 2 would not change we forced the producer to charge a single price to all consumers on and off the equilibrium. With this interpretation, however, we capture the idea that the producer offers one price "on the platform," to the participating consumers while interacting with the deviating consumer "offline." He then leverages the available market data to tailor the offline price.

The optimal level of common noise σ_ξ^* is strictly positive when the correlation coefficient of the consumers' willingness to pay $\alpha \triangleq \sigma_\theta^2/\left(\sigma_\theta^2+\sigma_{\theta_i}^2\right)$, or the number of consumers N are 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 consumer types are not sufficiently correlated, the intermediary can supplement the degree of coordination with additional common noise σ_ξ^*, for any structure of the initial noise terms β. However, as we establish in Proposition 6, no profitable intermediation is feasible for values of α less than a threshold that decreases with N. Finally, as α approaches this threshold, the optimal level of common noise grows without bound. Figure 5 shows the optimal variance in the additional common noise term.

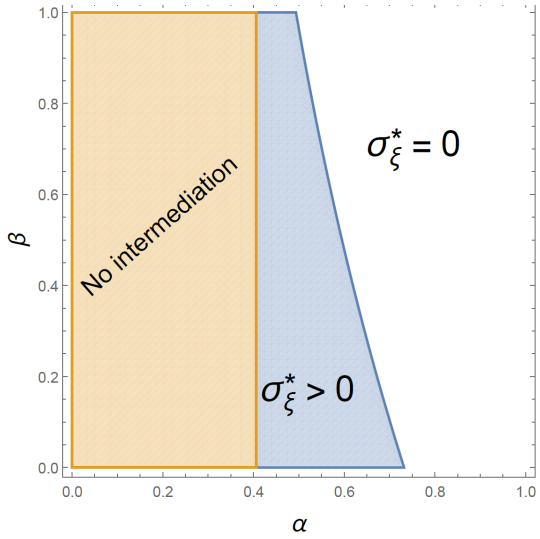


Figure 5: Optimal Additional Noise ($\sigma_w = \sigma_e = 1, N = 2$)

Proposition 6 (Profitability of Data Intermediation)
Under the optimal data policy, the intermediary's profits are strictly positive if and only if

$$\alpha > \frac{N(\sqrt{3}+1)-1}{2N\,(N+1)-1} \in (0,1)\,. \tag{39}$$

By introducing noise, the intermediary 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 averages the consumers' signals prior to re-selling them to the producer, it might appear surprising that the correlation structure in the additional noise terms plays such a critical role.

To gain intuition into the role of correlated noise, it might help to write the "regression coefficient" used by the producer to estimate the market demand (i.e., the average willingness to pay $\bar{w}$) from the aggregate demand signal $\bar{x}$ as in (35):

$$\frac{\text{cov}\left[\bar{w},\bar{x}\right]}{\text{var}\left[\bar{x}\right]}=\frac{\sigma_{\theta}^{2}+\sigma_{\theta_i}^{2}/N}{\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2}+\sigma_{\xi}^{2}+\left(\sigma_{\theta_i}^{2}+\sigma_{\varepsilon_i}^{2}+\sigma_{\xi_i}^{2}\right)/N}. \tag{40}$$

Multiple combinations of σ_{ξ}^{2} and $\sigma_{\xi_i}^{2}$ yield identical regression coefficients. The reason for choosing the degenerate combination $(\sigma_{\xi}^{2},\sigma_{\xi_i}^{2})=\left(\sigma_{\xi}^{*},0\right)$ is again the data externality. In particular, consider what happens to the producer's inference problem when consumer i does not sell her data to the intermediary. The producer wishes to estimate w_i from the aggregate signal $\bar{x}_{-i}$. In this new regression, the signal $\bar{x}_{-i}$ receives the following weight:

$$\frac{\text{cov}\left[w_i,\bar{x}_{-i}\right]}{\text{var}\left[\bar{x}_{-i}\right]}=\frac{\sigma_{\theta}^{2}}{\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2}+\sigma_{\xi}^{2}+\left(\sigma_{\theta_i}^{2}+\sigma_{\varepsilon_i}^{2}+\sigma_{\xi_i}^{2}\right)/\left(N-1\right)}. \tag{41}$$

Comparing (40) and (41), we observe that the common noise term enters identically into both, while the idiosyncratic noise $\sigma_{\xi_i}^{2}$ increases the variance in $\bar{x}_{-i}$ relatively more than the variance in $\bar{x}$. This finding means that the aggregate demand signal in the absence of consumer i's participation is a relatively worse predictor of w_i (and of $\bar{w}$) when the intermediary uses idiosyncratic noise, rather than correlated noise. Therefore, by loading all the noise on the common term ξ, the intermediary can hold constant the information content of the signal sold to the producer and reduce the cost of acquiring the information.

Intuitively, each consumer is performing two tasks under idiosyncratic noise: contributing to the estimation of market demand and reducing the noise of the aggregate signal (by averaging out the noise terms ξ_i over a larger sample size). The latter effect is absent when only common noise is used. Common noise renders consumer i less valuable to the producer and reduces her compensation.

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—interact with one another in a rich fashion. In particular, the value of common noise is deeply linked to that of aggregate data: in a constrained problem in which the intermediary must offer individual data intermediation, adding idiosyncratic noise becomes optimal when the consumers' initial signals s_i are sufficiently precise.[15]

Intuitively, if the intermediary is restricted to individual data inflow policies, it enables personalized price discrimination. We can then informally describe the producer's problem

[15] See Proposition 7 in Bergemann, Bonatti, and Gan (2020).

as consisting of two tasks: estimating the common component of market demand θ from the average signal $\bar{x}$; and then estimating the idiosyncratic taste shocks θ_i from the difference between average and individual signals. However, the difference between $\bar{x}$ in (35) and x_i in (37) can be written as

$$\bar{x} - x_i = \frac{1}{N} \sum\nolimits_{j=1}^{N} \left(\theta_j + \varepsilon_j + \xi_j\right) - \theta_i - \varepsilon_i - \xi_i.$$

This difference is an unbiased signal of $\bar{w} - w_i$, where the common noise terms ε and ξ drop out. Therefore, common noise without data aggregation reduces the value of information to the producer without protecting individuals from harmful price discrimination.

Having clarified why aggregation and common noise are integral parts of the optimal intermediation policy, we now turn to the implications of our results for large databases.

5 Value of Large Databases

In this section, we leverage our model to explain the "thirst" for data shown by large platforms. In particular, we examine the roles of data precision, of the number of consumers, and of consumer heterogeneity. We begin with the comparative statics of the optimal intermediation policy with respect to the precision of the consumers' signals and the number of consumers. We then turn to the intermediary's choice to enable finer segmentation of the product market through information provision.

5.1 Data Precision

A defining feature of data markets is the rapid increase in data sources and data services. For example, Facebook Connect allows Facebook to track consumers across additional services, such as Instagram, Snapchat, Facebook Open Graph, and Facebook Groups. Similarly, Google offers a number of services, such as Gmail, Google Maps, and YouTube. These services extend the number of sources of information about each consumer.

Our aim in this section is to capture the precision of information associated with multiple sources in a parsimonious manner. We therefore consider the comparative statics of the optimal data policy with respect to the total noise variance σ_e^2, holding the degree of correlation in the noise terms (β) constant.[16]

[16] This formulation assumes that each consumer learns more about her own preferences by using multiple services and is able to convey this information to the intermediary. We could alternatively fix the consumer's initial information and introduce noise in the communication channels, with similar results.

Greater signal precision has a direct effect on the data policy since it increases the value of information for the producer. In addition, to the extent that the intermediary can acquire more information from all consumers, it might be able to decrease the total compensation for a given data policy.

Proposition 7 (Precision of Information Collection)

1. *There exists a threshold* $\beta^*(\alpha, N) \in (0,1)$ *such that* σ_ξ^* *is increasing in* σ_e^2 *if* $\beta < \beta^*$*, and decreasing in* σ_e^2 *otherwise.*

2. *The threshold* $\beta^*(\alpha, N)$ *is strictly decreasing in both arguments.*

3. *The intermediary's profit is decreasing and convex in* σ_e^2 *whenever* $\sigma_\xi^* > 0$*.*

This result captures the idea that the intermediary supplements the amount of correlation in the consumers' signals: the optimal amount of additional common noise σ_ξ^* increases with the errors in the initial signals when these errors are independent, and it decreases when they are correlated. In particular, greater signal precision (lower σ_e^2) strengthens the correlation in the signals s_i (because the fundamentals w_i are themselves correlated) when the noise terms ε_i are independent. The intermediary can then reduce the amount of common noise. Conversely, when the noise terms are strongly correlated, any increase in precision can reduce the correlation in the initial signals. The intermediary then supplements these signals with an additional noise term σ_ξ^*.

Finally, greater signal precision improves the data intermediary's profits. This can occur through two channels. First, the potential information (and total surplus) gains increase. Second, information precision worsens the consumer's bargaining position due to the data externality. In particular, for sufficiently low σ_e^2, all of the signal structures exhibit substitutes, and the data externality is unambiguously negative. This effect *reduces* the compensation required by the individual consumer and explains the increasing returns (i.e., convex profits) that the intermediary obtains when data precision improves.

5.2 Value of Social Data

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 large number of (potential) participants, data sources, and services. We now pursue the implications of having a large number of participants and data sources for the social efficiency of data markets and the price of data.

We first consider what happens when the number of consumers becomes large. 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, in which the potential revenue increases without bound, while individual consumers make a small marginal contribution to the precision of aggregate data.

Theorem 4 (Large Markets)

1. *As $N \to \infty$, the individual consumer's compensation goes to zero, and the total compensation converges to*
$$\lim_{N\to\infty} N m_{i^*} = \frac{3}{4} \frac{\sigma_w^2 \alpha (1-\alpha)}{\alpha + \beta \sigma_e^2/\sigma_w^2}.$$
2. *For sufficiently large α, the total compensation is asymptotically decreasing in N.*
3. *As $N \to \infty$, 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 only paid for her marginal contribution, which explains how the total payments $\sum_{i=1}^{N} m_i$ 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 6 illustrates such an instance, in which it can be less expensive for the intermediary to acquire a larger dataset than a small one.

Finally, 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 (aggregate) data are freely available.

In practice, the results in Theorem 4 can shed light on the *digital privacy paradox* of Athey, Catalini, and Tucker (2017). Specifically, when aggregate information is collected, the incremental contribution of each individual consumer to the estimation of the average willingness to pay is close to nil. Therefore, independent of the final use of information, each consumer is willing to accept a negligible compensation (even on aggregate) to give up her private information. Far from being a paradox, this type of behavior reflects the market value of their information, which depends both on the correlation structure of their willingness to pay and on the intermediary's equilibrium data policy.

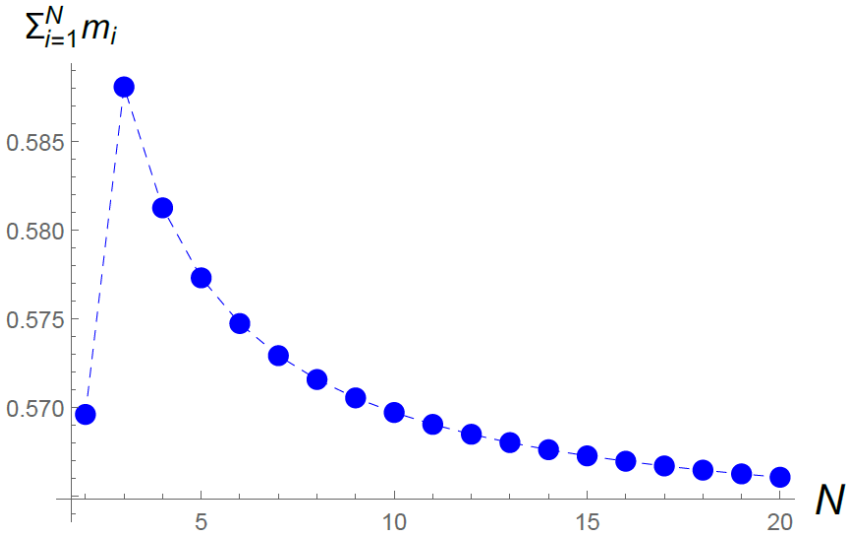


Figure 6: Total Consumer Compensation ($\sigma_w = 1, \sigma_e = 0$)

5.3 Market Segmentation and Data

We have so far used the assumption of ex ante homogeneity among consumers to produce some of the central implications of social data. A more complete description of consumer demand should consider additional characteristics that introduce heterogeneity across certain groups of consumers. These characteristics might include location, demographics, income, and wealth, among others.

We now explore how these additional characteristics influence the information policy and the profits of the data intermediary. To this end, we augment the description of consumer demand by splitting the population into subsets according to the common component of their willingness to pay,

$$w_{ij} = \theta_j + \theta_{ij}, i = 1, 2, ..., N_j, \ \ j = 1, ..., J. \tag{42}$$

Thus, each member of group j has the same common component. The J groups are identical ex ante: the common components θ_j are drawn from a normal distribution with mean μ and variance σ_θ^2. The idiosyncratic component θ_{ij} is normally distributed with zero mean and variance $\sigma_{\theta_i}^2$, and all of the random variables are independent. For the remainder of the analysis, we consider two consumer groups of equal size, i.e., $J = 2$ and $N_1 = N_2 = N$.

The intermediary's data policy space is now potentially richer. In particular, the intermediary must choose how to aggregate the consumers' signals both across groups and within each group. However, an identical argument to Theorem 2 establishes that it is always optimal to aggregate all signals within each homogeneous group, i.e., to sell at most two distinct

signals ($\bar{s}_1$ and $\bar{s}_2$) to the producer.

Corollary 2 (No Discrimination within Groups)
The optimal data policy aggregates all signals within each group $j = 1, 2$.

We then ask whether and under what conditions 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 send the simple average of all signals across groups to the producer, thus forcing the producer to offer a single price only. Alternatively, the intermediary could allow the producer to discriminate between two groups of consumers by transmitting the group-level means $\bar{s}_j$. 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 8 sheds light on the optimal resolution of this trade-off. In this result, we restrict attention to the case of noiseless signals ($\sigma_e^2 = 0$) and noiseless intermediation $\left(\sigma_\xi^2 = 0\right)$. We expect these results to extend to a noisy signal environment.

Proposition 8 (Segmentation)
Let $\sigma_e^2 = \sigma_\xi^2 = 0$.

1. *There exists $\overline{N}$ such that the data intermediary induces group pricing for all $N > \overline{N}$.*
2. *When $N \geq 3$, the threshold $\overline{N}$ is decreasing in the within-group correlation α.*

While the earlier Theorem 2 stated that the intermediary will not reveal any information about consumer identity, the present result shows that, if the market is sufficiently large, then the intermediary will convey limited identity information, i.e., each group's identity and average willingness to pay. This policy allows the producer to price discriminate across, but not within, groups. Conversely, if the producer faces a small number of consumers, and their types are not highly correlated, then pooling all signals reduces the cost of sourcing the data.

The limited amount of price discrimination, which optimally operates at the group level rather than at 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 use price discrimination based on location and time, as well as other dimensions that effectively capture group characteristics.

The result in Proposition 8 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 in a more refined way, 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). Proposition 9 formalizes this result, and Figure 7 illustrates it.

Proposition 9 (Marginal Consumer)
The profitability of an additional consumer for the intermediary is higher under group pricing than under uniform pricing (as long as either profit level is strictly positive).

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.[17]

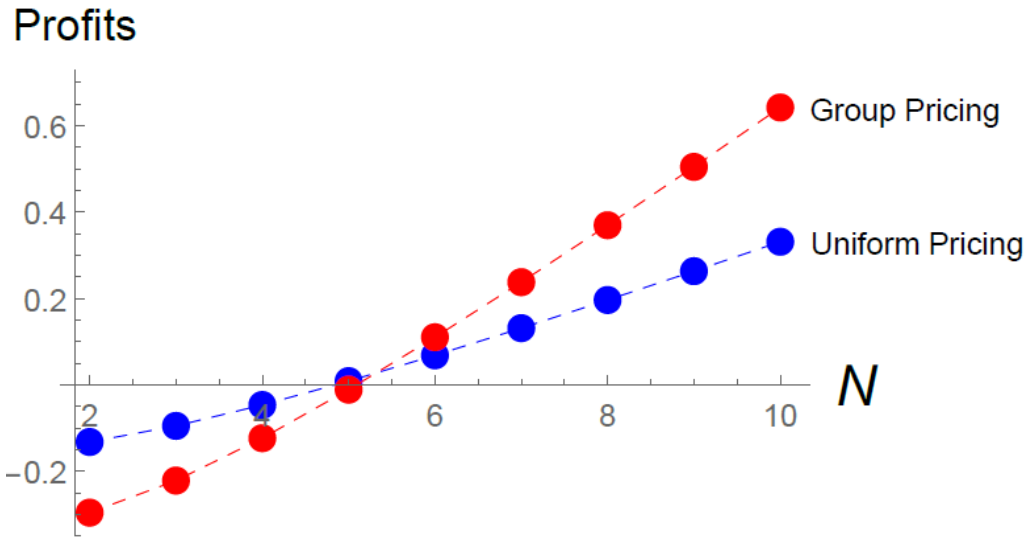


Figure 7: Marginal Value of an Additional Consumer ($\sigma_w = 1, \sigma_e = 0$)

[17] 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.

6 Intermediation and Value Creation

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 sufficiently strong 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 briefly describe two such extensions. The first one (quality pricing) generalizes our data aggregation result and traces its limits. The second one (recommender systems) shows the conditions under which information about idiosyncratic fundamentals is more likely to be traded than market level information.

6.1 Quality Pricing

This generalization of our framework allows the producer to charge a unit price p_i and to offer a quality level y_i to each consumer. Consumers are heterogeneous in their willingness to pay for the product, but they all value quality uniformly,

$$u_i\left(w_i, q_i, p_i, y_i\right) = \left(w_i + \gamma y_i - p_i\right) q_i - q_i^2/2.$$

The parameter $\gamma \in [0, \sqrt{2})$ is best viewed as a firm-level characteristic denoting the marginal value of quality.[18] (The case of $\gamma = 0$ yields the baseline model of price discrimination only.) The producer faces a constant marginal cost of quantity provision and a fixed cost of quality production, i.e.,

$$\pi = \sum\nolimits_{i=1}^{N} \left(p_i q_i - c q_i - y_i^2/2\right).$$

We assume that consumers perfectly observe their willingness to pay w_i at the onset, and we consider the effects of revealing all such information to the producer. In particular, we find that the producer benefits from information for all values of γ, while consumers benefit from collectively sharing information only if $\gamma > 1$, i.e., if the resulting targeted quality provision is sufficiently valuable. It also follows that there exists a unique $\gamma^* \in (0, 1)$ such that total surplus is higher when the producer observes all w_i than when he does not.

[18] Argenziano and Bonatti (2020) studied the role of privacy-protection regulations in a dynamic version of this model without data intermediation.

When the intermediary must procure the data from the consumers, a sufficiently strong data externality (e.g., large α or N) once again allows for the profitable intermediation of information. Whether the outcome improves or diminishes total surplus depends on the use of the data, which is represented here by the value of quality γ. Thus, equilibrium forces in the market for data do not prevent the diffusion of socially detrimental information in sufficiently large markets.

However, we can revisit the optimal aggregation policy in this model and establish the following generalization of Theorem 2.

Proposition 10 (Value of Quality and Aggregation)
The optimal data policy collects aggregate data if information reduces total surplus (i.e., if $\gamma < \gamma^*$) and individual data otherwise.

In other words, while the effect of information on total surplus imposes no restrictions on whether the data are traded in equilibrium, it does discipline the equilibrium level of data aggregation, which is provided in a socially efficient way.

An example consistent with these results provides a close B2B interpretation of our model. Consider a business that owns data about its customers and seeks to advertise on a large platform. In exchange for targeting benefits and a liquid supply of advertising space, this business shares data with the intermediary about the effectiveness of the ads. This performance data inform the demand for advertising in the future, as well as the platform-optimal (reserve) prices. In practice, the provision of advertising space and reserve prices in ad auctions are targeted at the advertiser level. In this case, our model suggests that this form of price and quality discrimination might be socially efficient; hence, the information that enables it is provided at the individual advertiser level.

6.2 Recommender System

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

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

with w_i denoting the consumer's willingness to pay and t_i the consumer's ideal location or product characteristic. Both the willingness to pay $w_i \in \mathbb{R}$ and the location $t_i \in \mathbb{R}$ of each consumer i are the sum of a component that is *common* to all consumers in the market

and an *idiosyncratic* component that reflects her individual taste shock. In particular, each consumer i's location is given by

$$t_i \triangleq \tau + \tau_i.$$

Because the distance $(y_i - t_i)^2$ reduces the intercept of the consumer's demand function, the case $\sigma_\tau^2 = \sigma_{\tau_i}^2 = 0$ yields the baseline model of price discrimination only. The producer has a constant marginal cost of quantity provision that we normalize to zero and can freely set the product's characteristic.

We examine the data intermediary's optimal data inflow policy, allowing for separate aggregation policies for willingness to pay and location information. We assume that the gains from trade under no information sharing are sufficiently large.[19] Finally, we assume that consumer i perfectly observes (w_i, t_i), but the extension to noisy signals is immediate. We then obtain another generalization of Theorem 2.

Proposition 11 (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 t_i.

Therefore, the recommender system enables the producer to offer targeted product characteristics that perfectly match y_i to t_i, but it does not allow for personalized pricing. The logic is once again given by the intermediary's sources of profits. Since the data externalities do not depend on the level of aggregation, the intermediary chooses to aggregate only the dimension of consumer data that would reduce total surplus if transmitted to the producer.

7 Discussion

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 the data intermediary has regarding the use of the data, as well as 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 any given data output.

Nonetheless, it is useful to consider the implications of the data intermediary's ability to commit to a certain data policy. In the working paper Bergemann, Bonatti, and Gan (2020), we formally analyzed the equilibrium outcome under this stronger commitment assumptions. We learned from Proposition 3 that the unrestricted use of data leads to a decrease in social

[19] This corresponds to assuming $\mu_\theta + \mu_{\theta_i}$ is sufficiently large relative to $\sigma_\tau^2 + \sigma_{\tau_i}^2$.

surplus. This finding suggests that the data intermediary could realize a higher profit with a data policy that limits the diffusion of information *on the equilibrium path*. Specifically, the intermediary could ask each consumer to share her data and commit to not passing it along to the downstream producer. In exchange for this commitment, the data intermediary requests compensation from the consumer. Proposition 14 in Bergemann, Bonatti, and Gan (2020) established that the revenue-maximizing data policy is indeed to acquire all consumer data and to never forward the data to the downstream producer.[20]

Unique Implementation Our analysis has characterized the intermediary's most preferred equilibrium. An ensuing question is whether the qualitative insights 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)).

Proposition 16 in Bergemann, Bonatti, and Gan (2020) showed that the intermediary can sequentially approach consumers and offer compensation conditional on all earlier consumers having accepted their offers. In this scheme, the first consumer receives compensation equal to her entire surplus loss, guaranteeing her acceptance regardless of the other consumers' decisions. More generally, consumer i receives the optimal compensation level in the baseline equilibrium when $N = i$. Thus, the cost of acquiring the consumers' data is strictly higher than in the intermediary's most preferred equilibrium. Nonetheless, as N grows without bound, the intermediary's profit per capita converges to the profit per capita when aggregate data are available "in the wild."

Competing Intermediaries In Bergemann, Bonatti, and Gan (2020), we asked whether competition among data intermediaries can restore the socially efficient data policy. We assumed that each intermediary $j \in \{1, ..., J\}$ collects a noisy signal about each consumer i,

$$r_{i,j} = w_i + \zeta_{i,j}.$$

As in Section 5, the noise term $\zeta_{i,j}$ represents limitations in the communication channel between the consumers and the intermediary. We assume that each shock $\zeta_{i,j}$ is independently

[20] 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; thus, an interim perspective is adopted. The shared insight is that the intermediary *with* commitment power might be able to extract a rent without any further influence on the efficiency of the allocation.

drawn from a normal distribution with zero mean and variance $\sigma_\zeta^2 > 0$. The timing of the game is otherwise unchanged: consumers and producer choose the subset of offers to accept. Proposition 18 in Bergemann, Bonatti, and Gan (2020) compared the equilibrium outcome under competition with the monopoly outcome. In the equilibrium outcome, competing intermediaries collect data in the same way as a monopolist would. Each intermediary's profit decreases due to a loss of bargaining power against the producer. Relative to monopoly, the producer obtains an additional portion of the surplus extracted from consumers, and consumer surplus decreases even further due to the multiple sources of information. This result underscores the possibility that competition in the data market may reduce the consumer surplus rather than protecting and increasing it.

8 Conclusion

We have explored the terms of information trading between data intermediaries with market power and multiple consumers with correlated preferences. The data externality that we have uncovered strongly suggests that data ownership is insufficient to bring about the efficient use of information. In large markets, arbitrarily small levels of compensation can induce an individual consumer to relinquish precise information about her preferences.

A possible solution to alleviate this problem—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 like large online platforms. In our baseline model, this policy proposal would restore the first best. However, consumer unionization would pose serious implementation challenges, both from theoretical and practical perspectives. Within the confines of our model a richer specification of consumer heterogeneity, e.g., one that allows for different marginal valuations of quality as in Section 6 could be considered. Then, revealing information to a given producer might improve some consumers' welfare but hurt others'. In this scenario, a consumer union would need to aggregate consumer preferences prior to negotiating compensation with the intermediary, which would inevitably lead to distortions in the allocation of information.

On a more constructive note, 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., Staples, Amazon). Thus, our analysis points to the significant welfare effects of group-based price discrimination and of

uniform prices that react in real time to changes in market-level demand.[21]

Of course, there are dimensions along which the data generate surplus, including some that do not interact with consumers' decisions to reveal their information: 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 also other welfare-reducing effects, such as spillovers of consumer data to other markets, including B2B markets. For example, if Amazon and Google use the information revealed by consumers to extract more surplus from advertisers, then consumers will also pay a higher price, depending on the pass-through rate of the marginal cost of advertising. In other words, the data externality that we identify is pervasive, starting with consumers but often extending to other economic agents whose decisions are informed by consumer data.

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.[22] Many such 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. A distinguishing feature of social data platforms is that they typically trade individual consumer information for services, rather than for money. The data externality then expresses itself in the level of service (i.e., in quantity and/or quality) and in the amount of consumer engagement, rather than in the level of monetary compensation.

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

[22] Product data platforms, such as Amazon, Uber and Lyft, acquire individual data from the consumer through the 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, mostly in the form of targeted advertising space. In terms of our model, the difference between the data intermediary and a product data platform is that the platform combines the roles of data intermediation and product pricing.

9 Appendix

The Appendix collects the proofs of all the results in the paper. Throughout, we assume without loss of generality that $\mu_{\theta_i} = c = 0$ and $\sigma_w^2 = 1$, and we let $\lambda \triangleq \sigma_e^2/\sigma_w^2$

Proof of Proposition 1. Denote consumer i's initial information (σ-algebra) as S_i, and suppose the intermediary provides a data outflow T. (This information need not be the complete sharing policy.) Consumer i demands the following quantity:

$$q_i = \mathbb{E}[w_i|T, S_i] - p_i,$$

and the producer sets the following price

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

The profit of the producer is:

$$\begin{aligned}\Pi_i(T) &= \mathbb{E}\left[\frac{\mathbb{E}[w_i|T]}{2}\left(\mathbb{E}[w_i|T, S_i] - \frac{\mathbb{E}[w_i|T]}{2}\right)\right], \\ &= \frac{\mathbb{E}\left[(\mathbb{E}[w_i|T])^2\right]}{4} = \frac{[\ \mathbb{E}[w_i|T]] + \mu_\theta^2}{4},\end{aligned} \tag{43}$$

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

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

The expected consumer surplus is given by

$$\begin{aligned}U_i(T) = \mathbb{E}\left[\left(w_i - \frac{\mathbb{E}[w_i|T]}{2}\right)\left(\mathbb{E}[w_i|T, S_i] - \frac{\mathbb{E}[w_i|T]}{2}\right) - \frac{1}{2}\left(\mathbb{E}[w_i|T, S_i] - \frac{\mathbb{E}[w_i|T]}{2}\right)^2\right], \\ = \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|T, S_i])^2 - \frac{3}{4}(\mathbb{E}[w_i|T])^2], \\ = \frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|T, S_i] - \mathbb{E}[w_i|T])^2 + \frac{1}{4}(\mathbb{E}[w_i|T])^2].\end{aligned} \tag{44}$$

Therefore the consumer surplus impact of complete data sharing is

$$U_i(S) - U_i(\varnothing) = \frac{1}{8}\,\text{var}[\mathbb{E}[w_i|S]] - \frac{1}{2}\,\text{var}[\mathbb{E}[w_i|S_i]],$$

which completes the proof. ■

Proof of Lemma 1. Recall $\lambda \triangleq \sigma_e^2/\sigma_w^2$ and write the posterior expectation of w_i as

$$\begin{aligned}
\mathbb{E}[w_i|s_i] &= \frac{\sigma_w^2}{\sigma_w^2+\sigma_e^2}s_i + \frac{\sigma_e^2}{\sigma_w^2+\sigma_e^2}\mu_\theta = \frac{1}{1+\lambda}s_i + \frac{\lambda}{1+\lambda}\mu_\theta, \\
\mathbb{E}[w_i|s] &= As_i + B\Sigma_{j\neq i}s_j + (1-A-B)\mu_\theta, \qquad (45) \\
A &= \frac{(\alpha-1)N(\alpha+\beta\lambda) - \alpha(\alpha+\beta\lambda-2) + (2\beta-1)\lambda - 1}{(\alpha+(\beta-1)\lambda-1)(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)}, \\
B &= \frac{\lambda(\beta-\alpha)}{(\alpha+(\beta-1)\lambda-1)(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)}.
\end{aligned}$$

When $\alpha = \beta$, we have $B = 0$ and $\mathbb{E}[w_i|s_i] = \mathbb{E}[w_i|s]$. Recalling the definition of G,

$$G \triangleq \text{var}\left[\hat{w}_i\left(s\right)\right] - \left[\hat{w}_i\left(s_i\right)\right],$$

we then have

$$G(\alpha,\alpha) = 0 \leq G(\alpha,\beta) \quad \forall \alpha,\beta.$$

Because $\mathbb{E}[w_i|s_i]$ is unchanged when varying α and β, we focus on $\text{var}[\mathbb{E}[w_i|s]]$, which equals

$$\frac{\alpha^2((1-\lambda)N+\lambda-1) + \alpha(2\beta\lambda N - 2\beta\lambda - N + 2) - \beta\lambda N + 2\beta\lambda - \lambda - 1}{(\alpha+(\beta-1)\lambda-1)(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)}.$$

Taking the derivative with respect to α, we have:

$$-\frac{\lambda^2(N-1)(\alpha-\beta)(N(\alpha(2\beta-1)+\beta(2(\beta-1)\lambda-1)) - 2(\beta-1)(\alpha+(\beta-1)\lambda-1))}{(\alpha+(\beta-1)\lambda-1)^2(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}.$$

The numerator is a quadratic function of α. Denote it as $f(\alpha)$, and notice that it satisfies

$$\begin{aligned}
f(\beta) &= 0, \\
f(1) &= (-1+\beta)^2\lambda^2(N-1)(2\lambda(1+\beta(-1+N))+N) > 0, \\
f(0) &= \beta\lambda^2(N-1)\left(2\beta^2\lambda(N-1) - \beta(2\lambda+1)(N-2) - 2(\lambda+1)\right) < 0.
\end{aligned}$$

Therefore the derivative is negative in $[0,\beta)$ and positive in $(\beta,1]$. Thus, G decreases with α in $[0,\beta]$ and increases with α in $[\beta,1]$. Similarly, taking the derivative of $\text{var}[\mathbb{E}[w_i|s]]$ with respect to β we have:

$$\frac{\lambda^2(N-1)(\alpha-\beta)(N(\alpha(2\beta-1)\lambda+2(\alpha-1)\alpha-\beta\lambda) - 2(\alpha-1)(\alpha+(\beta-1)\lambda-1))}{(\alpha+(\beta-1)\lambda-1)^2(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}.$$

The numerator is a quadratic function of β. Denote it as $g(\beta)$, and notice that

$$\begin{aligned} g(\alpha) &= 0, \\ g(1) &= (\alpha-1)^2\lambda^2(N-1)(2\alpha(N-1)+\lambda N+2) > 0, \\ g(0) &= \alpha\lambda^2(N-1)(2(\alpha-1)(-\alpha+\lambda+1)-\alpha N(-2\alpha+\lambda+2)) < 0. \end{aligned}$$

So G decreases with β in $[0,\alpha]$ and increases with β in $[\alpha,1]$.

Therefore we know G reaches maximum at either $(0,1)$ or $(1,0)$. Since

$$\begin{aligned} G(0,1) &= \frac{\lambda(N-1)+1}{8\lambda N+8}, \\ G(1,0) &= \frac{N}{8(\lambda+N)}, \end{aligned}$$

we conclude that $(0,1)$ is the maximizer if $\lambda > 1$ while $(1,0)$ is the maximizer if $\lambda < 1$. ■

Proof of Proposition 2. The effect of complete data sharing on social surplus can be written as

$$W_i(S) - W_i(\varnothing) = \frac{3}{8}G(\alpha,\beta) - \frac{1}{8}\operatorname{var}\left[\hat{w}_i(s_i)\right].$$

By Lemma 1, we know that $G(\alpha,\alpha) = 0$, and hence whenever $\alpha = \beta$, social welfare is decreasing with complete data sharing. This observation and Lemma 1 imply

$$\begin{aligned} \underline{\beta}(\alpha) < \alpha < \overline{\beta}(\alpha), \\ \underline{\alpha}(\beta) < \beta < \overline{\alpha}(\beta). \end{aligned}$$

Notice that the proposition allows for the possibility that all these four thresholds are the boundary ($\overline{\beta}(\alpha)$, $\overline{\alpha}(\beta)$ equals 1 or $\underline{\beta}(\alpha)$, $\underline{\alpha}(\beta)$ equals 0).

Finally we prove the third part of the statement. Suppose that $\alpha' > \alpha$: if either threshold is on the boundary the result is immediate. For interior thresholds, we know that

$$G(\alpha,\underline{\beta}(\alpha)) = G(\alpha',\overline{\beta}(\alpha')) = \frac{1}{8(1+\lambda)}.$$

By Lemma 1 we know that

$$\begin{aligned} G(\alpha',\underline{\beta}(\alpha)) > G(\alpha,\underline{\beta}(\alpha)) = \frac{1}{8(1+\lambda)}, \\ G(\alpha,\overline{\beta}(\alpha')) > G(\alpha',\overline{\beta}(\alpha')) = \frac{1}{8(1+\lambda)}. \end{aligned}$$

The first line implies $\underline{\beta}(\alpha') > \underline{\beta}(\alpha)$, and the second line implies $\overline{\beta}(\alpha') > \overline{\beta}(\alpha)$. The result for $\overline{\alpha}$ and $\underline{\alpha}$ can be proved similarly. ■

Proof of Proposition 3. As we have shown in the proof of Lemma 1, $\text{var}[\hat{w}_i(s_i)]$ does not depend on α and β for any fixed σ_e^2 and σ_w^2. Therefore, an increase in G means an increase in $\text{var}[\hat{w}_i(s)]$. By Proposition 1, all three surplus levels increase.

The positive impact of data sharing on producer surplus is immediate. We thus turn to consumer surplus. According to Proposition 1 and the equation (45) in the proof of Lemma 1, we know that

$$U_i(S) - U_i(\varnothing) = \frac{A_1 + B_1\lambda + C_1\lambda^2}{-8(\lambda+1)(\alpha + (\beta-1)\lambda - 1)(\alpha(N-1) + \beta\lambda(N-1) + \lambda + 1)},$$
$$A_1 = 3(-1+\alpha)(1+\alpha(-1+N)) < 0,$$
$$B_1 = \alpha(6\beta(N-1) - 3N + 6) - 3\beta(N-2) - 6 < 0,$$
$$C_1 = \alpha^2(N-1) - 2\alpha\beta N + 2\alpha\beta + 4\beta^2(N-1) - 3\beta(N-2) - 3.$$

(The formula does not hold in the degenerate case $\alpha = \beta = 1$, in which case the consumer surplus and social surplus are always negative so that both thresholds are infinite.) Therefore when $C_1 < 0$, the numerator is always negative. When $C_1 > 0$, there exists a finite threshold $\overline{\overline{\sigma}}_e^2$ such that the numerator is positive if and only if $\lambda > \overline{\overline{\sigma}}_e^2$.

For social surplus, we obtain

$$W_i(S) - W(\varnothing) = \frac{A_2 + B_2\lambda + C_2\lambda^2}{-8(\lambda+1)(\alpha + (\beta-1)\lambda - 1)(\alpha(N-1) + \beta\lambda(N-1) + \lambda + 1)},$$
$$A_2 = (\alpha-1)(\alpha(N-1) + 1) < 0,$$
$$B_2 = \alpha(2\beta(N-1) - N + 2) - \beta(N-2) - 2 < 0,$$
$$C_2 = 3\alpha^2(N-1) - 6\alpha\beta(N-1) + \beta(4\beta(N-1) - N + 2) - 1.$$

Thus when $C_2 < 0$, the numerator is always negative. When $C_2 > 0$, there exists a finite threshold $\overline{\sigma}_e^2$. Since $W_i(S) - W(\varnothing) > U_i(S) - U_i(\varnothing)$, clearly we have $\overline{\sigma}_e^2 < \overline{\overline{\sigma}}_e^2$ whenever $\overline{\sigma}_e^2$ is finite.

Finally, the last part of the proposition follows from the fact that for $\alpha < 1$, $\beta = 1$,

$$C_1 = (\alpha-1)^2(N-1) > 0,$$
$$C_2 = 3(\alpha-1)^2(N-1) > 0.$$

This completes the proof. ■

Before we proceed to the proof of Theorem 1, we establish a bound of α above which is intermediation with complete data sharing is profitable.

Lemma 2 *If the intermediary collects positive profit under complete data sharing, then:*

$$\alpha \geq \frac{1}{\sqrt{2N+1}} \tag{46}$$

Proof. Using the expression 43 and 44 in the proof of Proposition 1, the intermediary's profit under complete data sharing can be written as

$$R(S) = \frac{N}{8}\left(3\,\mathrm{var}\left[\hat{w}_i\left(s_{-i}\right)\right] - \mathrm{var}\left[\hat{w}_i\left(s\right)\right]\right),$$
$$= \frac{3(N-1)N\sigma_\theta^4}{8((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N-1)\sigma_{\theta_i}^4}{8(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}.$$

Now suppose we have:

$$\alpha < \frac{1}{\sqrt{2n+1}}, \quad \text{Or equivalently} \quad \sigma_\theta^2 < \frac{1+\sqrt{2N+1}}{2N}\sigma_{\theta_i}^2.$$

We will prove the profit is always negative in this case. Notice that,

$$\frac{\partial}{\partial N}\frac{3(N-1)\sigma_\theta^4}{8(\sigma_\varepsilon^2(N-1)+\sigma_{\varepsilon_i}^2+(N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)},$$
$$= \frac{3\sigma_\theta^4(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}{8(\sigma_\varepsilon^2(N-1)+\sigma_{\varepsilon_i}^2+(N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2} > 0.$$

We have a upper bound of the profit:

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

The numerator is linear function of $\sigma_{\varepsilon_i}^2$. Under **??**, we know the linear term is negative:

$$2N\sigma_\theta^4 < 2\sigma_\theta^2\sigma_{\theta_i}^2 + \sigma_{\theta_i}^4$$

To complete our proof, it then suffices to show that the constant term is also negative.

Plugging the above inequality into the expression for the constant term:

$$\begin{aligned}
&-N\sigma_{\theta_i}^2\left(N\sigma_{\theta_i}^2(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\theta_i}^2(-\sigma_\varepsilon^2+\sigma_\theta^2+\sigma_{\theta_i}^2)-2N\sigma_\theta^4\right),\\
<&-N\sigma_{\theta_i}^2\left(N\sigma_{\theta_i}^2(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\theta_i}^2(-\sigma_\varepsilon^2+\sigma_\theta^2+\sigma_{\theta_i}^2)-(2\sigma_\theta^2\sigma_{\theta_i}^2+\sigma_{\theta_i}^4)\right),\\
=&-N\sigma_{\theta_i}^2(N-1)\sigma_{\theta_i}^2(\sigma_\varepsilon^2+\sigma_\theta^2)<0,
\end{aligned}$$

which completes the proof. ■

Proof of Theorem 1. Rewrite the expression of R in the language of α, β and λ:

$$\begin{aligned}
R(S)=&\frac{3\alpha^2(N-1)N}{8(\alpha(N-1)-\alpha+\beta\lambda(N-1)+(1-\beta)\lambda+1)}\\
-&\frac{(\alpha N-\alpha+1)^2}{8(N(\alpha+\beta\lambda)-\alpha+(1-\beta)\lambda+1)}-\frac{(1-\alpha)^2(N-1)}{8(-\alpha+(1-\beta)\lambda+1)}.
\end{aligned}$$

We first prove that R is strictly increasing in α when $R>0$. Since the third term in the expression is strictly increasing in α, it is sufficient to prove that the first two terms are increasing in α when $R>0$. By Lemma 2 in the Appendix, when the intermediary collects positive profits, we have

$$\alpha\geq\frac{1}{\sqrt{2N+1}}.$$

The derivative of the first two terms with respect to α is:

$$g=\frac{3\alpha N(\alpha(N-2)+2(\beta\lambda(N-2)+\lambda+1))}{(\alpha(N-2)+\beta\lambda N-2\beta\lambda+\lambda+1)^2}-\frac{(\alpha(N-1)+1)(\alpha(N-1)+2\lambda(\beta(N-1)+1)+1)}{(\alpha(N-1)+\beta\lambda N-\beta\lambda+\lambda+1)^2}.$$

When $N=2$, direct calculation shows

$$g=\frac{2(\beta+1)\lambda^2(\alpha(6\beta+5)-1)+(\alpha+1)\lambda(\alpha(24\beta+23)-2\beta-3)+(12\alpha-1)(\alpha+1)^2}{(\lambda+1)(\alpha+\beta\lambda+\lambda+1)^2}.$$

the numerator is increasing in α, and is positive when $\alpha=1/\sqrt{5}$ which yields the result.

When $N\geq 3$, we have $1/N<1/\sqrt{2N+1}$, and we can bound α from below using the weaker condition $\alpha>1/N$. We then obtain

$$\begin{aligned}
g>&\frac{3\alpha N(\alpha(N-2)+2(\beta\lambda(N-2)+\lambda+1))}{(\alpha(N-1)+\beta\lambda N-\beta\lambda+\lambda+1)^2}-\frac{(\alpha(N-1)+1)(\alpha(N-1)+2\lambda(\beta(N-1)+1)+1)}{(\alpha(N-1)+\beta\lambda N-\beta\lambda+\lambda+1)^2},\\
=&\frac{\alpha^2(2(N-2)N-1)+\lambda(\beta(2\alpha(2(N-2)N-1)-2(N-1))+4\alpha N+2\alpha-2)+4\alpha N+2\alpha-1}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}.
\end{aligned}$$

Every term in the numerator above is positive, so $g>0$ when $N\geq 3$.

We complete the proof of the first part of the theorem by showing that

$$R(\alpha=0) = -\frac{N(\beta\lambda(N-2)+\lambda+1)}{8(1+(1-\beta)\lambda)(\beta\lambda(N-1)+\lambda+1)} < 0,$$
$$R(\alpha=1) = \frac{N(\lambda(\beta(2(N-2)N+3)+2N-3)+2(N-1)N)}{8(\beta\lambda N-2\beta\lambda+\lambda+N-1)(\beta\lambda N-\beta\lambda+\lambda+N)} > 0.$$

To prove the second part of the theorem, we first show that the $\partial R/\partial\beta$ is negative when $R=0$. This guarantees that the threshold α^* is increasing in β. Up to a multiplicative constant, we have

$$\frac{\partial R}{\partial\beta} = -\frac{3\alpha^2(N-1)N(\lambda(N-2))}{(\alpha(N-1)-\alpha+\beta\lambda(N-1)+(1-\beta)\lambda+1)^2} - \frac{(1-\alpha)^2\lambda(N-1)}{(-\alpha+(1-\beta)\lambda+1)^2} + \frac{(\alpha N-\alpha+1)^2(\lambda(N-1))}{(N(\alpha+\beta\lambda)-\alpha+(1-\beta)\lambda+1)^2}.$$

Use $R=0$ to eliminate the second term we obtain

$$\begin{aligned} 8\frac{1-\alpha+(1-\beta)\lambda}{1+\lambda}\frac{\partial R}{\partial\beta} &= -\frac{3\alpha^2(N-1)^2N}{(\alpha(N-2)+\beta\lambda(N-2)+\lambda+1)^2} \\ &\quad + \frac{(\alpha(N-1)+1)^2}{(N-1)(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2} \\ &< \frac{-3\alpha^2(N-1)^2N+(\alpha(N-1)+1)^2/(N-1)}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}. \end{aligned}$$

Simplifying the numerator, it suffices to show that

$$1+2(N-1)\alpha(1-(N-1)\alpha) < 0$$

whenever $\alpha > 1/\sqrt{2N+1}$. Both conditions are met, and hence $\partial R/\partial\beta < 0$ when $N \geq 6$.

As for the fixed point, it is easy to calculate the system of equations:

$$R(S)=0, \quad \alpha=\beta$$

has β_0 as its unique solution ($\beta \in [0,1]$). To see this, plug $\alpha=\beta$ in the expression of R:

$$R(S) = \frac{N(\beta(3\beta(N-1)-N+2)-1)}{8(\lambda+1)(\beta(N-2)+1)} = 0,$$
$$\Rightarrow \beta = \beta_0(N) \triangleq \frac{\sqrt{N^2+8N-8}+N-2}{6(N-1)}.$$

From Lemma 2 in the Appendix we know $\alpha^*(0) > 0$, so α^* crosses the diagonal line only

once, and from above. This completes the proof of the second part of the theorem.

Finally, we show that $\partial R/\partial\lambda < 0$ when $\alpha < \beta$ and $R = 0$ while $\partial R/\partial\lambda > 0$ when $\alpha > \beta$ and $R = 0$, which proves the last part of the theorem.

$$\frac{\partial R}{\partial \lambda} = -\frac{3\alpha^2(N-1)N(\lambda(N-2))}{8(\alpha(N-1)-\alpha+\beta\lambda(N-1)+(1-\beta)\lambda+1)^2} - \frac{(1-\alpha)^2\lambda(N-1)}{8(-\alpha+(1-\beta)\lambda+1)^2} + \frac{(\alpha N-\alpha+1)^2(\lambda(N-1))}{8(N(\alpha+\beta\lambda)-\alpha+(1-\beta)\lambda+1)^2}.$$

Using $R = 0$ to eliminate the second term we obtain

$$\begin{aligned} 8\frac{1-\alpha+(1-\beta)\lambda}{\beta-\alpha}\frac{\partial R}{\partial \lambda} &= -\frac{3\alpha^2(N-1)^2N}{(\alpha(N-2)+\beta\lambda(N-2)+\lambda+1)^2} \\ &\quad + \frac{N(\alpha(N-1)+1)^2}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}, \\ &< \frac{-3\alpha^2(N-1)^2N + (\alpha(N-1)+1)^2/\,(N-1)}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2} < 0 \end{aligned}$$

If $\alpha > \beta$ and inequality (46) holds, we have

$$\begin{aligned} &\frac{\partial}{\partial N}\frac{(N-1)^2N}{(\alpha(N-2)+\beta\lambda(N-2)+\lambda+1)^2} \\ &= \frac{(N-1)(\alpha((N-5)N+2)+\lambda(\beta((N-5)N+2)+3N-1)+3N-1)}{(\alpha(N-2)+\beta\lambda(N-2)+\lambda+1)^3} > 0. \end{aligned}$$

Thus we conclude that

$$\begin{aligned} 8\frac{1-\alpha+(1-\beta)\lambda}{\alpha-\beta}\frac{\partial R}{\partial \lambda} &> -\frac{N(\alpha(N-1)+1)^2}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2} \\ &\quad + \frac{3\alpha^2N^2(N+1)}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}, \\ &= \frac{N(\alpha(\alpha(N(2N+5)-1)-2N+2)-1)}{(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)^2}. \end{aligned}$$

Notice that $\alpha(\alpha(N(2N+5)-1)-2N+2)-1$ is increasing in α and N when equation (46) holds. Plugging in $N = 2$ and equation (46) we obtain:

$$\alpha(\alpha(N(2N+5)-1)-2N+2)-1 \approx 1.50557 > 0.$$

Therefore $\partial R/\partial\lambda > 0$ in this case. ■

Proof of Corollary 1. Because the information gain $G = 0$ when $\lambda = 0$ (i.e., when consumers know their value perfectly) we know that complete data sharing always decreases social surplus in that case. Therefore $\underline{\alpha}(\beta) = 0$ for all β (the meaning of the notation is the same as in Proposition 2). By continuity there exists a $\bar{\lambda}$ such that for all $\lambda \in [0, \bar{\lambda}]$

$$\underline{\alpha}(\beta, \lambda) \leq \underline{\alpha}(1, \lambda) \leq \underline{\alpha}(1, \bar{\lambda}) < \frac{N(\sqrt{3}+1)-1}{2N\left(N+1\right)-1}.$$

While by Proposition 6, the intermediary's profit is non-negative only when

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

which completes the proof. ■

Proof of Proposition 4. Define $D(\alpha, \beta)$ as the difference

$$\text{var}\left[\hat{w}_i\left(s\right)\right] - \text{var}\left[\hat{w}_i\left(s_{-i}\right)\right] - \text{var}\left[\hat{w}_i\left(s_i\right)\right]$$

when the data structure is (α, β). Signals are substitutes (complements) if $D \leq (\geq) 0$. We will prove that for any β, there exists a threshold $\tilde{\alpha}$ such that $D(\alpha, \beta) \leq 0$ iff $\alpha \geq \tilde{\alpha}$.

When $\alpha = 0$, we know:

$$\text{var}\left[\hat{w}_i\left(s_{-i}\right)\right] = 0.$$

Therefore by definition:

$$D(0, \beta) = \text{var}\left[\hat{w}_i\left(s\right)\right] - \left[\hat{w}_i\left(s_i\right)\right] \geq 0.$$

When $\alpha = \beta$, from the proof of Lemma 1 we know that:

$$\begin{aligned}\text{var}\left[\hat{w}_i\left(s\right)\right] &= \text{var}\left[\hat{w}_i\left(s_i\right)\right],\\ D(\beta, \beta) &= -\text{var}\left[\hat{w}_i\left(s_{-i}\right)\right] \leq 0.\end{aligned}$$

We then calculate the following quantities:

$$\begin{aligned}\text{var}\left[\hat{w}_i\left(s\right)\right] &= \frac{(\alpha-1)(\alpha(\lambda-1)-2\beta\lambda+\lambda+1) - N\left(\alpha^2(\lambda-1)-2\alpha\beta\lambda+\alpha+\beta\lambda\right)}{(\alpha+(\beta-1)\lambda-1)(\alpha(N-1)+\beta\lambda(N-1)+\lambda+1)},\\ \text{var}\left[\hat{w}_i\left(s_{-i}\right)\right] &= \frac{\alpha^2(N-1)}{\alpha(N-2)+\beta\lambda(N-2)+\lambda+1},\\ \text{var}\left[\hat{w}_i\left(s_i\right)\right] &= \frac{1}{\lambda+1}.\end{aligned}$$

Clearly var $[\hat{w}_i(s_{-i})]$ is strictly increasing in α, and according to Lemma 1, we know var $[\hat{w}_i(s)]$ is strictly decreasing in α when $\alpha \in [0, \beta]$. Therefore we know $D(\alpha, \beta)$ is strictly decreasing in α in $[0, \beta]$, and there exists a $\tilde{\alpha} \in [0, \beta]$ such that $D(\alpha, \beta) \geq 0$ if $\alpha \leq \tilde{\alpha}$ and $D(\alpha, \beta) \leq 0$ if $\tilde{\alpha} \leq \alpha \leq \beta$.

What remains to be proved is $D(\alpha, \beta) \leq 0$ for $\alpha \in [\beta, 1]$. To verify this, we calculate an expression for $D(\alpha, \beta)$:

$$\frac{A\beta^3 + B\beta^2 + C\beta + E}{(\lambda+1)(-\alpha+(1-\beta)\lambda+1)(\alpha(N-2)+\beta\lambda(N-2)+\lambda+1)(\alpha n-\alpha+\beta\lambda N+(1-\beta)\lambda+1)},$$

$$A = \lambda^3(N-2)(N-1) > 0,$$

$$B = \lambda^2(N-1)(\alpha(\alpha(\lambda+1)(N-1)-2\lambda(N-2)+N-2)+\lambda+1),$$

$$C = \alpha\lambda(N-1)\left(2\alpha^2(\lambda+1)(N-1)-\alpha(4\lambda+1)(N-2)-2\lambda(\lambda+1)\right),$$

$$E = \alpha^2(N-1)\left(\alpha^2(\lambda+1)(N-1)-\alpha(2\lambda+1)(N-2)-(\lambda+1)(2\lambda+1)\right) < 0.$$

Since the denominator is positive, we focus on the numerator $f(\alpha, \beta) \triangleq A\beta^3 + B\beta^2 + C\beta + E$. We complete the proof by showing that $f(\alpha, \beta) \leq 0$ for any $\beta \in [0, \alpha]$. From the previous analysis, we know $f(\alpha, \alpha) \leq 0$, and since $E < 0$ we know $f(\alpha, 0) < 0$. Now notice that

$$\frac{\partial^2 f}{\partial \beta^2}(\alpha, \beta) = 6A\beta + 2B.$$

Since $A > 0$, either f is convex in β, or f is concave for small β and convex for large β.

If $B \geq 0$ and f is convex in β, we are done, since:

$$f(\alpha, \beta) \leq \max\{f(\alpha, 0), f(\alpha, \alpha)\} \leq 0 \quad \forall \beta \in [0, \alpha].$$

If $B < 0$ and f is first concave then convex, we look at the first order derivative at 0:

$$\frac{\partial f}{\partial \beta}(\alpha, 0) = C.$$

If $C < 0$, then f is decreasing in the concave region, and we are done. At last, we prove that it is impossible to have $B < 0$ and $C > 0$ simultaneously. If so:

$$\frac{2B}{\lambda^2(N-1)} = 2\alpha^2(\lambda+1)(N-1) - 2\alpha(2\lambda-1)(N-2) + 2\lambda + 2 < 0,$$

$$\frac{C}{\alpha\lambda(N-1)} = 2\alpha^2(\lambda+1)(N-1) - \alpha(4\lambda+1)(N-2) - 2\lambda(\lambda+1) > 0.$$

Subtracting the second line from the first, we obtain:

$$-3\alpha\left(N-2\right)-2\left(1+\lambda\right)^{2}<0,$$

a contradiction. ■

Proof of Proposition 5. Suppose the intermediary collects a data inflow X, while consumer i has initial information structure S_i. The intermediary chooses an outflow policy, namely the signal $T = T(X)$ sent to the producer and the signal $T_i = T_i(X)$ 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 extracts through the fixed fee m_0.

The proof of this proposition is organized as follows. First, we show that a Perfect Bayesian Equilibrium (PBE) always exists under any information structure (T_i, T). Second, we argue that without loss of generality we can focus on information structures where T is measurable with respect to T_i, i.e., where consumer i observes all information signal sent to the producer. Third, we prove it is optimal to provide full public information, $T = T_i = X$.

When $\sigma_e = 0$, consumers know their value perfectly, and there is an essentially equilibrium under any information structure. It is easy to show using equation (43) that, in this case, providing full information to the producer is optimal. Thus from now on we will focus on the case where $\sigma_e > 0$.

Our first lemma establishes equilibrium existence by constructing a PBE under any information structure (T_i, T).

Lemma 3 *When $\sigma_e > 0$, under any information structure (T_i, T), there exists a PBE.*

Proof. For an arbitrary, noisy information structure (T_i, T), consumer i 's posterior $\mu(S_i, T_i)$ has full support almost surely (a.s.). Now consider the following pooling strategy profile:

$$p_i^*(T) = \frac{1}{2}\mu$$

$$q_i^*(p_i, T_i) = \begin{cases} \mathbb{E}[w_i|T_i, S_i] - p_i, & \text{if } p_i = \frac{1}{2}\mu, \\ 0 & \text{otherwise.} \end{cases}$$

The optimality of the consumer's demand is guaranteed by assuming that the consumer will believe $w_i = p_i$ whenever $p_i \neq \mu/2$. Such off-path belief is part of an equilibrium if

$\mu(S_i, T_i(S))$ a.s. has full support. Note that for any non-zero measure (with respect to Lebesgue measure) subset $E \subset \mathbb{R}$, we have:

$$\begin{aligned}\mu(S_i, T_i)(E) &= \mathbb{E}[1_{w_i \in E} | S_i, T_i] \\ &= \mathbb{E}\Big[\mathbb{E}[1_{w_i \in E} | S_i, S] \Big| S_i, T_i\Big] > 0 \ a.s.\end{aligned}$$

The inequality in the last line follows from the assumption

$$\mathbb{E}[1_{w_i \in E} | S_i, S] = \mu(S_i, S)(E) > 0 \ a.s,$$

which completes the proof. ■

Next, we show that introducing asymmetric information is always suboptimal.

Lemma 4 *It is without loss of generality to consider information structures where consumer i has superior information compared to the producer. Thus, $T(X)$ is measurable with respect to $T_i(X)$.*

Proof. For any information structure (T, T_i), denote an induced signalling equilibrium as $\sigma(T, T_i) = (q_i^*, p^*)$. We will prove there exists an equilibrium $\sigma(p^* \circ T, (T_i, p^* \circ T))$ under information structure $(T, (T_i, p^* \circ T))$ that brings the producer a weakly higher ex-ante payoff. In this new information structure, instead of revealing T to the producer, the intermediary only reveals what is necessary for the equilibrium pricing strategy in $\sigma(T, T_i)$, and tells consumer i both T_i and what the producer knows.

On the equilibrium path of $\sigma(T, T_i)$, consumer i updates her posterior $\mu(T_i, S_i, p^*(T))$ using T_i, her own private signal S_i, and the pricing signal $p^*(T)$. Thus the consumer demand is given by:

$$q_i\Big(p^*(T), \mu(T_i, S_i, p^*(T))\Big) = \mathbb{E}[w_i | T_i, S_i, p^*(T)] - p^*(T).$$

The ex-ante payoff of the producer is:

$$\mathbb{E}\left[p^*(T) q_i\Big(p^*(T), \mu(T_i, S_i, p^*(T))\Big)\right].$$

Now consider the new information structure $(T, (T_i, p^* \circ T))$. In this structure, since consumer i knows everything that the producer knows, the price has no signalling effect. There is a natural equilibrium where consumer i forms her demand using the signal $(T_i, p^* \circ T)$ from the intermediary and her own signal S_i. The demand (both on the path and off the path) will be:

$$q_i\Big(p, \mu(T_i, S_i, p^*(T))\Big) = \mathbb{E}[w_i | T_i, S_i, p^*(T)] - p.$$

Knowing this, the producer maximizes his ex ante payoff by choosing a pricing strategy $\hat{p}(\cdot)$ as a function of his signal: $p^* \circ T(w_i)$. Thus the producer's equilibrium profit is:

$$\max_{\hat{p}} \hat{p}\big(p^*(T)\big) q_i\Big(\hat{p}\big(p^*(T)\big), \mu(T_i, S_i, p^*(T))\Big).$$

Clearly $\hat{p}(p) = p$ is a feasible strategy and it will bring the same payoff as the equilibrium payoff of $\sigma(T, T_i)$. Consequently, the optimal value is weakly higher than what the producer could obtain in $\sigma(T, T_i)$. ■

So far, we have shown that we could assume without loss of generality that the producer receives a signal T, and consumer receives a signal (S_i, T). Thus, we can focus on equilibria where prices have no signalling effect. These equilibria coincide with those described in the proof of Proposition 1. As we have calculated there, the profit of the producer is:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{E}[w_i|T]}{2}\left(\mathbb{E}[w_i|T \cup S_i] - \frac{\mathbb{E}[w_i|T]}{2}\right)\right], \\
&= \frac{\mathbb{E}\left[(\mathbb{E}[w_i|T])^2\right]}{4} = \frac{[\,\mathbb{E}[w_i|T]] + \mu_\theta^2}{4}.
\end{aligned}$$

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

Proof of Theorem 2. Denote the non-aggregated inflow on and off the equilibrium path as X and X_{-i}, respectively. Similarly, denote the aggregated inflows as $\bar{X}$ and $\bar{X}_{-i}$. By Proposition 5, the information outflow coincides with inflow, and using the expressions in (43) and (44) in the proof of Proposition 1, the profit of the intermediary under non-aggregate information is:

$$\begin{aligned}
\frac{R\,(X)}{N} &= \Pi((S_i, X_i), X) - \Pi(S_i, \varnothing) - U_i((S_i, X_{-i}), X_{-i}) + U_i((S_i, X), X), \\
&= \frac{1}{2}\mathbb{E}\left[(\mathbb{E}[w_i|X \cup S_i] - \mathbb{E}[w_i|X])^2 + \frac{1}{4}(\mathbb{E}[w_i|X])^2\right] + \frac{[\mathbb{E}[w_i|X]]}{2}, \\
&\quad -\frac{1}{2}\mathbb{E}[(\mathbb{E}[w_i|X_{-i} \cup S_i] - \mathbb{E}[w_i|X_{-i}])^2 + \frac{1}{4}(\mathbb{E}[w_i|X_{-i}])^2].
\end{aligned}$$

However, notice that $X \cup S_i = X_{-i} \cup S_i$, and so we have

$$\begin{aligned}
\frac{R(X)}{N} &= \frac{1}{2}\mathbb{E}\left[-(\mathbb{E}[w_i|X])^2 + \frac{1}{4}(\mathbb{E}[w_i|X])^2\right] \\
-\frac{1}{2}\mathbb{E}\left[-(\mathbb{E}[w_i|X_{-i}])^2 + \frac{1}{4}(\mathbb{E}[w_i|X_{-i}])^2\right] &+ \frac{\text{var}[\mathbb{E}[w_i|X]]}{2}, \\
= -\frac{3}{8}\,\text{var}[\mathbb{E}[w_i|X]] + \frac{3}{8}\,\text{var}[\mathbb{E}[w_i|X_{-i}]] &+ \frac{\text{var}[\mathbb{E}[w_i|X]]}{2}, \\
= -\frac{1}{8}\,\text{var}[\mathbb{E}[w_i|X]] + \frac{3}{8}\,\text{var}[\mathbb{E}[w_i|X_{-i}]]&.
\end{aligned}$$

Under the aggregate scheme, since we assumed consumer i knows her own report, the profit of the intermediary is:

$$\frac{1}{N}R\left(\bar{X}\right) = \Pi((S_i, \bar{X}), \bar{X}) - \Pi(, \varnothing) - U_i((S_i, \bar{X}_{-i}), \bar{X}_{-i}) + U_i((S_i, \bar{X}, X_i), \bar{X}).$$

Notice that $\bar{X} \cup S_i \cup X_i = \bar{X}_{-i} \cup S_i$. By the same derivation as above, we obtain:

$$\begin{aligned}
\frac{1}{N}R\left(\bar{X}\right) &= -\frac{1}{8}\,\text{var}[\mathbb{E}[w_i|\bar{X}]] + \frac{3}{8}\,\text{var}[\mathbb{E}[w_i|\bar{X}_{-i}]], \\
&= -\frac{1}{8}\,\text{var}[\mathbb{E}[w_i|\bar{X}]] + \frac{3}{8}\,\text{var}[\mathbb{E}[w_i|X_{-i}]], \\
&\geq -\frac{1}{8}\,\text{var}[\mathbb{E}[w_i|X]] + \frac{3}{8}[\mathbb{E}[w_i|X_{-i}]], \\
&= \frac{1}{N}R(X).
\end{aligned}$$

To prove the inequality is strict, and to give a explicit expression of the profit, we have

$$\begin{aligned}
R\left(\bar{X}\right) &= -\frac{N}{8}\,\text{var}[\mathbb{E}[w_i|\bar{X}]] + \frac{3N}{8}\,\text{var}[\mathbb{E}[w_i|\bar{X}_{-i}]], \\
&= -\frac{N}{8}\,\text{var}\left[\mathbb{E}\left[\frac{\Sigma_i w_i}{N}\,\middle|\bar{X}\right]\right] + \frac{3N}{8}\,\text{var}[\ \ \mathbb{E}[w_i|X_{-i}]], \\
= \frac{3(N-1)N\sigma_\theta^4}{8((N-1)(\sigma_\varepsilon^2+\sigma_\xi^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\xi_i}^2+\sigma_{\theta_i}^2)} &- \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8(N(\sigma_\varepsilon^2+\sigma_\xi^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\xi_i}^2+\sigma_{\theta_i}^2)},
\end{aligned}$$

where the second line comes from the fact that $\mathbb{E}[w_i|X_{-i}]$ places the same linear weight on all x_j so that $\mathbb{E}[w_i|X_{-i}] = \mathbb{E}[w_i|\bar{X}_{-i}]$. (Note that this property only uses the symmetry of the joint distribution, not normality.)

In these expressions, the "true" noise ε and the additional noise ξ are equivalent in terms of the profit. With some abuse of notation, we denote $\underline{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2$, $\sigma_\varepsilon^2 = \sigma_\varepsilon^2 + \sigma_\xi^2$. The expression

of the profit can be rewritten as:

$$R\left(\bar{X}\right)=\frac{3(N-1)N\sigma_{\theta}^{4}}{8((N-1)(\sigma_{\varepsilon}^{2}+\sigma_{\theta}^{2})+\sigma_{\varepsilon_i}^{2}+\sigma_{\theta_i}^{2})}-\frac{(N\sigma_{\theta}^{2}+\sigma_{\theta_i}^{2})^{2}}{8(N(\sigma_{\varepsilon}^{2}+\sigma_{\theta}^{2})+\sigma_{\varepsilon_i}^{2}+\sigma_{\theta_i}^{2})}. \tag{47}$$

The true noise $\underline{\sigma}_{\varepsilon}^{2}$ places a lower bound on the total noise σ_{ε}^{2}. ■

Proof of Theorem 3. We first prove that the optimal additional idiosyncratic noise level is zero: $\sigma_{\xi_i}^{*}=0$. To show this result, suppose $\sigma_{\varepsilon_i}^{*}>\underline{\sigma}_{\varepsilon_i}$. 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 $\bar{\sigma}_{\varepsilon_i}^{2}\triangleq\sigma_{\varepsilon_i}^{2}-(N-1)\delta^{2}\geq\underline{\sigma}_{\varepsilon}^{2}$, the profits R will strictly increase. Too see this, notice that

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

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 6. We establish, equivalently, that the intermediary can obtain positive profits if 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:

$$\begin{aligned}R&=\frac{3n\sigma_{\theta}^{4}}{8(\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2})}\left(1-\frac{\sigma_{\theta_i}^{2}+\sigma_{\varepsilon_i}^{2}}{(N-1)(\sigma_{\varepsilon}^{2}+\sigma_{\theta}^{2})+\sigma_{\varepsilon_i}^{2}+\sigma_{\theta_i}^{2}}\right)\\&\quad-\frac{(n\sigma_{\theta}^{2}+\sigma_{\theta_i}^{2})^{2}}{8n(\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2})}\left(1-\frac{\sigma_{\theta_i}^{2}+\sigma_{\varepsilon_i}^{2}}{N(\sigma_{\varepsilon}^{2}+\sigma_{\theta}^{2})+\sigma_{\varepsilon_i}^{2}+\sigma_{\theta_i}^{2}}\right)\\&\leq\left(\frac{3n\sigma_{\theta}^{4}}{8(\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2})}-\frac{(n\sigma_{\theta}^{2}+\sigma_{\theta_i}^{2})^{2}}{8n(\sigma_{\theta}^{2}+\sigma_{\varepsilon}^{2})}\right)\left(1-\frac{\sigma_{\theta_i}^{2}+\sigma_{\varepsilon_i}^{2}}{N(\sigma_{\varepsilon}^{2}+\sigma_{\theta}^{2})+\sigma_{\varepsilon_i}^{2}+\sigma_{\theta_i}^{2}}\right)<0.\end{aligned} \tag{48}$$

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

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

Therefore the intermediary can obtain a positive profit R by setting σ_{ε}^{2} sufficiently large. ■

Proof of Proposition 7. We first give an explicit expression for the optimal additional noise. By Proposition 6, we can focus on the case where

$$N\left(\sqrt{3}-1\right)\sigma_\theta^2-\sigma_{\theta_i}^2>0, \tag{49}$$

i.e., where the intermediary can obtain positive profits. Straightforward algebra then yields

$$\frac{\partial R}{\partial\sigma_\varepsilon^2}=\frac{A\sigma_\varepsilon^4+B\sigma_\varepsilon^2+C}{8\left(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2\right)^2\left((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2\right)^2},$$

where

$$A=-(N-1)^2N\left(2N^2\sigma_\theta^4-2N\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right).$$

Therefore the numerator is a quadratic function of σ_ε^2 with a negative quadratic term. It has two roots, and the smaller one is given by

$$\begin{aligned}&\frac{-2(N-1)N^2\sigma_\theta^6-(\sqrt{3}-1)N\sigma_\theta^4\sigma_{\theta_i}^2+(3N-1)\sigma_\theta^2\sigma_{\theta_i}^4-\sqrt{3}\sigma_\theta^2\sigma_{\theta_i}^4+\sigma_{\theta_i}^6}{(N-1)\left(2N^2\sigma_\theta^4-2N\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right)}\\&+\frac{\left(-N(2N-3)\sigma_\theta^4+2N\sigma_\theta^2\sigma_{\theta_i}^2-\sqrt{3}\sigma_\theta^2(N\sigma_\theta^2+\sigma_{\theta_i}^2)+\sigma_{\theta_i}^4\right)}{(N-1)\left(2N^2\sigma_\theta^4-2N\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right)}\sigma_{\varepsilon_i}^2.\end{aligned}$$

When (49) holds, the denominator is positive while the nominator is negative so that the smaller root is always negative. Therefore the optimal σ_ε^2 (for fixed $\sigma_{\varepsilon_i}^2$) is either $\underline{\sigma_\varepsilon^2}$ or the larger root of the quadratic function, which can be simplified to

$$\begin{aligned}\sigma_\varepsilon^{2*}(\sigma_{\varepsilon_i}^2)=\max\Bigg\{&\underline{\sigma_\varepsilon^2},\frac{-2(N-1)N^2\sigma_\theta^6-(\sqrt{3}-1)N\sigma_\theta^4\sigma_{\theta_i}^2+(3N-1)\sigma_\theta^2\sigma_{\theta_i}^4-\sqrt{3}\sigma_\theta^2\sigma_{\theta_i}^4+\sigma_{\theta_i}^6}{(N-1)\left(2N^2\sigma_\theta^4-2N\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right)}+\\&\frac{\left(-N(2N-3)\sigma_\theta^4+2N\sigma_\theta^2\sigma_{\theta_i}^2-\sqrt{3}\sigma_\theta^2(N\sigma_\theta^2+\sigma_{\theta_i}^2)+\sigma_{\theta_i}^4\right)}{(N-1)\left(2N^2\sigma_\theta^4-2N\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right)}\underline{\sigma_{\varepsilon_i}^2}\Bigg\},\\=\max\Big\{&\underline{\sigma_\varepsilon^2},A'\underline{\sigma_{\varepsilon_i}^2}+B'\Big\}.\end{aligned} \tag{50}$$

Then clearly $\sigma_\xi^{*2}=\sigma_\varepsilon^{*2}-\underline{\sigma_\varepsilon^2}$ is weakly increasing with respect to σ_e if $A'\geq\beta/(1-\beta)$ and weakly decreasing otherwise. Since the right hand side is increasing in β while the right hand side does not depend on β, this proves the first part of the result.

To prove the second part, we simply calculate the derivative:

$$\frac{\partial A'}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left(\frac{\alpha\left(\alpha(N-1)\left(-2N+\sqrt{3}-1\right)+2N+\sqrt{3}-2\right)+1}{(N-1)(\alpha(\alpha(2N(N+1)-1)-2N+2)-1)}\right),$$
$$= -\frac{\alpha\left(\left(4\sqrt{3}+6\right)\alpha N^2+2\left(\sqrt{3}+3\right)(1-\alpha)N+\sqrt{3}(\alpha-2)\right)+\sqrt{3}}{(N-1)(\alpha(-2\alpha N(N+1)+\alpha+2(N-1))+1)^2} < 0.$$

Similarly, we compute the derivative of the term A' with respect to N. Up to positive multiplicative terms, we have

$$\frac{\partial A'}{\partial N} \propto (\alpha-1)^2\left(\left(\sqrt{3}-2\right)\alpha(2\alpha-1)-1\right)-4\alpha^4N^4+4\left(\sqrt{3}\alpha+\alpha+2\right)\alpha^3N^3$$
$$-2\left(3\sqrt{3}\alpha-2\sqrt{3}+3\right)\alpha^3N^2+4(\alpha-1)\left(\alpha\left(\alpha+\sqrt{3}-2\right)+1\right)\alpha N.$$

Finally, one can show that this expression is negative for all values of (α, N) for which the solution $\beta^*(\alpha, N)$ to the equation $A' = \beta/(1-\beta)$ satisfies $\beta^* \in [0,1]$.

For the third part of the proposition, note first that the optimal profits of the intermediary R are weakly decreasing in σ_e^2 by construction: the data intermediary can always add noise terms σ_ξ^2 and $\sigma_{\xi_i}^2$ to supplement the initial noise levels. Thus, a larger σ_e^2 cannot be strictly more profitable than a lower one.

Next, we show that profits are convex in $\sigma_{\varepsilon_i}^2$ when $R > 0$. Indeed, we have

$$\frac{\partial^2 R}{\partial(\sigma_{\varepsilon_i}^2)^2} = \frac{3(N-1)n\sigma_\theta^4}{4(\sigma_\varepsilon^2(N-1)+\sigma_{\varepsilon_i}^2+(N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^3} - \frac{(n\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{4(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^3},$$
$$\geq \frac{1}{(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^2}\left(\frac{3(N-1)n\sigma_\theta^4}{4(\sigma_\varepsilon^2(N-1)+\sigma_{\varepsilon_i}^2+(N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)} - \frac{(n\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{4(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\right)$$
$$= \frac{2}{(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^2}R > 0.$$

Now we turn to the intermediary's profits under the optimal data policy. Recall it is always optimal to set $\sigma_{\xi_i}^* = 0$ according to Theorem 3. Denote $R(\sigma_e^2, \sigma_\varepsilon)$ as the profit when the initial noise level is σ_e^2 while the final common noise (after information design) is $\sigma_\varepsilon^2 = \beta\sigma_e^2 + \sigma_\xi^{*2}$ Consider any point $\sigma_{e'}^2$ such that $\sigma_\xi^*(\sigma_{e'}^2) > 0$. By continuity, there exists an open interval $I = (\underline{\sigma}_e^2, \overline{\sigma}_e^2)$ around σ_e^2 such that for all $\sigma_e^2 \in I$, $\sigma_\varepsilon^*(\sigma_e^2) > \overline{\sigma}_e^2 > \sigma_e^2$. Now by a standard argument we can prove the profit under the optimal information design $R(\sigma_e^2, \sigma_\varepsilon^*(\sigma_e^2))$ is convex in I. Consider $\sigma_{e_1}^2$, $\sigma_{e_2}^2$, and $\sigma_{e_3}^2$ in I such that:

$$\sigma_{e_3}^2 = \mu\sigma_{e_1}^2 + (1-\mu)\sigma_{e_2}^2.$$

we have:

$$\begin{aligned}
&\mu R(\sigma_{e_1}^2, \sigma_\varepsilon^*(\sigma_{e_1}^2)) + (1-\mu) R(\sigma_{e_2}^2, \sigma_\varepsilon^*(\sigma_{e_2}^2)), \\
&\geq \mu R(\sigma_{e_1}^2, \sigma_\varepsilon^*(\sigma_{e_3}^2)) + (1-\mu) R(\sigma_{e_2}^2, \sigma_\varepsilon^*(\sigma_{e_3}^2)), \\
&> R(\sigma_{e_3}^2, \sigma_\varepsilon^*(\sigma_{e_3}^2)),
\end{aligned}$$

where the first inequality comes from the fact that $\sigma_\varepsilon^*(\sigma_{e_3}^2)$ is feasible by the construction of I, and the second inequality comes from the convexity of R. Given other parameters fixed, $R(\sigma_e^2, \sigma_\varepsilon^*(\sigma_e^2))$ is clearly second-order differentiable with respect to σ_e^2 in the region where $\sigma_\xi^* > 0$, and this region is an interval since σ_ξ^* is decreasing in σ_e^2 . Therefore $R(\sigma_e^2, \sigma_\varepsilon^*(\sigma_e^2))$ is convex when $\sigma_\xi^* > 0$. ■

Proof of Theorem 4. It is easy to see from equation (50) for the optimal common noise level that for N sufficiently large, it is optimal to not add any noise: $\sigma_{\xi_i}^* = \sigma_\xi^* = 0$. Therefore when N is large, the total compensation owed to consumers under the optimal policy can be written as

$$\begin{aligned}
&\sum_{i=1}^N m_i = \sum_{i=1}^N U_i((S_i, \bar{X}_{-i}), \bar{X}_{-i}) - U_i((S_i, \bar{X}), \bar{X}), \\
&= \frac{3}{8}\left(\frac{(\alpha(N-1)+1)^2}{\alpha(N-1)+\beta\lambda(N-1)+\lambda+1} - \frac{\alpha^2(N-1)N}{\alpha(N-2)+\beta\lambda(N-2)+\lambda+1}\right).
\end{aligned}$$

The limit of the compensation is positive but finite:

$$\lim_{N\to\infty} \sum_{i=1}^N m_i = \frac{3(1-\alpha)\alpha}{4(\alpha+\beta\lambda)},$$

Furthermore, the derivative is asymptotically negative if α is sufficiently large:

$$\frac{\partial}{\partial N}\sum_{i=1}^N m_i|_{\alpha=1} = -\frac{3(1-\beta)\lambda\left((\beta\lambda N+N)^2 + (1-\beta)\lambda((2\beta-1)\lambda+1)\right)}{8(\beta\lambda(N-2)+\lambda+N-1)^2(\beta\lambda(N-1)+\lambda+N)^2}.$$

And finally, profits grow linearly, because we have

$$\lim_{N\to\infty} \frac{R}{N} = \frac{\alpha^2}{4(\alpha+\beta\lambda)}.$$

This completes the proof. ■

Proof of Proposition 8. Consider two groups of N consumers. We first derive the expression of the profit in the two cases: uniform pricing and group pricing. In the case of uniform pricing, the producer does not have any identification information. He can thus

only provide one price for every consumer:

$$p_{ij}(S) = \frac{1}{4N}\mathbb{E}[\sum_{j=1}^{2}\sum_{i=1}^{N} w_{ij} \mid S] = \frac{1}{4N}\sum_{j=1}^{2}\sum_{i=1}^{N} s_{ij}.$$

The last equation holds since we only consider noiseless signals in this section. Off the equilibrium path, the intermediary will use the remaining $2N-1$ signals to estimate the willingness to pay of the deviating consumer, since he does not know which group this consumer is coming from:

$$p_{ij}(S_{-ij}) = \frac{1}{4}\mathbb{E}[\theta_1 + \theta_2 | S_{-ij}] = \frac{(2N-1)\sigma_\theta^2}{4((2N^2-2N+1)\sigma_\theta^2 + (2N-1)\sigma_{\theta_i}^2)}\sum_{i'j' \neq ij} s_{i'j'}.$$

The revenue of the intermediary is given by

$$\begin{aligned}\underline{R}(S) &= -N\,\mathrm{var}[p_{ij}(S)] + \frac{3}{2}\sum_{ij}\mathrm{var}[p_{ij}(S_{-ij})],\\ &= -\left(\frac{N}{8}\sigma_\theta^2 + \frac{1}{8}\sigma_{\theta_i}^2\right) + \frac{3N}{16}\frac{(2N-1)^2\sigma_\theta^4}{(2N^2-2N+1)\sigma_\theta^2 + (2N-1)\sigma_{\theta_i}^2}. \end{aligned} \tag{51}$$

On the other hand, in the group pricing scheme, the producer knows which group the consumer comes from, and hence the analysis is as in the baseline model. We immediately obtain the expression of the intermediary's profits:

$$\overline{R}(S) = \frac{3(N-1)N\sigma_\theta^4}{4((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{4(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}. \tag{52}$$

We then compare the derivatives of (51) and (52) with respect to N:

$$\begin{aligned}\frac{\partial(\overline{R}-\underline{R})}{\partial N} &= -\frac{3(2N-1)\sigma_\theta^2((2N(N(2N-3)+3)-1)\sigma_\theta^2 + 2N(4N-3)\sigma_{\theta_i}^2+\sigma_{\theta_i}^2)}{(2(N-1)N\sigma_\theta^2+(2N-1)\sigma_{\theta_i}^2+\sigma_\theta^2)^2} - 2\\ &\quad + \frac{12\sigma_\theta^2\left((N-1)^2\sigma_\theta^2+(2N-1)\sigma_{\theta_i}^2\right)}{((N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2},\\ &\geq -\frac{3(2N-1)\sigma_\theta^2((2N(N(2N-3)+3)-1)\sigma_\theta^2 + 2N(4N-3)\sigma_{\theta_i}^2+\sigma_{\theta_i}^2)}{((2N-1)((N-1)\sigma_\theta^2+\sigma_{\theta_i}^2))^2} - 2\\ &\quad + \frac{12\sigma_\theta^2\left((N-1)^2\sigma_\theta^2+(2N-1)\sigma_{\theta_i}^2\right)}{((N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2},\\ &= \frac{(2N(4(N-4)N+11)-7)\sigma_\theta^4 + (2N-1)(8N-5)\sigma_\theta^2\sigma_{\theta_i}^2 + 2(1-2N)\sigma_{\theta_i}^4}{(2N-1)((N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2}.\end{aligned}$$

Note that $4N\sigma_\theta^2 > \sigma_{\theta_i}^2$ is sufficient for the last expression to be positive. Furthermore, when

$4N\sigma_\theta^2 \leq \sigma_{\theta_i}^2$, neither pricing scheme yields a positive profit to the intermediary. This is true for the case of uniform pricing, since

$$\begin{aligned}\underline{R} &= -(\frac{N}{8}\sigma_\theta^2 + \frac{1}{8}\sigma_{\theta_i}^2) + \frac{3N}{16}\frac{(2N-1)^2\sigma_\theta^4}{(2N^2-2N+1)\sigma_\theta^2 + (2N-1)\sigma_{\theta_i}^2}, \\ &\leq -\frac{5N}{8}\sigma_\theta^2 + \frac{3N}{16}\frac{(2N-1)^2\sigma_\theta^2}{(2N^2-2N+1)+4N(2N-1)}, \\ &< -\frac{5N}{8}\sigma_\theta^2 + \frac{3N}{16}\frac{2N-1}{N-1+4N}\sigma_\theta^2 < 0.\end{aligned}$$

It is also true for group pricing, because Proposition 6 showed that, when $\left(\sqrt{3}-1\right)N\sigma_\theta^2 < \sigma_{\theta_i}^2$, group pricing is not profitable, and this condition is implied by $4N\sigma_\theta^2 \leq \sigma_{\theta_i}^2$.

Finally, imposing $4N\sigma_\theta^2 = \sigma_{\theta_i}^2$ we have

$$\begin{aligned}\overline{R} - \underline{R} &= \frac{N(N(40(7-11N)N-53)+1)\sigma_\theta^2}{16(5N-1)(2N(5N-3)+1)} < 0, \text{ and} \\ \lim_{N\to\infty}\frac{\overline{R}-\underline{R}}{N} &= \frac{\sigma_\theta^2}{4} > 0.\end{aligned}$$

Therefore there exists at most a single threshold $\overline{N}$ such that group pricing is more profitable if and only if $N > \overline{N}$.

Now we prove the second part of the result:

$$\begin{aligned}\overline{R} - \underline{R} &= \frac{1}{16}\left(\alpha N\left(-\frac{3\alpha(1-2N)^2}{2\alpha(N-1)^2+2N-1} + \frac{12\alpha(N-1)}{\alpha(N-2)+1} - 2\right) + 2(\alpha-1)\right), \\ \frac{\partial\overline{R}-\underline{R}}{\partial\alpha} &= \frac{2\left(\alpha^4(N-2)(N-1)^2(N(N(8(N-2)N+9)+10)-8)+(N-1)\left(-(1-2N)^2\right)\right)}{(\alpha N-2\alpha+1)^2\left(2\alpha N^2-4\alpha N+2\alpha+2N-1\right)^2} \\ &+\frac{2\left(\alpha^3(N(4N-9)+4)(N(N(8(N-2)N+9)+10)-8)\right)}{(\alpha N-2\alpha+1)^2\left(2\alpha N^2-4\alpha N+2\alpha+2N-1\right)^2} \\ &+\frac{2\left(\alpha^2(N(N(2N(6N(3N-8)+31)+67)-87)+24)\right)}{(\alpha N-2\alpha+1)^2\left(2\alpha N^2-4\alpha N+2\alpha+2N-1\right)^2} \\ &+\frac{2\left(\alpha\left(N\left(8N^3-36N+33\right)-8\right)\right)}{(\alpha N-2\alpha+1)^2\left(2\alpha N^2-4\alpha N+2\alpha+2N-1\right)^2}.\end{aligned}$$

The numerator is increasing in N because its partial derivative with respect to N is:

$$\begin{aligned}&2\left((4\alpha^3(N-2)(N-1)^2+3\alpha^2(N(4N-9)+4))(N(N(8(N-2)N+9)+10)-8)\right) \\ &+2\left(2\alpha(N(N(2N(6N(3N-8)+31)+67)-87)+24)+N\left(8N^3-36N+33\right)-8\right) > 0.\end{aligned}$$

Thus it is sufficient to prove the numerator is positive when $N(4\alpha+1) = 1$, because we

can focus on the region where $4N\sigma_\theta^2 > \sigma_{\theta_i}^2$. Plugging in $N(4\alpha+1)=1$ and simplifying, we obtain that the numerator is positive for all $N \geq 3$. This completes the proof. ■

Proof of Proposition 9. We have shown in the proof of Proposition 8 above that, as long as $4N\sigma_\theta^2 > \sigma_{\theta_i}^2$, we have

$$\frac{\partial(\overline{R}-\underline{R})}{\partial N} > 0.$$

Furthermore, when $4N\sigma_\theta^2 \leq \sigma_{\theta_i}^2$, neither pricing scheme yields positive profits. Therefore when either of the schemes is profitable, we must have $4N\sigma_\theta^2 > \sigma_{\theta_i}^2$, and thus the derivatives with respect to N are ranked the same for all N. ■

Proof of Proposition 10. Since $\sigma_e = 0$, i.e., consumers know their own type, the only relevant information is the information transmitted to the producer. Thus in the following two propositions, we will simply use S (instead of (S,T)) to represent the information structure. The consumer's demand under quality and price discrimination is given by

$$q_i = w_i + \gamma y_i - p_i,$$

where we restrict attention to the case where the complete-information outcome is well-defined, i.e., $\gamma < \sqrt{2}$. Under data outflow S, the producer offers the following price and quantity levels:

$$p_i(S) = \frac{1}{2-\gamma^2}\mathbb{E}[w_i|S],$$
$$y_i(S) = \frac{\gamma}{2-\gamma^2}\mathbb{E}[w_i|S].$$

The producer's profit is given by:

$$\Pi(S) = \sum\nolimits_{i=1}^{N} \mathbb{E}[(w_i+\gamma y_i - p_i)p_i - y_i^2/2] = \frac{1}{2(2-\gamma^2)}\sum\nolimits_{i=1}^{N}\mathrm{var}[w_i|S] + \Pi(\varnothing),$$

which is increasing in the amount of information conveyed by S. The surplus of consumer i is given by

$$U_i(S) = \frac{1}{2}\mathbb{E}\left[(w_i+\gamma y_i - p_i)^2\right] = -\frac{(3-\gamma^2)(1-\gamma^2)}{2(2-\gamma^2)^2}\mathrm{var}[w_i|S] + U_i(\varnothing).$$

For $0 \leq \gamma < 1$, information reduces consumer surplus, while for $1 < \gamma < \sqrt{2}$, information

increases consumer surplus. Finally, the social surplus is given by

$$W(S) = \left(\frac{1}{2(2-\gamma^2)} - \frac{(3-\gamma^2)(1-\gamma^2)}{2(2-\gamma^2)^2} \right) \sum_{i=1}^{N} \text{var}[w_i|S] + W(\varnothing).$$

It is straightforward to verify that information increases social surplus if and only if $\gamma > \gamma^*$, where $\gamma^* = \sqrt{(3-\sqrt{5})/2} < 1$. The profit of intermediation is:

$$R(S) = W(S) - \Pi(\varnothing) - \sum_{i=1}^{N} U_i(S_{-i}).$$

By the same argument as the proof of Theorem 2, anonymization does not affect $\Pi(\varnothing)$ and $U_i(S_{-i})$, and anonymization increases $W(S)$ if and only if $\gamma < \gamma^*$. Therefore, anonymization increases R if and only if $\gamma < \gamma^*$. ■

Proof of Proposition 11. Each consumer demands

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

This means the producer's profit is given by

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

Therefore, under any information structure S, the producer offers

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

where the second line relies on the fact that the underlying random variables are normal so

that $t_i - \mathbb{E}[t_i|S]$ is independent of S . The consumer's surplus is then given by

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

Therefore the difference is:

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

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

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

Similarly we have:

$$\Pi_i(S) = \mathbb{E}\left[\left(w_i - (t_i - \mathbb{E}[t_i \mid S])^2 - \frac{\mathbb{E}[w_i \mid S] - \mathbb{E}[(t_i - \mathbb{E}[t_i \mid S])^2]}{2}\right)\frac{\mathbb{E}[w_i \mid S] - \mathbb{E}[(t_i - \mathbb{E}[t_i \mid S])^2]}{2}\right]$$

$$= \frac{1}{4}\mathbb{E}\left[\left(\mathbb{E}[w_i \mid S] - \mathbb{E}\left[(t_i - \mathbb{E}[t_i \mid S])^2\right]\right)\right]$$

$$= \frac{1}{4}\mathbb{E}\left[\mathbb{E}[w_i \mid S]^2 - 2\mathbb{E}[w_i \mid S]\,\mathbb{E}\left[(t_i - \mathbb{E}[t_i \mid S])^2\right] + \mathbb{E}\left[(t_i - \mathbb{E}[t_i \mid S])^2\right]^2\right].$$

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

$$W_i(S) - W_i(\varnothing) = U_i(S) - U_i(\varnothing) + \Pi_i(S) - \Pi_i(\varnothing),$$

$$U_i(S) - U_i(\varnothing) - \frac{1}{8}\operatorname{var}[\mathbb{E}[w_i \mid S]] + \frac{3}{4}\mu_\theta \operatorname{var}[\mathbb{E}[t_i \mid S]] + \frac{11}{8}\left(\operatorname{var}[\mathbb{E}[t_i \mid S]]^2 - 2\operatorname{var}[\mathbb{E}[t_i \mid S]](\sigma_\tau^2 + \sigma_{\tau_i}^2)\right).$$

Therefore we can write the difference as a quadratic function of the variance of the conditional expectation $x \triangleq \operatorname{var}[\mathbb{E}[t_i|S]]$. In particular, we let

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

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

Finally, as in the proof of Propositions 6 and 10, aggregating w_i increases $W_i(S)$ while keeps $\Pi(\varnothing)$ and $U_i(S_{-i})$ unchanged. Also not aggregating t_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 t_i. ■

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, N., G. Lewis, and S. Vasserman (2019): "Voluntary Disclosure and Personalized Pricing," Discussion paper, Pennsylvania State University.

Argenziano, R., and A. Bonatti (2020): "Information Revelation and Privacy Protection," Discussion paper, MIT.

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.

Bergemann, D., A. Bonatti, and T. Gan (2020): "The Economics of Social Data," Discussion Paper 2203R, Cowles Foundation for Research in Economics, Yale University.

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

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

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.

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.

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.

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.