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

Recommended Citation

Bergemann, Dirk; Bonatti, Alessandro; and Gan, Tan, "The Economics of Social Data" (2019). *Cowles Foundation Discussion Papers*. 46.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/46

This Discussion Paper is brought to you for free and open access by the Cowles Foundation at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Cowles Foundation Discussion Papers by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

THE ECONOMICS OF SOCIAL DATA

By

Dirk Bergemann, Alessandro Bonatti, and Tan Gan

September 2019

Revised March 2020

COWLES FOUNDATION DISCUSSION PAPER NO. 2203R



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§

March 2, 2020

Abstract

A data intermediary pays consumers for information about their preferences and sells the information so acquired to firms that use it to tailor their products and prices. The social dimension of the individual data—whereby an individual's data are predictive of the behavior of others—generates a *data externality* that reduces the intermediary's cost of acquiring information. We derive the intermediary's optimal data policy and show that it preserves the privacy of the consumers' identities while providing precise information about market demand to the firms. This enables the intermediary to capture the entire value of information as the number of consumers grows large.

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

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, Enrico Moretti, Stephen Morris, Denis Nekipelov, Asu Özdağlar, Fiona Scott-Morton, and Glen Weyl for helpful discussions. We thank the audiences at ASSA 2019, ESSET 2019 and the ACM-EC 2019 Plenary Lecture for 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 Internet platforms–such as Facebook, Google, and Amazon in the US, and of similar large entities in China, such as JD, Tencent and Alibaba–has led to an unprecedented collection and commercial use of individual data. The ever-increasing user bases of these platforms generate massive amounts of data about individual consumers: their preferences, their locations, their friends, their political views, and almost all other facets 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]

The recent disclosures on the use and misuse of social data by Internet platforms have prompted regulators to limit the hitherto largely unsupervised use of individual data by these companies. As a result, nearly all proposed and enacted regulation to date is centered on ensuring that consumers retain control over their data. The idea is that, because consumer data must be acquired, aggregated, packaged, and sold, the allocation of control rights impacts the ultimate use of the data.[2] In particular, by assigning ownership and control to individual consumers, regulators hope to enable an efficient use of information or at least to grant individual consumers the appropriate compensation for the data they reveal.[3] Despite the appeal of the Coase theorem implicitly invoked in the regulation, it is far from evident that the resulting allocation of information is efficient due to multiple potential market failures. In this paper, we explore the implications of externalities across consumers.

A central feature of the data collected from individuals is its social aspect. Namely, the data captured from an individual user are not only informative about that specific individual but also about appropriately defined nearby individuals. Thus, these *individual data* are really *social data*. The social nature of the data generates a *data externality*, the sign and magnitude of which depend on the ultimate use of the information so gained. In the context of geolocation data, for instance, an individual user conveys information about the traffic conditions for nearby drivers. In the context of shopping data, an individual's purchases convey information about the willingness to pay for a given product among consumers with similar purchase histories.

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

[2] The aggregation of consumers into targeting categories is especially important on large platforms that intermediate the exchange of information about tastes for products between consumers and advertisers.

[3] The Stigler Committee on Digital Platforms (2019) notes that "Many technology platforms are distinctive because they provide valued services to consumers without charging a monetary price. Instead, consumers barter their attention and data to the platforms in exchange for these services. The platforms use that attention and data to generate monetary payments from advertisers."

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, advertisers, and large Internet 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 they provide in equilibrium about individual consumers.

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

In the product market, each consumer (she) determines the quantity she wishes to purchase, and a single producer (he) sets the unit price at which to offer a product to the consumers. The product market features demand uncertainty, and 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 in some possibly aggregated or noisy version to the producer. The data intermediary can choose how much information to buy from the consumers and how much information to resell to the producer. The (upstream) data market and the (downstream) product market are linked by the pricing decision of the producer. The consumers' willingness to sell their personal data depends on how the pricing policy of the producer responds to the data so acquired.

Knowledge of the demand data has distinct implications for consumers and the producer. The demand information allows the producer to engage in price discrimination, which occurs at the individual or at the market level, depending on the data aggregation level chosen by the intermediary. For any information structure, however, the producer gains revenue by tailoring his price to the demand. Conversely, in our model, the consumer loses in utility due to the distortion in her consumption entailed by a more responsive price.[4] Consequently, the social value of information is also negative in our setting.

[4]It is well known (e.g., Bergemann, Brooks, and Morris (2015)) that price discrimination can be beneficial to consumers. In this case, we deliberately choose a model where information reduces total surplus (and, hence, a fortiori, consumer surplus) to highlight how the data externality enables profitable intermediation to take place.

The Value and Price of Social Data The social dimension of the data—whereby a consumer's data are also predictive of the behavior of others—is critical to understand the consumer's incentives to share her data with a large platform. A naive argument suggests that if a consumer is empowered to take control of her data and anticipates any negative consequences of revealing her information, she will demand compensation from the intermediary (e.g., through the quality of the services received). In turn, the need to compensate consumers will disrupt the intermediaries' business model if the transmission of information to the downstream producers reduces total surplus.

However, this argument ignores the social aspect of the data. The consumer's choice to provide information is guided only by her private benefits and costs, i.e., the externality generated by the data she provides is not considered in her decision making. Thus, the intermediary has to 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 for any changes in her welfare caused by information revealed by others. Consequently, the cost of acquiring individual data can be substantially below the value of information to the platform.

We can now see how social data drive a wedge between the efficient and profitable uses of information. While 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. The presence of the data externality thus indicates that the mere allocation of ownership and control to consumers does not ensure efficient information trade in these markets. In our model, the difference between the large revenues associated with selling the information about many consumers and the small compensation necessary to acquire that information means the intermediary can profitably trade data even if information is socially harmful.

The presence of a data externality can therefore provide a novel explanation for the *digital privacy paradox* experimentally documented by Athey, Catalini, and Tucker (2017), whereby small monetary incentives have large effects on subjects' willingness to relinquish their private data. In practice, this force also likely drives the extraordinary appetite of Internet platforms to gather information.[5]

What restrictions, if any, does the allocation of control rights to consumers over their data then impose on the equilibrium trading of information? While the data externality facilitates the diffusion of *some* information about consumers, we find that *how* the informa-

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

tion is collected depends on the fundamental characteristics of both the upstream and the downstream market. In Section 4, we show that when consumers are homogeneous ex ante and producers use information to set prices, the intermediary prefers to collect aggregate, market-level information. This means that the intermediary does not enable the producer to set personalized prices: the data are transmitted fairly precisely but disconnected from the users' personal profiles. In other words, although from a technological standpoint the data are freely available to the intermediary, the fundamental role of social data is more subtle in determining the modality of information acquisition and use.[6]

A more nuanced version of this result emerges when we extend the model to accommodate heterogeneous groups of consumers in Section 5. Indeed, we find that data are aggregated at least at 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 one can interpret as discriminatory on observable characteristics, such as location) has welfare consequences in between those of complete privacy and price personalization. Finally, we extend our aggregation result by allowing for heterogeneous uses of information, some of which increase total surplus. In Section 6, we find that aggregate (market-level) information is collected if and only if it reduces total surplus. In other words, even if data *transmission* is socially detrimental (as in the case of price discrimination downstream), the equilibrium level of data *aggregation* is socially efficient.

We conclude the paper by commenting on the policy implications of these findings in the context of large Internet platforms and business-to-business (B2B) markets.

Our analysis is related to, among others, the model of selling information in Bergemann, Bonatti, and Smolin (2018). Relative to our earlier work, the framework in Section 2 introduces the problem of sourcing information from an individual consumer who makes her participation decision *ex ante*. In other words, the consumer makes a single decision regarding whether to use a platform's services, rather than trying to influence a data intermediary's perception of her willingness to pay. Thus, our work is related to the analysis of the welfare effects of third-degree price discrimination in Bergemann, Brooks, and Morris (2015) and to the model of market segmentation with second-degree price discrimination in Haghpanah and Siegel (2019). In contrast, a recent literature studies settings in which data intermediaries collect information about consumers' actions, rather than about their preferences. For example, Ball (2020) analyzes the design of a quality score that aggregates multiple dimensions of an agent's performance, and Bonatti and Cisternas (2019) examine the optimal

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

aggregation of consumer purchase histories.

Ichihashi (2019) studies competing data intermediaries that acquire perfect information from consumers. His model predicts multiple equilibria, in some of which competition harms consumers to an equal extent as monopoly—see also Westenbroek, Dong, Ratliff, and Sastry (2019) on this point. In contrast, our competition model assumes that the information collected by each intermediary contains noise, which means that every intermediary can add value. This yields a unique equilibrium outcome in terms of the information transmitted and the players' payoffs, and competition has no effect on protecting privacy. Choi, Jeon, and Kim (2019) consider a model of privacy in which data collection requires consumers' consent. They emphasize the information externalities and coordination failures among users as drivers of an excessive loss of privacy.

Finally, independent work by Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019) also studies an environment with data externalities. Their work is different from and largely complementary to ours. In particular, they analyze a network economy in which asymmetric users with exogenous and heterogeneous privacy concerns trade information with a data platform. They derive conditions under which the equilibrium allocation of information is (in)efficient and examine the welfare effects of competition among intermediaries. In contrast, we endogenize privacy concern in a market setting, which enables us to quantify the downstream welfare impact of data intermediation. Our framework also accommodates questions around information design, such as when will privacy be preserved, and what are possible sources of inefficiency other than privacy violations.

2 Model

We consider a trading environment with a single intermediary in the data market, a single producer in the product market, and many consumers. The environment is meant to capture the choices of consumers on large online platforms. In later sections, we generalize the analysis to allow for competition in the data 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):

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

Each consumer i has a realized willingness to pay for the product denoted by:

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

The willingness to pay $w_i \in W = \mathbb{R}$ of consumer i is the sum of a component θ that is *common* to all consumers in the market and an *idiosyncratic* component θ_i that reflects her individual taste shock. Throughout, we assume that all random variables are normally distributed and thus described by a mean vector and a variance-covariance matrix:

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

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 from the data intermediary, the producer does not know the realized demand shocks prior to setting his price. As a consequence, in the absence of additional information from the data intermediary, it is optimal for the producer to offer a uniform unit price to all consumers.

2.2 Data Market

The data market is run by a single data intermediary (it). The data intermediary can acquire demand information from an individual consumer, package this information, and then sell it to the producer. We consider bilateral contracts between the data intermediary and individual consumers, as well as between the data intermediary and the producer. The data intermediary offers these bilateral contracts *ex ante*, that is, before the realization of any demand shocks. The contracts are offered simultaneously to all consumers. Each bilateral contract determines a *data price* and a *data policy*.

The data price determines the fee for the transfer of information. The data policy determines the inflow of information to and the outflow of information from the data intermediary.

As a monopolist market maker, the data intermediary decides how to collect the information from the consumers and how to transmit it to the producer. Thus, the data intermediary faces both an information design and a pricing problem. The data and product markets are summarized in Figure 1 below.

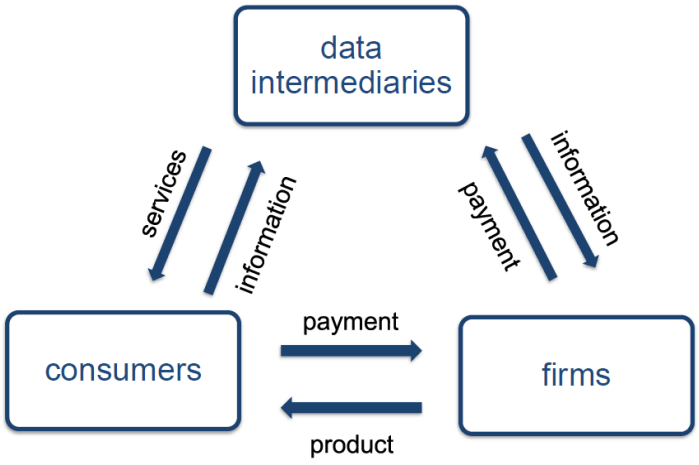


Figure 1: Data Market and Product Market

The data intermediary can transmit all information about each consumer i or some, possibly noisy, statistic of individual and market demand. The two key dimensions of the data intermediary's information policy are *noise* and *aggregation*. In particular, if consumer i participates, the intermediary collects a linear, differentially private signal of her willingness to pay:

$$s_i \triangleq \sum_{j=1}^{N} \alpha_{ij} \left(w_j + \varepsilon + \varepsilon_j \right) a_j, \tag{5}$$

where the weights $\alpha_{ij} \in \mathbb{R}$ for all i, j determine the influence that the willingness to pay of consumer j has on the signal about i, and $a_j \in \{0, 1\}$ represents the participation decision of consumer j. 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.

The weights $\alpha_{ij} \in \mathbb{R}$ allow the data intermediary to fully or partially aggregate the consumers' signals. The leading cases are $\alpha_{ij} = \mathbf{1}_{i=j}$, in which case the intermediary transmits individual demand information, and $\alpha_{ij} = 1/N$, in which case the intermediary transmits market demand (average) information.

In either case, the *data inflow* from consumer i about her private information w_i may be subject to common and idiosyncratic noise terms, ε and ε_i, with variance, σ_ε^2 and $\sigma_{\varepsilon_i}^2$, respectively. The choice of the variance terms σ_ε^2 and $\sigma_{\varepsilon_i}^2$ is an instrument available to the data intermediary to guarantee some degree of informational privacy to consumer i. We

refer to the inflow data policy described by the signals s_i as an information structure

$$S : \mathbb{R}^N \to \Delta\mathbb{R}^N.$$

In turn, the outflow data policy is given by any information structure

$$T : \mathbb{R}^N \to \Delta\mathbb{R}^N$$

that is weakly less informative than the inflow policy S. The outflow data policy can be chosen, exactly as the inflow data policy, with aggregation and noise. In particular, if every consumer participates, the data intermediary reports a vector of (potentially noisy) signals $t \in \mathbb{R}^N$. The resulting information structure in the data market is given by (S, T).

The data intermediary makes a bilateral offer to each consumer i that specifies an inflow data policy S_i and a fee m_i to be paid to the consumer. Similarly, the data intermediary offers an outflow data policy T and a fee m_0 to be paid by the producer. The data intermediary maximizes the net revenue it receives from the consumers and the producer:[7]

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

2.3 Equilibrium and Timing

The game proceeds sequentially. First, the terms of trade on the data market are determined, 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, S_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, T) to the producer. The producer accepts or rejects the offer.

3. The data w and the information flows (s, t) 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 realized demand w_i.

[7] A priori, the respective monetary fees are unrestricted and can be positive or negative. We anticipate the results here and describe the transfers from the point of view of the data intermediary. Thus, we implicitly assume that there is a price of data,or $m_0 > 0$ and an expense for data, or $m_i > 0$.

We analyze the Perfect Bayesian Equilibrium of the game. At the contracting stage, the information is imperfect but symmetric. The analysis proceeds by backward induction. Given an established data policy (S, T), the optimal pricing policy of the producer is informed by the data outflow,

$$p^* : S \times T \to \mathbb{R}^N,$$

where logically, the dependence is on both S and T through the realized data policy, but the price must be measurable with respect to T. The optimal price is thus a vector of potentially distinct and personalized prices $p^*(t) \in \mathbb{R}^N$. The resulting net revenue of the producer is:

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

By contrast, if the producer chooses not enter into a contract with the data intermediary, then he must price under the common prior distribution only. We denote the monopoly price in the absence of information by $\bar{p}$, the net revenue of the producer is:

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

If the producer does not receive any information from the data intermediary, he will not have to pay for any data. If the producer accepts the proposal of the data intermediary, then he must pay the fee m_0. The participation constraint for the producer is thus

$$\Pi(S, T) - \Pi(\varnothing, \varnothing) \geq m_0. \tag{7}$$

Proceeding backwards, the data intermediary, having implemented an inflow data policy S, offers an outflow data policy T, possibly as a function of S, thus $T(S)$ and an associated data price $m_0(T(S))$ that maximizes its overall data revenue (6). Finally, each consumer receives an offer for a data contract (m_i, S_i). We denote the gross expected utility of consumer i from data sharing by

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

Each pair (m_i, S_i) must satisfy the consumer's participation constraint. Holding fixed the decision of the remaining consumers, by rejecting her contract, consumer i would not receive compensation for the data but would also not participate in data sharing. This may affect the realized data outflow policy $T(\varnothing, S_{-i})$ and the resulting pricing policy of the producer. In particular, if consumer i does not accept the intermediary's offer, the producer

can personalize the price on the basis of any data he acquires about other consumers.[8] The consumer's surplus from rejecting the intermediary's offer is:

$$U_i(\varnothing, S_{-i}) \triangleq \mathbb{E}\left[u_i(w_i, q_i, p_i) \,|\varnothing, S_{-i}, T(\varnothing, S_{-i})\right].$$

The data intermediary has to choose a data policy (m_i, S_i) that satisfies the participation constraint of each consumers i:

$$U_i(S) - U_i(\varnothing, S_{-i}) \geq m_i, \text{ for all } i. \tag{8}$$

We emphasize that the participation constraint of every consumer i and the producer are required to hold at the *ex ante* level. Thus, the consumer (and the producer) agree to the data policy before the realization of any particular demand realization w_i. The choice of the *ex ante* participation constraint is meant to capture the prevailing condition where the consumer and the producer accept the "terms of use agreement" or "terms of service" before any particular consumption choice or search event. This reflects the practice, for instance when using Facebook, Amazon, or any search engine, where an account is established before the realization of any particular event. In particular, the consumer evaluates the consequence of sharing her data flow ex ante, i.e., she requires a level of compensation that allows her to profitably share the information on average. Conditional upon agreeing to share the information, there are no further incentive compatibility constraints.

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

$$\left\{(S^*, T^*, m^*);\ p^*(S, T);\ a^*\right\}, \tag{9}$$

where

$$a_0^* : S \times T \times \mathbb{R} \to \{0, 1\}, \qquad a_i^* : S_i \times \mathbb{R} \to \{0, 1\}, \tag{10}$$

such that (i) the producer maximizes his expected profits, (ii) the intermediary maximizes its expected revenue, and (iii) each consumer maximizes her net utility.

[8] This assumption captures the idea that the data intermediary is often a large online platform that directly enables the interaction between the producer and the consumers. Under this interpretation, even if a consumer rejects the intermediary's offer $(m_i, \mathcal{S}_i)$, she still interacts with the producer, who can use all the information available to him.

3 Data in the Wild

We begin the equilibrium analysis by characterizing the value of exogenous information in our model. To do so, we abstract from the data prices and fix an arbitrary information structure that is freely available to the producer. In terms of the data policy (S, T), it is thus as if we were to assume that the data would flow without friction from consumer i to the producer:

$$t_i = s_i, \forall i.$$

We therefore refer to an information structure simply as S. For the subsequent analysis, we shall assume that the information structure S carries some information about the willingness to pay by requiring that the variance of the conditional expectation is positive, or:

$$\text{var}\left[\mathbb{E}\left[w_i \left| S\right.\right]\right] > 0, \text{ for all } i.$$

Thus, the conditional expectation of the willingness to pay under the information structure S carries information about the demand. The realized demand function of each consumer i is given by

$$q_i = w_i - p_i, \tag{11}$$

and hence, the optimal personalized price $p_i^*(S)$ is a linear function of the posterior mean regarding each consumer i's type:

$$p_i^*(S) = \frac{\mathbb{E}\left[w_i \mid S\right] + c}{2}. \tag{12}$$

From (1) and (4), the ex ante expectation of consumer i's surplus and of the producer's profit can be written in terms of the personalized price $p_i^*(S)$ in (12) as

$$U_i(S) = \frac{1}{2}\mathbb{E}\left[\left(w_i - p_i^*(S)\right)^2\right], \text{ and} \tag{13}$$

$$\Pi_i(S) = \mathbb{E}\left[\left(p_i^*(S) - c\right)\left(w_i - p_i^*(S)\right)\right]. \tag{14}$$

Because prices and quantities are linear functions of the producer's conditional expectation of w_i, the ex ante mean of both prices and quantities are constant across all information policies. Consequently, the welfare implications of a given information policy S are captured by the variance and covariance of the equilibrium actions and the consumers' types. In particular, the ex ante surplus of consumer i in (13) is given by

$$U_i(S) = -\text{cov}\left[w_i, p_i^*(S)\right] + \frac{1}{2}\text{var}\left[p_i^*(S)\right] + U_i(\varnothing), \tag{15}$$

where $U_i(\varnothing)$ represents the consumer surplus under no demand data.

Intuitively, any ability of the producer to correlate the price p_i^* to the realized willingness to pay w_i is detrimental to consumer i's welfare, as shown by the covariance term $\text{cov}[w_i, p_i^*]$. This effect is partially offset by the variance of the price $\text{var}[p_i^*]$, which allows the consumer to adjust her demand q_i in her linear optimization problem. The monopoly price p_i^* in (12) is itself linear in the producer's expectation of w_i. The properties of the Gaussian distribution imply that $\text{var}[p_i^*] = \text{cov}[w_i, p_i^*]/2$, so the overall effect of information on consumer welfare is negative.

Similarly, the net revenue of the producer in (14) can be written as

$$\Pi_i(S) = \text{cov}[w_i, p_i^*(S)] - \text{var}[p_i^*(S)] + \Pi_i(\varnothing), \tag{16}$$

where the covariance term captures the intuition that adapting prices to willingness to pay improves profits, while the variance term reflects the loss from varying the choice variable in a concave optimization problem. On net, because $\text{var}[p_i^*] = \text{cov}[w_i, p_i^*]/2$, the effect of information is positive for the producer. Using the expressions in (15) and (16), Proposition 1 contrasts the equilibrium welfare levels when the demand data are publicly available to when the producer only has the common prior information about the demand.

Proposition 1 (Value of Demand Data)
The profits of the producer are higher and the consumer and social surplus are lower under any information structure S than under no information.

To gain some intuition, suppose that perfectly informative demand data are freely available to the producer, who then uses the data to the maximal extent possible in his pricing policy. Thus, the producer pursues a *personalized* pricing policy towards each individual consumer. As the producer adapts his pricing policy to the willingness to pay w_i of consumer i, he increases the monopoly price $p_i^* = w_i/2$ in response to an increase in demand. The equilibrium quantity q_i^* also increases with the willingness to pay w_i but at half the rate it would if the producer only had the common prior information about demand, in which case the consumer would face a constant price. This reduced responsiveness of the level of trade to the gains from trade is what reduces social welfare.[9]

[9] This result is related to the classic environment of third-degree price discrimination with linear demand studied by Robinson (1933) and Schmalensee (1981), with some subtle differences. In the current model, each consumer has a linear demand function, the intercept of which is given by her willingness to pay. Price discrimination then occurs across different *realizations* of the willingness to pay. Indeed, the results in Proposition 1 remain valid if the producer has access to aggregate market-level demand data only and sets a single price. In contrast, in Robinson (1933) and Schmalensee (1981), price discrimination occurs across different markets. 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.

The complete availability of the demand data is admittedly an extreme benchmark in at least two respects. First, the producer may only have access to some noisy or aggregate version of the demand data. Second, with many consumers, the intermediary may observe only a subsample of the consumer demand data. Both elements are critical to the profitability of intermediation, but neither changes the fact that the transmission of *any* amount of information is welfare-reducing in our model.

Thus, we are starting (intentionally) in an economic environment where, absent market imperfections, the consumer data should not be shared and transmitted at all. In Section 6, we consider more general game forms in the downstream interaction between consumers and the producer, in some of which information increases total surplus.

In the presence of social data, the welfare results above hold not only in levels but also on the margin, whereby each individual consumer revealing information contributes to part of the surplus loss. 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 other consumers.

We can now formally define our notion of *data externality*.

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

$$DE_i(S) \triangleq U_i(\varnothing, S_{-i}) - U_i(\varnothing) < 0.$$

The data externality $DE_i(S)$ is strictly negative for any nontrivial information structure. This follows immediately if one applies Proposition 1 to the information structure $S \triangleq (\varnothing, S_{-i})$. Looking ahead to the amount of compensation owed to each consumer i, Proposition 1 formalizes the intuition that the ability to make inferences about w_i from observing signals s_{-i} has negative implications for consumer i's surplus.

Corollary 1 (Marginal Value of Information)
The marginal welfare effect of consumer i's data flow is smaller than the total effect,

$$0 < U_i(\varnothing, S_{-i}) - U_i(S) < U_i(\varnothing) - U_i(S). \tag{17}$$

The second inequality in (17) suggests that the data externality across consumers facilitates the profitability of data intermediation. Thus, the overall effect of the data flow S on consumer i's surplus is partly due to the data flow S_i from her and partly due to the information flow S_{-i} from the other consumers.

4 Data and Aggregation

In contrast to the preceding analysis, where the consumer data were available "in the wild," we now explicitly assign each consumer the ownership of her demand data. Thus, unless the consumer explicitly accepts an arrangement to share her data with a data intermediary, the data will remain private information to her.

4.1 Data Flow

A critical driver of the consumer's decision to share the data is her ability to anticipate the intermediary's use of the information so gained. We therefore begin with a preliminary result regarding the nature of the interaction between the data intermediary and the producer: given the data inflow S that the intermediary has acquired, we establish that the subsequent interaction between the intermediary and the producer is efficient. In other words, the intermediary implements a data outflow policy that maximizes the producer's gross surplus, which it then extracts through the fee m_0.

Proposition 2 (Data Outflow Policy)
The data intermediary offers a data policy of complete sharing, $T^*(S) = S$, for all S. The data policy $T^*(S) = S$ maximizes the gross revenue of the producer among all feasible outflow data policies given S.

This result remains valid as long as we do not introduce multiple producers who compete with one another. Because the interaction between the data intermediary and the producer is frictionless, the presence of external contracting of the data does not matter for the structure of the optimal data outflow policy. Consequently, we denote a data policy simply by S. Furthermore, because we know that *any* nontrivial information structure S reduces total surplus, we obtain an immediate corollary of Propositions 1 and 2 on the impossibility of profitable intermediation in the presence of a single (large) consumer.

Corollary 2 (Unprofitable Data Intermediation)
If $N = 1$, the data intermediary cannot generate strictly positive profits.

We must therefore turn to the social dimension of the data to understand why data intermediation can be profitable and, in particular, to study the comparative statics of the value of information as more consumers sell their data to the intermediary. In what follows, we focus on the intermediary-optimal Perfect Bayesian Equilibrium (henceforth, the equilibrium). Intuitively, in this equilibrium, all consumers accept the intermediary's offer along the path of play.

Because we are looking to maximize the intermediary's equilibrium profits, the consumer's participation constraint (8) must be binding. The intermediary's fees are given by

$$\begin{aligned} m_i^* &= U_i(\varnothing, S_{-i}) - U_i(S) \\ &= \underbrace{U_i(\varnothing, S_{-i}) - U_i(\varnothing)}_{DE_i(S)<0} - \underbrace{(U_i(S) - U_i(\varnothing))}_{\Delta U_i(S)<0} \end{aligned}$$

The optimal intermediary fee can thus be written as the difference between two terms, both of which are negative: the total change $\Delta U_i(S)$ in consumer i's surplus associated with information structure S and the data externality $DE_i(S)$ imposed on i by consumers $-i$ when they sell their own data to the intermediary. Thus, the data externality reduces the compensation that the intermediary owes consumers and creates the possibility of profitable information intermediation.

Similarly, the optimal fee that the intermediary charges the producer is given by

$$m_0^* = \Pi(S) - \Pi(\varnothing).$$

With the notion of a data externality, we then can write the intermediary's profit (6) as

$$R(S) = m_0 - \sum\nolimits_{i=1}^{N} m_i(S) = \Delta TS(S) + \sum\nolimits_{i=1}^{N} |DE_i(S)|, \tag{18}$$

where ΔTS denotes the variation in total surplus resulting from information policy S. Thus, the intermediary's profits are given by the sum of two terms: the variation in total surplus associated with a data policy S and the total data externality across N agents. The expression in (18) further clarifies why with only one consumer (or with independent consumer types), intermediation is not profitable and why the intermediary's objective diverges from the social planner's.

4.2 Data Aggregation

We now explore the intermediary's decision to aggregate the individual consumers' demand data through the weights α_{ij} as in (5). At one extreme, the intermediary can collect and transmit individual demand data ($\alpha_{ij} = \mathbf{1}_{i=j}$), thereby allowing the producer to match the demand data to each specific consumer. This would be the optimal policy if the data were freely available. At the other extreme, the intermediary can collect market demand data ($\alpha_{ij} = 1/N$). In this case, the producer observes a vector of identical (possibly noisy) signals

of market demand:

$$s_i = \frac{1}{N}\sum\nolimits_{j=1}^{N}(w_j + \varepsilon + \varepsilon_j) \triangleq \bar{s}, \text{ for all } i. \tag{19}$$

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.[10]

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 is then a feasible way of reducing the data acquisition costs by *de facto* anonymizing the consumers' information. Proposition 3 shows that, perhaps surprisingly, aggregating the consumers' data as in (19) is *always* optimal for the intermediary.

Proposition 3 (Optimality of Data Aggregation)
For any noise level in the signals $\left(\sigma_{\varepsilon}^2, \sigma_{\varepsilon_i}^2\right)$*, the profit-maximizing aggregation weights for the intermediary are given by* $\alpha_{ij}^* = 1/N$ *for all* i, j*.*

Within the confines of our policies, the data intermediary always finds it advantageous to not attempt to elicit the identity of the consumer at all. An important implication is that the producer will not offer personalized prices. Instead, he will offer variable prices that adjust to the realized information about market demand. While this form of third-degree price discrimination is detrimental to social surplus, it represents a milder distortion than personalized pricing.

This finding suggests why we may 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 may interpret the restraint in 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. Suppose that the intermediary acquires individual data, i.e.,

$$s_i = w_i + \varepsilon + \varepsilon_i, \tag{20}$$

with fixed noise levels $\left(\sigma_{\varepsilon}^2, \sigma_{\varepsilon_i}^2\right)$. Now consider the data externality imposed on consumer i

[10] This would also be the case if, equivalently, the producer had access to individual data, including identifying information about the consumer, but could not identify consumers when he offers his product.

by all other consumers $-i$:

$$DE_i(S_i) \triangleq U_i(\varnothing, S_{-i}) - U_i(\varnothing) = -\operatorname{cov}\left[w_i, p_i^*(S_{-i})\right] + \frac{1}{2}\operatorname{var}\left[p_i^*(S_{-i})\right]. \tag{21}$$

In the above expression (21), $p_i^*(S_{-i})$ denotes the monopoly price charged to consumer i when she rejects the intermediary's contract, i.e., the optimal personalized price for consumer i off the equilibrium path.

If the producer observes the $N-1$ signals s_{-i} only, he is restricted in his ability to offer a personalized price to consumer i. The lack of data about any given individual can, however, be partially compensated with data about other consumers. In particular, as 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. This price is given by a convex combination of the prior mean and the *average* of all the other consumers' signals,

$$p_i^*(S_{-i}) = \frac{\mathbb{E}\left[w_i \mid S_{-i}\right] + c}{2} = \frac{1}{2}\left(c + x\frac{\Sigma_{j\neq i} s_j}{N-1} + (1-x)\,\mu\right) \tag{22}$$

for some weight $x \in (0,1)$. In this sense, one can view the average signal $\bar{s}_{-i}$ as a differentially private signal of s_i. The extent to which the average $\bar{s}_{-i}$ is informative about s_i then depends on the variance of the common shock σ_θ^2 and on the variance of 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 as in (19). Along the path of play, the producer offers the same price to all consumers. If one consumer does not participate in the contract, however, the producer will be able to offer two prices: a single price for the $N-1$ participating consumers about whom the producer has aggregate and another price to the deviating "anonymous" consumer. The latter price is again based on market data and, in fact, is equal to $p_i^*(S_{-i})$ in (22). In other words, with individual signals, the producer optimally aggregates the available data to form the best predictor of the missing data point. Therefore, the sale of aggregate data has no impact on the pricing policy off the equilibrium path. Specifically, all data policies with the same noise levels generate identical levels of the data externality, independent of the degree of data aggregation.[11]

At this point, it is immediately clear why it is optimal for the intermediary to aggregate

[11] The result in Proposition 3 would not change if we forced the producer to charge a single price to all consumers off the equilibrium path when aggregate data are collected. With this interpretation, however, we want to capture the idea that the producer offers one price "on the platform," to the participating consumers but will still interact with the deviating consumer "offline" and will be able to leverage what he has learned from the available market data to tailor his offline price.

the individual information of all consumers: fix any given noise levels, and start from any symmetric matrix of weights $\alpha_{ij} \neq 1/N$. Relative to this set of weights, aggregating the resulting signals (such that the effective weights become $\alpha = 1/N$) leaves the DE_i terms in (18) unchanged but reduces the amount of information conveyed to the producer in equilibrium. Therefore, aggregate data policies minimize the loss of total surplus and hence maximize the intermediary's profits.

Finally, Proposition 3 enables us to contrast the market outcome when consumer data are freely available with the outcome under data ownership. When consumers own their data, the presence of intermediaries might induce socially inefficient information transmission and third-degree price discrimination, but the contractual outcome will now preserve the personal identity of the consumer. Below, we will qualify the result in Proposition 3 by means of two extensions. In Section 5.3, we consider heterogeneous consumers and endogenous market segmentation. In Section 6, we consider more general models in which information may increase total surplus. We then show that the data aggregation decision is made efficiently in equilibrium, while the provision of information may nonetheless be socially excessive.

4.3 Basic Data Intermediation

Having shown that the intermediary prefers to acquire and transmit aggregate information, we now establish whether the intermediary can make positive profits at all. We first consider the case of basic data intermediation, in which the intermediary collects a signal about the average willingness to pay $\bar{s}$ without adding any noise. In the next section, we refine our result by introducing the optimal level of noise.

In a sense, Proposition 3 shows that assigning to the consumers the control rights over their data imposes a constraint on the amount of private information traded in the market when those data are later used for price discrimination. Conversely, Proposition 4 identifies conditions under which the data market enables trading of surplus-reducing information.

Proposition 4 (Data Intermediation and Size of the Database)
Suppose that the intermediary collects noiseless, aggregate information about the consumers' willingness to pay (i.e., $\sigma_{\varepsilon}^2 = \sigma_{\varepsilon_i}^2 = 0$ and $\alpha_{ij} = 1/N$ for all i, j).

1. *There exists a threshold $\bar{N} > 1$ such that the intermediary obtains positive profits if and only if $N \geq \bar{N}$.*

2. *The intermediary's profit is increasing in σ_{θ}^2 and decreasing in $\sigma_{\theta_i}^2$.*

Thus, if the data externality is significant and information is sufficiently valuable to the producer, then aggregate demand data is profitably transmitted by the intermediary.

Because multiple Gaussian signals are substitute inputs into the producer's learning technology, the data externality is larger—and the individual compensation owed to consumers is lower—when more consumers are present.[12] Similarly, if the common shock σ_θ^2 is large, then the informational externality will allow the data intermediary to offer favorable terms of trade earlier, in terms of the threshold size of the market. By contrast, if the idiosyncratic shock $\sigma_{\theta_i}^2$ is large, then the cost of compensating the consumer will be large and prevent profitable intermediation. This leads to a higher threshold $\overline{N}$ when the trade can arise.

A similar result would hold if we forced the intermediary to transmit individual consumer data without aggregation. However, by transmitting only the aggregate data, the intermediary can operate profitably with a smaller number of consumers, for any given level of $\sigma_{\theta_i}^2$ and σ_θ^2. More important, the aggregate data policy generates smaller losses in total surplus, and as we have seen, higher revenues for the intermediary for a given market size.

5 Optimal Data Intermediation

In this section, we show that a more sophisticated information policy than considered previously can increase the revenue of the data intermediary. In particular, we allow the data intermediary to choose the variance levels of the additional noise terms, σ_ε^2 and $\sigma_{\varepsilon_i}^2$. We then consider how the optimal data policy changes with the number of consumers and how it adapts to a richer demand specification with consumer heterogeneity.

5.1 Information Design

We begin with the information structure that maximizes the intermediary's profits. We leverage the optimality of trading market-level demand data (Proposition 3), and focus on the optimal noise levels $(\sigma_\varepsilon, \sigma_{\varepsilon_i})$ associated with the aggregation weights $\alpha_{ij} = 1/N$.

Proposition 5 (Optimal Data Intermediation)
The intermediary's optimal data policy includes

1. *no idiosyncratic noise,* $\sigma_{\varepsilon_i}^2 = 0$,
2. *(weakly) positive aggregate noise* $\sigma_\varepsilon^2 = \max\{0, \sigma^*\}$, *where*

$$\sigma^* = \frac{\sigma_{\theta_i}^2 + N\sigma_\theta^2}{N-1} \frac{\sigma_{\theta_i}^2 - (N-1)\left(\sqrt{3}-1\right)\sigma_\theta^2}{N\left(\sqrt{3}-1\right)\sigma_\theta^2 - \sigma_{\theta_i}^2}.$$

[12] This result is closely connected to the submodularity property of information in Acemoglu, Makhdoumi, Malekian, and Ozdaglar (2019).

The optimal level of common noise σ^* is strictly positive when the ratio $\sigma^2_{\theta_i}/\sigma^2_\theta$ is large enough and when N is small enough. When positive, σ^* is strictly increasing in the ratio $\sigma^2_{\theta_i}/\sigma^2_\theta$ and decreasing in N. Our characterization in Proposition 5 yields a necessary and sufficient condition for the profitability of data intermediation.

Proposition 6 (Profitability of Data Intermediation)
The data intermediary's profits are strictly positive if and only if

$$N\left(\sqrt{3}-1\right)\sigma^2_\theta > \sigma^2_{\theta_i}. \tag{23}$$

Figure 2 shows the optimal variance of the additional common noise term: if the consumers' preferences are sufficiently correlated (or if the market is large enough), the intermediary does not add any noise. Conversely, if the consumer types are not sufficiently correlated—meaning that condition (23) is close to binding—the optimal level of common noise grows without bound. In particular, for values $\sigma_{\theta_i}/\sigma_\theta$ higher than the threshold $(\sqrt{3}-1)/N$, no profitable intermediation is feasible. Therefore, the amount of information traded in equilibrium is continuous in both σ_θ and σ_{θ_i}.

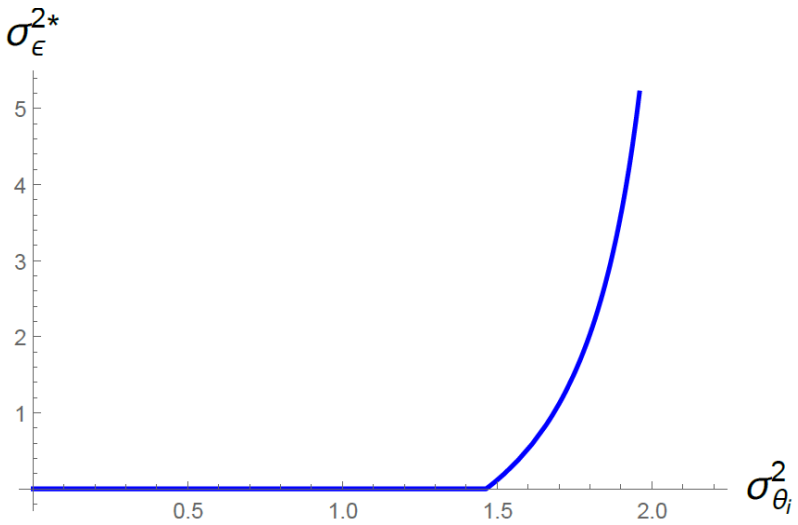


Figure 2: Optimal Common Noise ($N = 3, \sigma^2_\theta = 1$)

By introducing additional 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 reselling them to the producer, it may appear surprising that the correlation structure in the additional noise terms plays such a critical role.

To gain intuition for the role of correlated noise, it may 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 $\bar{s}$ as in (19). This weight is given by

$$\frac{\text{cov}\left[\bar{w}, \bar{s}\right]}{\text{var}\left[\bar{s}\right]} = \frac{\sigma_\theta^2 + \sigma_{\theta_i}^2/N}{\sigma_\theta^2 + \sigma_\varepsilon^2 + \left(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2\right)/N}. \tag{24}$$

Clearly, one could obtain the same coefficient with many combinations of σ_ε^2 and $\sigma_{\varepsilon_i}^2$.

The reason for choosing a unique (degenerate) combination $\left(\sigma_\varepsilon^2, \sigma_{\varepsilon_i}^2\right) = (\sigma^*, 0)$ is again linked to 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 $\bar{s}_{-i}$. In this new regression, the signal $\bar{s}_{-i}$ receives the following weight

$$\frac{\text{cov}\left[w_i, \bar{s}_{-i}\right]}{\text{var}\left[\bar{s}_{-i}\right]} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2 + \left(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2\right)/(N-1)}. \tag{25}$$

Comparing (24) and (25), we observe that the common noise term enters identically in both, while the idiosyncratic noise $\sigma_{\varepsilon_i}^2$ increases the variance of $\bar{s}_{-i}$ relatively more than the variance of $\bar{s}$. This 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 ε, 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 error terms ε_i over a larger sample size). The latter effect is absent when only common noise is used. Common noise makes 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, the two key elements of the information design, aggregation and noise, interact with one another in a rich way. In particular, the value of common noise is deeply linked to that of aggregate data. To highlight the contrast, suppose that the intermediary were constrained to offering personal data intermediation (i.e., the weights $\alpha_{ij} = \mathbf{1}_{i=j}$). While Proposition 3 shows that idiosyncratic noise is never optimal in the full problem, it becomes optimal in the constrained problem.

Proposition 7 (Noise and Personalized Prices)
The intermediary's optimal data policy when restricted to personalized prices includes

1. *no common noise,* $\sigma_\varepsilon^2 = 0$,

2. *weakly positive idiosyncratic noise* $\sigma_{\varepsilon_i}^2$ *that satisfies*

$$\lim_{N\to\infty} \sigma_{\varepsilon_i} = \infty \text{ and } \lim_{N\to\infty} \frac{\sigma_{\varepsilon_i}}{N} = 0.$$

These results help us understand the economic forces involved in the interaction between price discrimination and information design. In particular, correlated noise terms are not part of this restricted optimal information design. Intuitively, when the intermediary offers personalized signals, it enables personalized price discrimination. We can then informally describe the producer's problem as consisting of two tasks: estimate the common component of market demand θ from the average signals $\bar{s}$, and then estimate the idiosyncratic taste shocks θ_i from the difference between average and individual signals. However, the difference between $\bar{s}$ in (19) and s_i in (20) can be written as

$$\bar{s} - s_i = \Sigma_{j=1}^{N} \left(\theta_j + \varepsilon_j\right)/N - \theta_i - \varepsilon_i,$$

which is an unbiased signal of the difference $\bar{w} - w_i$, where the common noise term ε drops out. Therefore, common noise with personalized signals reduces the value of information to the producer without protecting individuals from harmful price discrimination.

Finally, larger amounts of idiosyncratic noise become optimal as the number of consumers grows: as $N \to \infty$, the intermediary can perfectly estimate the common demand component θ for free. In the presence of personalized pricing, noise then becomes necessary to prevent the producer from fully reacting to the consumers' private information.

Having clarified that common noise is an important and integral part of the optimal intermediation, we now turn to the comparative statics of the information design and of the intermediary's profits as the number of consumers grows.

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 and the large number of 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 in the market becomes large. Each additional consumer presents an additional opportunity for trade in the downstream market. Thus, the feasible social surplus is linear in the number of consumers. In addition, with every additional consumer, the intermediary can obtain additional information about the idiosyncratic and aggregate demand components. These two effects suggest that intermediation becomes increasingly profitable in larger markets, where the consumers impose severe data externalities on one another, while the value of information to the producer grows without bound. Proposition 8 formalizes this intuition.

Proposition 8 (Large Markets)

1. *As $N \to \infty$, the individual consumer's compensation goes to zero, and the total compensation converges to a finite positive value.*

2. *The total compensation is asymptotically decreasing in N if and only if $\sigma_\theta^2 > \sigma_{\theta_i}^2$.*

3. *As $N \to \infty$, the intermediary's revenue and profit grow linearly in N.*

As the optimal data policy only estimates the common demand shock, each additional signal contributed by 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. In particular, this convergence can occur *from above:* the decrease in the marginal contribution can be strong enough to offset the increase in the number of consumers, which means that it can be less expensive to acquire a larger dataset than a smaller dataset. Figure 3 illustrates such an instance.

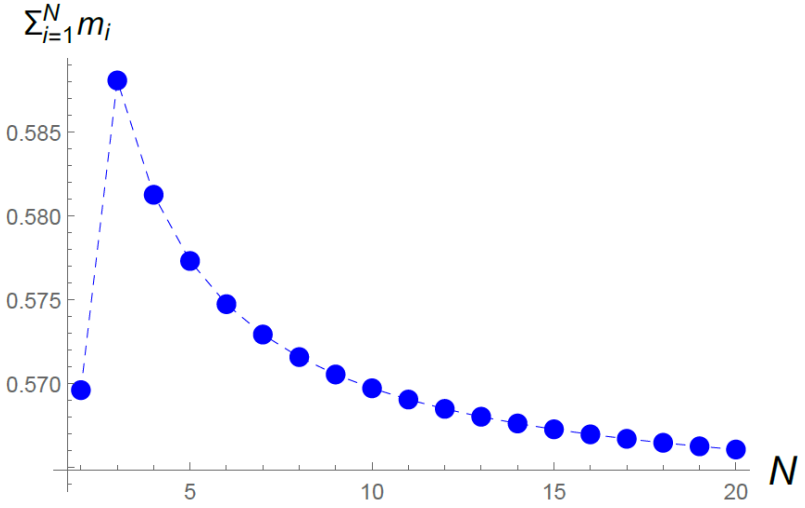


Figure 3: Total Consumer Compensation $\left(\sigma_\theta^2 = 1, \sigma_{\theta_i}^2 = 3/4\right)$

Finally, the revenue that the data intermediary can extract from the producer is linear in the number of consumers. As the market size grows without bound, our model therefore implies that the per capita profit of the data intermediary converges to the per capita profit when market-level (aggregate) data are freely available.

In practice, the results in Proposition 8 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, independently of the final use of information, each consumer is willing to accept a negligible compensation (even on aggregate) to give up her private information. In this sense, our results qualify the sentiment that "consumers are very much willing to consent for their data to be used for their benefit."[13] Instead, we show that consumers will consent to *any* use of their data for very little aggregate compensation. Far from being a paradox, this type of behavior reflects the market value of their information, which depends both on the correlation structure and on the intermediary's equilibrium data policy.

5.3 Market Segmentation and Data

Thus far, we have defined the demand of every individual consumer in terms of a common and an idiosyncratic component. We then used the assumption of ex ante homogeneity to produce some of the central implications of social data. Nonetheless, a more complete description of consumer demand should account for additional characteristics that introduce heterogeneity across certain groups of consumers. These might include characteristics such as location, demographics, income, and wealth.

In this section, we explore how these additional characteristics may influence the value of intermediation and the information policy of the data intermediary. Towards 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,

$$w_{ij} = \theta_j + \theta_{ij}, i = 1, 2, ..., N_j, \;\; j = 1, ..., J. \tag{26}$$

Thus, each member of group j has the same common component. Across groups, the common components are drawn from a normal distribution with mean μ and variance σ_θ^2—the two groups are identical ex ante. The idiosyncratic component θ_{ij} remains normally distributed with zero mean and variance $\sigma_{\theta_i}^2$, and all random variables are independent. For

[13] The quote is from Sasan Goodarzi, CEO of the financial software company Intuit, upon acquiring the financial application Credit Karma. See The New York Times (2020).

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 Proposition 3 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 3 (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, by sending the sample average of all signals across groups to the producer, the intermediary forces the producer to offer a single price. 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 allows the intermediary to charge a higher fee to the producer but also increases the compensation owed to consumers.

Proposition 9 sheds light on the optimal resolution of this trade-off. In this result, we restrict attention to the case of noiseless signals ($\sigma_\varepsilon = \sigma_{\varepsilon_i} = 0$). We expect these results to extend to a noisy signal environment.

Proposition 9 (Segmentation)
Let $\sigma_\varepsilon = \sigma_{\varepsilon_i} = 0$.

1. *There exists $\overline{N}$ such that the data intermediary induces group pricing for all $N > \overline{N}$.*
2. *The threshold $\overline{N}$ is decreasing in the variance of the common component σ_θ^2 and increasing in the variance of the idiosyncratic component $\sigma_{\theta_i}^2$.*

Thus, while the earlier Proposition 3 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., the groups' identity. This will allow the producer to price discriminate across but not within groups.

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 benefits of pooling signals *across* groups arise in similar conditions as those of the additional noise in the case of homogeneous consumers. If the producer faces a small number of consumers and their types are not highly correlated, then pooling reduces the cost of sourcing the data.

The result in Proposition 9 is perhaps the sharpest manifestation of the value of big data. By enabling the producer to adopt a richer pricing model, a larger database allows the intermediary to extract more surplus. Our result also clarifies the appetite of the platforms for large datasets: as 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 10 formalizes this result, and Figure 4 illustrates it.

Proposition 10 (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.

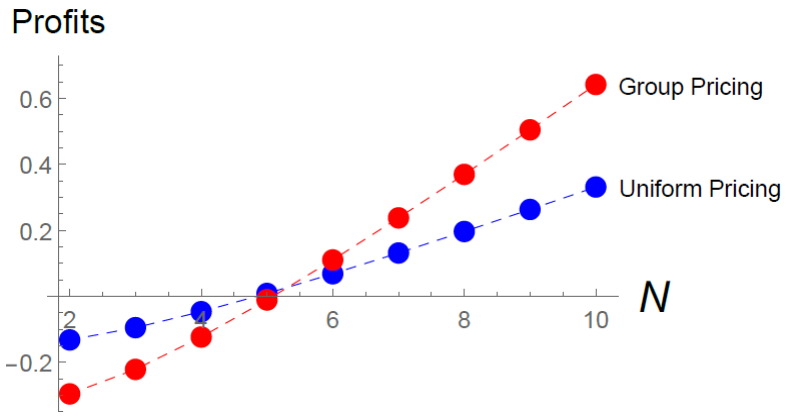


Figure 4: Uniform Pricing vs. Group Pricing $\left(\sigma_\theta^2 = 3/10, \sigma_{\theta_i}^2 = 1\right)$

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 of acquiring more precise information from consumers. As the data externality

grows sufficiently strong, acquiring the data becomes cheaper, and the intermediary exploits the richer structure of consumer demand.[14]

5.4 Multiple Services

Another 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. This service extends the number of sources of information about each consumer.[15]

Our aim in this section is to capture the precision of information associated with multiple sources in a parsimonious way. Thus, we assume that each consumer observes (or is able to transmit) only a noisy signal of her willingness to pay,

$$r_i \triangleq w_i + \zeta_i, \tag{27}$$

where ζ_i is an exogenous Gaussian noise term with variance $\sigma^2_{\zeta_i}$. (One could easily derive (27) from a model with K different sources of information, each with its own idiosyncratic noise term $\zeta_{i,k}$.)

We now consider the role of augmenting the precision of information collection, i.e., of reducing $\sigma^2_{\zeta_i}$. This has a direct effect on the data policy, as it increases the value of information. In addition, to the extent that the intermediary can acquire more information from all consumers, it may be able to lower the total compensation for a given data policy.

Proposition 11 (Precision of Information Collection)

1. *The optimal amount of common noise σ^* is increasing in $\sigma^2_{\zeta_i}$.*
2. *The intermediary's profit is decreasing and convex in $\sigma^2_{\zeta_i}$.*

This result captures the idea that more sources and more services yield more data transmission. In particular, we have assumed that with more information sources, the noise in each individual observation $\sigma^2_{\zeta_i}$ decreases. This greater precision strengthens the correlation in the signals r_i, because the fundamentals w_i are themselves correlated, while the noise terms ζ_i are independent. Therefore, the informational externality increases as $\sigma^2_{\zeta_i}$ decreases,

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

[15] Similarly, Google offers a number of services, such as Gmail, Google Maps, and YouTube, that gather information about a consumer from many different sources. Such "super profiles" have recently attracted regulatory attention in the EU.

which lowers the compensation required by the individual consumer. Furthermore, greater correlation in the underlying signals r_i allows the intermediary to further reduce the amount of additional correlated noise, which explains the convexity in profits: the intermediary obtains increasing returns by reducing $\sigma^2_{\zeta_i}$. We leverage this model with exogenously noisy observations when we introduce competing intermediaries in Section 7.3.

6 Intermediation and Value Creation

In our baseline model, the consumer's information is only used to set prices. This is, in a sense, the worst-case scenario for the intermediary: as data transmission reduces total surplus, no intermediation is profitable without a sufficiently strong data externality. In practice, consumer data can also be used in surplus-enhancing ways, e.g., to facilitate the provision of products and quality levels targeted to the consumer's tastes.

In this section, we develop a generalization of our framework that 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,

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

with $\beta \geq 0$ being the consumer's marginal valuation of quality.[16] (The case of $\beta = 0$ yields the baseline model of price discrimination only, while the case of $\beta = 1$ is a useful benchmark in which the consumer values quality and money equally.) The producer has a constant marginal cost of quantity provision and a fixed cost of quality production per consumer, i.e.,

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

We first revisit our basic results on the value of exogenous information for all the market participants. In particular, we obtain the following extension of Proposition 1.

Proposition 12 (Value of Exogenous Information)
Fix any nontrivial information structure S.

1. *Producer profits are higher under S than under no information for all $\beta \geq 0$.*

2. *Consumer surplus is higher under S than under no information for all $\beta \geq 1$.*

[16] Argenziano and Bonatti (2019) provide a treatment of signaling and ratchet effects in a dynamic version of this model under a fixed information structure, i.e., without data intermediation.

3. *There exists a unique* $\beta^* \in (0,1)$ *such that total surplus is higher under* S *than under no information for all* $\beta \geq \beta^*$.

When the intermediary must procure the data from the consumers, the data externality once again allows for the profitable intermediation of information. Furthermore, for sufficiently strong data externalities, the intermediary will transmit all the data. Whether the outcome improves or diminishes total surplus depends on the use of the data, which is represented here by the value of quality. 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 Proposition 3.

Proposition 13 (Value of Aggregation)
Fix any noise levels $(\sigma_\varepsilon, \sigma_{\varepsilon_i})$. *The intermediary collects aggregate data* $(\alpha_{ij} = 1/N)$ *if information reduces total surplus (i.e., if* $\beta < \beta^*$*), and individual data* $(\alpha_{ij} = \mathbf{1}_{ij})$ *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 the 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, and hence, the information that enables it is provided at the individual advertiser level.

7 Further Extensions

7.1 Data Intermediation with Commitment

In our analysis, the data intermediary maintains complete control over the use of the acquired data. In particular, given the data inflow, the data intermediary chooses the sequentially optimal data policy to be offered to the producer. The sequential optimality reflects the substantial control that the data intermediary has regarding the use of the data. It also

reflects the opacity in how the data outflow is linked to the data inflow. In other words, it is difficult to establish how any given data input has informed 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 this subsection, we return to our baseline model to describe what would happen under stronger commitment assumptions. We learned from Proposition 1 that the unrestricted use of data leads to a decrease in social surplus. This suggests that the data intermediary could realize a higher profit with a data policy that limits the diffusion of information *on the equilibrium path.*

In particular, 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 1 established that each consumer would prefer that the demand information not be transferred to the producer. Thus, the remaining question is the amount of compensation that the data intermediary could extract from each consumer. This depends on the threat that the data intermediary can impose on the individual consumer should she fail to sign up with the data intermediary. In the absence of an agreement with consumer i, the data intermediary could forward its estimate about the demand of consumer i based on the information from all of the remaining consumers, and this would indeed be the least favorable outcome for consumer i.

Proposition 14 (Commitment Solution)
Suppose that the intermediary can commit to a data outflow policy. Then, the revenue-maximizing data policy is to acquire all consumer data and to never forward the data to the downstream producer.

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.[17] The shared insight is that the intermediary *with* commitment power may be able to extract a rent without any further influence on the efficiency of the allocation.

In both the commitment solution and the sequentially optimal solution, we did not impose any restrictions on the sign or size of the monetary payments. In the sequentially optimal solution, every consumer receives compensation for her marginal damage. In the commitment

[17] The distinction between ex ante and interim contracting may disappear in the setting of in Lizzeri (1999), where the intermediary has a testing technology available that allows him to verify the private information of the agent. Thus, one might be able to decentralize (or distribute) the ex ante payment over the interim state in such a way that in expectation the ex ante and interim contracts are payoff equivalent.

solution, every consumer pays a fee to avoid information disclosure. We might then ask what the scope of data policy is if the data intermediary can neither reward nor punish the consumer. In other words, we restrict the data intermediary to offering a data policy that does not involve a monetary transaction with the consumer, or $m_i = 0$, for all i.

Proposition 15 (Commitment Solution without Monetary Compensation)
Suppose that the intermediary cannot use monetary transfers with the consumers and thus that $m_i = 0$, for all i. Then, the optimal data policy is to collect all demand data and to enable the producer to offer each consumer i a personalized price that does not rely on data provided by consumer i.

Thus, in the absence of monetary transfers between the data intermediary and consumers, the data intermediary still acquires the demand data from all the consumers but exercises some restraint in their use. In particular, the data intermediary forwards only the data that will enable the producer to offer a personalized price to consumer i, where the personalized price is computed without reference to the demand information provided by consumer i.[18]

7.2 Unique Implementation

In our baseline model, the profit and compensation are calculated in the intermediary's most preferred equilibrium. A natural question is to ask whether the ability of the intermediary to extract full surplus in per capita terms will be preserved if we consider the intermediary's least preferred equilibrium in the subgame following its offers to consumers.

To answer this question, we could consider a "divide-and-conquer" scheme whereby the intermediary will sequentially approach consumers and offer compensation conditioned on all earlier consumers having accepted their offers. In this scheme, the first consumer receives compensation equal to her entire surplus loss, which guarantees 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 first consumer accepts her offer in all equilibria, and the second consumer will accept regardless of the remaining consumers' decisions. By a contagion argument, we can guarantee the unique implementation of the equilibrium, which gives rise to a lower bound on the profit of the data intermediary.

Proposition 16 (Divide and Conquer)
Suppose that the intermediary uses the divide-and-conquer compensation scheme. 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."

[18] In contrast, without commitment, there is no trade of information without monetary transfers.

Unlike in the baseline model, the total compensation owed to consumers does not converge but rather diverges at rate $O(\log(N))$. However, the individual compensation level decreases fast enough that the linear growth in revenue drives the asymptotics, and the profits of the intermediary converge to the complete information level.

7.3 Competing Intermediaries

We now ask whether competition promotes privacy protection. As a first step, we extend our model to accommodate $J \geq 2$ competing intermediaries. We introduce heterogeneity across these intermediaries in a way that is pertinent to information markets. Specifically, 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 5.4, the noise term $\zeta_{i,j}$ represents limitations in the communication channel between consumer and intermediary. We assume that each shock $\zeta_{i,j}$ is independently drawn from a normal distribution with zero mean and variance $\sigma_\zeta^2 > 0$. Thus, we treat σ_ζ^2 as a lower bound on the variance of the idiosyncratic noise specific to each intermediary.

The timing of the game is as in Section 2.3, except for the competing intermediaries:

1. The intermediaries simultaneously offer data price and data inflow policies $(m_{i,j}, S_{i,j})$ to all consumers, and each consumer chooses a subset of offers to accept.

2. Given the realized data inflows, the intermediaries simultaneously offer a data price and data outflows $(m_{0,j}, T_j)$ to the producer.

3. The producer decides which data offers to accept and sets prices for the consumers.

Mirroring our baseline analysis, we first consider the subgame beginning after the consumers' acceptance decisions. Denote by S_j the realized data inflow to intermediary j and by S the total information collected by all intermediaries. The following proposition characterizes the unique equilibrium.

Proposition 17 (Unique Pricing Subgame)
Let $\sigma_\zeta^2 > 0$, and fix any realized data inflow $(S_j)_{j=1}^J$. There exists a unique equilibrium in which intermediary j sells $T_j = S_j$ for a fee of $m_{0,j} = \pi(S) - \pi(S \backslash S_j)$, and the monopolist accepts all intermediaries' offers.

Note that the uniqueness result would break down if $\sigma_\zeta^2 = 0$, i.e., if each intermediary could collect perfect information from consumers. In this extreme case, once one intermediary

obtains one copy of a report from consumer i, any other report about i collected by the other intermediary is useless for the producer. Therefore, in the pricing subgame, once a report of consumer i is sold, the other intermediary becomes indifferent about whether to sell i's information, which breaks the unique equilibrium price in that subgame. This causes multiplicity of equilibria, as in Ichihashi (2019).

For the remainder of this section, we focus on the case of perfectly correlated types $w_i = \theta$ for the sake of tractability. In the absence of noise terms $\zeta_{i,j}$, this setting is one in which a monopolist data intermediary would be able to source information about the common component θ at zero cost. The presence of competition and differentiation makes the analysis here as rich as the baseline model with imperfectly correlated types. We restrict attention to equilibria in which each consumer chooses to accept the maximal number of offered data contracts (subject to her participation constraint). We now compare the equilibrium outcome under competition with the monopoly outcome.

Proposition 18 (Competing Data Intermediaries)
In every pure and maximal equilibrium with competing intermediaries, the following hold.

1. *The producer's surplus, the consumer surplus, the intermediaries' total profits, and the information structure are unique.*
2. *The total amount of information transmitted to the producer by J competing intermediaries with noise levels σ_ζ^2 is equal to the monopoly case with noise level σ_ζ^2/J.*
3. *Consumer surplus is decreasing in J and increasing in σ_ζ^2.*
4. *The producer's profit is increasing while the intermediaries' total profit is asymptotically decreasing in the number of intermediaries J.*
5. *The total profit of intermediaries converges to a constant as N becomes large.*

This result highlights the possibility that, instead of protecting consumers' privacy, competition may make things worse. In the unique equilibrium outcome, competing intermediaries collect data in exactly the same way as a monopolist would. Each intermediary's profit decreases, which is purely due to a loss of bargaining power against the producer in the product market. This loss is most clearly seen in the welfare properties of equilibrium as the number of consumers grows large: while a monopolist intermediary asymptotically obtains the entire value of information for the producer (which grows without bound), an imperfectly competitive industry's profits remain bounded.

Relative to the case of 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. Apart from the analysis of entry, our model also offers predictions in terms of mergers: the equilibrium noise level is the same whether the firms merge or compete, and consumer surplus and social surplus remain unchanged.

Thus, competition in the data market *per se* may not be a very effective tool to correct the data externality. In fact, the loss in consumer surplus indicates that, under any kind of friction in the data market, competition could be detrimental to social welfare. Suppose, for example, that the data intermediaries did not share the same consumer dataset but had only partially overlapping sets. Then, the inference problem of the producer becomes harder, and the value of the data declines because of omissions and repetitions. However, due to the data externality, the data price would fail to reflect the loss of information, and hence fail to correct for the loss in informational efficiency.

8 Conclusion

We have explored the terms of information trade between data intermediaries with market power and multiple consumers whose preferences are correlated. The data externality we have uncovered strongly suggests that data ownership is insufficient to bring about the efficient use of information. Indeed, our baseline model focused on the case in which information harms consumers. In large markets, we have shown that arbitrarily small levels of compensation can induce an individual consumer to relinquish precise information about her preferences.

A simple 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 face serious implementation challenges, both from a theoretical and a practical perspective. Within the confines of our model, one could consider a richer specification of consumer heterogeneity, e.g., one that allows for different marginal valuations of quality in Section 6. Thus, 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 on the aggregation of consumer information further suggest that privacy regulations need to 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 level (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.[19]

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 make 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 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 consumer and producer. By contrast, the consumer can often access a given producer only through a data platform. In fact, our results apply directly to a large class of data platforms.[20] Many such platforms can then be thought of as selling access (often through an auction) to the consumer. The data platform provides the bidding producer with additional information that firms 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. We explore the collection of social data on these platforms in an ongoing project, Bergemann, Bonatti, and Gan (2020).

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

[20] These include product data platforms, such as Amazon, Uber and Lyft, that acquire individual data from the consumer through the purchase of services and products. In addition to these, social data platforms such as Google and Facebook offer data services to individual users and sell the information, mostly in the form of advertising placement, to third parties. Relative to the basic model that we analyze, the difference between the data intermediary and the product platform is that the product platform combines in a single-decision maker the roles of data intermediation and product pricing.

9 Appendix

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

Proof of Proposition 1. Given data policy S, the profit of the producer is given by

$$\Pi(S) = \Sigma_i(\text{cov}[w_i, p_i] - \text{var}[p_i]) + \Pi(\varnothing)$$
$$\text{where } p_i = \frac{1}{2}\mathbb{E}\left[w_i \,|S\right].$$

Because both w_i and p_i are Gaussian random variables, the profit of the producer can also be written as

$$\Pi\left(S\right) = \sum\nolimits_{i=1}^{N} \text{var}[p_i] + \Pi(\varnothing) = \frac{1}{4}\sum\nolimits_{i=1}^{N} \text{var}[\mathbb{E}\left[w_i \,|S\right]] + \Pi(\varnothing),$$

which means profits are increasing with more information, as measured by the variance of the producer's posterior mean. Similarly, the expected surplus of consumer i is given by

$$U_i(S) = -\,\text{cov}[w_i, p_i] + \frac{1}{2}\,\text{var}[p_i] + U_i(\varnothing) = -\frac{3}{8}\,\text{var}[\mathbb{E}\left[w_i \,|S\right]] + U_i(\varnothing),$$

which is decreasing with more information. Finally, the total value generated by data policy S (net of the surplus without information $U_i(\varnothing) + \Pi(\varnothing)$) is given by

$$-\frac{1}{8}\sum\nolimits_{i=1}^{N} \text{var}[\mathbb{E}\left[w_i \,|S\right]],$$

which is also decreasing with more information. ■

Proof of Proposition 2. Fix a realized data inflow S. Because the intermediary holds all the bargaining power, it internalizes the value of information for the producer, which it then extracts through the fee m_0. Thus, the intermediary seeks to maximize the producer's profit, which is increasing in the amount of available information by Proposition 1. Finally, the complete sharing data outflow policy $T = S$ dominates all other policies in the Blackwell order, which establishes the claim. ■

Proof of Proposition 3. Fix a data policy S offered by the intermediary, and normalize the marginal cost of production c to 0. We denote by $p_i\left(S\right)$ the optimal price charged to consumer i under complete information transmission, and by $p_i\left(S_{-i}\right)$ the price when i does

not report her signal. Given the data policy S, these optimal prices are given by

$$p_i(S) = \frac{1}{2}\mathbb{E}[w_i|S]$$
$$p_i(S_{-i}) = \frac{1}{2}\mathbb{E}[\theta|S_{-i}] + \frac{1}{2}\mathbb{E}[\theta_i] = \frac{1}{2}\mathbb{E}[\theta|S_{-i}]$$

The total value of information for the producer when all consumers accept the intermediary's offer is given by

$$m_0 = \frac{1}{4}\sum\nolimits_{i=1}^{N} \text{var}\left[\mathbb{E}[w_i|S]\right] = \sum\nolimits_{i=1}^{N} \text{var}[p_i(S)].$$

The compensation owed to consumer i is:

$$\begin{aligned} m_i &= \frac{1}{8}\left(\text{var}\left[\mathbb{E}[w_i|S_{-i}]\right] - \text{var}\left[\mathbb{E}\left[w_i|S\right]\right]\right) - \frac{1}{2}\,\text{cov}\left[w_i, \mathbb{E}[w_i|S_{-i}] - \mathbb{E}\left[w_i|S\right]\right] \\ &= \frac{3}{8}(\text{var}\left[\mathbb{E}\left[w_i|S\right]\right] - \text{var}\left[\mathbb{E}[w_i|S_{-i}]\right])) \\ &= \frac{3}{2}(\text{var}[p_i(S)] - \text{var}[p_i(S_{-i})]). \end{aligned}$$

Thus, we can calculate the profit of the intermediary as

$$R(S) \triangleq m_0 - \sum\nolimits_{i=1}^{N} m_i = -\frac{1}{2}\sum\nolimits_{i=1}^{N} \text{var}[p_i(S)] + \frac{3}{2}\sum\nolimits_{i=1}^{N} \text{var}[p_i(S_{-i})] \tag{28}$$
$$= -\frac{1}{8}\sum\nolimits_{i=1}^{N} \text{var}[\mathbb{E}[w_i|S]] + \frac{3}{8}\sum\nolimits_{i=1}^{N} \text{var}[\mathbb{E}[\theta|S_{-i}]]. \tag{29}$$

The first (negative) term is the effect of information on the realized social surplus on path, while the second (positive) term corresponds to subtracting the consumers' surplus off path.

Now consider any symmetric weights α_{ij}, where $\alpha_{ii} = \bar{\alpha}$ for all i, and $\alpha_{ij} = \hat{\alpha}$ for all $i \neq j$. We will argue that it is always more profitable for the intermediary to aggregate the N signals so to obtain a single average signal $\bar{s}$. This operation is itself equivalent to setting $\alpha_{ij} = 1/N$ for all i, j from the beginning. Denote the original data policy (with personalized prices) by $\underline{S}$ and the aggregated data policy by $\bar{S}$. By definition any variables that are measurable with respect to $\overline{S}$ are measurable with respect to $\underline{S}$. Therefore

$$\text{var}[\mathbb{E}[w_i|\underline{S}]] \geq \text{var}[\mathbb{E}[w_i|\overline{S}]]$$

Furthermore, aggregation causes no loss in precision in the estimation of w_i off the equilibrium path. This is because the optimal price $p_i(S_{-i})$ is a function of the average of the other $N-1$ agents' reports

$$\mathbb{E}[\theta|\overline{S}_{-i}] = \mathbb{E}[\theta|\underline{S}_{-i}] = A\Sigma_{j\neq i}s_i,$$

and

$$\text{var}[\mathbb{E}[\theta|\underline{S}_{-i}]] = \text{var}[\mathbb{E}[\theta|\overline{S}_{-i}]].$$

Consequently, by expression (29), the intermediary's profits from $\bar{S}$ are always larger than those from $\underline{S}$. ■

Proof of Proposition 4. Fix a data policy S that fully aggregates the consumers' signals and introduces noise terms with variance σ_ε^2 and $\sigma_{\varepsilon_i}^2$. Recall from expression (28) that the intermediary's profit can be written as

$$R(S) = -\frac{1}{2}\sum\nolimits_{i=1}^{N} \text{var}[p_i(S)] + \frac{3}{2}\sum\nolimits_{i=1}^{N} \text{var}[p_i(S_{-i})],$$

where the optimal prices are given by

$$\begin{aligned} p_i(S) &= \frac{1}{2}\mathbb{E}\left[\theta + \frac{1}{N}\sum\nolimits_{j=1}^{N}\theta_j \mid S\right] \text{ for all } i, \text{ and} \\ p_i(S_{-i}) &= \frac{1}{2}\mathbb{E}\left[\theta \mid S_{-i}\right]. \end{aligned}$$

and

$$p_i' = \frac{1}{2}\mathbb{E}[\theta|S] + \frac{1}{2}\mathbb{E}[\theta_i] = \frac{1}{2}\mathbb{E}[\theta|S].$$

By Bayes' rule, we obtain

$$\begin{aligned} p_i(S) &= \frac{1}{2}\frac{N\sigma_\theta^2 + \sigma_{\theta_i}^2}{N^2(\sigma_\varepsilon^2+\sigma_\theta^2) + N(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\sum\nolimits_{j=1}^{N} s_j, \\ p_i(S_{-i}) &= \frac{1}{2}\frac{(N-1)\sigma_\theta^2}{(N-1)^2(\sigma_\varepsilon^2+\sigma_\theta^2) + (N-1)(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\sum\nolimits_{j\neq i} s_j, \end{aligned}$$

and the resulting expected profits for the intermediary

$$R(S) = \frac{3(N-1)N\sigma_\theta^4}{8((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}. \tag{30}$$

Now impose $\sigma_\varepsilon^2 = \sigma_{\varepsilon_i}^2 = 0$. We obtain

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

To prove the existence of a threshold of $\bar{N}$ above which intermediation profits are positive, we first prove that R/N is strictly increasing with respect to N. Up to a positive multiplicative

constant, we have

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

Therefore, R is itself increasing with respect to N. Finally, we have

$$R(1) = -\frac{(\sigma_\theta^2 + \sigma_{\theta_i}^2)}{8} < 0, \text{ and}$$
$$\lim_{N\to\infty} \frac{R(N)}{N} = \frac{\sigma_\theta^2}{4} > 0,$$

which establishes the existence of a threshold $\bar{N}$.

Now we prove the monotonicity in terms of σ_θ^2 and $\sigma_{\theta_i}^2$. In particular, we have

$$\frac{\partial R}{\partial \sigma_\theta^2} = \frac{N\left(2(N-1)^2\sigma_\theta^4 + 4(N-1)\sigma_\theta^2\sigma_{\theta_i}^2 - \sigma_{\theta_i}^4\right)}{2((N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)^2} > 0$$

so that R is increasing in σ_θ^2.

To prove that R decreases with $\sigma_{\theta_i}^2$, it is enough to compute the derivative

$$\frac{\partial R}{\partial \sigma_{\theta_i}^2} = -\frac{3(N-1)N\sigma_\theta^4}{2((N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)^2} - \frac{1}{2} < 0,$$

which establishes the desired result. ■

The following Lemma is instrumental in the proofs of Proposition 5 and 6.

Lemma 1 (Optimality of Symmetric Noise)
Fix a data policy S with full aggregation. It is optimal for the intermediary to choose a constant level of the variance of the idiosyncratic noise term, $\sigma_{\varepsilon_i}^2 = \bar{\sigma}$ for all i.

Proof of Lemma 1. Fix the variance of the common noise term σ_ε^2, and allow the variance of the idiosyncratic terms $\sigma_{\varepsilon_i}^2$ to depend on i. The optimal price levels are then given by

$$p_i(S) = \sum\nolimits_{i=1}^{N} \frac{1}{2N} \frac{N\sigma_\theta^2 + \sigma_{\theta_i}^2}{1 + \sum_{j=1}^{N} \frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_j}^2}} \frac{s_i}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2},$$
$$p_i(S_{-i}) = \sum\nolimits_{j\neq i} \frac{1}{2} \frac{\sigma_\theta^2}{1 + \sum_{k\neq i} \frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_k}^2}} \frac{s_j}{\sigma_{\theta_j}^2 + \sigma_{\varepsilon_j}^2}.$$

The revenue of the intermediary can therefore be expressed as:

$$R(S) = -\frac{N}{2}\sum_{i=1}^{N}\text{var}[p_i(S)] + \frac{3}{2}\sum_{i=1}^{N}\text{var}[p_i(S_{-i})]$$

$$= -\frac{1}{8N^2}\frac{\left(N\sigma_\theta^2+\sigma_{\theta_i}^2\right)^2}{\left(1+\sum_{i=1}^{N}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}\right)^2}\sum_{i=1}^{N}\frac{1}{\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}\left(1+\sum_{i=1}^{N}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}\right)+\frac{3}{2}\sum_i\text{var}[p_i(S_{-i})]$$

$$= -\frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8N}\frac{1}{1+\sum_{i=1}^{N}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}\sum_{i=1}^{N}\frac{1}{\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}+\sum_{i=1}^{N}\frac{3\sigma_\theta^4}{8}\frac{1}{1+\sum_{j\neq i}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}\sum_{j\neq i}\frac{1}{\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2}$$

$$= -\frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8N}\frac{\sum_{i=1}^{N}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}{1+\sum_{i=1}^{N}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}+\sum_{i=1}^{N}\frac{3\sigma_\theta^4}{8}\frac{\sum_{j\neq i}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_j}^2+\sigma_{\varepsilon_j}^2}}{1+\sum_{j\neq i}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_j}^2+\sigma_{\varepsilon_j}^2}}.$$

We now argue that it is optimal to set $\sigma_{\varepsilon_i}^2$ to the same level for all i. To do so, we first change variables

$$x_i \triangleq \frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2} \in \left(0, \frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2}\right],$$

so that optimizing R over $\sigma_{\theta_i}^2$ is equivalent to optimizing R over x_i. In particular, we have

$$R(S) = -\frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8N}\frac{\sum_{i=1}^{N}x_i}{1+\sum_{i=1}^{N}x_i}+\sum_i\frac{3\sigma_\theta^4}{8}\frac{\sum_{j\neq i}x_j}{1+\sum_{j\neq i}x_j}$$

$$= \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8N}\frac{1}{1+\sum_{i=1}^{N}x_i}-\sum_i\frac{3\sigma_\theta^4}{8}\frac{1}{1+\sum_{j\neq i}x_j}+C$$

$$= A\frac{1}{1+\sum_{i=1}^{N}x_i}-\sum_i B\frac{1}{1+\sum_{j\neq i}x_j}+C$$

Since $B>0$, it is easy to verify that for any $\{x_i\}$ such that $x_1 \neq x_2$, then setting $x_1' = x_2' = (x_1+x_2)/2$ will strictly increase profits. ■

Proof of Proposition 5. Having established in Lemma 1 that it is optimal to use a symmetric noise scheme, we now prove that the optimal idiosyncratic noise level is nil, $\sigma_{\varepsilon_i}^* = 0$. To show this result, suppose $\sigma_\varepsilon^2 \geq 0$ and $\sigma_{\varepsilon_i}^2 > 0$. 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 $\bar{\sigma}_{\varepsilon_i}^2 \triangleq \sigma_{\varepsilon_i}^2 - (N-1)\delta^2 \geq 0$ the profits R will strictly increase. Too see this, notice that

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

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

Next we turn to the expression of the optimal common noise. First we use first order condition to calculate the optimal σ_ε^2 for any fixed $\sigma_{\varepsilon_i}^2$ (for future reference). By Proposition 6 (which we prove below), we can focus on the case where

$$N\left(\sqrt{3}-1\right)\sigma_\theta^2-\sigma_{\theta_i}^2>0, \tag{31}$$

which corresponds to case in which 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 negative quadratic term. It has two roots, and the smaller one is given by

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

When (31) holds, the denominator is positive while the nominator is negative: the smaller root is always negative. Therefore the optimal σ_ε^2 (for fixed $\sigma_{\varepsilon_i}^2$) is either 0 or the larger root of the quadratic function, which can be simplified to

$$\sigma_\varepsilon^{2*}(\sigma_{\varepsilon_i}^2)=\max\left\{\frac{\sigma_{\theta_i}^2+N\sigma_\theta^2}{N-1}\frac{\sigma_{\theta_i}^2-(N-1)\left(\sqrt{3}-1\right)\sigma_\theta^2}{N\left(\sqrt{3}-1\right)\sigma_\theta^2-\sigma_{\theta_i}^2}\right. \tag{32}$$
$$\left.+\frac{\left(3+\sqrt{3}-2N\right)N\sigma_\theta^4+\left(\sqrt{3}+2N\right)\sigma_\theta^2\sigma_{\theta_i}^2+\sigma_{\theta_i}^4}{(N-1)\left(2N^2\sigma_\theta^4-2N\sigma_\theta^2\sigma_{\theta_i}^2-\sigma_{\theta_i}^4\right)}\sigma_{\varepsilon_i}^2\right\}.$$

In particular, since it is optimal to set $\sigma_{\varepsilon_i}^2=0$, we have,

$$\sigma_\varepsilon^{2*}=\max\left\{0,\frac{\sigma_{\theta_i}^2+N\sigma_\theta^2}{N-1}\frac{\sigma_{\theta_i}^2-(N-1)\left(\sqrt{3}-1\right)\sigma_\theta^2}{N\left(\sqrt{3}-1\right)\sigma_\theta^2-\sigma_{\theta_i}^2}\right\}, \tag{33}$$

which completes the proof. ■

Proof of Proposition 6. When condition (31) is not satisfied, using expression (30) in the proof of Proposition 4 (set $\sigma_{\varepsilon_i}^2 = 0$), we have:

$$R = \frac{3N\sigma_\theta^4}{8(\sigma_\theta^2+\sigma_\varepsilon^2)}\left(1-\frac{\sigma_{\theta_i}^2}{(N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\theta_i}^2}\right) - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8N(\sigma_\theta^2+\sigma_\varepsilon^2)}\left(1-\frac{\sigma_{\theta_i}^2}{N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\theta_i}^2}\right), \tag{34}$$

and thus

$$R \le \left(\frac{3N\sigma_\theta^4}{8(\sigma_\theta^2+\sigma_\varepsilon^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8N(\sigma_\theta^2+\sigma_\varepsilon^2)}\right)\left(1-\frac{\sigma_{\theta_i}^2}{N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\theta_i}^2}\right) \le 0.$$

Conversely, when (31) holds, we can rewrite R as

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

Thus, the intermediary receives a positive profit as long as σ_ε^2 is sufficiently large. ■

Proof of Proposition 7. We first show that it is asymptotically optimal to set $\sigma_\varepsilon^2 = 0$ and $\sigma_{\varepsilon_i}^2 > 0$ that grows sub-linearly with N. Similar to the proof of Proposition 4, we can compute the intermediary's profit under a data policy S that enables personalized pricing, i.e., when the weights are $\alpha_{ij} = \mathbf{1}_{i=j}$:

$$R(S) = \frac{3(N-1)N\sigma_\theta^4}{8(\sigma_\varepsilon^2(N-1)+\sigma_{\varepsilon_i}^2+(N-1)\sigma_\theta^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{8(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N-1)\sigma_{\theta_i}^4}{8(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}. \tag{35}$$

The right-hand side of (35) contains an additional negative term, compared with the profit under data aggregation (30), which echoes the result in Proposition 3. The additional term does not depend on σ_ε^2, thus we can use the result of Proposition 5 to find the optimal common noise for given $\sigma_{\varepsilon_i}^2$ (32). (Because the profit with personalized pricing is strictly smaller than under uniform pricing, we can focus on the case $(\sqrt{3}-1)N\sigma_\theta^2 \ge \sigma_{\theta_i}^2$.)

If N is sufficiently large, both linear term (of $\sigma_{\varepsilon_i}^2$) and constant term is negative. Thus it is optimal to set common noise to zero (for any $\sigma_{\varepsilon_i}^2$). Then

$$\lim_{N\to\infty}\frac{R}{N} \le \frac{\sigma_\theta^2}{4} - \lim_{N\to\infty}\frac{\sigma_{\theta_i}^4}{8(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} \le \frac{\sigma_\theta^2}{4},$$

and the upper bound can be attained if the idiosyncratic noise $\sigma_{\varepsilon_i}^2$ grows without bound but sub-linearly in N.

Indeed, if $\sigma_{\varepsilon_i}^2$ grows faster than linearly

$$\lim_{N\to\infty} \frac{\sigma_{\varepsilon_i}^2}{N} = \infty$$
$$\text{then } \lim_{N\to\infty} \frac{R}{N} = 0$$

which is not optimal. Letting $\sigma_{\varepsilon_i}^2$ grow at linear speed is also not optimal:

$$\lim_{N\to\infty} \frac{\sigma_{\varepsilon_i}^2}{N} = a > 0$$
$$\text{then } \lim_{N\to\infty} \frac{R}{N} = \frac{\sigma_\theta^4}{4(a+\sigma_\theta^2)} < \frac{\sigma_\theta^2}{4}.$$

Therefore the optimal asymptotic solution must involve setting $\sigma_\varepsilon^2 = 0$ and $\sigma_{\varepsilon_i}^2$ growing less than linearly in N.

Finally we want to prove that it is always optimal to set $\sigma_\varepsilon^2 = 0$. We first rule out an interior optimum, i.e., $\sigma_\varepsilon^2 > 0$ and $\sigma_{\varepsilon_i}^2 > 0$. This is done by showing that there is no solution $(\sigma_\varepsilon, \sigma_{\varepsilon_i})$ to the system of two first-order conditions $\partial R/\partial\sigma_{\varepsilon_i} = 0$ and $\partial R/\partial\sigma_\varepsilon = 0$. Thus the optimal noise structure must involve a corner solution. We then rule out that $\sigma_\varepsilon^2 > 0$ and $\sigma_{\varepsilon_i}^2 = 0$ at the optimum by showing that, when $\sigma_{\varepsilon_i}^2 = 0$, the two inequalities $\partial R/\partial\sigma_\varepsilon > 0$ and $R > 0$ cannot hold at the same time. ■

Proof of Proposition 8. Using the calculations in the proof of Proposition 3, the total compensation owed to consumers can be written as

$$\sum\nolimits_{i=1}^{N} m_i = \Sigma_i \frac{3}{2}(\text{var}[p_i(S)] - \text{var}[p_i(S_{-i})])$$
$$= \frac{3N(N\sigma_\theta^2 + \sigma_{\theta_i}^2)^2}{8\left(N^2(\sigma_\varepsilon^2+\sigma_\theta^2) + N(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)\right)} - \frac{3(N-1)^2 N\sigma_\theta^4}{8\left((N-1)^2(\sigma_\varepsilon^2+\sigma_\theta^2) + (N-1)(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)\right)}.$$

In the proof of Proposition 5, we have shown that for N sufficiently large, but finite, it is optimal to set $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2 = 0$. We can simplify the expression for large N to

$$\sum\nolimits_{i=1}^{N} m_i = \frac{3\sigma_{\theta_i}^2((2N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)}{8((N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)}.$$

The limit of the compensation is positive but finite:

$$\lim_{N\to\infty} \sum\nolimits_{i=1}^{N} m_i = \frac{3\sigma_{\theta_i}^2}{4}.$$

Furthermore, the derivative is asymptotically negative if and only if $\sigma_\theta^2 > \sigma_{\theta_i}^2$:

$$\frac{\partial}{\partial N}\sum\nolimits_{i=1}^{N} m_i = \frac{3\sigma_\theta^2\sigma_{\theta_i}^2(\sigma_{\theta_i}^2 - \sigma_\theta^2)}{8((N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)}.$$

And finally, profits grow linearly, because we have

$$\lim_{N\to\infty}\frac{R}{N} = \frac{\sigma_\theta^2}{4}.$$

This ends the proof. ■

Proof of Proposition 9. Consider two groups of N consumer each. We first derive the expression of the profit in the two cases of uniform pricing and group pricing. In the case of uniform pricing, the producer charges does not have any identification information. He can thus only provide one price for every consumer:

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

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:

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

The revenue of the intermediary is given by

$$\begin{aligned}\underline{R}(S) &= -\frac{1}{2}2N\,\mathrm{var}[p_{ij}(S)] + \frac{3}{2}\sum\nolimits_{ij}\mathrm{var}[p_{ij}(S_{-ij})] \\ &= -\left(\frac{N}{8}\sigma_\theta^2 + \frac{1}{8}\sigma_{\theta_i}^2\right) + \frac{3N}{16}\frac{(2N-1)^2\sigma_\theta^4}{(2N^2-2N+1)\sigma_\theta^2 + (2N-1)\sigma_{\theta_i}^2}. \end{aligned} \tag{36}$$

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

$$\overline{R}(S) = \frac{3(N-1)N\sigma_\theta^4}{4((N-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)} - \frac{(N\sigma_\theta^2+\sigma_{\theta_i}^2)^2}{4(N(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}. \tag{37}$$

We then compare the derivatives of (36) and (37) 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 $4N\sigma_\theta^2 = \sigma_{\theta_i}^2$ yields

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

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

Proof of Proposition 10. We have shown in the proof of Proposition 9 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 for all N. ■

Proof of Proposition 11. We first prove the monotonicity of σ_ε^* with respect to $\sigma_{\zeta_i}^2$. From equation (32) in the proof of Proposition 5, we obtain that the optimal common noise is given by $\sigma_\varepsilon^{2*} = \max\{0, \sigma^*\left(\sigma_{\zeta_i}^2\right)\}$, where

$$\sigma^*\left(\sigma_{\zeta_i}^2\right) = \frac{\sigma_{\theta_i}^2 + N\sigma_\theta^2}{N-1}\frac{\sigma_{\theta_i}^2 - (N-1)\left(\sqrt{3}-1\right)\sigma_\theta^2}{N\left(\sqrt{3}-1\right)\sigma_\theta^2 - \sigma_{\theta_i}^2} + \frac{\left(3+\sqrt{3}-2N\right)N\sigma_\theta^4 + \left(\sqrt{3}+2N\right)\sigma_\theta^2\sigma_{\theta_i}^2 + \sigma_{\theta_i}^4}{(N-1)\left(2N^2\sigma_\theta^4 - 2N\sigma_\theta^2\sigma_{\theta_i}^2 - \sigma_{\theta_i}^4\right)}\sigma_{\zeta_i}^2.$$

So for fixed N, σ_θ^2 and $\sigma_{\theta_i}^2$, σ^* is a linear function of $\sigma_{\zeta_i}^2$ When the coefficient on $\sigma_{\zeta_i}^2$ is positive we are done. When it is not, we prove that $\sigma^* < 0$ so that the optimal common noise σ_ε^* is constant and equal to zero. If the coefficient is negative, because denominator is always positive when $N\left(\sqrt{3}-1\right)\sigma_\theta^2 > \sigma_{\theta_i}^2$, we must have

$$N(3-2N)\sigma_\theta^4 + 2N\sigma_\theta^2\sigma_{\theta_i}^2 + \sqrt{3}\sigma_\theta^2(N\sigma_\theta^2 + \sigma_{\theta_i}^2) + \sigma_{\theta_i}^4 < 0.$$

The constant term is also negative (using the fact that $N\left(\sqrt{3}-1\right)\sigma_\theta^2 > \sigma_{\theta_i}^2$), so

$$\begin{aligned}
&-2(N-1)N^2\sigma_\theta^6 + \left(\sqrt{3}+1\right)N\sigma_\theta^4\sigma_{\theta_i}^2 + \left(3N+\sqrt{3}-1\right)\sigma_\theta^2\sigma_{\theta_i}^4 + \sigma_{\theta_i}^6 \\
&< N\sigma_\theta^2\left(N(3-2N)\sigma_\theta^4 + 2N\sigma_\theta^2\sigma_{\theta_i}^2 + \sqrt{3}\sigma_\theta^2(N\sigma_\theta^2 + \sigma_{\theta_i}^2) + \sigma_{\theta_i}^4\right) < 0.
\end{aligned}$$

Therefore $\sigma^* < 0$, and the optimal common noise is constant (and nil).

Finally we prove that the intermediary's profit is convex in $\sigma_{\zeta_i}^2$. To do this, we simply calculate the second-order derivative,

$$\begin{aligned}
\frac{\partial^2 R}{\partial(\sigma_{\zeta_i}^2)^2} &= \frac{3(N-1)N\sigma_\theta^4}{4(\sigma_\varepsilon^2(N-1) + \sigma_{\zeta_i}^2 + (N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)^3} - \frac{(N\sigma_\theta^2 + \sigma_{\theta_i}^2)^2}{4(N(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\zeta_i}^2 + \sigma_{\theta_i}^2)^3} \\
&\geq \frac{1}{(N(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\zeta_i}^2 + \sigma_{\theta_i}^2)^2}\frac{3(N-1)N\sigma_\theta^4}{4(\sigma_\varepsilon^2(N-1) + \sigma_{\zeta_i}^2 + (N-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)} \\
&\quad - \frac{1}{(N(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\zeta_i}^2 + \sigma_{\theta_i}^2)^2}\frac{(N\sigma_\theta^2 + \sigma_{\theta_i}^2)^2}{4(N(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\zeta_i}^2 + \sigma_{\theta_i}^2)} \\
&= \frac{1}{(N(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\zeta_i}^2 + \sigma_{\theta_i}^2)^2}R,
\end{aligned}$$

which is therefore positive whenever positive profits are feasible. ■

Proof of Proposition 12. The demand of each consumer in the case of quality and price discrimination is given by

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

where we restrict attention to the case where the complete-information outcome is well-defined, i.e., $\beta < \sqrt{2}$. Under data policy S, the producer offers the following price and quantity levels (again normalizing the marginal cost of quantity $c = 0$):

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

The producer's profit is given by:

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

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

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

For $0 \leq \beta < 1$, information reduces consumer surplus, while for $1 < \beta < \sqrt{2}$, information increases consumer surplus. Finally, the social surplus is given by

$$\left(\frac{1}{2(2-\beta^2)} - \frac{(3-\beta^2)(1-\beta^2)}{(2-\beta^2)^2}\right)\sum\nolimits_{i=1}^{N} \text{var}[w_i|S] + \sum\nolimits_{i=1}^{N} U_i(\varnothing) + \Pi(\varnothing)$$

It is straightforward to verify that information increases social surplus if and only if $\beta > \beta^*$, where $\beta^* = (7-\sqrt{17})/4 < 1$. ■

Proof of Proposition 13. The argument mirrors the proof of Proposition 3. Any symmetric weighting scheme α leads to the same off-path prices $p_i(S_{-i})$, which are a function of the average signal of consumers $j \neq i$ only. Therefore, in order to maximize profits, the intermediary chooses the weights that maximize the social value of information on path. Let $\underline{S}$ denote the data policy with $\alpha_{ij} = \mathbf{1}_{i=j}$, and $\overline{S}$ the data policy with $\alpha_{ij} = 1/N$; also let S denote an arbitrary data policy with symmetric weights. Clearly, we have $\overline{S} \subset S \subset \underline{S}$. (Any symmetric weighting signals can be constructed from perfect signals, and the uniform weighting signals can be constructed by averaging the signals with any symmetric weights.)

Therefore we have:

$$\text{var}[\mathbb{E}[w_i|\underline{S}]] \geq \text{var}[\mathbb{E}[w_i|S]] \geq \text{var}[\mathbb{E}[w_i|\overline{S}]]$$

Then for $0 \geq \beta < \beta^*$, information decreases social surplus, thus $\alpha_{ij} = 1/N$ is optimal; for $\beta^* < \beta < \sqrt{2}$, information increases social surplus, thus $\alpha_{ij} = \mathbf{1}_{ij}$ is optimal. ■

Proof of Proposition 14. The revenue of the intermediary is given by

$$R(S) = m_0 - \sum_{i=1}^{N} m_i = -\frac{1}{8}\sum_{i=1}^{N} \text{var}[\mathbb{E}[w_i|S]] + \frac{3}{8}\sum_{i=1}^{N} \text{var}[\mathbb{E}[\theta|S_{-i}]].$$

Now our problem becomes:

$$\max_{S,S_i} \left[-\sum_{i=1}^{N} \frac{1}{8} \text{var}\left[\mathbb{E}\left[w_i\,|S\right]\right] + \frac{3}{8}\sum_{i=1}^{N} \text{var}\left[\mathbb{E}\left[w_i\,|S_{-i}\right]\right] \right].$$

Since by definition the information off-path is constructed by $(s_j)\ j \neq i$, $S_{-i} \subset \mathcal{F}(s_1, ..., \hat{s}_i, ...s_N)$, we have:

$$R \leq -0 + \frac{3}{8}\sum_{i=1}^{N} \text{var}(\mathbb{E}[w_i|\mathcal{F}(s_1, ..., \hat{s}_i, ...s_N)])$$

where the inequality binds (the optimality is reached) when on the path S contains no information and off the path S_{-i} use all the information from signals $(s_j)\ j \neq i$. ■

Proof of Proposition 15. Recall that

$$R(S) = -\frac{1}{8}\sum_{i=1}^{N} \text{var}[\mathbb{E}[w_i|S]] + \frac{3}{8}\sum_{i=1}^{N} \text{var}[\mathbb{E}[\theta|S_{-i}]].$$

and the consumer surplus is:

$$\Delta U_i = -\frac{3}{8}(\text{var}\left[\mathbb{E}\left[w_i\,|S\right]\right] - \text{var}\left[\mathbb{E}\left[w_i\,|S_{-i}\right]\right]).$$

Since no transfer is allowed, ΔU_i must be non-negative (consumers must be weakly better off on the path). Now our problem becomes:

$$\max_{S,S_i} -\sum_{i=1}^{N} \frac{1}{8} \text{var}\left[\mathbb{E}\left[w_i\,|S\right]\right] + \frac{3}{8}\sum_{i=1}^{N} \text{var}\left[\mathbb{E}\left[w_i\,|S_{-i}\right]\right]$$
$$\text{s.t.} \quad \text{var}\left[\mathbb{E}\left[w_i\,|S\right]\right] - \text{var}\left[\mathbb{E}\left[w_i\,|S_{-i}\right]\right] \leq 0.$$

Clearly the maximum of this program is smaller than the following relaxed program:

$$\max_{S,S_i} \frac{1}{4}\sum\nolimits_{i=1}^{N} \text{var}\left[\mathbb{E}\left[w_i \left| S_{-i}\right.\right]\right]$$
$$\text{s.t.} \quad \text{var}\left[\mathbb{E}\left[w_i \left| S\right.\right]\right] - \text{var}\left[\mathbb{E}\left[w_i \left| S_{-i}\right.\right]\right] \geq 0.$$

Then any data policy such that $S_{-i} = \mathcal{F}(s_1, ..., \hat{s}_i, ...s_N)$ and $\text{var}\left[\mathbb{E}\left[w_i \left| S\right.\right]\right] = \text{var}\left[\mathbb{E}\left[w_i \left| S_{-i}\right.\right]\right]$, i.e., a policy that uses all the off-path information for personalized price and keeps consumers at the same surplus level on path, solves the original problem. ■

Proof of Proposition 16. We consider a "divide and conquer" compensation scheme for the intermediary. The payment to consumer $k = 1, \ldots N$ compensates her for her incremental loss in surplus, given that only $k-1$ other consumers reveal their information. We have already calculated these quantities in the baseline model where everyone receives their surplus loss given that $N-1$ other consumers reveal their information. Therefore the total compensation is easily written as

$$\begin{aligned}
\sum\nolimits_{i=1}^{N} m_i &= \sum\nolimits_{k=1}^{N} \frac{3(k\sigma_\theta^4 + \sigma_{\theta_k}^4 + 2k\sigma_\theta^2\sigma_{\theta_k}^2)}{8\left(k^2(\sigma_\varepsilon^2 + \sigma_\theta^2) + k(\sigma_{\varepsilon_k}^2 + \sigma_{\theta_k}^2)\right)} \\
&- \sum\nolimits_{k=1}^{N} \frac{3(k-1)^2\sigma_\theta^4}{8\left((k-1)^2(\sigma_\varepsilon^2 + \sigma_\theta^2) + (k-1)(\sigma_{\varepsilon_k}^2 + \sigma_{\theta_k}^2)\right)} \\
&\approx \sum\nolimits_{k=1}^{N} \frac{6k\sigma_\theta^2\sigma_{\theta_k}^2}{8\left(k^2(\sigma_\varepsilon^2 + \sigma_\theta^2) + k(\sigma_{\varepsilon_k}^2 + \sigma_{\theta_k}^2)\right)} \\
&\approx \sum\nolimits_{k=1}^{N} \frac{3\sigma_\theta^2\sigma_{\theta_k}^2}{4k(\sigma_\varepsilon^2 + \sigma_\theta^2)} \approx \frac{3\sigma_\theta^2\sigma_{\theta_k}^2}{4(\sigma_\varepsilon^2 + \sigma_\theta^2)}\log(N).
\end{aligned}$$

Therefore the average compensation also goes to 0 as N goes to infinity. ■

Proof of Proposition 17. We first verify that the candidate equilibrium indeed constitutes an equilibrium. Abusing notation, let J denote the set of intermediaries. Suppose purchasing data from intermediaries $D \subsetneq J$ were optimal for the producer. If $J\backslash D = \{j\}$, then by construction, purchasing J is equally profitable for the producer and thus optimal. If $J\backslash D$ contains two intermediaries, suppose j is one of them. In our Gaussian-linear model where all the noise is additive, the marginal value for information is decreasing in the number of sources. Since signals are imperfect (recall that we have assumed $\sigma_\zeta^2 > 0$), we have

$$\Pi(\{j\} \cup D) - \Pi(D) > \Pi(\{j\} \cup (J\backslash\{j\})) - \Pi(J\backslash\{j\}) = m_{0,j},$$

where $\Pi(D)$ denotes the producer's profit when purchasing S_D. Thus purchasing D is worse

than $D \cup \{j\}$ which is a contradiction.

On the intermediaries' side, fix the other intermediaries pricing $m_{0,\neg j}$. Suppose it is optimal for intermediary j to charge $m'_{0,j} > m_{0,j}$. Then suppose that, under the new price, the producer would purchase data from D. By construction of $m_{0,j}$ we know $D \neq J$, otherwise it is better not to purchase from j. For the same reason as in the last paragraph, it is also impossible for $J \backslash D$ to contains two intermediaries. Therefore it is only possible that $\{j'\} = J \backslash D$. However, since (denote $\bar{\Pi}$ as the profit of the producer under new price),

$$\begin{aligned}
\bar{\Pi}(J\backslash\{j\}) - \bar{\Pi}(D) &= (\Pi(J\backslash\{j\}) - \Sigma_{k\in J\backslash\{j\}} m_{0,k}) - (\Pi(D) - \Sigma_{k\in D} m_{0,k}) \\
&= -(\Pi(J) - \Pi(J\backslash\{j\})) + m'_{0,j} + \Pi(J) - \Pi(D) - m_{0,j'} \\
&= -(\Pi(J) - \Pi(J\backslash\{j\})) + m'_{0,j} > 0.
\end{aligned}$$

Therefore D is not optimal. This contradiction completes the verification of the equilibrium.

Now we proceed to uniqueness. We first argue that in any equilibrium, intermediary j cannot obtain a payoff lower than $\Pi(J) - \Pi(J\backslash\{j\})$. Suppose toward a contradiction that j only obtains $R_j < \Pi(J) - \Pi(J\backslash\{j\})$ in an equilibrium. It could always sell its data at $m_{0,j}$ such that $R_j < m_{0,j} < \Pi(J) - \Pi(J\backslash\{j\})$ and the producer would buy its database regardless of his other purchasing decisions, because of the decreasing returns to information sources.

Now suppose that in equilibrium, one intermediary j sells its database at $m_{0,j} > \Pi(J) - \Pi(J\backslash\{j\})$. From the last paragraph we know $m_{0,j'} \geq \Pi(J) - \Pi(J\backslash\{j'\})$. Therefore, on the equilibrium path, the producer should not purchase all databases, because buying all databases is strictly worse than buying from every intermediary except j. This is a contradiction, because the rejected intermediary will obtain zero profits, whereas we have shown in the last paragraph it can guarantee a strictly positive payoff. ■

Before proceeding to the proof of Proposition 18, we introduce an equilibrium refinement.

Definition 2 (Maximal Equilibrium)
A Perfect Bayesian Equilibrium is maximal if, for given offers by the intermediaries $\{(m_{i,j}, \sigma_{ij}, \sigma_j)\}_{j\in J}$, *the accepting sets of consumers* $\{A_i\}$ *are maximal. That is, there is no other acceptance set* $\{A'_i\}$ *that is an equilibrium of the subgame induced by* $\{(m_{i,j}, \sigma_{ij}, \sigma_j)\}_{j\in J}$ *such that* $A_i \subset A'_i$ *for all* i *and* $A_i \subsetneq A'_i$ *for some* i.

Lemma 2 (Unique Equilibrium for Accepting Game)
For given offers $\{(m_{i,j}, \sigma_{ij}, \sigma_j)\}_{j\in J}$, *there exists a unique maximal equilibrium.*

Proof of Lemma 2. Suppose there are two maximal accepting sets $\{A_i\} \neq \{\bar{A}_i\}$. We will construct accepting sets $\{A_i^\infty\}$ such that $A_i^\infty \supset A_i \cup \bar{A}_i \ \forall i$ which yields a contradiction.

We will use an iterated expansion approach to complete the construction. Denote $A_i^0 = A_i \cup \bar{A}_i$. Denote also $U_i(A_i, A_{\neg i})$ as consumers' payoff gross of the monetary transfers when the accepting sets are $\{A_i\}$. Since $\{A_i\}$ and $\{\bar{A}_i\}$ are equilibrium choices, we have:

$$\begin{aligned} U_i(A_i, A_{\neg i}) - U_i(A_i \backslash B_i, A_{\neg i}) + \Sigma_{j \in B_i} m_{i,j} &\geq 0 \quad \forall B_i \subset A_i, \\ U_i(\bar{A}_i, \bar{A}_{\neg i}) - U_i(\bar{A}_i \backslash B_i, \bar{A}_{\neg i}) + \Sigma_{j \in B_i} m_{i,j} &\geq 0 \quad \forall B_i \subset \bar{A}_i. \end{aligned}$$

Therefore, because of decreasing compensation, we will have:

$$U_i(A_i^0, A_{\neg i}^0) - U_i(A_i^0 \backslash B_i, A_{\neg i}^0) + \Sigma_{j \in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i, \text{or } \forall B_i \subset \bar{A}_i.$$

For $B_i \cup \bar{B}_i \subset A_i^0$ where $B_i \subset A_i$, $\bar{B}_i \subset \bar{A}_i$ and $B_i \cap \bar{B}_i = \varnothing$ we have:

$$\begin{aligned} & U_i(A_i^0, A_{\neg i}^0) - U_i(A_i^0 \backslash (B_i \cup \bar{B}_i), A_{\neg i}^0) + \Sigma_{j \in (B_i \cup \bar{B}_i)} m_{i,j} \\ = \; & U_i(A_i^0, A_{\neg i}^0) - U_i(A_i^0 \backslash B_i, A_{\neg i}^0) + \Sigma_{j \in \bar{B}_i} m_{i,j} \\ & + U_i(A_i^0 \backslash B_i, A_{\neg i}^0) - U_i(A_i^0 \backslash (B_i \cup \bar{B}_i), A_{\neg i}^0) + \Sigma_{j \in \bar{B}_i} m_{i,j} \geq 0. \end{aligned}$$

In conclusion we have:

$$U_i(A_i^0, A_{\neg i}^0) - U_i(A_i^0 \backslash B_i, A_{\neg i}^0) + \Sigma_{j \in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i^0.$$

If for each i, there exists a non-empty $B_i \subset (A_i^0)^c$ such that

$$U_i(A_i^0 \cup B_i^0, A_{\neg i}^0) - U_i(A^0, A_{\neg i}^0) + \Sigma_{j \in B_i^0} m_{i,j} \leq 0,$$

then we end the construction and denote $A_i^0 = A_i^\infty$. Otherwise, let $A_i^1 = A_i^0 \cup B_i^0$. Again because of decreasing compensation, we keep the following property within $\{A_i^1\}$:

$$U_i(A_i^1, A_{\neg i}^1) - U_i(A_i^1 \backslash B_i, A_{\neg i}^1) + \Sigma_{j \in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i^1.$$

This iterated process ends in finitely many steps, and the final set $\{A_i^\infty\}$ satisfies

$$\begin{aligned} & U_i(A_i^\infty, A_{\neg i}^\infty) - U_i(A_i^\infty \backslash B_i, A_{\neg i}^\infty) + \Sigma_{j \in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i^\infty, \\ & U_i(A_i^\infty \cup B_i, A_{\neg i}^\infty) - U_i(A_i^\infty, A_{\neg i}^\infty) + \Sigma_{j \in B_i} m_{i,j} < 0 \quad \forall \varnothing \neq B_i \subset (A_i^\infty)^c. \end{aligned}$$

To verify that $\{A_i^\infty\}$ is indeed an equilibrium for accepting game, we still need to check

possible deviation to $B_i \cup C_i$ where $B_i \subset A_i^\infty$, $C_i \subset (A_i^\infty)^c$:

$$\begin{aligned}
& U_i(A_i^\infty, A_{\neg i}^\infty) + \Sigma_{j\in A_i^\infty} m_{i,j} - U_i(B_i \cup C_i, A_{\neg i}^\infty) - \Sigma_{j\in B_i\cup C_i} m_{i,j} \\
= \; & U_i(A_i^\infty, A_{\neg i}^\infty) - U_i(A_i^\infty \cup C_i, A_{\neg i}^\infty) - \Sigma_{j\in C_i} m_{i,j} \\
& + U_i(A_i^\infty \cup C_i, A_{\neg i}^\infty) - U_i(B_i \cup C_i, A_{\neg i}^\infty) + \Sigma_{j\in A_i^\infty \setminus B_i} m_{i,j} \\
\geq \; & -(U_i(A_i^\infty \cup C_i, A_{\neg i}^\infty) - U_i(A_i^\infty, A_{\neg i}^\infty) + \Sigma_{j\in C_i} m_{i,j}) \\
& + U_i(A_i^\infty, A_{\neg i}^\infty) - U_i(B_i, A_{\neg i}^\infty) + \Sigma_{j\in A_i^\infty \setminus B_i} m_{i,j} \\
> \; & 0.
\end{aligned}$$

Given that other consumers choose A_{-i}^∞, consumer i choosing A_i^∞ is then indeed optimal. Therefore, $\{A_i^\infty\}$ is a equilibrium with strictly larger accepting set, which is a contradiction that completes the proof. ■

Proposition 19 (Unique Equilibrium Prediction)
In every pure-strategy maximal equilibrium, each intermediary collects information from every consumer without additional noise. The consumers are indifferent between accepting all offers and rejecting all of them.

Proposition 19 thus establishes the first bullet point of Proposition 18 holds. Before we proceed to prove Proposition 19, however, we introduce three lemmas.

Denote $\{\bar{A}_i\}$ as the the maximal accepting set given offers $\{(m_{i,j}, \sigma_{ij}, \sigma_j)\}_{j\in J}$. Considering any larger accepting set (which is not an equilibrium for the accepting game), we next show that there must be one consumer who strictly prefers a smaller accepting set.

Lemma 3 (Want Less When Asked More)
For any $\{A_i\} \supsetneq \{\bar{A}_i\}$, $\exists$ i and $B_i \subsetneq A_i$ such that:

$$U_i(A_i, A_{\neg i}) - U_i(A_i \setminus B_i, A_{\neg i}) + \Sigma_{j\in B_i} m_{i,j} < 0.$$

Proof of Lemma 3. Suppose not, then we have

$$U_i(A_i, A_{\neg i}) - U_i(A_i \setminus B_i, A_{\neg i}) + \Sigma_{j\in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i.$$

Since $\{A_i\}$ is strictly larger than maximal accepting set, it is not an equilibrium accepting set, therefore $\exists$ i and A_i' such that $A_i'^1$ is more profitable than A_i. Choose A_i^1 such that A_i^1 is the minimal one, i.e. none of its real subsets is more profitable than A_i. Then it is easy

to see that, due to decreasing returns, that A_i^1 preserve the property of A_i that it is better than its every subset:

$$U_i(A_i^1, A_{\neg i}) - U_i(B_i, A_{\neg i}) + \Sigma_{j \in A_i^1 \setminus B_i} m_{i,j} \geq 0.$$

Then we can use the iterated expansion again, and get $\{A_i^\infty\} \supsetneq \{A_i\}$ such that $\{A_i^\infty\}$ is a equilibrium accepting set, which yields a contradiction. ■

Lemma 4 (Every Intermediary Buys from Every Consumer)
In every pure-strategy maximal equilibrium, every intermediary in the market collects data from every consumer.

Proof of Lemma 4. Suppose that, on the equilibrium path, intermediary j' does not collect data from consumer i'. Denote the equilibrium accepting set, which is maximal, as $\{A_i\}$. We will construct a profitable deviation for j' towards a contradiction. In our construction, we will hold fixed the level of common noise.

As we saw in the proof of Lemma 2, we know that facing the equilibrium prices, consumers strictly prefer A_i to any larger set. Equivalently, the following strict inequality holds,

$$U_i(A_i, A_{\neg i}) - U_i(B_i, A_{\neg i}) - \Sigma_{j \in B_i \setminus A_i} m_{i,j} > 0 \quad \forall B_i \supsetneq A_i.$$

From Lemma 3 we know that for any strictly larger $\{\bar{A}_i\}$, there exist i and $\bar{B}_i$ such that consumer i prefers $\bar{B}_i$ over $\bar{A}_i$:

$$U_i(\bar{A}_i, \bar{A}_{\neg i}) - U_i(\bar{A}_i \setminus \bar{B}_i, \bar{A}_{\neg i}) + \Sigma_{j \in \bar{B}_i} m_{i,j} < 0.$$

We rewrite the above system of inequality of all possible B_i and $\bar{A}_i$ as:

$$\begin{aligned} U_i(A_i, A_{\neg i}, \sigma^2_{\varepsilon_{i'j'}} = \infty) - U_i(B_i, A_{\neg i}, \sigma^2_{\varepsilon_{i'j'}} = \infty) - \Sigma_{j \in B_i \setminus} m_{i,j} &> 0 \quad \forall B_i \supsetneq A_i, \\ U_i(\bar{A}_i, \bar{A}_{\neg i}, \sigma^2_{\varepsilon_{i'j'}} = \infty) - U_i(\bar{A}_i \setminus \bar{B}_i, \bar{A}_{\neg i}, \sigma^2_{\varepsilon_{i'j'}} = \infty) + \Sigma_{j \in \bar{B}_i} m_{i,j} &< 0. \end{aligned}$$

Here, $U_i(A_i, A_{-i}, \sigma^2_{\varepsilon_{i'j'}})$ is the hypothetical gross utility for consumer i from final goods market if all consumers' accepting sets are $\{A_i\}$ and in addition, consumer i' report data to intermediary j' with idiosyncratic noise $\varepsilon_{i'j'}$. The equilibrium path could be represented by setting $\sigma^2_{\varepsilon_{i'j'}} = \infty$.

Now we have a finite system of strict inequalities that are continuous in $\sigma^2_{\varepsilon_{i'j'}}$. Thus

intermediary j' could find a $\sigma^2_{\varepsilon^*_{i'j'}} < \infty$ such that all these equations still hold:

$$U_i(A_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - U_i(B_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - \Sigma_{j\in B_i\setminus A_i} m_{i,j} > 0 \quad \forall B_i \supsetneq A_i, \tag{38}$$

$$U_i(\bar{A}_i, \bar{A}_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - U_i(\bar{A}_i\setminus\bar{B}_i, \bar{A}_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) + \Sigma_{j\in \bar{B}_i} m_{i,j} < 0. \tag{39}$$

Moreover, from the equilibrium condition we have:

$$U_i(A_i, A_{\neg i}, \infty) - U_i(B_i, A_{\neg i}, \infty) + \Sigma_{j\in A_i} m_{i,j} - \Sigma_{j\in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i.$$

Therefore by decreasing compensation, we know:

$$U_i(A_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - U_i(B_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) + \Sigma_{j\in A_i} m_{i,j} - \Sigma_{j\in B_i} m_{i,j} \geq 0 \quad \forall B_i \subset A_i. \tag{40}$$

Now consider a deviation from intermediary j' that leaves $\sigma_{j'}$, $\sigma_{ij'}$, $m_{i,j'}$ $\forall i \neq i'$ unchanged, sets $\sigma_{i'j'} = \sigma^*_{i'j'}$, and offers a price $m^*_{i',j'}$ to consumer i' that keeps i' indifferent between $A_{i'}$ and $A_{i'} \cup \{j'\}$:

$$U_{i'}(A_{i'}, A_{\neg i'}, \sigma^2_{\varepsilon^*_{i'j'}}) - U_{i'}(A_{i'}, A_{\neg i'}, \infty) + m^*_{i',j'} = 0.$$

Now we verify first that $A_i \cup \{j'\}$ and $A_{\neg i}$ are indeed optimal choices for each consumer, and that they are maximal. From these two conclusions, we can ensure $A_i \cup \{j'\}$ and $A_{\neg i}$ will be the accepting sets of the subgame, if intermediary j' makes such a deviation.

First, from inequality 38, we know that, given other consumers choose $\{A_i \cup \{j'\}, A_{\neg i}\}$, consumer $i \neq i'$ prefers A_i over any larger set, and consumer i' prefers $A_{i'} \cup \{j'\}$ over any larger set. From inequality 40, we know all consumers prefer A_i over any smaller set. Moreover, since by construction of $m^*_{i',j'}$, i' is indifferent between A_i and $A_i \cup \{j'\}$, then we know that $A_i \cup \{j'\}$, and that $A_{\neg i}$ is indeed an optimal choice for each consumer.

Second, for any set $\{\bar{A}_i\} \supsetneq \{A_i\}$ such that $j' \notin \bar{A}_{i'}$, inequality (39) holds. Therefore, $\{\bar{A}_i \cup \{j'\}, \bar{A}_{\neg i}\}$ cannot be equilibrium choices for consumers. Thus no set larger than $\{A_i \cup \{j'\}, A_{\neg i}\}$ can be an equilibrium choice. One last step toward a contradiction is to verify that such deviation is profitable. From the previous analysis, we know that the information collected by any other intermediary j would remain unchanged. Thus we can denote their signal simply by $s_j = \theta + \delta_j$, where δ_j is a independent normal random variable. Because we have assumed a strictly positive lower bound on the noise $\sigma^2_\zeta > 0$, we must have $\sigma^2_{\delta_j} > 0$. Denote the signal sent by j' before deviating as $\theta + \delta_{j'}$ and the signal sent after deviating by $\theta + \hat{\delta}_{j'}$, where $\sigma^2_{\delta_{j'}} > \sigma^2_{\hat{\delta}_{j'}}$. By construction, we then have

$$m_{i,j} = \hat{U}_i(A_i) - \hat{U}_i(A'_i) = \frac{3}{8}\,\mathrm{var}\left[\mathbb{E}\left[\theta|s_{\neg j'}\hat{s}_{j'}\right]\right] - \frac{3}{8}\,\mathrm{var}\left[\mathbb{E}\left[\theta|s_{\neg j'}s_{j'}\right]\right].$$

On the other hand, the extra fee intermediary j' could charge to the producer, according to Proposition 17 is:

$$\Pi(\hat{s}_{j'}, s_{\neg j}) - \Pi(s_{\neg j}) - (\Pi(s_{j'}, s_{\neg j}) - \Pi(s_{\neg j})) = \frac{N}{4} \operatorname{var}\left[\mathbb{E}\left[\theta | s_{\neg j'} \hat{s}_{j'}\right]\right] - \frac{N}{4} \operatorname{var}\left[\mathbb{E}\left[\theta | s_{\neg j'} s_{j'}\right]\right].$$

Thus as long as $N \geq 2$, such a deviation is always profitable. ■

Proof of Proposition 19. First suppose there exist a consumer i and an intermediary j' such that in equilibrium $(\sigma^2_{\varepsilon_{ij'}}, \sigma^2_{\varepsilon_{j'}}) > (\sigma^2_\zeta, 0)$. This means one intermediary adds extra noise of either kind. From Lemma 4 we know that, in any equilibrium, $A_i = J$ for any i. Consider a deviation from intermediary j': it changes $(\sigma^2_{\varepsilon_{ij'}}, \sigma^2_{\varepsilon_{j'}})$ to $(\sigma^2_\zeta, 0)$ and increases compensation to $m^*_{i,j'} + \delta$, for some small $\delta > 0$, where

$$m^*_{i,j'} = m_{i,j'} + U_i(J, \sigma^2_{\varepsilon^*_{ij'}}) - U_i(J, \sigma^2_{\varepsilon_{ij'}}).$$

From original equilibrium condition we know:

$$U_i(J, J, \sigma^2_{\varepsilon_{ij'}}, \sigma^2_{\varepsilon_{j'}}) - U_i(B_i, J, \sigma^2_{\varepsilon_{ij'}}, \sigma^2_{\varepsilon_{j'}}) + \Sigma_{j \in J \setminus B_i} m_{i,j} \geq 0 \quad \forall B_i.$$

Because of decreasing compensation and the fact that $\hat{\sigma}^2_{\varepsilon_i} > 0$, we immediately have:

$$U_i(J, J, \hat{\sigma}^2_{\varepsilon_i}, \hat{\sigma}^2_\varepsilon) - U_i(B_i, J, \hat{\sigma}^2_{\varepsilon_i}, \hat{\sigma}^2_\varepsilon) + \Sigma_{j \in J \setminus B_i} m_{i,j} > 0 \quad \forall B_i \supset \{j'\}.$$

By the construction of $m_{i,j}$ we also have:

$$\begin{aligned} & U_i(J, \sigma^2_{\varepsilon^*_{ij'}}) - U_i(B_i, \sigma^2_{\varepsilon^*_{ij'}}) + \Sigma_{j \in J \setminus (B_i \cup \{j'\})} m_{i,j} + m^*_{i,j'} + \delta \\ \geq \; & U_i(J, \sigma^2_{\varepsilon_{ij'}}) - U_i(B_i, \sigma^2_{\varepsilon_{ij'}}) + \Sigma_{j \in J \setminus B_i} m_{i,j} + \epsilon > 0 \quad \forall B_i \subset \{j'^c\} \end{aligned}$$

Therefore J will still be consumer i's optimal choice given other all accept J. other consumers on the other hand prefer J even more because more information is revealed by i. Therefore every consumers accepting every intermediary's offer is indeed a equilibrium outcome, and it is the outcome induced by the deviation in our setting since it is clearly the largest. A similar argument as in Proposition 4 then shows that this deviation is profitable, which leads to a contradiction.

Finally, we will establish the indifference condition from consumers' side. Suppose consumer i strictly prefers reporting to all intermediaries to none, then denote $C = \{C_1, C_2, ...\}$ as the set of all optimal accepting choices for i given others all accept J ($J \in C$). Note that C is complete under set inclusion, because if $C_1 \not\subset C_2$, $C_2 \not\subset C_1$, and they are equally good,

then by decreasing return $C_1 \cup C_2$ is strictly better. Thus we could assume $C_1 \not\subset C_2 \not\subset ... \not\subset J$. By assumption we know $C_1 \neq \varnothing$.

Then consider a deviation for $j' \in C_1$. This intermediary could deviate by making $m_{i,j'}$ slightly smaller. Under this deviation, and assuming other consumers all choose J, every accepting choice in C brings less utility to consumer i equally and is still strictly better than the rest, thus it is still optimal for i to choose J. So all consumers accepting all offers is still an equilibrium outcome under this deviation and (since it is clearly maximal) it is our selected equilibrium. This deviation is profitable to j', which is a contradiction. Therefore the indifference conditions for all consumers must hold. Since the noise structure is uniquely pinned down (no noise is added), and consumers are indifferent between selling to all intermediaries and selling to none, consumers surplus, producer surplus, total profit of the intermediaries are also uniquely determined. ■

Proof of Proposition 18. Proposition 18.1 was established by Proposition 19. Proposition 18.2 is a direct corollary of Proposition 18.1. We now compute the equilibrium payoff of all the players. On the equilibrium path, each consumer i will send intermediary j a noisy signal $\theta + \varepsilon_{ij}$. Therefore on the path, the producer will get a noisy signal $\theta + \delta$ where δ has variance $\sigma_\zeta^2/(NJ)$. If one consumer i deviates by rejecting all offers, the producer will get a noisy signal $\theta + \delta_{-i}$ where δ_{-i} has variance $\sigma_\zeta^2/((N-1)J)$. If there is one intermediary not selling its database to the producer, the producer will get a noisy signal $\theta + \delta_{-j}$ where δ_{-j} has variance $\sigma_\zeta^2/(N(J-1))$.

As we have shown before, the value of information for the producer and the consumer surplus are given by

$$\Pi(S) = \frac{N}{4}\operatorname{var}[\mathbb{E}[\theta|S]] = \frac{N}{4}\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\delta^2}$$
$$U_i(S) = -\frac{3}{8}\operatorname{var}[\mathbb{E}[\theta|S]] = -\frac{3}{8}\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\delta^2}$$

The equilibrium consumer surplus is increasing in σ_ζ^2 and decreasing in J:

$$U_i(S_{-i}) = -\frac{3}{8}\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/((N-1)J)}$$

The producer surplus is increasing in J:

$$\Pi(S) - J(\Pi(S) - \Pi(S_{-j})) = \frac{N}{4}\Big(\frac{J\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/(N(J-1))} - \frac{(J-1)\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/(NJ)}\Big)$$
$$= \frac{N\sigma_\theta^4}{4}\frac{\sigma_\theta^2}{(\sigma_\theta^2 + \sigma_\zeta^2/(N(J-1)))(\sigma_\theta^2 + \sigma_\zeta^2/(NJ))}$$

The total profit of intermediaries is:

$$R = J(\Pi(S) - \Pi(S_{-j})) + N(U_i(S) - U_i(S_{-i}))$$
$$= \frac{N}{4}\Big(\frac{J\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/(NJ)} - \frac{J\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/(N(J-1))}\Big) + \frac{3N}{8}\Big(\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/((N-1)J)} - \frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\zeta^2/(NJ)}\Big)$$
$$\lim_{J\to\infty}\frac{\partial R}{\partial J}J^2 = -\frac{(2N-5)\sigma_\zeta^2}{8(N-1)} < 0$$
$$\lim_{N\to\infty} R = \frac{\sigma_\zeta^2}{4(J-1)}.$$

This completes the proof. ■

Propositions 18 and 19 provide a characterization of the equilibrium outcome, but do not establish the existence of an equilibrium.

Proposition 20 (Symmetric Equilibrium)
For any $N \geq 3$, $J \geq 2$, and $\sigma_\theta^2 > 0$, there exists $x > 0$ such that, if $\sigma_\zeta^2 \geq x$, a symmetric equilibrium exists in which all intermediaries offer the same compensation to every consumer.

As we have just argued, in the symmetric candidate equilibrium, the potential deviation for the intermediary is to deviate to some information structure with higher noise such that he could persuade consumers not to provide any information to other intermediaries. The following lemma characterizes "the most profitable deviation."

Lemma 5 (Best Candidate Deviation)
The best deviation for intermediary j' consists of offering a data policy with $\sigma^2_{\varepsilon^*_{ij'}} = 0$, $\sigma_{\varepsilon^*_{j'}} > 0$ and price $m^*_{i,j'}$ such that:

1. *conditional on every other consumer reporting her information only to j', consumer i is indifferent between reporting to no one and only to j';*

2. *conditional on every other consumer reporting her information to all intermediaries, consumer i is indifferent between reporting only to j' and to all intermediaries.*

Proof of Lemma 5. Suppose there is a profitable deviation for intermediary j', and denote it as $\{\sigma^*_{ij'}, \sigma^*_{j'}, m^*_{i,j'}\}$, and the accepting set it induces as $\{A^*_i\}$. If such deviation does not reduce the information that other intermediaries collect, i.e. $A^*_i = A_i = J \ \forall i \neq i'$, then we know such deviation is not profitable for sure.

Now we will argue that $A^*_i \subset \{j'\}$ for every profitable deviation. Suppose not, then there must be some consumer i_0 report to some other intermediary $j_0 \neq j'$. Note by previous paragraph we know at least one consumer i_1 does not report to some other intermediary $j_1 \neq j'$. Therefore we have the following two inequalities:

$$\begin{aligned} U_i(A^*_{i_0}, A^*_{\neg i_0}) - U_i(A^*_{i_0} \backslash \{j_0\}, A^*_{\neg i_0}) + m_{i,j} &\geq 0, \\ U_i(A^*_{i_1} \cup \{j_1\}, A^*_{\neg i_1}) - U_i(A^*_{i_1}, A^*_{\neg i_1}) + m_{i,j} &\leq 0. \end{aligned}$$

Note the other intermediary's compensation $m_{i,j}$ is the same since we are considering the symmetric candidate. But this two inequalities contradict with decreasing return and imperfect information transmission. (This is a stronger form of decreasing return.)

Now since $A^*_i \subset \{j'\}$, we could assume $A^*_i = \{j'\}$ and put $\sigma^2_{\varepsilon_{ij'}} = \infty$, $m_{i,j'} = 0$ when j' is actually not collecting data from i.

Up to now, we have argued that if a profitable deviation exist, it must induce $A^*_i = \{j'\}$, what remaining is to set noise level and compensation optimally under the constraint. It is clear that to support $A^*_i = \{j'\}$ under maximal accepting assumption, we need at least these two conditions:

1. $m^*_{i,j'}$ need to at least make consumer indifferent between reporting to j' and to none, conditioned on other consumers only report to j';

2. the noise shall be large enough so that every consumer reporting to every intermediary is not an equilibrium outcome.

Since we are just considering deviation from symmetric candidate equilibrium, if consumer i find it attractive to report to $j \neq j'$, by decreasing return, she shall find it also attractive to report to any other $j \neq j'$. Therefore the second requirement is equivalent to: the noise shall be large enough so that, even when all other consumers report to all intermediaries, reporting only to j' is better than reporting to all.

On the other hand, because of symmetry again, once these two conditions are satisfied, we know that every consumer reporting only to j' is the maximal accepting set: now that the consumer do not want to report to $j \neq j'$ even when all other consumers reporting to all intermediaries, they will never want to do so when less information is transmitted.

Now we need to pin down the noise level. By rescaling and aggregating the noisy reports with the proper weights, we can represent the signal collected by intermediary j' on the equilibrium path as $\theta + \delta$, while the signal without report from consumer i as $\theta + \delta_i$, where

$$\sigma_\delta^2 = \frac{1}{\sigma_{\varepsilon_{j'}}^2} + \frac{1}{\sum_i \frac{1}{\sigma_{\varepsilon_{ij'}}^2}} \qquad \sigma_{\delta_i}^2 = \frac{1}{\sigma_{\varepsilon_{j'}}^2} + \frac{1}{\Sigma_{i \neq i'} \frac{1}{\sigma_{\varepsilon_{ij'}}^2}}$$

Therefore, by reduce $\sigma_{\varepsilon_{ij'}}^2$ (or increase $1/\sigma_{\varepsilon_{ij'}}^2$ if you worry about infinity) and increase $\sigma_{\varepsilon_{j'}}^2$ correspondingly to keep σ_δ^2 unchanged, we could always reduce $\sigma_{\delta_i}^2$. This is profitable because it weakens the consumers' bargaining power and reduces the necessary compensation. Thus we know it is optimal to set $\sigma_{\varepsilon_{ij'}}^2 = 0$.

What is left to pin down is the level of common noise. The requirement that $A_i^* = \{j'\}$ gives a system of lower bounds for the common noise, the most stringent one of which is

$$U_i(\{j'^*_{\neg i}) - U_i(J, A^*_{\neg i}) + (|J| - 1)m_{i,j} > 0.$$

The intermediary j' would always want to decrease noise level as long as the above inequality is satisfied. Thus the payoff from an optimal deviation is obtained by providing an information structure with zero idiosyncratic noise and a level of common noise that is arbitrarily close to making this constraint bind. ■

Having identified the best deviation, we then calculate the associated deviation payoff to verify the existence of a symmetric equilibrium. It is then easy to show that the deviation is not profitable if we take $x \to \infty$ (with $N \geq 3$). This ends the proof of Proposition 20.

References

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

Argenziano, R., and A. Bonatti (2019): "Information Revelation and Consumer Privacy," Discussion paper, MIT.

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.

Ball, I. (2020): "Scoring Strategic Agents," Discussion paper, Yale University.

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): "Data for Service," Discussion paper, Yale University.

Bergemann, D., A. Bonatti, and A. Smolin (2018): "The Design and Price of Information," *American Economic Review*, 108, 1–45.

Bergemann, D., B. Brooks, and S. Morris (2015): "The Limits of Price Discrimination," *American Economic Review*, 105, 921–957.

Bonatti, A., and G. Cisternas (2019): "Consumer Scores and Price Discrimination," *Review of Economic Studies*, forthcoming.

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

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

Haghpanah, N., and R. Siegel (2019): "Consumer-Optimal Market Segmentation," Discussion paper, Pennsylvania State University.

Ichihashi, S. (2019): "Non-competing Data Intermediaries," Discussion paper, Bank of Canada.

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

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.

Stigler Committee on Digital Platforms (2019): "Final Report," Discussion paper, University of Chicago.

The New York Times (2020): "Intuit to Buy Credit Karma to Create Financial Data Giant," February 24.

Westenbroek, T., R. Dong, L. Ratliff, and S. Sastry (2019): "Competitive Statistical Estimation with Strategic Data Sources," *IEEE Transcaction on Automatic Control*.

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