THE ECONOMICS OF SOCIAL DATA

By

Dirk Bergemann, Alessandro Bonatti, and Tan Gan

September 2019

COWLES FOUNDATION DISCUSSION PAPER NO. 2203

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§

September 25, 2019

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 product offers and prices. The social dimension of the individual data—whereby an individual's data is predictive of the behavior of others—generates a *data externality* that reduces the intermediary's cost of acquiring information. We derive the data intermediary's optimal information policy, and show that it preserves privacy over the identity of the consumers, but provides precise information about market demand to the firms.

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

JEL CLASSIFICATION: D44, D82, D83.

*We thank Daron Acemoglu, Steve Berry, Nema Haghpanah, Nicole Immorlica, Scott Kominers, Annie Liang, Roger McNamee, Enrico Moretti, Stephen Morris, Fiona Scott-Morton and Glen Weyl for helpful discussions. We thank the audience at ESSET 2019 and 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 the Internet The rise of large Internet platforms, such as Facebook, Google, and Amazon in the US, and similar large entities in China, such as JD, Tencent and Alibaba has lead to an unprecedented collection and commercial use of individual data. The ever increasing user base of these platforms generates 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. Bergemann and Bonatti (2019) provide a recent introduction.

A central feature of the data collected from the individuals is its social aspect. Namely, the data captured from an individual user is not only informative about that specific individual, but also about (appropriately defined) nearby individuals. Thus, the *individual data* is really *social data*. The social nature of the data generates a *data externality*. In the context of geolocation data, 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 of consumers with similar purchase histories.

The recent disclosures on the use and misuse of social data by Internet platforms indicate the need to reflect about the largely unsupervised and unregulated use of individual data by these companies. To the extent that individual users provide most of the original data in their interaction with these platforms, it is important to understand the nature of the trade between Internet platforms and their users, and whether individual consumers receive the appropriate compensation for their data.[1]

This question gains importance in the presence of the externality generated by the social data. We know that in the presence of economic externalities, such as the environmental externality of carbon emissions, the market by itself rarely guarantees the socially efficient outcome. In the case of data markets, this means that granting individuals control rights over their personal data is likely an insufficient intervention.

We analyze three critical aspects of the economics of social data. First, how the collection of individual data changes the terms of trade between consumers, advertisers, and large Internet platforms. Second, how the social dimension of the data magnifies the value of

[1] 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."

individual data for the platforms. Third, how the indirect sale of data (e.g., through the provision of targeted advertising) changes the information available in equilibrium about individual consumers.

How Data is Used The business models of large Internet platforms share some important structural similarities. To a first approximation, companies such as Amazon, Facebook, and Google are technology platforms that facilitate matches. The collection of data about the matching partners is then critical for the success of these platforms: a larger database about the characteristics of potential partners increases the quality and quantity of matches (see Bergemann, Bonatti, and Smolin (2018)). These companies monetize the value of their matching services mostly through advertising revenues. Thus, the information about an individual consumer impacts the volume of trade and the level of surplus generated on a platform (see Bergemann and Morris (2019)).

However, the very same information that is valuable to form a good match will typically also impact the way the resulting surplus is distributed. In particular, advertisers value the information that the Internet platforms collect, as it enables to segment the market by tailoring their advertising and pricing decisions (see Bergemann and Bonatti (2015)). For example, if a seller had even partial information about buyers, it could offer different product varieties to different subsets of consumers, or charge different prices to different segments of the population, on the basis of some observable characteristics. In all these cases, information allows sellers to change the terms of trade. Thus, information not only affects the amount of surplus generated online; it can significantly change the way in which that surplus is shared between buyers and sellers.

The value of information can then be positive for one party and negative for the other party. Indeed, the ability to tailor the terms of trade clearly benefits the sellers, who can reach more customers and offer them tailored products at prices that are closer to their willingness to pay. In contrast, the impact of the additional data on consumer surplus is less clear cut. While *perfect* price discrimination is clearly harmful for the consumer, whether other forms of market segmentation are harmful or beneficial is a priori ambiguous—see Bergemann, Brooks, and Morris (2015) for a general statement and result. Therefore, the *overall* effect of information markets on consumer welfare is entirely an empirical matter. However, a key open question for economic theory is whether the market for social data leads to an efficient allocation of information.

A Model of Data Intermediation We develop a model of data intermediation with 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) decides upon the quantity she wishes to purchase and a single firm (he) decides the unit price at which it offers the product to the consumer. In the product market, there is 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 can acquire demand information from the individual consumers and then sell the data in some, possible aggregated and bundled version to the firm. The data intermediary can choose how much information to buy from the consumers and how much information to sell to the firm.

We refer to the data market as the upstream market and the product market as the downstream market. First, the data is exchanged in the data market. Second the producer offers prices in the product market. It is the pricing decision of the producer that links the product market and the data market. As the pricing policy of the firm responds to the data acquired, it provides the interaction between the data market and the product market.

The Value and Price of Social Data The social dimension of the data—whereby an individual's data is also possibly 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, as consumers become better informed and empowered to take control of their data, firms will need to compensate them for their information, which will disrupt their business model. Indeed, if a consumer anticipates any negative consequences of revealing her information, she may demand compensation, e.g., through the quality of the services received.

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 individual data he provides is not part of her decision making. It follows that a platform has to compensate the 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, nor for any changes in her welfare caused by information revealed by others. In consequence, the cost of acquiring the individual data can be substantially below the value of information to the platform.

The resulting difference between the possible revenue gain in the interaction with many consumers and the small compensation necessary to acquire that information likely drives the extraordinary appetite of the Internet platforms to gather information.[2] We can now see

[2]The Furman report identifies "the central importance of data as a driver of concentration and barrier to

how social data drives a wedge between the efficient and the profitable uses of information. While many uses of consumer information exhibit positive externalities (e.g., real-time traffic information for driving directions), very little stands in the way of the platform trading data for profitable uses that are, in fact, harmful to consumers. The presence of an informational externality thus indicates that the standard argument for competitive prices to establish efficient trade does not necessarily operate in these markets.

Data Intermediary - Data Platform - Data Aggregator The model of data intermediation suggested here most immediately reflects the business model of data brokers. But in fact our results apply directly to a large class of data platforms. It might be useful to distinguish between three different types of data platforms: (*i*) data brokers who buy and sell information; (*ii*) product data platforms; and (*iii*) social data platforms. By product data platform, we refer to data platforms such as Amazon, Uber and Lyft that acquire individual data from the consumer through the purchase of services and products. By social data platform, we refer to the internet platforms like Google and Facebook which offer data services to individual users, offer them digital services, 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 role of data intermediation and product pricing. As we will see, the transfer of information between the data intermediary and the product firm maximizes the joint surplus of the two entities, and thus our results will apply directly to the analysis of the product platform. A distinguishing feature of social data platforms is that these platforms 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) rather than in the level of monetary compensation. In addition, the information is frequently sold item-by-item in an auction format.[3]

Related Literature 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

competition in digital markets" (Digital Competition Expert Panel (2019)). The social dimension of data helps explain these forces.

[3]The distinction in the data gathering model translates naturally into different kinds of social data being collected. The data gathered on facebook, instgram, snapchat most directly corresponds to social data that maps into a social network. Data product platform establish social networks through revealed preference in terms of similar purchase behavior, similar demographics. In terms, data brokers buy and sell both direct and indirect social data. To a first approximation, the differences in the data can be represented in terms of their informativeness about the demand of the consumer.

consumer who makes her participation decision *ex ante*. In other words, the consumer makes a single decision as to whether to use a platform's services, rather than trying to influence a data broker's perception of her type, as in Bonatti and Cisternas (2019). Our work is also 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).

Ichihashi (2019) studies competing data intermediaries who acquire perfect information from consumers. His model predicts multiple equilibria, in some of which competition hurts consumers equally as monopoly—see also Westenbroek, Dong, Ratliff, and Sastry (2019). In contrast, our competition model assumes the information collected by each broker contains noise, which means information collected by each broker will never be valueless. This leads to uniqueness of equilibrium. In this equilibrium the consumers surplus are worse off than in the monopoly environment. 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 for 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 where asymmetric users with exogenous privacy concerns trade information with a data platform, and they derive conditions under which the provision of information is (in)efficient.

2 Model

We initially consider the case of a single data intermediary in the data market, and a single firm in the product market. In later sections, we generalize the analyze and 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 chooses a quantity level q_i to maximize her net utility given a unit price p_i offered by the firm to consumer i:

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

Each consumer i has a true willingness-to-pay for the firm's product that is given by:

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

The willingness-to-pay w_i of consumer i is the sum of a *common* demand shock θ and an *idiosyncratic* demand shock θ_i. Thus, the demand of each consumer has a component θ that is common to all consumers in the market, and an idiosyncratic component θ_i that reflects the idiosyncractic taste. 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 offer his product to each consumer i. The producer has a linear production cost

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

The producer seeks to maximizes his profit:

$$\pi \triangleq \sum_i (q_i - c)\, p_i.$$

The producer knows the structure of the demand, and thus the common prior distribution given by (3). But absent any additional information from the data intermediary, the firm does now know the realized demand shocks prior to setting his price. As a consequence, in the absence of any additional information from the data intermediary, it is optimal for the producer to offer a uniform unit price to all consumers.

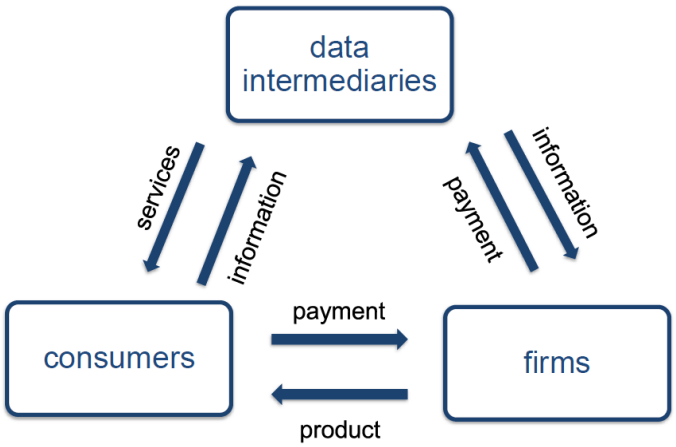
2.2 Data Market

The data market is run by a monopolist data intermediary. The data intermediary can acquire demand information from the individual consumers and then packages the demand information and sell it to the firm. 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 the bilateral contracts *ex-ante*, that is before the realization of any demand shocks. 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 flow of information. The data policy allows the data intermediary to control the

in-flow as well as the out-flow of information. Thus, the data intermediary can control the price as well as the design of the information (flow). As the market maker, the data intermediary decides how to collect the information from the consumers, and how to transmit it to the firm. Thus, the data intermediary faces an information design problem.

The data and product markets are summarized in Figure 2.2 below.



Data Intermediation

With respect to the information flows, we maintain the restriction to normally distributed random variables. Each consumer observes a noisy signal of her willingness-to-pay:

$$r_i \triangleq w_i + \zeta + \zeta_i. \tag{4}$$

The noise can have common and idiosyncratic component, ζ and ζ_i, respectively. The noise variables are independent with zero mean and variance σ_ζ^2 and $\sigma_{\zeta_i}^2$. The signal r_i may reflect the data-producing activity of the consumer. For example, we may interpret r_i as the search term that consumer i enters into a search engine like Google or her activity on a social network like Facebook. In addition to the value of r_i, the identity of consumer i is itself also privately known by i.

The (symmetric) data policies with respect to the consumers are exhaustively described by the following class of random variables:

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

Thus, the *data in-flow* from consumer i, which is her noisy private information r_i, may be

subject to a common and an idiosyncratic noise terms, ε and ε_i, with variance, σ_ε^2 and $\sigma_{\varepsilon_i}^2$, respectively. We refer to the in-flow data described by s_i as an information structure S_i.

In turn, the out-flow data policy is given by a vector of signals about each consumer i:

$$t_i \triangleq \sum_j \alpha_{ij} w_j + \eta + \eta_i, \tag{6}$$

with weights $\alpha_{ij} \in \mathbb{R}$. The *data out-flow* may again be subject to a common and an idiosyncratic noise terms, η and η_i, with variances given by σ_η^2 and $\sigma_{\eta_i}^2$, correspondingly. We refer to the out-flow data described by $t = (t_1, .., t_i, ..., t_I)$ as the information structure T.

The data intermediary makes a bilateral offer to each consumer i with the terms under which the consumer will share her information. Thus, the data intermediary offers an information structure and a monetary fee $m_i(T)$ to consumer i for the transmitted information,

$$m_i : S_i \to \mathbb{R}. \tag{7}$$

The information structure S_i being transmitted can be the entire information of consumer i or some, possibly noisy, statistic of her information, as described above.

Similarly, the data intermediary offers to share its information about the consumers with the firm, and in exchange asks for a transfer fee

$$m_0 : T \to \mathbb{R}. \tag{8}$$

The objective function of the data intermediary is to maximize the net revenue it receives from the consumers and the firm:

$$R \triangleq \sum_{i=0}^{I} m_i.$$

A priori, the respective monetary fees can be positive or negative. The data structure in the data market is given by (S, T).

2.3 Equilibrium and Timing

The game proceeds sequentially. First, the terms for trade on the data market are determined, and then the terms of trade in the product market are established. Thus, we sometimes refer to the data and product market as upstream and downstream markets, respectively.

The data intermediary can commit to bilateral contracts with the consumers and to a bilateral contract with the firm. The contract or mechanism determines how much informa-

tion gets transmitted and how the information is used. Given the information received, the firm then chooses to optimally employ its information in the interaction with the consumers. The firm therefore does not commit itself to any particular use of the information received. This allows the firm to use the information opportunistically.

At the contracting stage, the information is imperfect but symmetric, and hence it suffices to consider the subgame perfect equilibrium of the game. The analysis thus proceeds by backward induction.

Given an agreed data policy (S,T), the pricing policy of the firm is informed by the data out-flow that it receives, thus

$$p^{*}:T\rightarrow\mathbb{R}^{N}$$

The optimal price can be a vector of individualized prices, thus $p^{*}(t)\in\mathbb{R}^{N}$.

The resulting net-revenue is given by:

$$\mathbb{E}\left[\sum_{i}q_{i}\left(p_{i}-c\right)|S,T\right].$$

Proceeding backwards, if the firm accepts the proposal of the data intermediary then its profit is given by:

$$\Pi_{i}\left(S,T\right)\triangleq\mathbb{E}\left[\sum_{i}q_{i}\left(p_{i}-c\right)|S,T\right]-m_{0}\left(T\right).$$

By contrast, if the firm chooses not enter into a contract with the data intermediary, then the profit of the firm is given by:

$$\Pi_{i}\left(S,\varnothing\right)\triangleq\mathbb{E}\left[\sum_{i}q_{i}\left(p_{i}-c\right)|S,\varnothing\right].$$

If the firm does not receive any information from the data intermediary, it will not have pay for any data either. The resulting profit is given by

$$\mathbb{E}\left[\sum_{i}q_{i}\left(p_{i}-c\right)|S,\varnothing\right]=\mathbb{E}\left[\sum_{i}q_{i}\left(p_{i}-c\right)\right].$$

The participation constraint for the firm is thus

$$\Pi_{i}\left(S,T\right)\geq\Pi_{i}\left(S,\varnothing\right). \tag{9}$$

Proceeding backwards, the data intermediary, having put a data policy S into place, offers a data policy T, possibly as a function of S, thus $T\left(S\right)$ and an associated data price $m_{0}\left(T\left(S\right)\right)$

that maximizes its overall data revenue:

$$R\left(S, T\left(S\right)\right) = m_0\left(T\left(S\right)\right) + \sum_{i=1}^{I} m_i\left(S\right). \tag{10}$$

Proceeding backwards on step further, each consumer receives an offer for a data contract $(S_i, m_i(S_i))$. We denote the expected utility of consumer i from data sharing by

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

Each data contract $(S_i, m_i(S_i))$ has to satisfy the participation constraint for each consumer. The alternative to the proposed data contract is to decline the contract $(S_i, m_i(S_i))$. The resulting payoff, keeping the decision of the remaining consumers unchanged, then traces the payoff consequences. Consumer i would not receive the compensation for the data, but also not participate in the data sharing. Subsequently this may affect the optimal data policy $T(\varnothing, S_{-i})$ and the resulting pricing policy, thus

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

In equilibrium the data intermediary chooses a data policy $(S, T(S))$ that satisfies the participation constraint of the firm (9) and the participation constraint of every consumer i:

$$U_i\left(S, T\left(S\right)\right) \geq U_i\left(\varnothing, S_{-i}, T\left(\varnothing, S_{-i}\right)\right), \text{ for all } i. \tag{11}$$

A recursive equilibrium is then given by a triple of in-flow data policy, out-flow data policy and pricing policy:

$$\left\{\left(\left(S^*, m^*\left(S\right)\right); \left(T^*\left(S\right), m^*\left(T^*\left(S\right)\right)\right); p^*\left(T, S\right)\right)\right\}. \tag{12}$$

We emphasize that the participation constraint of every consumer i and the firm are required to hold at the *ex-ante* level. Thus, the consumer (and the firm) are agreeing to the data policy before the realization of any specific willingness-to-pay w_i. The choice of the *ex-ante* participation constraint is meant to capture a situation where the consumer and the firm accepts the "terms of use agreement" or "terms of service" before any particular consumption choice or search event. This is similar to using Facebook, or amazon, or a search engine, where the account is established before the realization of any particular event. In particular, the consumer evaluates the consequence of the data-flow from an ex-ante point of view, and requires a level of compensation that allows her to share the information. Conditional

upon agreeing to share the information, there is no further incentive compatibility constraint which would guarantee the interim optimality of sharing the information.

We note that the recursive structure of the equilibrium implies that the data intermediary, once it has agreement about the data in-flow is at liberty to choose the data outflow so as to maximize its revenue. The only restriction on the data policy is a measurability condition, namely that the intermediary cannot sell more information than it has bought.

We can now summarize the exact timing of exogenous and endogenous variables, that is data realization and strategic choices as follows:

1. The data intermediary offers a fee m_i to each consumer i for the data acquisition. The consumers simultaneously accept or reject the intermediary's offer.
2. The data intermediary offers a fee m_0 to the firm. The firm accepts or rejects the offer.
3. The data r and the information flow (s, t) are realized and transmitted according to the terms of the bilateral contracts.
4. The firm set a unit price p_i for each consumer i who makes a purchase decision q_i.

3 Data in the Wild

We begin the analysis with a review of what the market outcome would be if the demand information of the consumers would be available to the firm. In turn, the firm would use the information to the extent possible in its pricing policy. In terms of the data policy it is thus as if we were to assume that the data would flow without friction from the consumer to the firm, and thus:

$$s_i = t_i = w_i, \forall i. \tag{13}$$

With all the available information, the firm will pursue a personalized pricing policy towards each individual consumer. As the true demand function of each consumer is given by:

$$q_i = w_i - p_i,$$

the optimal personalized price is p_i:

$$p_i^* = \frac{w_i + c}{2}, \tag{14}$$

and the realized demand is given by:

$$q_i^* = \frac{w_i - c}{2}. \tag{15}$$

The demand data thus allows the producer to engage in *personalized* pricing. Namely, the producer adapts his pricing policy to the willingness-to-pay w_i of consumer i. The producer increases his price p_i^* in response to an increase in demand. In response the equilibrium quantity q_i^* increases with the willingness-to-pay, but at half the rate it would if the consumer where to face a constant price.

The knowledge of the demand data has distinct implications for consumers and producers. In particular, we can compare the equilibrium outcome when the firm does not have any demand data beyond the prior distribution and when the demand data is available to the firm. The demand information allows the firm to engage in third degree price discrimination—the producer gains in revenue from a more tailored price. However, the consumer loses due to the distortion in her consumption that comes with a more responsive price. Consequently, the social value of information is negative—the sum of the consumer and producer surplus is declining with the data flow.

Proposition 1 (Third Degree Price Discrimination)
The demand data increases the profit of the producer, decreases the consumer surplus, and decreases the social surplus.

This comparison brings us to the seminal result of Robinson (1933) and Schmalensee (1981) regarding the impact of third-degree price discrimination in markets with linear demand (and all markets served). To highlight the connection, consider a given information structure R. Ex ante, the surplus of consumer i is given by

$$CS_i = \frac{1}{2}\mathbb{E}\left[(w_i - p)^2\right] = -\operatorname{cov}\left[w_i, p\right] + \frac{1}{2}\operatorname{var}\left[p\right] + \text{constant terms.} \tag{16}$$

Thus, the ability of the seller to tailor the price p to the willingness-to-pay is detrimental to the consumer. The profit of the seller is a function of the price:

$$\pi = N \cdot \mathbb{E}\left[p\right]^2 + N \operatorname{var}\left[p\right].$$

The complete availability of the demand data is admittedly an extreme benchmark in at least two respects. First, the firm may only have access to some noisy version of the demand data, and second it may only observe a sample of the consumer data. Thus, let us next suppose the data flow would be given by:

$$s_i = t_i = w_i + \varepsilon + \varepsilon_i, \text{ for } i = 1, ..., k,$$

for some $\sigma_\varepsilon^2, \sigma_{\varepsilon_i}^2 > 0$, and and where k be strictly smaller than I.

As the demand data becomes noisy, the ability of the producer to tailor the price to the demand of the consumer weakens. Moreover, to the extent that the producer only observes a subsample, he is restricted in his ability to offer personalized prices. The lack in the precision of the individual data can, however, be partially compensated with aggregate data. In this case, the firm's pricing policy involves *third-degree* price discrimination. As the demand shock of each consumer has an idiosyncratic as well as common component, the producer can use the demand data from his entire sample to estimate the demand of any specific consumer. The extent to which the aggregate demand data is informative about the individual demand data depends on the variance of the common shock, denoted earlier by σ_θ^2 and the variance of the idiosyncratic shock, denoted by $\sigma_{\theta_i}^2$. This suggests that there are two limiting case of interest: (i) when there is only common demand uncertainty, and hence $\sigma_{\theta_i}^2 = 0$; (ii) when there is only idiosyncratic demand uncertainty.

In these two limiting case, the impact that the additional information has on the revenue, is always positive, but of a distinct qualitative nature. In this case of common demand uncertainty, every additional sample allows the producer to reduce the common demand uncertainty by a decreasing among in terms of the conditional variance. The seller will maintain a uniform price towards the consumers , and the revenue function is concave in the sample size k. By contrast if there is only idiosyncratic uncertainty, then each additional sample unlocks the possibility of a personalized prices. It also help to reduce the common noise ε in the observation and thus refine the personalized price on all the other consumer. Thus, the revenue function is convex in the sample size k.

Proposition 2 (Noisy Individual Information)
Given the noisy individual demand data of k consumers, the profit of the firm is increasing in k, and

1. *it is convex in k if $\sigma_{\theta_i}^2 = 0$;*
2. *it is concave in k if $\sigma_\theta^2 = 0$.*

So far, we assumed that the firm could gain access to individual demand data, and thus could identify the demand data with a specific individual consumer. Frequently, the firm might have noisy demand data but it may not be able to link the demand data to any particular consumer. In this case, the demand data from a sample of k consumers simply constitute aggregate demand data. As a consequence, the demand data will inform the firm in its pricing policy towards the entire market, but it cannot provide support for personalized pricing anymore.

Proposition 3 (Noisy Aggregate Information)
Given the noisy aggregate demand data of k consumers:

1. *the profit of the firm is increasing and concave in k;*
2. *the consumers' surplus is decreasing and convex in k;*
3. *the social surplus is decreasing and convex in k.*

Thus, the aggregate demand data still allows the firm to perform third-degree price discrimination but limits its ability to extract surplus from the individual consumer. The above result would not change if the firm would have access to individual data, including identifying information about the consumer, but could distinguish between the consumers when it offers his product.

In the current model, the downstream interaction between the consumer and the firm is represented by a model of third-degree price discrimination. Moreover, each consumer has a linear demand given its expected willingness to pay and given any constant unit price chosen by the firm. The individual (and the aggregate demand) are thus as in the classic environment studied by Robinson (1933) and Schmalensee (1981). The central result in this setting is that while the average demand will not change, social welfare is lower while the firm's profit is higher under more information (and segmentation). In Robinson (1933) and Schmalensee (1981), the linear demand in each segment was implicitly assumed to arise from an aggregation of individual consumer demand in each segment, whereas we take the linear demand as coming from each consumer separately.

We briefly discuss a few implications that the choice of the downstream game has in the larger context of this paper.

First, the results of In Robinson (1933) and Schmalensee (1981) establish that the social value of information sharing are negative, as information that supports third-degree price discrimination leads to a lower social welfare. Thus, we are starting (intentionally) in an economic environment where absent market imperfection the information should not be shared and transmitted. Alternatively, we could have considered other game forms in the downstream interaction. To the extent, that the informational externality would have the same sign, we expect the welfare result to be similar.

We restrict our attention to second degree price discrimination with linear price tariffs. We could extend the analysis to consider a larger class of tariffs, say two-parts tariffs or general non-linear price tariffs. The linear-quadratic specification of consumer and firm means that many of results of Maskin and Riley (1984) apply here. In particular, the average price per unit decreases with q, and that the optimal menu can be implemented

with a menu of two-part tariffs (see also Tirole (1988)). The disadvantage of the nonlinear analysis is that the resulting indirect utility function will not be linear quadratic anymore, and thus less amenable to the analysis of information design.

4 Data Rights and Data Intermediation

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

We begin with a preliminary result regarding the nature of the interaction between data intermediary and firm. For the moment, we are considering the interaction between a single data intermediary and firm. Given the in-flow data policy S that the intermediary has established, the subsequent interaction between the intermediary and the single seller is efficient in the sense that the intermediary will implement a data outflow policy that maximize the gross profits of the firm.

Proposition 4 (Data Outflow Policy)
The data intermediary will offer a complete information sharing data policy, $T^*(S) = S$, for all S. The data policy $T^*(S) = S$ maximizes the gross revenue of the firm among all feasible outflow data policies given S.

This preliminary result remains valid until we introduce either competition among intermediaries or among firms. Until then, that is until we arrive at Section 8, the above result directs us towards the analysis of the in-flow data policy. This result also informs us that the subsequent result not only apply to the tripartite setting of consumers, intermediary and firm, but remain valid when intermediary and firm are acting as a single unit in pursuit of maximizing revenue. This will then describe platforms that both collect data from consumers and directly price services or goods to the consumers. The absence of any intermediation is relevant for integrated product platforms such as Amazon or Uber. In other words, since the interaction between data intermediary and firm occurs without friction (provided there is only a single intermediary and downstream firm), the presence of external contracting of the data does not matter for the structure of the optimal data policy.

4.1 Basic Data Intermediation

We begin the analysis with a basic version of data intermediation. The data broker collects and aggregates the individual information of all the consumers. Subsequently, the data

broker transmits the total demand information to the producer. In response, the producer tailors the price to the individual consumer using the total demand information. We then ask in Proposition 5 under what conditions a profitable market for data intermediation arises.

Proposition 5 (Basic Data Intermediation)
Suppose the broker transmits all information received.

1. *The profit is increasing in σ_θ^2 and decreasing in $\sigma_{\theta_i}^2$.*

2. *The profit is negative if $n = 1$, and is asymptotically positive if and only if*

$$\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\varepsilon^2} > \frac{1}{2}\frac{\sigma_{\theta_i}^4}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2}.$$

3. *There exists a threshold $\bar{n}$ such that the broker obtains positive profits if and only if the number of consumers is $n > \bar{n}$. (Need some condition, like $\sigma_\theta^2 \geq \sigma_\varepsilon^2$)*

A first step towards a more sophisticated data policy is anonymize the data and transmit only the aggregate demand data. The aggregate demand is the average of the individual demand. To the extent that the individual information transmitted by the consumer affects the estimate, the consumer requires a compensation for the information. But the demand of each individual consumer is coming from two sources, the idiosyncratic shock and the common shock. While each consumer has a monopoly over the transmission of the idiosyncratic shock, the firm can learn about the common shock not only from consumer i, but from all other consumers. This opens a wedge between the benefit of the information for the firm, and the cost of the information for the individual consumer. The informational externality is then at the source of the profitable trade of information for the data broker, even in the absence of socially valuable information.

Proposition 6 (Basic Aggregate Information Intermediation)
Suppose the broker transmits only the aggregate demand information:

1. *there exists a threshold $\overline{\overline{n}}$ such that the broker obtains positive profits if and only if the number of consumers is*

$$n > \overline{\overline{n}};$$

2. *the threshold $\overline{\overline{n}}$ is (weakly) increasing in $\sigma_{\theta_i}^2$ and is decreasing in σ_θ^2.*

Within the confines of these elementary policies, we can then ask what is the preferred information policy for the data intermediary. By comparing the revenues across these two different information policies, we find that the data broker always prefers to simply transmit the aggregate information. The intermediary thus does not attempt to elicit the identity of the consumer at all.

Proposition 7 (Optimality of Aggregate Intermediation)

1. *The thresholds satisfy the inequality:*

$$\bar{n} \geq \overline{\overline{n}},$$

   for all $\sigma^2_{\theta_i}$ *and* σ^2_{θ}.

2. *The aggregate information transmission always generates a larger profit than the individual information transmission.*

By transmitting only the aggregate policy, the data market can operate profitably with a smaller number of consumers, for any given constellation of $\sigma^2_{\theta_i}$ and σ^2_{θ}. More importantly, the aggregate policy generates larger revenues for the intermediary for any size of the consumer market. An important implication of this comparison is that the firm will not offer personalized prices but rather adjust prices to the level of aggregate demand. This in itself is an interesting finding as it suggests why we may see personalized prices in fewer settings than initially anticipated.

For example, the merchandise platform amazon and the transportation platform uber engage rarely in personalized pricing. Yet, the price of every single good or service is subject to substantial variation. In light of the above result, we may interpret the restraint use of personalized pricing in the presence of aggregate price volatility as the optimal trade-off in the use of consumer information.

The information externality also informs us when trade of information is more likely to arise. If the size of the idiosyncratic shock is large, or $\sigma^2_{\theta_i}$ is large, then the cost of compensating the consumer is going to be large, and stand in the way of profitable intermediation. It leads to a larger threshold $\bar{n}$ when the trade can arise. By contrast, if the common shock is large, or σ^2_{θ} is large, then the informational externality will allow the data broker to offer favorable terms of trade earlier, in terms of the size of the market.

Proposition 7 has significant implications for the data ownership. We can contrast the market outcome when the consumer data is freely available, thus there is no data ownership, with the outcome under data ownership. With the data ownership, there might be still be

socially inefficient price discrimination, but the contractual outcome will now preserve the personal identity of the consumer. As an implication, there will not be personalized pricing but rather variable pricing that adjust to the demand information that the merchant has.

4.2 Data Flow and Uncertainty

As we consider multivariate normally distributed uncertainty, the data structure will ultimately resolve in a description of mean and variance of the willingness-to-pay. As the expectation of the mean will remain the same across all feasible data structures, a complete data structure reduces to a description of the conditional variance of the willingness-to-pay given the data policy. Thus, a more compact description of the data policy S or T is given by the matrix of the variance of the conditional expectation. If we consider symmetric policies, the equilibrium variance matrix can then be described by a (2×2) matrix:

$$\overline{\Sigma} = \begin{bmatrix} \overline{\sigma}_i^2 & \rho_i \overline{\sigma}_i \overline{\sigma} \\ \rho_i \overline{\sigma}_i \overline{\sigma} & \overline{\sigma}^2 \end{bmatrix} \tag{17}$$

and the complete data policy configuration, in and out of equilibrium is given by $(I+1) \times (I+1)$ dimensional matrix that represent the entire variance-covariance structure of the in-flow (and out-flow data):

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \rho_{1I}\sigma_1\sigma_I & \rho_1\sigma_1\sigma \\ \vdots & \ddots & & \vdots \\ \rho_{1I}\sigma_1\sigma_I & \cdots & \sigma_I^2 & \rho_I\sigma_I\sigma \\ \rho_1\sigma_1\sigma & \cdots & \rho_I\sigma_I\sigma & \sigma^2 \end{bmatrix}. \tag{18}$$

Thus, we could alternatively describe the indirect utility functions-the value functions-in terms of the variance of the conditional expectation. This representation will be valuable when we describe a larger class of linear-quadratic problems.

5 Information Design and Data Intermediation

We established the basic result regarding the possibility of trade in a very simple setting. We restricted the informational policy of the intermediary in two important dimensions: (i) the data broker was forwarding all the information, and (ii) the data broker only provided a single aggregate estimate of the market. We now show that a more sophisticated information policy can increase the revenue of the data broker. In particular, the data intermediary can

choose to acquire the identity of consumer i and transmit it to the firm. The variance levels of the additional noise terms, σ_η^2 and $\sigma_{\eta_i}^2$, as well as the decision to (de)-anonymize the data, are the strategic variables of the data intermediary. When consumer i makes her purchase decision q_i, we assume that consumer i will have learned w_i.

5.1 Optimal Information Design

Now we fix $\left(\hat{\sigma}_{\varepsilon_i}^2, \hat{\sigma}_\varepsilon^2\right)$ as errors in transmission (or equivalently, noise in the consumers' own initial estimates). These are minimum noise levels that must be included by construction in the broker's information structure, i.e., $\left(\sigma_{\varepsilon_i}^2, \sigma_\varepsilon^2\right) \geq \left(\hat{\sigma}_{\varepsilon_i}^2, \hat{\sigma}_\varepsilon^2\right)$.

We now turn to the information structure that maximizes the broker's profits.

Proposition 8 (Necessary and Sufficient Condition for Data Intermediation) *The data intermediary profits are strictly positive if and only if*

$$n\sigma_\theta^2 > \sigma_{\theta_i}^2.$$

The optimal common noise σ_ε^2 is positive for $\sigma_{\theta_i}^2$ large enough and n small enough.

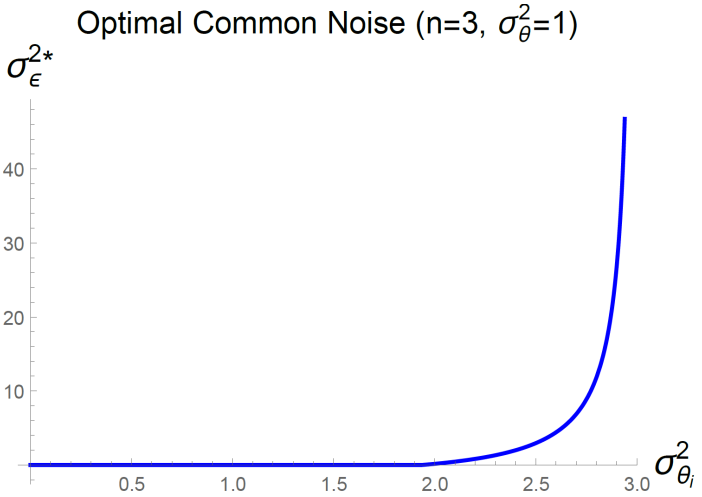
Proposition 9 (Optimal Data Intermediation) *The broker's optimal information structure is symmetric and*

1. *without aggregate noise:* $\sigma_{\varepsilon_i}^2 = \hat{\sigma}_\varepsilon^2$;
2. *with idiosyncratic noise:* $\sigma_\varepsilon^2 = \max\left\{\hat{\sigma_\varepsilon^2}, \sigma^*\right\}$, *where*

$$\begin{aligned}\sigma^* \;=\; & \frac{2n^2\sigma_\theta^4 - 2n^3\sigma_\theta^4 + n\sigma_{\theta_i}^2\left(\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right) + \sigma_{\theta_i}^2\left(-\sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)}\right)}{2n\left(n-1\right)\left(n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)} \\ & + \frac{-2n^2\sigma_\theta^2 - \sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)} + n\left(3\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right)}{2n\left(n-1\right)\left(n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)}\sigma_{\varepsilon_i}^2\end{aligned}$$

We note that the symmetry of the optimal information structure is established as part of the argument. To do so, we fix the revenue of the broker (the total precision of the information collected) and we show that symmetric payments minimize the cost of acquiring the information. Intuitively, one can leverage the externality when buying from multiple consumers, so symmetric schemes should be better.

Figure 5.1 shows the optimal variance of the additional common noise term: if the consumers' preferences are not sufficiently correlated, the broker does not trade any information.



5.2 Intermediation with Commitment

A maintained feature in our analysis is that the data intermediary maintains complete control over the use of the acquired data. In particular, given the data acquired, the data intermediary chooses the sequentially optimal data policy to be offered to the downstream merchant. 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-out flow is linked to the data-in flow. In other words, it is difficult to establish how a given data input has informed a given data output.

Nonetheless, it is informative to consider the implications of an enhanced ability by the data intermediary to commit to a certain data policy vis-a-vis the consumers. Thus, in this subsection we shall briefly describe what would happen under strong commitment assumptions. Now, from Proposition 1 we learned that the unrestricted use of data leads to a decrease in the consumer surplus as well as in the social surplus. This suggest that the data intermediary could support a different data policy and realize a larger social surplus. In fact, it could ask each consumer to share her data with the intermediary but commit not to pass it along to the downstream firm. In exchange for this commitment, the data intermediary would request a compensation from the consumer. Proposition 1 establishes that each consumer would prefer that the demand information is not transferred to the firm. Thus the remaining question is how large a compensation the data intermediary could extract from the seller. 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 10 (Commitment Solution)
Suppose the intermediary can commit to a data out-flow policy. Then the revenue-maximizing data policy is to acquire the entire consumer data, and to never forward the data to the downstream data.

This environment with commitment is related to the analysis in Lizzeri (1999), yet has a number of different features. First, in Lizzeri (1999), the private information is held by a single agent and there are multiple downstream firms that 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.[4] Yet, the shared insight is that the intermediary *with* commitment power maybe able to extract the information rent without any further influence on the efficiency of the allocation relative to the equilibrium outcome under uncertainty.

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 a compensation for his marginal damage. In the commitment solution, every consumer pays a fee to avoid information disclosure. We might then ask what is the scope of data policy if the data intermediary can neither reward nor punish the consumer. Thus, we are restricting the data intermediary to offer a data policy that does not involve a monetary transaction with the consumer, or $m_i = 0$, for all i.

Proposition 11 (Commitment Solution without Monetary Compensation)
Suppose the intermediary cannot use monetary transfers with the consumers, thus or $m_i = 0$, for all i. Then the optimal data policy is to collect all demand data and to enable the firm to offer personalized price recommendation for each consumer i that does not use data provided by consumer i.

Thus, in the absence of monetary transfers between data intermediary and consumer, the data intermediary still acquires the demand data from all the consumers, but exercises some restraint in its use. In particular, the data intermediary forwards only the data that

[4] 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 which 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 the interim contract are pay-off equivalent.

will enable the firm to offer a personalized price to consumer i, where the personalized price is computed without reference to the demand information provided by consumer i.[5]

5.3 Intermediation and Value Creation

In our baseline model, the consumer's information is only used to set prices. As we have seen this is, in a sense, the worst-case scenario for the intermediary. In particular, as data transmission reduces total surplus, no intermediation is profitable without a sufficiently strong data externality. In practice, consumer data can be used in surplus-enhancing ways as well. For example, information facilitates the provision of products and quality levels targeted to the consumer's tastes.

An immediate extension of our framework allows the firm to charge a unit price p and to offer a quality level y to each consumer. (Argenziano and Bonatti (2019) provide a full treatment of this model under a fixed information structure, i.e., without intermediation.) Consumers are heterogeneous in their willingness to pay for the product, but they all value quality uniformly,

$$\begin{aligned} u &= (w_i + by - p)\, q_i - q^2/2, \\ b &\in \left[0, \bar{b}\right], \end{aligned}$$

with $\bar{b} > 1$. (The case $b = 0$ yields the baseline model.) The firm has a constant marginal cost of quantity provision, and a fixed cost per consumer of quality production, thus

$$\pi = pq - cq - y^2/2.$$

For this model, one can then show the following: (i) consumer surplus is decreasing in the precision of the information transmitted by the intermediary if and only if $b < 1$; (ii) information transmission increases total surplus if $b > \hat{b}$ for some $\hat{b} < 1$; (iii) for any $b \geq 0$ there exists a critical $\bar{n}$ such that all information is transmitted by the intermediary if $n \geq \bar{n}\,(b)$; finally, the intermediary discloses the consumers' identities for b sufficiently large.

In other words, the main message of Proposition 9 is that the data externality allows for the profitable intermediation of information in a number of heterogeneous markets. For sufficiently strong data externalities, the intermediary will transmit all the data; whether the outcome improves or diminishes total surplus, that depends on the use of the data. However, there are no market forces that prevent the diffusion of socially detrimental information in sufficiently large markets.

[5] By contrast, without commitment, there is no trade of information without monetary transfers.

6 Value of Social Data

Thus far, we have considered the optimal data policy for a given finite number of consumers, and a single source of information. Perhaps, *the* defining feature of data markets is the large number of (potential) participants and the large number of data sources and services. In this section we pursue the implications of a large number of participants and data sources for the social efficiency of data markets and the price of data.

6.1 Many Consumers

We first consider what happens when the number of consumer possibly participating in the data market and the product market grows. Thus we are considering what happens to prices and revenue as N grows large and without bounds.

With every additional consumer, we receive additional information about idiosyncratic and aggregate demand. In addition, each additional consumer presents an additional opportunity of trade in the downstream market. Thus, the feasible social surplus is additive in each consumer.

Proposition 12 (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:*

$$\frac{\sigma_{\eta_i}^2 \sigma_\theta^2 (3\sigma_\theta^2 - 4\sigma_{\theta_i}^2) + \sigma_{\theta_i}^2 \left(4\sigma_\eta^2 \sigma_{\theta_i}^2 + 3\sigma_\theta^4\right)}{8(\sigma_\eta^2 + \sigma_\theta^2)^2} > 0$$

3. *The total compensation is decreasing in n if $\sigma_{\eta_i}^2 = 0$.*

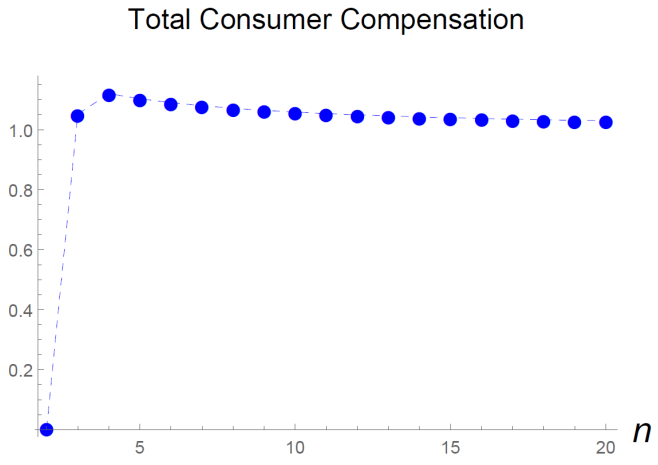
4. *As $n \to \infty$, the broker's revenue and profit grow linearly in n.*

5. *The per capita profit of the data broker, $m_0 - m_i$, will increasingly converge to the per capita profit when data is available "in the wild".*

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. As each consumer is only paid for her marginal contribution, the sum of the payments

$$\sum_{i=1}^{N} m_i$$

converges against a finite expressions. The decrease in the marginal contribution can be strong enough to offset the increase in the number of consumers, and thus it can be less expensive to acquire a larger dataset than a smaller dataset. Figure 6.1 illustrates.



As the data intermediary controls the access to the information, the revenue that it can extract from the downstream is linear in the number of consumer. Thus, while the revenue of the data intermediary is not convex over the entire range of the size of the consumer base, it can have convex components, and the per capita profit of the data broker is increasing in the number of the consumer population.

In our baseline model, the profit and the compensation are calculated in the intermediary's most preferred equilibrium. A natural question is to ask whether the ability of the broker to extract full surplus in per capita level will be preserved if we consider the intermediary's least preferred equilibrium. To answer this question, we could consider a "divide and conquer" scheme where the broker will approach consumers sequentially and gives a compensation conditioned on all consumers before have accepted their offers. In this scheme the first consumer will receive a compensation matching all her loss which guarantees her acceptance regardless of other consumers' decision. Given the first consumer will accept her offer in all equilibria, the second consumer will accept regardless of the remaining consumers' decision. And by a contagion argument we can guarantee the unique implementation of the equilibrium, which brings a lower bound of the profit that the broker could guarantee.

Proposition 13 (Profit Lower Bound using Divide and Conquer) *When the intermediary use a divide and conquer compensation scheme.*

- *The total compensation will grow at a speed of* $O(\log\ n)$*, and the compensation per capita will asymptotically decrease to* 0*.*

- *The per capita profit of the data broker, $m_0 - m_i$, will also asymptotically increase to the per capita profit when aggregate data is available "in the wild".*

6.2 More Services, More Data

The second defining features of data markets is the rapid increase in data sources and data services. For example, a service like Facebook connect allows the facebook to track consumers across many websites. Additional service like Instagram, Snapchat, Facebook Open Graph, Facebook Groups extends the number of sources at which information is collected from a single consumer.[6]

We now consider the role of the precision in the collection of information. Each additional source of information is creating an additional signal

$$s_{i,k} = w_i + \varepsilon_k + \varepsilon_{i,k}. \tag{19}$$

Thus, multiple sources of information generate additional signals about the willingness-to-pay of the consumer. We consider $k = 1, ..., K$ different sources of information. Multiple sources of information clearly reduce the noise about consumer i, so $\hat{\sigma}^2_{\varepsilon_i}$ is decreasing. This has a direct effect for the data policy, it increases the value of the information. In addition, to the extent that we are increasing information from all consumers, the data intermediary may be able to lower the total compensation of the data policy as the correlation among the signals is increasing, and therefore the informational externality has the beneficial effect of lowering total compensation necessary to acquire the data.

Proposition 14 (Precision in Information Collection)

1. *The optimal amount of common noise σ^*_ε is weakly increasing in $\hat{\sigma}^2_{\varepsilon_i}$.*
2. *The broker's profit is convex in $\hat{\sigma}^2_{\varepsilon_i}$*

This captures the idea of more sources, more services, more data, and explains the role of precision in acquiring idiosyncratic information. We could further break down costs and revenues (which are proportional to one another, unlike in the case of the number of consumers N).

With more information sources, the noise in each individual observation is decreasing. As this strengthen the correlation in the signals, the informational externality is increasing, this might lower the compensation necessary to the individual consumer.

[6] Similarly, Google offers a number of services, beginning with gmail, google maps, youtube, that gatehr infromation about a consumer from many different sources.

As we considered an increase in the data sources, we simply added more signals. In particular, we did not change the structure of the data that the intermediary gathered from the individual consumer. If the composition of the data would change as well, for example, a new source would generate a differently weighted signal in terms of the common and idiosyncratic shock, say

$$s_{i,k} = (1 + \lambda_k)\,\theta + (1 - \lambda_k)\,\theta_i + \varepsilon + \varepsilon_i,$$

then additional services would become even more valuable as they would generate more information about the common and idiosyncratic shock.

7 Segmentation and Data

So far, we have defined the demand of an individual consumer in terms of a common and an idiosyncratic component. The binary distinction was useful to bring about some of the central implications of social data. A more complete description of consumer demand should account for additional characteristics that are shared within a group of consumers, but not necessarily common among all consumers. This 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, we augment the description of the demand by the consumers to split the population into two subsets:

$$w_{ij} = \theta_j + \theta_{ij}, i = 1, 2, ..., N_j,\ \ j = 1, 2. \tag{20}$$

We maintain the assumption that the group characteristics, θ_1 and θ_2, are normal random variables, independently distributed with mean μ and variance σ^2_θ. The idiosyncratic component θ_{ij} remains normally distributed with zero mean and variance $\sigma^2_{\theta_i}$, and all random variables are independent. A more general model would have J groups of consumers, where the size of each group is given by N_j. For the remainder of the analysis, we will assume that each group has the same size, thus $N_1 = N_2 = N/2$.

Thus, each member in group j has the same common component. Across groups, the common component is drawn from the same distribution with the same mean, thus ex-ante the groups are identical. A natural extension would be to allow for different means and different variances across groups, thus μ_j and $\sigma^2_{\theta_j}$.

The identity of each consumer is now given by the pair (i, j), which determines his

personal identity as well as his group characteristics. We then ask whether and under what conditions the data intermediary may collect and transmit group characteristics.

7.1 Group Segmentation

By collecting information about the group characteristics, the intermediary will also influence the extent of price discrimination. For example, by choosing not to provide the group characteristic to the firm, the intermediary will induce the firm to offer a single price by sending the sample average of all signals. Alternatively the intermediary could allow the firm to discriminate between two groups of consumers by transmitting the group characteristics. By allowing price discrimination across groups, the intermediary is able to charge higher fee from the firm, but it also increase the compensation owed to consumers.

For the purpose of this section, we shall restrict attention to noiseless signals by the consumers, thus $r_i = w_i$.

Proposition 15 (Segmentation)
With noiseless signals, there exists $\overline{N}$, such that for all $N > \overline{N}$, the data intermediary will induce group pricing.

Thus, while the earlier Proposition 8 stated that the seller will not elicit any identity information, the present result shows that if the individual group is sufficiently large, then the intermediary will convey some limited identity information, and this will allow the seller to price discriminate across groups, but not within groups. We expect these results to extend to a noisy signal environment.

The benefits of segmentation along group identity arise from similar sources as in the aggregate market.

Proposition 16 (Comparative Statics of Segmentation)
With noiseless signals, 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 $\sigma_{\theta_i}^2$.

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, and other dimension that effectively capture group characteristics.[7]

[7] The discussion and the results here are somewhat sensitive to what we assume what happens if we allow personalized pricing yet do not use personal information on buyer i.

Finally, 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 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 takes advantage of the richer structure of consumer demand.[8]

7.2 Identification and Personalized Pricing

We established earlier in Proposition 8 that the data intermediary will not collect information that would allow for personalized prices at the level of the individual. This conclusion might not hold anymore if there is prior information that the consumer may belong to different groups, but importantly neither the consumer nor the intermediary know a priori which consumer belongs to which group.

Thus, suppose each consumer shares her characteristics with either one of the groups, but neither she nor the intermediary would know which group she belongs to. Then there are instances where the intermediary would ask for identity information as it would help to bring the information of similar consumers to bear on the demand estimation problem.

Proposition 17 restricts attention to the noiseless case.

Proposition 17 (Bayesian Identification and Price Discrimination)
Asymptotically as $N \to \infty$, personalized pricing is more profitable than uniform pricing if and only if:

$$\sigma_\theta^2 \geq \sigma_{\theta_i}^2 \tag{21}$$

The learning and classification problem that arises when we elicit information from the individual consumer, for example, the determination of the relevant characteristics is an issue that is an open question for which the current framework is too limited.

The group characteristics and the scale of the data could also interact here. Suppose the data intermediary receives more information from more services. Then this should support more group characteristics.

The model of group characteristics introduced above,

$$w_{ij} = \theta_j + \theta_{ij}, i = 1, 2, ..., N_j, \;\; j = 1, 2, \tag{22}$$

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

can be augmented in a number of ways. As the common component in the demand, it refers to group characteristics, but as we saw it does not have to refer to all agents. Suppose then there is a large number of group characteristics, then additional members are valuable because the increase the group size, and thus increase the informational externality. Then, if the group characteristic are drawn from the same distribution the inference becomes weaker with the number of participants, unless the intermediary can distinguish them.

In the above, we simply assumed that the group membership can be identified. A more plausible approach is that new data services allow the intermediary to gather the data to enable the distinction. This would mean that services would increases the power of inference.

8 Competitive Brokers

To study the question "does competition promote privacy protection?", we can look at the model with m brokers. Our model is able to articulate these policy relevant questions by introducing heterogeneity in a way that is pertinent to information markets.

Each broker $j \in \{1, 2, ..., m\}$ is characterized by an upper bound on the information it can collect. We specify a lower bound x_j on the variance of the idiosyncratic noise that is specific to that broker. The j-specific noise terms $\varepsilon_{i,j}$ are independent across both i and j.

The timing of the game is as follows:

1. The brokers simultaneously offer payments and information structures to all consumers.
2. Consumers choose a subset of offers to accept and transmit the corresponding signals.
3. Given the information collected, the brokers simultaneously decide what to sell and offer prices to the monopolist.
4. The monopolist decides which databases to purchase and sets prices for the consumers.

In our setting (though by no means in general) informative signals have decreasing returns. This can be shown by direct calculation using the formula for the residual variance of the monopolist's beliefs. Now we consider the equilibrium where each intermediary j sells access to an exogenous database consisting of reports s_{ij}, $i \in S_j$ from consumers. The following proposition characterizes the unique equilibrium for this subgame as long as the databases are imperfect.

Proposition 18 (Unique Pricing Subgame)
If $x_j > 0$ for all j, given the data collected, there exist a unique equilibrium in which intermediary j sells their database at a price $p_j = \pi(J) - \pi(J \backslash \{j\})$, and the monopolist purchases all databases.

It is worthwhile to point out that the uniqueness result will break down when $x = 0$, i.e., if the broker could collect perfect information from consumers. In this extreme case, once one broker gets one copy of report from consumer i, any other report about i collected by the other broker is useless for the firm. Therefore in the pricing subgame, once a report of consumer i is sold, the other broker becomes indifferent as to whether to sell i's information, which breaks the unique equilibrium price in that subgame. This causes multiplicity of equilibria, as in Ichihashi (2019).

To proceed our analysis in the data market, we focus on a particular class of equilibria where consumers choose the maximal accepting set.

Definition 1 (Equilibrium Refinement) *For any given offers $\{p_{ij}, \varepsilon_{ij}, \varepsilon_j\}$, 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 $\{p_{ij}, \varepsilon_{ij}, \varepsilon_j\}$ such that $A_i \subset A_i'$ for all i and $A_i \subsetneq A_i'$ for some i.*

This assumption is implicitly embedded in our baseline model of monopoly broker by focusing on broker-most-preferred equilibrium. And it is clear that we could never expect any uniqueness result without equilibrium selection in this contracting with externalities model where coordination among consumers is crucial.[9]

Proposition 19 (Unique Equilibrium Prediction)
Suppose $\sigma^2_{\theta_i} = 0$, then in every pure equilibrium each broker will collect information from every consumers, with both noises as small as possible: $\sigma^{2*}_{\varepsilon_{i,j}} = x_j$, $\sigma^{2*}_{\varepsilon_j} = 0$. Consumers will be indifferent between reporting to all brokers and keeping privacy.

The result is intuitive at the first glance since fixing the interaction of the consumers with other brokers, marginal information about the common shock θ will always bring positive profit to broker j when $\sigma^2_{\theta_i} = 0$. However, the formal proof is much more involved, because the interaction between consumers and other brokers will be affected by the offer between consumers and broker j itself. Consumers' willingness to accept offers from other brokers will increase as broker j collect more precise information. Therefore the broker might have a incentive to strategically collect less information from consumers in the hope to persuade them to reject offers from other brokers. By this deviation, the broker might gain more bargaining power in the downstream product market and sell his database at a higher price even with less information collected.

[9]The problem of consumers' coordination is different from the problem caused by perfect information collection. Even with maximal accepting assumption, multi equilibria will appear in perfect information collection setting because Proposition 18 does not hold.

Our prediction pins down the information structure, the total compensation to consumers and the transfer between brokers and downstream firm. Consequently, consumers surplus, producer surplus and total profit of brokers are uniquely determined. The only undetermined variable is the specific compensation paid by each broker. This comes from the fact that generically consumers' marginal loss from reporting to one broker, given she has reported to other firm, is smaller than the compensation provided by that broker. For example, if a consumer allows Facebook, Twitter, Amazon, and other countless other apps to access their phone's data, further exchanging data for the service of Google Map sounds like a great deal.

To establish existence and carry out some comparative statics, we focus on the symmetric equilibrium where each broker offers the same compensation to each consumer.

Proposition 20 (Symmetric Equilibrium)
Suppose $\sigma^2_{\theta_i} = 0$, then for any (n, m, σ^2_θ) , there exists x large enough such that, if $x_A = x_B \geq x$, there exists a symmetric equilibrium where each broker pays the same compensation to every consumers.

We now compare the equilibrium outcome under competition with that under monopoly.

Proposition 21 (Comparative Statics)

1. *The individual brokers' provision of information equals the monopoly level, and the total provision of information will increase in the number of brokers, m ($m = 1$ is the case of monopoly).*

2. *Consumer surplus is decreasing in m, and increasing in x.*

3. *Brokers' profit is decreasing in m, while the firm's profit is increasing in m.*

This result highlights the possibility that, instead of protecting consumers' privacy, competition may make things worse. Competing brokers in the unique equilibrium collect data exactly the same way as they will in the monopoly environment. Each broker's profit will decrease, which is purely because of a loss of bargaining power against the firm in the product market. Product firms get additional portion of surplus extracted from consumers, and consumers surplus decreases even further due to the multiple sourcing of information. Apart from the analysis of entry, our model also gives predictions in terms of mergers: the equilibrium noise level is the same whether the firms merge or compete and consumers surplus and social surplus remain unchanged. (Of course the price of data and profits will change.)

9 Discussion and Conclusion

Variance Reduction The data transfer allows the downstream merchant to better match the supply to the demand, this enables value creation. The data transfer also allows him to better match the price to the demand, this enables value extraction. The linear price discrimination environment enables both value creation and value extraction. This gives a particular combination of how valuable the variance reduction in their posterior estimate of the demand is for each of the participants. The net effect across all the market participants can have social positive or negative effects. For the pricing and the efficiency of the data market, it is the differential margins that matter. But clearly, the insights here generalize, and allows us to consider other market environments.

Personalization The additive model of idiosyncratic and common shock is a specific model by which value is generated. We could consider different linear, convex combination, both for the firm and the consumer. This in turn would enable different policies and different interests. For example, the firm could be able to tailor less, even offer separate products for common and idiosyncratic shocks, through personalization and other product policies.

State and Model In the present analysis, we considered a model of data sharing where the data intermediary and the firm know the structure of the model, the additive structure of idiosyncratic and common shock. But an important aspect of the data is that it allows the intermediary to learn/estimate the structure and then make the prediction for a specific data profile. This second aspect/value of data is something that is currently not in the model, and would give an additional reason/benefit to accumulate the data. This would suggest a dynamic model of learning and data acquisition.

Data Platform The data intermediary collected and redistributed the consumer data but played no role in the interaction between consumer and firm. By contrast, often the consumer access or can access the firm only trough a data platform. The data platform can then be thought of as selling the access (often through an auction) to the consumer among the highest bidders. The data platform provides the bidding firm with additional information that the firms can use to tailor their interaction with the consumers.

The data platform typically allows the consumer to access the service/site for free and possibly offer additional services/products to generate more information. Thus the interaction is monetized only among the competing firms. The competitive allocation of the access to the consumer through a bidding mechanism allows the platform to introduce an element

of "reverse price discrimination." Thus even if the firm sets the price uniform across all consumers, it bids vis-à-vis the platform in information/type dependent way. Thus, the price setting across all consumer interactions might be influenced by the bidding behavior even if the discriminatory aspect does not appear in the uniform price setting.[10]

[10]For example, Amazon does not, nor does it allow its vendors to use price discrimination. On the other hand, the advertising on the site is distributed and priced using consumer characteristics.

10 Appendix

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

Proof of Proposition 2. The demand function of consumer i is $q_i(w_i, p_i) = w_i - p_i$. Therefore, the optimization problem of the firm will be:

$$\max_{p_1,p_2,...,p_n \in \mathcal{I}} \Sigma_i p_i(w_i - p_i)$$

And thus the optimal pricing is simply:

$$p_i = \frac{1}{2}\mathbb{E}[w_i|\mathcal{I}]$$

For consumer $i = 1, 2, ..., k$ whose data and identity is known by the firm, the optimal price is:

$$p_i = \frac{1}{2}\frac{\sigma^2_{\theta_i}}{\sigma^2_{\theta_i} + \sigma^2_{\varepsilon_i} + \sigma^2_{\eta_i}} s_i + \frac{\sigma^2_\theta(\sigma_{\eta_i} + \sigma^2_{\varepsilon_i}) - \sigma^2_{\theta_i}(\sigma^2_\eta + \sigma^2_\varepsilon)}{2(k(\sigma^2_\theta + \sigma^2_\varepsilon + \sigma^2_\eta) + \sigma^2_{\theta_i} + \sigma^2_{\varepsilon_i} + \sigma^2_{\eta_i})(\sigma^2_{\theta_i} + \sigma^2_{\eta_i} + \sigma^2_{\varepsilon_i})} \Sigma^k_{j=1} s_j.$$

Since what matters is the second order moment, we have omitted the constant term which will equals $\mathbb{E}[p_i] = \mu/2$. Similarly for consumer $i = k+1, ..., n$ whose data and identity is not known, the optimal price is uniform:

$$p_i = \frac{1}{2}\frac{\sigma^2_\theta}{k(\sigma^2_\theta + \sigma^2_\varepsilon + \sigma^2_\eta) + \sigma^2_{\theta_i} + \sigma^2_{\varepsilon_i} + \sigma^2_{eta_i}} \Sigma^k_{j=1} s_j.$$

The value of information to the merchant, as a function of k, is:

$$g(k) = \Sigma_i \operatorname{var}[p_i],$$

with

$$\frac{\partial^2 g(k)}{\partial k^2} = \frac{(\sigma^2_\theta + \sigma^2_\varepsilon + \sigma^2_\eta)((\sigma^2_\varepsilon + \sigma^2_\eta - \sigma^2_\theta)\sigma^4_{\theta_i} - (\sigma^2_{\varepsilon_i} + \sigma^2_{\eta_i})n\sigma^4_\theta - \sigma^2_\theta(2(\sigma^2_{\varepsilon_i} + \sigma^2_{\eta_i}) + n\sigma^2_\theta)\sigma^2_{\theta_i})}{2(k(\sigma^2_\theta + \sigma^2_\varepsilon + \sigma^2_\eta) + \sigma^2_{\theta_i} + \sigma^2_{\varepsilon_i} + \sigma^2_{\eta_i})}.$$

If $\sigma^2_\theta = 0$, each individual data will unlock one additional personalized price. The additional data will also help to eliminate the common noise in existing consumers. Thus the data has a increasing return.

$$\frac{\partial^2 g(k)}{\partial k^2} = \frac{(\sigma^2_\varepsilon + \sigma^2_\eta)((\sigma^2_\varepsilon + \sigma^2_\eta)\sigma^4_{\theta_i}}{2(k(\sigma^2_\varepsilon + \sigma^2_\eta) + \sigma^2_{\theta_i} + \sigma^2_{\varepsilon_i} + \sigma^2_{\eta_i})} > 0.$$

If $\sigma_{\theta_i}^2 = 0$, the firm will always charge a uniform price, and he will use the additional data to provide a better estimation of the same common demand. Thus the data displays decreasing returns:

$$\frac{\partial^2 g(k)}{\partial k^2} = -\frac{(\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\eta^2)(\sigma_{\varepsilon_i}^2 + \sigma_{\eta_i}^2) n \sigma_\theta^4}{2(k(\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\eta^2) + \sigma_{\varepsilon_i}^2 + \sigma_{\eta_i}^2)}.$$

Now, $g(k)$ is clearly increasing with respect k. The remaining results directly follow the observation that the total consumers surplus and the social welfare could be represented by:

$$\begin{aligned} TCS(k) &= \mathbb{E}\Sigma_i \frac{1}{2}(w_i - p_i)^2 = TCS(0) - \frac{3}{2}\Sigma_i \\ \text{var}[p_i] &= TCS(0) - \frac{3}{2} g(k), \end{aligned}$$

and

$$SS(k) = TCS(k) + g(k) = SS(0) - \frac{1}{2} g(k),$$

which completes the proof. ■

Proof of Proposition 3. When the identity of the consumer is not known, then the optimal price is a uniform price:

$$p_i = \frac{1}{2n}\mathbb{E}[\Sigma_i w_i | \mathcal{I}] = \frac{1}{2} \frac{\sigma_\theta^2 + \sigma_{\theta_i}^2/n}{k(\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\eta^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + \sigma_{\eta_i}^2} \Sigma_{j=1}^k s_j,$$

where

$$g(k) = n \,\text{var}[p_i] = \frac{k(n\sigma_\theta^2 + \sigma_{\theta_i}^2)^2}{4n(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + \sigma_{\eta_i}^2 + k(\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\eta^2))},$$

and

$$\frac{\partial^2 g(k)}{\partial k^2} = -\frac{(\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\eta^2)(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + \sigma_{\eta_i}^2)(n\sigma_\theta^2 + \sigma_{\theta_i}^2)^2}{2n(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + \sigma_{\eta_i}^2 + k(\sigma_\theta^2 + \sigma_\varepsilon^2 + \sigma_\eta^2))^3} < 0,$$

which completes the proof. ■

Proof of Proposition 5. First we derive the expression of profit. Given the information $\mathcal{I}$, the optimal pricing in consumer research is:

$$p_i = \frac{1}{2}\mathbb{E}[w_i|\mathcal{I}] = \frac{1}{2}(\mathbb{E}[\theta|\mathcal{I}] + \mathbb{E}[\theta_i|\mathcal{I}]).$$

Moreover, denote p_i the price under null information $\mathcal{I}$, p_i' the price under full information $\mathcal{I}_{-i}$. The total value of information for the firm is:

$$g = \frac{1}{4}\Sigma_{i=1}^n \,\text{var}(\mathbb{E}[w_i|\mathcal{I}_1].$$

As for the marginal compensate for consumer, we have:

$$\begin{aligned}\Delta CS_i &= -\mathbb{E}[p_i' w_i - \frac{1}{2}(p_i'^2] + \mathbb{E}[p_i w_i - \frac{1}{2}(p_i)^2] = \frac{1}{2}\mathbb{E}[(p_i' - p_i)(p_i' + p_i - 2w_i)] \\ &= \frac{1}{8}(\text{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}]) - \text{var}(\mathbb{E}w_i|\mathcal{I}])) - \frac{1}{2}\,\text{cov}(w_i, (\mathbb{E}[w_i|\mathcal{I}_{-i}] - \mathbb{E}w_i|\mathcal{I}])) \\ &= \frac{3}{8}(\text{var}(\mathbb{E}[w_i|\mathcal{I}]) - \text{var}(\mathbb{E}w_i|\mathcal{I}_{-i}])).\end{aligned}$$

Then we can calculate the profit of the intermediary:

$$\begin{aligned}8g - 8\Sigma_i \Delta CS_i &= \Sigma_{i=1}^n - \text{var}(\mathbb{E}[w_i|\mathcal{I}]) + 3Var(\mathbb{E}[w_i|\mathcal{I}_{-i}]) \\ &= \Sigma_{i=1}^n \,\text{cov}(w_i, 3\mathbb{E}[w_i|\mathcal{I}_{-i}] - \mathbb{E}[w_i|\mathcal{I}]) \\ &= \Sigma_{i=1}^n \,\text{cov}(\theta + \theta_i, 3\mathbb{E}[\theta|\mathcal{I}_{-i}] - \mathbb{E}[\theta|\mathcal{I}] - \mathbb{E}[\theta_i|\mathcal{I}])\end{aligned}$$

Omitting the constant term, we could express the conditional expectations as follows::

$$\mathbb{E}[\theta|\mathcal{I}] = \frac{\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Sigma_i s_i,$$

$$\mathbb{E}[\theta|\mathcal{I}_{-i}] = \frac{\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Sigma_{i' \neq i} s_{i'},$$

and

$$\mathbb{E}[\theta_i|\mathcal{I}] = \frac{\sigma_{\theta_i}^2(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2))}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2)(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))} s_i - \Sigma_{i' \neq i} \frac{\sigma_{\theta_i}^2(\sigma_\theta^2 + \sigma_\varepsilon^2)}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2)(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))} s_{i'}.$$

The revenue of the broker is simply :

$$\begin{aligned}R(n) &= g - \Sigma_i \Delta CS_i \\ = \frac{n\sigma_\theta^2}{8}\Big(&\frac{3(n-1)\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{2\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Big) \\ &- \frac{n\sigma_{\theta_i}^4(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2))}{8(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2)(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))}.\end{aligned}$$

Now we are ready to prove the monotonicity in terms of σ_θ^2 and $\sigma_{\theta_i}^2$. Since

$$\begin{aligned}\frac{R(n)}{n} = \frac{\sigma_\theta^2}{8}\Big(&\frac{3(n-1)\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{2\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Big) \\ &- \frac{1}{8}\frac{\sigma_{\theta_i}^4}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2}\Big(\frac{n-1}{n} + \frac{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2)/n}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Big)\end{aligned}$$

To prove that $R(n)$ increases with σ_θ^2, we only need to prove that $h(\sigma_\theta^2, \sigma_{\theta_I}^2)$ is increasing, where:

$$h(\sigma_\theta^2, \sigma_{\theta_I}^2) = \frac{3(n-1)\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}$$
$$\frac{\partial h}{\partial \sigma_\theta^2} = \frac{3(n-1)(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)\sigma_\varepsilon^2)}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} - \frac{n(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n\sigma_\varepsilon^2)}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2}$$
$$\frac{\partial h}{\partial \sigma_\theta^2} \geq \frac{3(n-1)(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)\sigma_\varepsilon^2)}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} - \frac{n(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n\sigma_\varepsilon^2)}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} > 0.$$

Second, since we could also rewrite $R(n)$ as:

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

To prove $R(n)$ is decreasing with respect to $\sigma_{\theta_i}^2$ we only need to prove $h(\sigma_\theta^2, \sigma_{\theta_i}^2)$ is decreasing with respect to $\sigma_{\theta_i}^2$:

$$\frac{\partial h}{\partial \sigma_{\theta_i}^2} = -\frac{3(n-1)\sigma_\theta^2}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} + \frac{n\sigma_\theta^2}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2}$$
$$< -\frac{(2n-3)\sigma_\theta^2}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} < 0.$$

Finally we show that there exist a threshold of profitability. Notice that:

$$R(1) = -\frac{1}{8}\frac{\sigma_\theta^4 + 2\sigma_\theta^2\sigma_{\theta_i}^2 + \sigma_{\theta_i}^4}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + \sigma_\theta^2 + \sigma_\varepsilon^2},$$
$$\lim_{n\to\infty}\frac{R(n)}{n} = \frac{1}{2}\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\varepsilon^2} - \frac{1}{4}\frac{\sigma_{\theta_i}^4}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2}.$$

The per capital profit $R(n)/n$ is increasing with respect to n:

$$\frac{\partial R(n)/n}{\partial n} = \frac{3\sigma_\theta^4(\sigma_{\varepsilon_i}^2 + \sigma_{\theta_i}^2)}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{(\sigma_{\varepsilon_i}^2\sigma_\theta^2 - \sigma_\varepsilon^2\sigma_{\theta_i}^2)^2}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2)}$$
$$> \frac{(\sqrt{3}\sigma_\theta^2(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2) + \sigma_\theta^2\sigma_{\varepsilon_i}^2 - \sigma_{\theta_i}^2\sigma_\varepsilon^2)(\sqrt{3}\sigma_\theta^2(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2) - \sigma_\theta^2\sigma_{\varepsilon_i}^2 + \sigma_{\theta_i}^2\sigma_\varepsilon^2)}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2)}$$

If $\sigma_\theta^2 \geq \sigma_\varepsilon^2$ then the per capital profit will be increasing. ■

Proof of Proposition 6. We first derive the expression of profit. Given the information $\mathcal{I}$, since the firm has no identity information, it is optimal to set a uniform price (we will omit the constant term again):

$$p = \frac{1}{2n}\mathbb{E}[\Sigma_{i=1}^n w_i|\mathcal{I}] = \frac{1}{2}\mathbb{E}[\theta|\mathcal{I}] + \frac{1}{2n}\Sigma_{i=1}^n\mathbb{E}[\theta_i|\mathcal{I}].$$

Denote p_0 the price under null information $\mathcal{I}_0$, p_1 the price under full information $\mathcal{I}_1$, and p_{-j} the price under $\mathcal{I}_{-j}$ when information from all consumers except j is collected. Then the total value of information for the firm is:

$$g = \frac{1}{2}\ \mathbb{E}[(\mathbb{E}[\overline{w}|\mathcal{I}_1] - \mathbb{E}[\overline{w}|\mathcal{I}_0])(\Sigma_i w_i) - \frac{n}{2}(\mathbb{E}^2[\overline{w}|\ \mathcal{I}_1] - \mathbb{E}^2[\overline{w}|\mathcal{I}_0])] = \frac{n}{4}\,\mathrm{var}(\mathbb{E}[\overline{w}|\mathcal{I}_1]$$

As for the marginal compensate for consumer, we have:

$$\begin{aligned}\Delta CS_j &= -\mathbb{E}[p_{-j}w_j - \frac{1}{2}(p_{-j})^2] + \mathbb{E}[p_1 w_j - \frac{1}{2}(p_1)^2]\\ &= \frac{1}{2}\mathbb{E}[\ (p_{-j} - p_1)(p_{-j} + p_1 - 2w_j)\ ]\\ &= \frac{1}{2}\mathbb{E}[\ w_j(\mathbb{E}\bar{w}|\mathcal{I}_1 - \mathbb{E}\bar{w}|\mathcal{I}_{-j})\ ] - \frac{1}{8}(\mathrm{var}(\mathbb{E}\bar{w}|\mathcal{I}_1) - \mathrm{var}(\mathbb{E}\bar{w}|\mathcal{I}_j^2))\end{aligned}$$

Therefore the total profit for the broker is:

$$\begin{aligned}R &= g - \Sigma_j \Delta CS_j\\ &= \frac{3}{8}nVar(\mathbb{E}[\bar{w}|\mathcal{I}_1]) - \frac{1}{8}\Sigma_j\,\mathrm{var}(\mathbb{E}[\bar{w}|\mathcal{I}_{-j}]) - \frac{1}{2}n\mathbb{E}[\bar{w}(\mathbb{E}[\bar{w}|\mathcal{I}_1] - \mathbb{E}\bar{w})] + \frac{1}{2}\Sigma_j\mathbb{E}[w_j(\mathbb{E}[\bar{w}|\mathcal{I}_{-j}] - \mathbb{E}\bar{w})]\\ &= \frac{1}{2}\Sigma_j\,\mathrm{cov}(w_j, E[\bar{w}|\mathcal{I}_{-j}]) - \frac{1}{8}\Sigma_j\,\mathrm{var}(\mathbb{E}[\bar{w}|\mathcal{I}_{-j}]) - \frac{1}{8}nVar(\ \mathbb{E}[\bar{w}|\mathcal{I}_1])\\ &= \frac{1}{2}\Sigma_j\,\mathrm{cov}(\bar{w}, E[w_j|\mathcal{I}_{-j}]) - \frac{1}{8}\Sigma_j\,\mathrm{var}(\mathbb{E}[\bar{w}|\mathcal{I}_{-j}]) - \frac{1}{8}nVar(\ \mathbb{E}[\bar{w}|\mathcal{I}_1])\\ &= \mathrm{cov}(\theta + \frac{1}{n}\Sigma_i\theta_i, \frac{3}{8}\Sigma_j\ \mathbb{E}[\theta|\mathcal{I}_{-j}] - \Sigma_j\frac{1}{8n}\Sigma_{i\neq j}\ \mathbb{E}[\theta_i|\mathcal{I}_{-j}] - \frac{1}{8}\Sigma_i\mathbb{E}[\theta_i|\mathcal{I}_1] - \frac{n}{8}\mathbb{E}[\theta|\mathcal{I}_1])\\ &= \Big(\frac{3(n-1)\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n-1}{n}\frac{\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)}\\ &\qquad - \frac{\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Big)\frac{n\sigma_\theta^2 + \sigma_{\theta_i}^2}{8}.\end{aligned}$$

Denote $A = 8R/(n\sigma_\theta^2 + \sigma_{\theta_i}^2)$, whose sign will indicate the profitability of intermediation.

Notice that we have

$$A(1) = -\frac{\sigma_\theta^2 + \sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + \sigma_\theta^2 + \sigma_\varepsilon^2} < 0,$$
$$\lim_{n\to\infty} A(n) = 2\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} > 0.$$

To prove the existence of the threshold $\overline{\overline{n}}$, we calculate its derivative first:

$$\frac{\partial A}{\partial n} = \frac{\sigma_{\varepsilon_i}(2n^2\sigma_\theta^2 - \sigma_{\theta_i}^2) + ((2n(n-1)+1)\sigma_\varepsilon^2 + (2n(2n-1)+1)\sigma_\theta^2 - \sigma_{\theta_i}^2)\sigma_{\theta_i}^2}{n^2(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2}.$$

When $2n^2\sigma_\theta^2 < \sigma_{\theta_i}^2$, it is easy to see $A < 0$ from its expression, while when $2n^2\sigma_\theta^2 > \sigma_{\theta_i}^2$, A is strictly increasing with respect to n, thus the threshold $\overline{\overline{n}}$ exists.

To prove the monotonicity of profit with respect to σ_θ^2 and $\sigma_{\theta_i}^2$, we will prove A is increasing with respect to σ_θ^2 and decreasing with respect to $\sigma_{\theta_i}^2$. And consequently $\overline{\overline{n}}$ is (weakly) increasing in $\sigma_{\theta_i}^2$ and is decreasing in σ_θ^2.

$$A = \frac{3(n-1)\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{n-1}{n}\frac{\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2)} - \frac{\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2)}.$$

The last two term in the above equation is clearly increasing w.r.t. σ_θ^2, and we have proved in the proof of proposition 5 that the first two terms are increasing too. As for $\sigma_{\theta_i}^2$ we have:

$$\frac{\partial A}{\partial \sigma_{\theta_i}^2} = -\frac{1}{8}\frac{(n-1)((4n-1)\sigma_\theta^2 + (n-1)\sigma_\varepsilon^2 + \sigma_{\varepsilon_i}^2)}{n(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + (n-1)(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} - \frac{\sigma_{\varepsilon_i}^2 + n\sigma_\varepsilon^2}{(\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2 + n(\sigma_\theta^2 + \sigma_\varepsilon^2))^2} < 0,$$

which completes the proof. ■

Proof of Proposition 7. For any fixed parameter, denote $\overline{R}$ as the profit when the broker transmits only the demand information but not the identity information, while we

denote $\underline{R}$ as the profit when the broker transmits all information:

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

Therefore, for any fixed parameters, transmitting aggregate information only is more profitable for the intermediary. The first statement about the thresholds is a direct corollary of this result. ■

Proof of Proposition 8. Since now we are in a world of information design, $\sigma_{\varepsilon_i}^2$ might not be symmetric among different consumers, so we re-derive the profit of the broker:

$$\begin{aligned}
R &= g - \Sigma_j \Delta CS_j \\
&= \operatorname{cov}[\theta + \frac{1}{n}\Sigma_i\theta_i, \frac{3}{8}\Sigma_j\mathbb{E}[\theta|\mathcal{I}_{-j}] - \Sigma_j\frac{1}{8n}\Sigma_{i\neq j}\mathbb{E}[\theta_i|\mathcal{I}_{-j}] - \frac{1}{8}\Sigma_i\mathbb{E}[\theta_i|\mathcal{I}_1] - \frac{n}{8}\mathbb{E}[\theta|\mathcal{I}_1]] \\
&= \operatorname{cov}[\theta + \frac{1}{n}\Sigma_i\theta_i, \frac{1}{8n}\Sigma_j(\Sigma_{i\neq j}\frac{\frac{3n\sigma_\theta^2-\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}{1+\Sigma_{i\neq j}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}r_i) - \frac{1}{8}(\Sigma_i\frac{\frac{n\sigma_\theta^2+\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}{1+\Sigma_i\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}r_i)] \\
&= \Big(\frac{1}{8n}\Sigma_j(\Sigma_{i\neq j}\frac{\frac{3n\sigma_\theta^2-\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}{1+\Sigma_{i\neq j}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}) - \frac{1}{8}(\Sigma_i\frac{\frac{n\sigma_\theta^2+\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}{1+\Sigma_i\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}})\Big)(\sigma_\theta^2+\frac{1}{n}\sigma_{\theta_i}^2) \\
&= \Big(\frac{1}{8n}\Sigma_j\frac{3n\sigma_\theta^2-\sigma_{\theta_i}^2}{\sigma_\theta^2+\sigma_\varepsilon^2}(1-\frac{1}{1+\Sigma_{i\neq j}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}}) - \frac{1}{8}\frac{n\sigma_\theta^2+\sigma_{\theta_i}^2}{\sigma_\theta^2+\sigma_\varepsilon^2}(1-\frac{1}{1+\Sigma_i\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}})\Big)(\sigma_\theta^2+\frac{1}{n}\sigma_{\theta_i}^2) \\
&= \frac{\sigma_\theta^2+\frac{1}{n}\sigma_{\theta_i}^2}{8(\sigma_\theta^2+\sigma_\varepsilon^2)}\Big(2n\sigma_\theta^2-2\sigma_{\theta_i}^2+(n\sigma_\theta^2+\sigma_{\theta_i}^2)\frac{1}{1+\Sigma_i\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}} - \frac{3n\sigma\theta^2-\sigma_{\theta_i}^2}{n}(\Sigma_j\frac{1}{1+\Sigma_{i\neq j}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}})\Big)
\end{aligned}$$

Note that when $n\sigma_\theta^2 < \sigma_{\theta_i}^2$, the profit is always negative.

If $n\sigma_\theta^2 > \sigma_{\theta_i}^2$. We will first show that it is optimal to use a symmetric idiosyncratic noise scheme. To find the optimal idiosyncratic noise $\sigma_{\varepsilon_i}^2$ fixing other parameters, it is equivalently

to find optimal $x_i \in (0, \frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2}]$ maximizing:

$$A\frac{1}{1+\Sigma_i\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}} - B\Sigma_j\frac{1}{1+\Sigma_{i\neq j}\frac{\sigma_\theta^2+\sigma_\varepsilon^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2}} = A\frac{1}{1+\Sigma_i x_i} - B\Sigma_j\frac{1}{1+\Sigma_{i\neq j}x_j}.$$

Where $A, B > 0$. For any $\{x_i\}$ such that $x_1 \neq x_2$, then $x_1' = x_2' = (x_1 + x_2)/2$ will strictly increase the objective.

Now since we have shown the optimality of the symmetric noise scheme, we can reuse the expression of A in the previous analysis.

$$\begin{aligned} A &= \frac{3(n-1)\sigma_\theta^2 - \frac{n-1}{n}\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2+(n-1)(\sigma_\theta^2+\sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2+\sigma_{\theta_i}}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2+n(\sigma_\theta^2+\sigma_\varepsilon^2)} \\ &= \frac{2(n-1)(n\sigma_\theta^2-\sigma_{\theta_i}^2)\sigma_\varepsilon^2 + B(\sigma_\theta^2,\sigma_{\theta_i}^2,\sigma_{\varepsilon_i}^2)}{(\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2+(n-1)(\sigma_\theta^2+\sigma_\varepsilon^2))(\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2+n(\sigma_\theta^2+\sigma_\varepsilon^2))}. \end{aligned}$$

Note that the numerator is a positive affine function of σ_ε^2 thus the profit will eventually become positive as σ_ε^2 grows to infinity as long as $n\sigma_\theta^2 > \sigma_{\theta_i}^2$. ■

Proof of Proposition 9. In Proposition 8 we showed that it is optimal to use a symmetric noise scheme so that we could use the expression derived before safely. We first prove $\sigma_{\varepsilon_i}^* = \hat{\sigma}_{\varepsilon_i}$ simply by pointing out that, if $\sigma_{\varepsilon_i}^2$ does not reach the lower bound, by changing $\bar{\sigma}_\varepsilon^2 = \sigma_\varepsilon^2 + \sigma^2$, and $\bar{\sigma}_{\varepsilon_i}^2 = \sigma_{\varepsilon_i} - (n-1)\sigma^2$, A will strictly increase and so will R.

Next we turn to the expression of the optimal common noise. Since $n\sigma_\theta^2 < \sigma_{\theta_i}^2$, the intermediary will not enter the market, we focus on the case where $n\sigma_\theta^2 > \sigma_{\theta_i}^2$:

$$\begin{aligned} \frac{\partial R}{\partial \sigma_\varepsilon^2} &= \frac{1}{8}(n\sigma_\theta^2+\sigma_{\theta_i}^2)\left(\frac{n(n\sigma_\theta^2+\sigma_{\theta_i}^2)}{(n(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^2} - \frac{(n-1)^2(3n\sigma_\theta^2-\sigma_{\theta_i}^2)}{n((n-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^2}\right) \\ &= \frac{B(\sigma_{\varepsilon_i}^2,\sigma_\theta^2,\sigma_{\theta_i}^2)\sigma_\varepsilon^4 + C(\sigma_{\varepsilon_i}^2,\sigma_\theta^2,\sigma_{\theta_i}^2)\sigma_\varepsilon^2 + D(\sigma_{\varepsilon_i}^2,\sigma_\theta^2,\sigma_{\theta_i}^2)}{n((n-1)(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^2(n(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)^2} \\ B(\sigma_{\varepsilon_i}^2,\sigma_\theta^2,\sigma_{\theta_i}^2) &= -2(n-1)^2n^2(n\sigma_\theta^2-\sigma_{\theta_i}^2) \end{aligned}$$

Therefore the numerator is a quadratic function of σ_ε^2 with negative quadratic term. It has

two real zero point and the larger one is:

$$\sigma^* = \frac{2n^2\sigma_\theta^4 - 2n^3\sigma_\theta^4 + n\sigma_{\theta_i}^2\left(\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right) + \sigma_{\theta_i}^2\left(-\sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)}\right)}{2n\left(n-1\right)\left(n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)}$$
$$+ \frac{-2n^2\sigma_\theta^2 - \sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)} + n\left(3\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right)}{2n\left(n-1\right)\left(n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)}\hat{\sigma}_{\varepsilon_i}^2,$$

which completes the proof. ■

Proof of Proposition 10. As we are considering the general information design problem, we use the general formula derived in the case of personalized price[11], to express the profit as well as the necessary compensation to the consumers:

$$R = g - \Sigma_i \Delta CS_i = -\Sigma_{i=1}^n \frac{1}{8}\operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) + \frac{3}{8}\Sigma_{i=1}^n \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}]),$$

Now our problem becomes:

$$\max_{\mathcal{I},\mathcal{I}_i} -\Sigma_{i=1}^n \frac{1}{8}\operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) + \frac{3}{8}\Sigma_{i=1}^n \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}]).$$

Since by definition $\mathcal{I}_{-i} \subset \mathcal{F}(s_1, ..., \hat{s}_i, ...s_n)$, we have:

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

where the inequality binds when $\mathcal{I}$ contains no information and $\mathcal{I}_{-i}$ contains all signals collected. ■

Proof of Proposition 11. The revenue of the broker is given by:

$$R = g - \Sigma_i \Delta CS_i = -\Sigma_{i=1}^n \frac{1}{8}\operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) + \frac{3}{8}\Sigma_{i=1}^n \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}])$$

and the consumer surplus is:

$$\Delta CS_i = \frac{3}{8}(\operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) - \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}])).$$

[11] We can not use formula in the case of uniform price because there we directly use the fact that the broker does not collect id: $\mathbb{E}[\theta_{i'}|\mathcal{I}] = \mathbb{E}[\theta_i|\mathcal{I}]$.

Now our problem becomes:

$$\max_{\mathcal{I},\mathcal{I}_i} -\Sigma_{i=1}^n \frac{1}{8}\operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) + \frac{3}{8}\Sigma_{i=1}^n \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}])$$
$$s.t. \quad \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) - \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}]) \geq 0.$$

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

$$\max_{\mathcal{I},\mathcal{I}_i} \frac{1}{4}\Sigma_{i=1}^n \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}])$$
$$s.t. \quad \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) - \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}]) \geq 0.$$

Then any information structure such that $\mathcal{I}_{-i} = \mathcal{F}(s_1, ..., \hat{s}_i, ...s_n)$ and $\operatorname{var}(\mathbb{E}[w_i|\mathcal{I}]) = \operatorname{var}(\mathbb{E}[w_i|\mathcal{I}_{-i}])$, i.e. use all the off-path information for personalized price and keep consumers at the same surplus level on the path, will maximize our original problem. ■

Proof of Proposition 12. The total compensation is:

$$\Sigma_i m_i = \frac{(n\sigma_\theta^2 + \sigma_{\theta_i}^2)(4\sigma_\varepsilon^2(n-1)n\sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2(3n\sigma_\theta^2 + 4n\sigma_{\theta_i}^2 - \sigma_{\theta_i}^2) + (4n-1)\sigma_{\theta_i}^2(n\sigma_\theta^2 + \sigma_{\theta_i}^2))}{8n(\sigma_\varepsilon^2(n-1) + \sigma_{\varepsilon_i}^2 + (n-1)\sigma_\theta^2 + \sigma_{\theta_i}^2)(n(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\varepsilon_i}^2 + \sigma_{\theta_i}^2)},$$

and thus

$$\lim_{n\to\infty} \Sigma_i m_i = \frac{\sigma_\theta^2 \sigma_{\theta_i}^2}{2(\sigma_\epsilon^2 + \sigma_\theta^2)}.$$

Asymptotically the sign of derivative is decided by:

$$\lim_{n\to\infty} n^2 \frac{\partial \Sigma_i m_i}{\partial n} = \frac{\sigma_{\eta_i}^2 \sigma_\theta^2 (3\sigma_\theta^2 - 4\sigma_{\theta_i}^2) + \sigma_{\theta_i}^2 \left(4\sigma_\eta^2 \sigma_{\theta_i}^2 + 3\sigma_\theta^4\right)}{8(\sigma_\eta^2 + \sigma_\theta^2)^2}.$$

When $\sigma_{\varepsilon_i}^2 = 0$ (and of course $n\sigma_\theta^2 > \sigma_{\theta_i}^2$), the derivative is:

$$\begin{aligned}\frac{\partial \Sigma_i m_i}{\partial n} =& \frac{\sigma_{\theta_i}^2 \left(-(n\sigma_\theta^2 + \sigma_{\theta_i}^2)^2 \left(3n^2\sigma_\theta^4 - 2n\sigma_\theta^2\sigma_{\theta_i}^2 + \sigma_{\theta_i}^2(\sigma_\theta^2 - \sigma_{\theta_i}^2)\right)\right)}{8n^2((n-1)(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\theta_i}^2)^2(n(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\theta_i}^2)^2} \\ & \frac{\sigma_{\theta_i}^2 \left(-4\sigma_\varepsilon^6(n-1)^2n^2\sigma_{\theta_i}^2 - \sigma_\varepsilon^4 n \left(3n^3\sigma_\theta^4 + 2\left(4n^2 - 6n + 3\right)n\sigma_\theta^2\sigma_{\theta_i}^2 + (n(8n-11)+2)\sigma_{\theta_i}^4\right)\right)}{8n^2((n-1)(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\theta_i}^2)^2(n(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\theta_i}^2)^2} \\ & \frac{\sigma_{\theta_i}^2 \left(-(n\sigma_\theta^2 + \sigma_{\theta_i}^2)^2 \left(3n^2\sigma_\theta^4 - 2n\sigma_\theta^2\sigma_{\theta_i}^2 + \sigma_{\theta_i}^2(\sigma_\theta^2 - \sigma_{\theta_i}^2)\right)\right)}{8n^2((n-1)(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\theta_i}^2)^2(n(\sigma_\varepsilon^2 + \sigma_\theta^2) + \sigma_{\theta_i}^2)^2}\end{aligned}$$

The derivative is negative once we notice that $n\sigma_\theta^2 > \sigma_{\theta_i}^2$ and:

$$3n^2\sigma_\theta^4 - 2n\sigma_\theta^2\sigma_{\theta_i}^2 + \sigma_{\theta_i}^2(\sigma_\theta^2 - \sigma_{\theta_i}^2) \geq \sigma_\theta^2\sigma_{\theta_i}^2 > 0.$$

Finally we characterize the asymptotic profit. By Proposition 6, the profit $R = (n\sigma_\theta^2 + \sigma_{\theta_i}^2)A/8$. Since

$$\lim_{n\to\infty} A = 2\frac{\sigma_\theta^4}{\sigma_\varepsilon^2 + \sigma_\theta^2}$$

Thus asymptotically, it is optimal to set $\sigma_\varepsilon^2 = \hat{\sigma}_\varepsilon^2$, and the per capita profit equals the one gained when aggregate data is "in the wild" (see the proof of Proposition 3). ■

Proof of Proposition 13. When the intermediary provides a "divide and conquer" compensation scheme for the individual compensation, the total compensation would be:

$$\begin{aligned}
\Sigma_k(CS_{k,-i} - CS_k) &= \Sigma_k c(n)(\frac{\sigma_\theta^2}{2} - \frac{c(n)}{8n^2})(\frac{1}{v(k)} - \frac{1}{v(k-1)}) + \frac{c(n)\sigma_{\theta_i}^2}{2v(k)k} \\
&= c(n)(\frac{\sigma_\theta^2}{2} - \frac{c(n)}{8n^2})\frac{1}{v(n)} + \frac{c(n)\sigma_{\theta_i}^2}{2n^2}\Sigma_k \frac{1}{k(\sigma_\theta^2 + \sigma_\varepsilon^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2}.
\end{aligned}$$

Thus, the asymptotic compensation is:

$$\begin{aligned}
\Sigma_i m_i &= c(n)(\frac{\sigma_\theta^2}{2} - \frac{c(n)}{8n^2})\frac{1}{v(n)} + \frac{c(n)\sigma_{\theta_i}^2}{2n^2}\Sigma_k \frac{1}{k(\sigma_\theta^2 + \sigma_\varepsilon^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2} \\
&\approx c(n)(\frac{\sigma_\theta^2}{2} - \frac{c(n)}{8n^2})\frac{1}{v(n)} + \frac{c(n)\sigma_{\theta_i}^2}{2n^2(\sigma_\theta^2 + \sigma_\varepsilon^2)}\Sigma_k \frac{1}{k} \\
&\approx c(n)(\frac{\sigma_\theta^2}{2} - \frac{c(n)}{8n^2})\frac{1}{v(n)} + \frac{c(n)\sigma_{\theta_i}^2}{2n^2(\sigma_\theta^2 + \sigma_\varepsilon^2)}\log n \\
&\approx \frac{3}{8}\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\varepsilon^2} + \frac{\sigma_\theta^2\sigma_{\theta_i}^2}{2(\sigma_\theta^2 + \sigma_\varepsilon^2)}\log n.
\end{aligned}$$

Thus asymptotically the average compensation will decrease to 0. Now we calculate the total revenue for the intermediary:

$$\begin{aligned}
R &= \frac{c(n)^2}{4nv(n)} - c(n)(\frac{\sigma_\theta^2}{2} - \frac{c(n)}{8n^2})\frac{1}{v(n)} - \frac{c(n)\sigma_\theta^2}{2n^2}\Sigma_k \frac{1}{k(\sigma_\theta^2 + \sigma_\varepsilon^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2} \\
&= (\frac{1}{4} + \frac{1}{8n})\frac{(n\sigma_\theta^2 + \sigma_{\theta_i}^2)^2}{n(\sigma_\theta^2 + \sigma_\varepsilon^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2} - \frac{\sigma_\theta^2}{2}\frac{n\sigma_\theta^2 + \sigma_{\theta_i}^2}{n(\sigma_\theta^2 + \sigma_\varepsilon^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2} \\
&\quad - \frac{(n\sigma_\theta^2 + \sigma_{\theta_i}^2)\sigma_{\theta_i}^2}{2n}\Sigma_k \frac{1}{k(\sigma_\theta^2 + \sigma_\varepsilon^2) + \sigma_{\theta_i}^2 + \sigma_{\varepsilon_i}^2} \\
&\approx (\frac{n}{4} - \frac{3}{8})\frac{\sigma_\theta^4}{\sigma_\theta^2 + \sigma_\varepsilon^2} - \frac{\sigma_\theta^2\sigma_{\theta_i}^2}{2(\sigma_\theta^2 + \sigma_\varepsilon^2)}\log n,
\end{aligned}$$

where we define

$$c(k) \triangleq \text{cov}\left[(n/k)\,\Sigma_{i=1}^{k} s_i, \Sigma_{i=1}^{n} t_i\right] = n^2\sigma_\theta^2 + n\sigma_{\theta_i}^2$$
$$v(k) \triangleq \text{var}\left[(n/k)\,\Sigma_{i=1}^{k} s_i\right] = n^2\sigma_\theta^2 + \left(n^2/k\right)\sigma_{\theta_i}^2 + n^2\sigma_\varepsilon^2 + \left(n^2/k\right)\sigma_{\varepsilon_i}^2$$

Therefore it is asymptotically optimal to set $\sigma_\varepsilon^2 = \hat{\sigma}_\varepsilon^2$ and the per capita profit will increase to the one gained when aggregate data is "in the wild".\ ■

Proof of Proposition 14. We first prove the monotonicity of σ_ε^* w.r.t. $\hat{\sigma}_{\varepsilon_i}^2$. Recall the optimal common noise is $\max\{\hat{\sigma}_\varepsilon^2, \sigma^*\}$, where:

$$\sigma^* = \frac{2n^2\sigma_\theta^4 - 2n^3\sigma_\theta^4 + n\sigma_{\theta_i}^2\left(\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right) + \sigma_{\theta_i}^2\left(-\sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)}\right)}{2n\left(n-1\right)\left(n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)}$$
$$+ \frac{-2n^2\sigma_\theta^2 - \sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)} + n\left(3\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right)}{2n\left(n-1\right)\left(n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)}\hat{\sigma}_{\varepsilon_i}^2$$

So for fixed n, σ_θ^2 and $\sigma_{\theta_i}^2$, σ^* is a linear function of $\hat{\sigma}_{\varepsilon_i}^*$ When the linear parameter of $\hat{\sigma}_{\varepsilon_i}^*$ we are done. When it is not, we prove that $\sigma^* < 0$ so that $\sigma_\varepsilon^* = \hat{\sigma}_\varepsilon$ is a constant. In fact,

$$-2n^2\sigma_\theta^2 - \sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)} + n\left(3\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right) < 0,$$
$$-2n^2\sigma_\theta^2\sigma_{\theta_i}^2 - \sigma_{\theta_i}^4 + \sigma_{\theta_i}^2\sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)} + 2\sigma_{\theta_i}^4 + 3n\sigma_{\theta_i}^2\sigma_\theta^2 < 0.$$

Then we can show that the constant term is also negative:

$$2n^2\sigma_\theta^4 - 2n^3\sigma_\theta^4 + n\sigma_{\theta_i}^2\left(\sigma_\theta^2 + 2\sigma_{\theta_i}^2\right) + \sigma_{\theta_i}^2\left(-\sigma_{\theta_i}^2 + \sqrt{\left(3n\sigma_\theta^2 - \sigma_{\theta_i}^2\right)\left(n\sigma_\theta^2 + \sigma_{\theta_i}^2\right)}\right)$$
$$< (2n^2 - 2n^3)\sigma_\theta^4 - (2n - 2n^2)\sigma_\theta^2\sigma_{\theta_i}^2 = (2n - 2n^2)\sigma_\theta^2(n\sigma_\theta^2 - \sigma_{\theta_i}^2) < 0.$$

Finally we prove the profit is convex with respect to $\sigma^2_{\varepsilon_i}$

$$\begin{aligned}
\frac{\partial^2 R}{\partial(\sigma^2_{\varepsilon_i})^2} &= \frac{1}{8}(n\sigma^2_\theta+\sigma^2_{\theta_i})\left(\frac{2(n-1)(3n\sigma^2_\theta-\sigma^2_{\theta_i})}{n((n-1)(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})^3}-\frac{2(n\sigma^2_\theta+\sigma^2_{\theta_i})}{(n(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})^3}\right)\\
&> \frac{1}{8}\frac{n\sigma^2_\theta+\sigma^2_{\theta_i}}{(n(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})^2}\left(\frac{2(n-1)(3n\sigma^2_\theta-\sigma^2_{\theta_i})}{n((n-1)(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})}-\frac{2(n\sigma^2_\theta+\sigma^2_{\theta_i})}{(n(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})}\right)\\
&= \frac{1}{8}\frac{n\sigma^2_\theta+\sigma^2_{\theta_i}}{(n(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})^3 n((n-1)(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})}\Big(2(2\sigma^2_\varepsilon(n-1)n(n\sigma^2_\theta-\sigma^2_{\theta_i})\\
&\quad+\sigma^2_{\varepsilon_i}(n(2n-3)\sigma^2_\theta-2n\sigma^2_{\theta_i}+\sigma^2_{\theta_i})+(n\sigma^2_\theta+\sigma^2_{\theta_i})(2(n-1)n\sigma^2_\theta-2n\sigma^2_{\theta_i}+\sigma^2_{\theta_i}))\Big)\\
&> \frac{1}{8}\frac{(n\sigma^2_\theta+\sigma^2_{\theta_i})\frac{2(\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})\sqrt{(n-1)^2n^2(3n\sigma^2_\theta-\sigma^2_{\theta_i})(n\sigma^2_\theta+\sigma^2_{\theta_i})}}{(n-1)n}}{(n(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})^3 n((n-1)(\sigma^2_\varepsilon+\sigma^2_\theta)+\sigma^2_{\varepsilon_i}+\sigma^2_{\theta_i})}>0
\end{aligned}$$

Where the last equation holds when $\sigma_\varepsilon \geq \sigma^*$ and $n\sigma^2_\theta \geq \sigma^2_{\theta_i}$. Now since for fixed parameters profit is a convex function of $\sigma^2_{\varepsilon_i}$, the optimal profit, where we know $\sigma^2_{\varepsilon_i}$ always equals its lower bound, will also be a convex function of this lower bound. ■

Proof of Proposition 15. We first derive the expression of profit in these two cases. In the uniform pricing scheme, intermediary reports one signal about the total average demand $\mathbb{E}[\Sigma_{j,i}t_{ji}|\mathcal{I}]$.

$$\begin{aligned}
p &= \frac{1}{2}\mathbb{E}[\bar{t}|\mathcal{I}] = \frac{1}{2}\frac{\text{cov}[\bar{s},\bar{t}]}{\text{var}[\bar{s}]}\bar{s}\\
g &= 2nVar[p] = \frac{1}{8n}\frac{(2n^2\sigma^2_\theta+2n\sigma^2_{\theta_i})^2}{2n^2(\sigma^2_\theta+\sigma^2_\varepsilon)+2n(\sigma^2_{\theta_i}+\sigma^2_{\varepsilon_i})} = \frac{1}{4}\frac{(n\sigma^2_\theta+\sigma^2_{\theta_i})^2}{n(\sigma^2_\theta+\sigma^2_\varepsilon)+\sigma^2_{\theta_i}+\sigma^2_{\varepsilon_i}}\\
CS &= -\frac{3}{4n}g\\
CS_{-ji} &= -\text{cov}[t_{ji},p_{-ji}]+\frac{1}{2}\text{var}[p_{-ji}]\\
p_{-ji} &= \frac{1}{2}\frac{\text{cov}[\bar{s}_{-ji},\bar{t}]}{\text{var}[\bar{s}_{-ji}]}\bar{s}_{-ji} = \frac{1}{2}\frac{2n-1}{2n}\frac{n(2n-1)\sigma^2_\theta+(2n-1)\sigma^2_{\theta_i}}{(n^2+(n-1)^2)(\sigma^2_\theta+\sigma^2_\varepsilon)+(2n-1)(\sigma^2_{\theta_i}+\sigma^2_{\epsilon_i})}\bar{s}_{-ji}\\
CS_{-ji} &= \frac{n(2n-1)\sigma^2_\theta+(2n-1)\sigma^2_{\theta_i}}{4n((n^2+(n-1)^2)(\sigma^2_\theta+\sigma^2_\varepsilon)+(2n-1)(\sigma^2_{\theta_i}+\sigma^2_{\epsilon_i}))}(\frac{2n-1}{8n}\sigma^2_{\theta_i}-\frac{6n-7}{8}\sigma^2_\theta)
\end{aligned}$$

Therefore the revenue for the broker is:

$$\begin{aligned}
R_1 &= \Delta\pi + 2nCS_n - 2nCS_{-ji} \\
&= \frac{1}{16}(n\sigma_\theta^2 + \sigma_{\theta_i}^2)\left(-\frac{2(n\sigma_\theta^2+\sigma_{\theta_i}^2)}{n(\sigma_\varepsilon^2+\sigma_\theta^2)+\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2} + \frac{(2n-1)\left((6n-7)\sigma_\theta^2 - \left(2-\frac{1}{n}\right)\sigma_{\theta_i}^2\right)}{(2(n-1)n+1)(\sigma_\varepsilon^2+\sigma_\theta^2)+(2n-1)(\sigma_{\varepsilon_i}^2+\sigma_{\theta_i}^2)}\right) \\
\lim_{n\to\infty}\frac{R_1}{2n} &= \frac{1}{8}\frac{\sigma_\theta^4}{\sigma_\theta^2+\sigma_\varepsilon^2}
\end{aligned}$$

In the two price scheme, the intermediary reports two signal about the average demand of two natural group respectively:

$$\mathbb{E}[\Sigma_i t_{1i}|\ \mathcal{I}], \mathbb{E}[\Sigma_i t_{2i}|\mathcal{I}].$$

Since the whole market is divided into two separate parts, we can directly use the previous result.

$$\begin{aligned}
p_j &= \frac{1}{2}\mathbb{E}[\bar{t}_j|\mathcal{I}] = \frac{1}{2}\frac{\operatorname{cov}[\bar{s}_j,\bar{t}_j]}{\operatorname{var}[\bar{s}_j]}\bar{s}_j \\
g_j &= n\operatorname{var}[p_j] \\
R_2 &= \frac{n\sigma_\theta^2+\sigma_{\theta_i}^2}{4}\left(\frac{3(n-1)\sigma_\theta^2 - \frac{n-1}{n}\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2+(n-1)(\sigma_\theta^2+\sigma_\varepsilon^2)} - \frac{n\sigma_\theta^2+\sigma_{\theta_i}^2}{\sigma_{\theta_i}^2+\sigma_{\varepsilon_i}^2+n(\sigma_\theta^2+\sigma_\varepsilon^2)}\right) \\
\lim_{n\to\infty}\frac{R_2}{2n} &= \frac{1}{4}\frac{\sigma_\theta^4}{\sigma_\theta^2+\sigma_\varepsilon^2}
\end{aligned}$$

In noiseless case, the difference of profit is:

$$R_1 - R_2 = \frac{1}{16}(n\sigma_\theta^2+\sigma_{\theta_i}^2)\left(-\frac{4(n-1)(3n\sigma_\theta^2-\sigma_{\theta_i}^2)}{n((n-1)\sigma_\theta^2+\sigma_{\theta_i}^2)} - \frac{(2n-1)\left((7-6n)\sigma_\theta^2+\left(2-\frac{1}{n}\right)\sigma_{\theta_i}^2\right)}{(2(n-1)n+1)\sigma_\theta^2+(2n-1)\sigma_{\theta_i}^2} + 2\right)$$

Then we simply take the derivatives and find that (we focus on case where $n > 1$)

$$\frac{\partial(R_1-R_2)}{n} < 0.$$

Therefore the difference is monotonically decreasing in n for $n \geq 2$. Notice that from previous result, $R_2 < 0$ as long as $\sigma_{\theta_i}^2/\sigma_\theta^2 > n$. On the other hand, we could rewrite R_1 as:

$$R_1 = \frac{1}{16}(n\sigma_\theta^2+\sigma_{\theta_i}^2)\frac{n(8(n-2)n+5)\sigma_\theta^2+n(6-8n)\sigma_{\theta_i}^2-\sigma_{\theta_i}^2}{n(2(n-1)n\sigma_\theta^2+(2n-1)\sigma_{\theta_i}^2+\sigma_\theta^2)}.$$

The threshold of profitability is:

$$\frac{\sigma_{\theta_i}^2}{\sigma_\theta^2} < \frac{(8n^2-16n+5)n}{(8n-6)n+1} < \frac{8n^2-16n+5}{8n-5} < n.$$

Thus the one price scheme will be able to bring positive profit when group pricing could not, thus $R_1 - R_2 > 0$ in this case. On the other hand, since

$$\lim_{n\to\infty} \frac{R_1 - R_2}{n} = -\frac{1}{4}\sigma_\theta^2$$

Thus $\pi_1 - \pi_2 < 0$ when n is sufficiently large, which proves the existence of the threshold. ■

Proof of Proposition 16. We simply take the derivative with respect to σ_θ^2, for $n > 1$:

$$\begin{aligned}\frac{\partial R_1 - R_2}{\partial \sigma_\theta^2} =& -\frac{(n-1)^2n^2(2(n-1)n+1)\left(8n^2+3\right)\sigma_\theta^8 + 2(n-1)n^2\left(8n^2+3\right)(n(4n-5)+2)\sigma_\theta^6\sigma_{\theta_i}^2}{16n((n-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2(2(n-1)n\sigma_\theta^2+(2n-1)\sigma_{\theta_i}^2+\sigma_\theta^2)^2} \\ &- \frac{2(2n-1)(n(n(4n(4n-7)+21)-13)+3)\sigma_\theta^2\sigma_{\theta_i}^6 + (2n-3)(2n-1)^3\sigma_{\theta_i}^8}{16n((n-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2(2(n-1)n\sigma_\theta^2+(2n-1)\sigma_{\theta_i}^2+\sigma_\theta^2)^2} \\ &- \frac{(2n(n(n(n(48(n-2)n+83)-55)+27)-9)+3)\sigma_\theta^4\sigma_{\theta_i}^4}{16n((n-1)\sigma_\theta^2+\sigma_{\theta_i}^2)^2(2(n-1)n\sigma_\theta^2+(2n-1)\sigma_{\theta_i}^2+\sigma_\theta^2)^2} < 0.\end{aligned}$$

Therefore the threshold is decreasing with respect to σ_θ^2. Notice that in the noiseless case, it is the ratio of $\sigma_\theta^2/\sigma_{\theta_i}^2$ that decides the sign of $R_1 - R_2$, therefore the threshold is increasing with respect to $\sigma_{\theta_i}^2$. ■

Proof of Proposition 17. From the expression derived in the proof of proposition, the asymptotic per capita profit for extreme price discrimination in noiseless case is:

$$\frac{R}{2n} = \frac{\sigma_\theta^2}{4} - \frac{\sigma_{\theta_i}^2}{8}.$$

While from the expression derived in the proof of proposition, the asymptotic per capita profit for uniform pricing in noiseless case is:

$$\frac{R_1}{2n} = \frac{\sigma_\theta^2}{8},$$

which completes the proof. ■

Proof of Proposition 18. We first verify that the candidate equilibrium indeed constitutes an equilibrium. On the monopoly side, when it faces the equilibrium prices, suppose purchasing $D \subsetneq J$ is optimal for monopoly. If $J \backslash D = \{j\}$, then by construction, purchasing

J is equally well for monopoly and thus optimal. If $J\backslash D$ contains two intermediaries, suppose j is one of them. Since the value for information is decreasing, and the information is not perfect due to condition $\hat{\sigma}^2_{\epsilon_i} > 0$, we have:

$$\pi(\{j\} \cup D) - \pi(D) > \pi(J) - \pi(J\backslash\{j\}) = p_j.$$

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

On the intermediaries side, fixed other intermediaries pricing $p_{\neg j}$, suppose it is optimal for intermediary j to report $p'_j > p_j$. Then denote monopoly would purchase database from D. By construction we know $D \neq J$, otherwise it is better not to purchase from j. For the same reason as last paragraph is also impossible for $J\backslash D$ to contains two intermediaries. Therefore it is only possible that $\{j'\} = J\backslash D$. However, since

$$\begin{aligned}
\bar{\pi}(J\backslash\{j\}) - \bar{\pi}(D) &= (\pi(J\backslash\{j\}) - \Sigma_{k\in J\backslash\{j\}} p_k) - (\pi(D) - \Sigma_{k\in D} p_k) \\
&= -(\pi(J) - \pi(J\backslash\{j\})) + p'_j + \pi(J) - \pi(D) - p_{j'} \\
&= -(\pi(J) - \pi(J\backslash\{j\})) + p'_j > 0.
\end{aligned}$$

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 could get payoff higher than $\pi(J) - \pi(J\backslash\{j\})$. Suppose in contradiction j only get $u_j < \pi(J) - \pi(J\backslash\{j\})$ in a equilibrium, he could always sell his product at p_j such that $u_j < p_j < \pi(J) - \pi(J\backslash\{j\})$ and the monopoly would buy his database regardless of its purchasing plan, because of decreasing return of information.

Suppose in the equilibrium one intermediary j sells his database at $p_j > \pi(J) - \pi(J\backslash\{j\})$, from last paragraph we know $p_{j'} \geq \pi(J) - \pi(J\backslash\{j'\})$. Therefore on the equilibrium path, optimally reacted monopoly would not purchase all the database, but this is a contradiction since the rejected intermediary will only get 0 on this equilibrium path. ■

Before we proceed to prove Proposition 19, we prove three lemmas first.

Lemma 1 (Unique Equilibrium for Accepting Game)
For given offers $\{(p_{ji}, \varepsilon_{ij}, \varepsilon_j)\}$, the maximal accepting set is unique.

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

We will use a iterated expanding approach to complete the construction. Denote $A^0_i = A_i \cup \bar{A}_i$. Denote also $CS_i(A_i, A_{\neg i})$ as consumers gross pay off when accepting sets are $\{A_i\}$.

Since $\{A_i\}$ and $\{\bar{A}_i\}$ are equilibrium choices, we have:

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

Therefore, because of decreasing compensation, we will have:

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

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 $B_i \cap \bar{B}_i = \emptyset$ we have:

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

In conclusion we have:

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

If $\forall i$, $\nexists \emptyset \neq B_i \subset (A_i^0)^c$ such that

$$CS_i(A_i^0 \cup B_i^0, A_{\neg i}^0) - CS_i(A^0, A_{\neg i}^0) + \Sigma_{j \in B_i^0} p_{ji} \leq 0.$$

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 $\{A_i^1\}$:

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

Continuing this iterated process, in finite steps it will stop and the final set $\{A_i^\infty\}$ satisfies:

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

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 $B_i \subset A_i^\infty$, $C_i \subset (A_i^\infty)^c$:

$$\begin{aligned}
& CS_i(A_i^\infty, A_{\neg i}^\infty) + \Sigma_{j\in A_i^\infty} p_{ji} - CS_i(B_i \cup C_i, A_{\neg i}^\infty) - \Sigma_{j\in B_i\cup C_i} p_{ji} \\
= \; & CS_i(A_i^\infty, A_{\neg i}^\infty) - CS_i(A_i^\infty \cup C_i, A_{\neg i}^\infty) - \Sigma_{j\in C_i} p_{ji} \\
& + CS_i(A_i^\infty \cup C_i, A_{\neg i}^\infty) - CS_i(B_i \cup C_i, A_{\neg i}^\infty) + \Sigma_{j\in A_i^\infty \backslash B_i} p_{ji} \\
\geq \; & -(CS_i(A_i^\infty \cup C_i, A_{\neg i}^\infty) - CS_i(A_i^\infty, A_{\neg i}^\infty) + \Sigma_{j\in C_i} p_{ji}) \\
& + CS_i(A_i^\infty, A_{\neg i}^\infty) - CS_i(B_i, A_{\neg i}^\infty) + \Sigma_{j\in A_i^\infty \backslash B_i} p_{ji} \\
> \; & 0,
\end{aligned}$$

which completes the proof. ■

Note that the proof also shows that consumer strictly prefer maximal accepting set over other smaller accepting set.

Denote $\{\bar{A}_i\}$ as the the maximal accepting set given offers $\{(p_{ji}, \varepsilon_{ij}, \varepsilon_j)\}$. Considering any larger accepting set (which is not an equilibrium for the accepting game), we next show that there must be one consumer strictly prefer a smaller accepting plan.

Lemma 2 (Want Less When Asked More)
For any $\{A_i\} \supsetneq \{\bar{A}_i\}$, $\exists$ i and $B_i \subsetneq A_i$ such that:

$$CS_i(A_i, A_{\neg i}) - CS_i(A_i \backslash B_i, A_{\neg i}) + \Sigma_{j\in B_i} p_{ji} < 0.$$

Proof. Suppose not, then we have

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

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 subset is profitable than A_i. Then it is easy to see that according to decreasing return that A_i^1 preserve the property of A_i that it is better than its every subset:

$$CS_i(A_i^1, A_{\neg i}) - CS_i(B_i, A_{\neg i}) + \Sigma_{j\in A_i^1 \backslash B_i} p_{ji} \geq 0.$$

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

Lemma 3 (Every Broker Access To Every Consumer)
Suppose $(\hat{\sigma}_\varepsilon^2, \hat{\sigma}_{\varepsilon_i}^2) \neq (0, 0)$, then every intermediary entering the market will collect data from every consumers.

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

As we can see in the proof of the unique accepting equilibrium, we know that facing the equilibrium prices, consumers strictly prefer A_i then other larger set. Equivalently, the following inequality holds strictly:

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

From Lemma 2 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 $\bar{A}_i$:

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

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

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

Here, when followed with the third argument, CS_i 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 system of finite strict inequality which are continuous with respect to $\sigma^2_{\varepsilon_{i'j'}}$. Thus intermediary j' could find a $\sigma^2_{\varepsilon^*_{i'j'}} < \infty$ such that all these equations still hold:

$$CS_i(A_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - CS_i(B_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - \Sigma_{j\in B_i\setminus A_i} p_{ij} > 0 \quad \forall B_i \supsetneq A_i, \tag{23}$$

$$CS_i(\bar{A}_i, \bar{A}_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - CS_i(\bar{A}_i\setminus\bar{B}_i, \bar{A}_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) + \Sigma_{j\in \bar{B}_i} p_{ji} < 0. \tag{24}$$

Moreover, from the equilibrium condition we have:

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

And therefore by decreasing compensation, we know:

$$CS_i(A_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) - CS_i(B_i, A_{\neg i}, \sigma^2_{\varepsilon^*_{i'j'}}) + \Sigma_{j\in A_i} p_{ij} - \Sigma_{j\in B_i} p_{ij} \geq 0 \quad \forall B_i \subset A_i. \tag{25}$$

Now consider a deviation from intermediary j' such that he remain $\varepsilon_{j'}$, $\varepsilon_{ij'}$, $p_{j'i}$ $\forall i \neq i'$ unchanged, set $\varepsilon_{i'j'} = \varepsilon^*_{i'j'}$ and offer a price $p^*_{j'i'}$ to consumer i' which keep i' exactly indifferent among $A_{i'}$ and $A_{i'} \cup \{j'\}$:

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

Now we will verify first that $A_i \cup \{j'\}$ and $A_{\neg i}$ are indeed optimal choice for each consumer, and second that they are maximal. By this two conclusion, we can ensure it is the unique outcome if intermediary j' makes such deviation.

First, from inequality 23 we know given other consumers choose $\{A_i \cup \{j'\}, A_{\neg i}\}$, consumer $i \neq i'$ prefer A_i over any larger set and consumer i' prefer $A_{i'} \cup \{j'\}$ over any larger set. From inequality 25 we know all consumers prefer A_i over any smaller set. Moreover, since by construction of $p^*_{j'i'}$, i' is indifferent between A_i and $A_i \cup \{j'\}$. Then we know (with a trick used frequently before) that $A_i \cup \{j'\}$ and $A_{\neg i}$ is indeed optimal choice for each consumer.

Second, for any set $\{\bar{A}_i\} \supsetneq \{A_i\}$ such that $j' \notin \bar{A}_{i'}$, inequality 24, therefore $\{\bar{A}_i \cup \{j'\}, \bar{A}_{\neg i}\}$ can not be equilibrium choices for consumers. Thus no set larger than $\{A_i \cup \{j'\}, A_{\neg i}\}$ could be equilibrium choice. One last step to lead to the contradiction is to verify such deviation is profitable. From previous analysis we know that the information collected by other intermediary would remain unchanged, thus we can simply denote their signal simply by $s_j = \theta + \delta_j$, where δ_j is a independent normal random variable. Because of the assumption $(\hat{\sigma}^2_\varepsilon, \hat{\sigma}^2_{\varepsilon_i}) \neq (0, 0)$ that information could not be perfect, we must have $\sigma^2_{\delta_j} > 0$. The signal sent by j' before and after deviation are $\theta + \delta_{j'}$ and $\theta + \hat{\delta}_{j'}$ where $\sigma^2_{\delta_{j'}} > \sigma^2_{\hat{\delta}_{j'}}$. By construction we know:

$$p_{ij'} = \hat{CS}_i(A_i) - \hat{CS}_i(A'_i) = \frac{3}{8}\operatorname{var}[\mathbb{E}\theta | s_{\neg j'} \hat{s}_{j'}] - \frac{3}{8}\operatorname{var}[\mathbb{E}\theta | s_{\neg j'} s_{j'}].$$

On the other hand, the extra fee intermediary j' could charge to the monopoly, according to Proposition **??** is:

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

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

Proof of Proposition 19. First suppose $\exists i, j'$ such that in the equilibrium $(\sigma^2_{\varepsilon_{ij'}}, \sigma^2_{\varepsilon_{j'}}) > (\hat{\sigma}^2_{\varepsilon_i}, \hat{\sigma}^2_\varepsilon)$, so there is one broker add extra noise in either common part or idiosyncratic part. From Proposition 3 we know in any equilibrium $A_i = J$ for any i. Consider a deviation from intermediary j': he simply change $(\sigma^2_{\varepsilon_{ij'}}, \sigma^2_{\varepsilon_{j'}})$ to $(\hat{\sigma}^2_{\varepsilon_i}, \hat{\sigma}^2_\varepsilon)$ and increases compensation to $p^*_{j'i} + \epsilon$ where

$$p^*_{j'i} = p_{j'i} + CS_i(J, \sigma^2_{\varepsilon^*_{ij'}}) - CS_i(J, \sigma^2_{\varepsilon_{ij'}}).$$

From original equilibrium condition we know:

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

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

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

By the construction of p^*_{ij} we also have:

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

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 broker 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 3 then shows that this deviation is profitable, which leads to the contradiction.

At last we will prove 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 $C_1 \neq \emptyset$.

Then consider a deviation for $j' \in C_1$, he could deviate by making $p_{ij'}$ epsilon 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 every consumers accept all offers are still an equilibrium outcome under this deviation and (since it is clearly maximal) is our selected equilibrium. This deviation is profitable to j' which makes a contradiction. ■

Proof of Proposition 20 As we have just argue in the main text. In the symmetric candidate equilibrium, the potential deviation for the broker is to deviate to some information structure with higher noise such that he could persuade consumers not to provide any information to other brokers. The following lemma characterizes "the most profitable deviation":

Lemma 4 (The Maximal Deviation in Symmetric Candidate)
The optimal deviation for the broker can be calculated by assuming broker j' deviates symmetrically to a information structure with $\sigma^2_{\varepsilon^*_{ij'}} = 0$, $\sigma_{\varepsilon^*_{j'}} > 0$ and price $p^*_{ij'}$ such that:

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

2. *conditional on every other consumer report their information to all brokers, consumer is indifferent between reporting only to j' and to all brokers.*

Proof. Suppose there is a profitable deviation for broker j', and denote it as $\{\varepsilon^*_{ij'}, \varepsilon^*_{j'}, p^*_{ij'}\}$, and the accepting set it induces as $\{A^*_i\}$. If such deviation does not reduce the information that other brokers 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 broker $j_0 \neq j'$. Note by previous paragraph we know at least one consumer i_1 does not report to some other broker $j_1 \neq j'$. Therefore we have the following two inequalities:

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

Note the other broker's compensation p_{ij} 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 $\sigma^2_{\varepsilon_{ij'}} = \infty$, $p_{ij'} = 0$ 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. $p^*_{ij'}$ 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 broker 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 brokers, 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 every consumer report 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 brokers, 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 all the noisy reports with proper weight, we can represent the signal collected by broker j' on the equilibrium path as $\theta + \delta$, while the signal without report from consumer i as $\theta + \delta_i$. We know:

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

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 left is the level of common noise. The requirement that $A_i^* = \{j'\}$ gives a system of lower bounds for common noise, and according to decreasing return, the most stringent one is:

$$CS_i(\{j'^*_{\neg i}) - CS_i(J, A^*_{\neg i}) + (|J| - 1)p_{ij} \geq 0.$$

The broker j' would always want to decrease noise level as long as the above inequality is satisfied. Thus the 'optimal deviation' is to provide an information structure with zero idiosyncratic noise and some common noise which make this constraint binding. [12]. ■

[12] It's not *the optimal* one because when the constraint is binding, the maximal accepting assumption would lead to $A_i^* = J$. But the noise level could be as close to this as possible.

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.

Bergemann, D., and A. Bonatti (2015): "Selling Cookies," *American Economic Journal: Microeconomics*, 7, 259–294.

——— (2019): "Markets for Information: An Introduction," *Annual Review of Economics*.

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.

Bergemann, D., and S. Morris (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57, 44–95.

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.

Maskin, E., and J. Riley (1984): "Monopoly with Incomplete Information," *Rand Journal of Economics*, 15, 171–196.

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

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.

Tirole, J. (1988): *The Theory of Industrial Organization*. MIT Press, Cambridge.

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