The Economics of Data

David Easley Shiyang Huang Liyan Yang Zhuo Zhong*

July 10, 2018

Abstract

We analyze the economic consequences of selling consumer data to oligopoly producers. Without data sales, producers keep secret their private consumer data, leading to efficiency loss and in some cases, to a prisoners dilemma for producers. In the presence of an independent data vendor who maximizes its own profits with smart contracts, data sales causes producers to effectively share their consumer data in equilibrium, thereby improving total surplus. This setting is consistent with a situation in which data is owned by consumers and analyzing such a setting provides a way to quantify the economic value of consumer data. When data is owned by producers, a data vendor a la a trade association is likely to maximize the total profits of producers, and its presence can address the prisoners dilemma for producers. Our analysis provides implications for the debates about data ownership and privacy.

JEL Classification: D61; G14; M41

Keywords: Data, Industry organization, Welfare, Prisoners' dilemma, Ownership

*David Easley (dae3@cornell.edu) is at Cornell University; Shiyang Huang (huangsy@hku.hk) is at University of Hong Kong; Liyan Yang (liyan.yang@rotman.utoronto.ca) is at University of Toronto; Zhuo Zhong (zhuo.zhong@unimelb.edu.au) is at University of Melbourne.

1 Introduction

The debate about firms or platforms using, sharing or selling consumer data typically focuses on privacy concerns related to data generated by individuals' online behavior, see for example, Acquisti (2015). Data about individual behavior is obviously valuable as it enables advertisers to target ads and sellers to target product offers to specific individuals. Indications of the aggregate value of this ability to use consumer data are reflected in the enormous profits of firms such as Facebook and the emerging market for the sale of data about consumer behavior. Consumers often say that they are concerned about privacy, and privacy or the lack of it is a frequent topic addressed in the media, but the effort the individuals seem to be willing to undertake to protect privacy or the amount that they are willing to pay for it are minimal (see the Economist [2018]).

In this paper we analyze an economy in which the value that individuals' place on privacy is derived from the economic consequences of ownership of data about their behavior. We do not assume that consumers have an innate value for their data by doing something as crude as putting privacy in their utility function; instead, our consumers can be harmed or helped by firms' knowledge of their behavior and this generates an indirect value of their data to them. We, of course, don't deny that individuals may have other more fundamental values for privacy, and that a direct payoff to privacy may lead to an additional reason for consumers to be concerned about it. But we show that consumers should be concerned about who owns their data and what they can do with it even without a direct value for it.

To enable us to focus on consumers' indirect value of data we examine markets in which firms care about aggregated consumer data as this data allows them to better predict future consumer demand for their products and to adjust prices accordingly. Who owns the data and what they can do with it affects consumer surplus, firm profits and efficiency. In this world, privacy is less of a direct concern than it is in the Facebook world, as here the data that firms want is an aggregation (which can be anonymized in large markets) of individual consumer data. Sharing of data about aggregate consumer behavior is always socially valuable in the

economy we analyze, although who gains and who loses from data sharing depends on the market structure. Nonetheless, consumers should care about the ownership of the data they and firms jointly produce.

We analyze an economy in which several firms sell similar products in multiple markets with each firm participating in some, but not necessarily all of the markets. If consumer demand is correlated over time and across markets, then each firm could better estimate the demand for its products if it had information about sales in markets in which it does not participate and so does not have sales data. Sharing of this information could potentially harm or benefit consumers as it could allow firms to better exploit any market power or it could allow firms to better tailor their output to actual demand. Thus, even if consumers have no direct value for privacy, they may have indirect derived economic value for keeping data about their purchases private as whether firms have the data may affect prices and thus consumer welfare. Firms too may prefer that data about their sales be shared or that it be kept private.

We are interested in two types of questions. First, who gains and who loses if consumer data is shared, and what happens to total surplus? Second, if there is private ownership of consumer data how does the initial ownership of the data affect sharing and thus consumer welfare, firm profit and total surplus? There are several obvious possible initial owners of consumer data: consumers, firms, and in some industries platforms on which firms sell their products. How society assigns property rights to consumer data clearly matters for the payoffs to individual players (the consumers, firms and platforms), but we show that it also matters for total surplus.

Suppose first that firms own the data. Any individual firm obviously benefits from having information about other firms sales as this information may be useful in predicting demand for its own output. Total firm profits may also increase if all firms have access to all data. However, it is well known that the data sharing game can be a prisoners' dilemna, see Darrough (1993) in which it is not beneficial for any firm to share its sales data with other

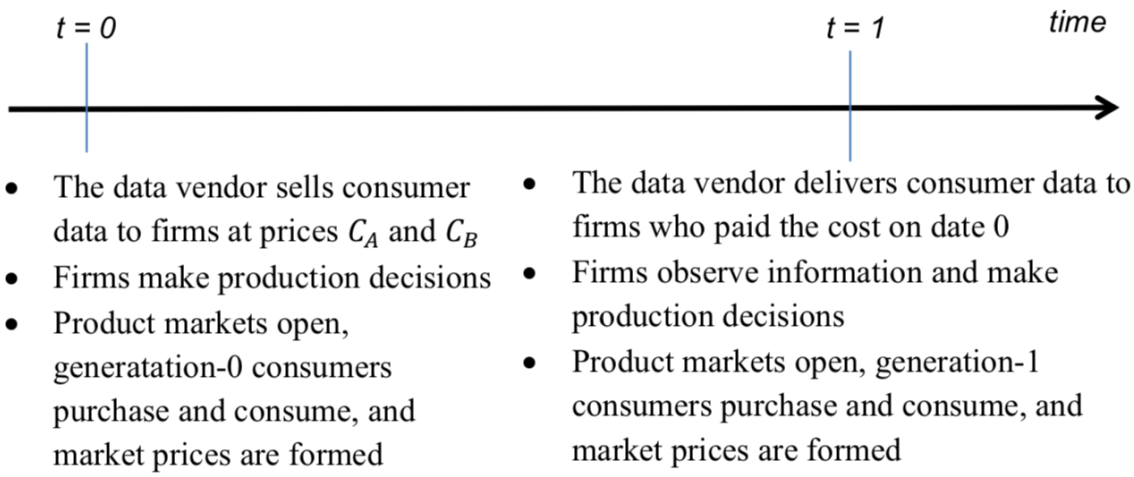
firms. So even if total profits for the firms would increase were sales data to be shared, voluntary data sharing may not occur.

Alternatively, firms may sell their products via an online platform which could own the data. The platform, of course, wants to sell the data and we show that a profit maximizing platform will sell the socially optimal amount of data. This occurs because the platform can extract all of the surplus that data sales generate, so its in the platforms best interest to generate maximal surplus. We show that this sale of data will increase consumer surplus. Whether it also increases firms' profits depends on market structure. If the firms have monopoly power in sufficiently many markets then the extra profit generated by better data in these markets can more than offset the reduced profit resulting from better data in a competitive market.

Finally, consumers could own their own data. If they can costlessly organize to create a data vendor then they too will maximize total surplus for the same reason that a platform owning the data would maximize social surplus. That is, as long as the data is owned by someone other than the firms who want to use it, social surplus is maximized. Of course, the division of this surplus is affected by ownership. But most important is that if firms own the data they face an tradeoff between selling all of the data and thus maximizing total surplus, and the increased competition they face in the product market from data sales, and this can lead to a loss of social surplus.

The paper is organized as follows. Our model of market structure, consumer demand and use of data is set out in Section 2. Section 3 provides an analysis of the value of consumer data and the sale of it by firms. Section 4 considers a data vendor and shows that the data vendor sells the socially optimal amount of data. In Section 5 we consider possible ownership structures for the data vendor and the consequences of ownership. We offer our conclusions in Section 6.

Figure 1: Timeline



2 The model

The economy lasts for two periods, $t = 0, 1$. The timeline of the economy is described by Figure 1. At the beginning of date 0, a data vendor designs contracts to sell consumer data to two firms, labeled as A and B. Firms then make production decisions and the product markets clear. The date-0 equilibrium product prices generate the consumer data, which is observable to the data vendor at the end of date 0. At the beginning of date 1, the data vendor delivers the promised consumer data to firms. Then, firms make optimal production decisions based on their information sets. Finally, the product markets clear, consumers purchase goods and consume, and firms realize their profits. We next describe in greater detail the product markets and the data market, and then define the equilibrium concept.

2.1 Consumers and product demand

There are two generations of consumers, each living for one period and consuming goods produced by firms. Within each generation, there are three types of consumers: A-type, B-type, and AB-type. An A-type consumer buys goods only from firm A; a B-type consumer buys goods only from firm B; and an AB-type consumer buys either from firm A or from firm

B. We interpret A-type consumers and B-type consumers respectively as each firm's "local" market consumers, and AB-type consumers as firms' "global" market consumers. Firms behave as monopolists in their local markets and compete in the global market. There are M local markets for each firm, where M is a positive integer. In each local market, there is one representative consumer. In the global market, there exist $N \geq 1$ representative AB-type consumers (and thus the global market is relatively larger than a typical local market). We label local markets as X-markets and the global market as the Y-market.

For illustration, imagine that firms A and B each sell to their own customers either through their respective brick-and-mortar stores (the X-markets) or through Amazon.com (the Y-market). The customers visiting brick-and-mortar stores of one firm do not consult the other firm. Perhaps these customers are unaware of the other firm, search costs make finding the other firm's price too difficult, or simply these customers do not have the habit of checking Amazon. For those customers who often visit Amazon.com, they see the products available from both firms and so in this market the two firms compete. This same story works for other web sites that consolidate markets. Amazon-like sites including Taobao and eBay. Another class of example are travel sites (expedia, travelocity, and hotels.com) for airlines, hotel rooms and rental cars.

We denote by U_A, U_B, and U_{AB} the utilities for type-A, type-B, and type-AB consumers, respectively. Consumers derive utility from consuming the products produced by firms according to the following quasi-linear form:[1]

$$U_A(x^t_{A,i}) = \tilde{s}^t_{A,i} x^t_{A,i} - \frac{(x^t_{A,i})^2}{2} - p^t_{A,i} x^t_{A,i},\ i = 1, 2, ..., M; \tag{1}$$

$$U_B(x^t_{B,j}) = \tilde{s}^t_{B,j} x^t_{B,j} - \frac{(x^t_{B,j})^2}{2} - p^t_{B,j} x^t_{B,j},\ j = 1, 2, ..., M; \tag{2}$$

$$U_{AB}(y^t_k) = \tilde{s}^t_{AB,k} y^t_k - \frac{(y^t_k)^2}{2} - p^t_y y^t_k,\ k = 1, 2, ..., N. \tag{3}$$

Here, variables $x^t_{A,i}$ and $x^t_{B,j}$ are the the quantities consumed by the generation-t consumers

[1]This preference specification is common in the industrial-organization literature (e.g., Singh and Vives, 1984). Consumers' preference takes a quasi-linear utility form, $u(x, w) = \tilde{s}x - \frac{x^2}{2} + w$, where x and w are the quantities of goods produced by the firms and of the numeraire good. Combining with the budget constraint, $w + px = W$ with W being the total wealth, we can see that the function form in equations (1)–(3) is the reduced form with the budget constrain substituted into the consumers' original preference.

in firm A's ith X-market and in firm B's jth X-market, respectively. Variables $p^t_{A,i}$ and $p^t_{B,j}$ represent the product prices in these local markets. Similarly, variable y^t_k represents the demand of a typical generation-t consumer k in the global Y-market, and p^t_y is the product price at the Y-market in period t.

Variables $\tilde{s}^t_{A,i}$, $\tilde{s}^t_{B,j}$, and $\tilde{s}^t_{AB,k}$ capture preference shocks. They contain two random components, a time invariant common component $\tilde{\theta}$ and an idiosyncratic component $\tilde{\varepsilon}$:

$$\tilde{s}^t_{A,i} = \tilde{\theta} + \tilde{\varepsilon}^t_{A,i}, i = 1, 2, ..., M,$$

$$\tilde{s}^t_{B,j} = \tilde{\theta} + \tilde{\varepsilon}^t_{B,j}, j = 1, 2, ..., M,$$

$$\tilde{s}^t_{AB,k} = \tilde{\theta} + \tilde{\varepsilon}^t_{AB,k}, k = 1, 2, ..., N,$$

where $\tilde{\theta} \sim \mathcal{N}\left(\bar{\theta}, \tau_\theta^{-1}\right)$, $\tilde{\varepsilon}^t_{A,i} \sim \mathcal{N}\left(0, \tau_\varepsilon^{-1}\right)$, $\tilde{\varepsilon}^t_{B,j} \sim \mathcal{N}\left(0, \tau_\varepsilon^{-1}\right)$, and $\tilde{\varepsilon}^t_{AB,k} \sim \mathcal{N}\left(0, \tau_\varepsilon^{-1}\right)$ with $\bar{\theta} \geq 0$, $\tau_\theta > 0$, and $\tau_\varepsilon > 0$. We assume that $\{\tilde{\theta}, \left\{\tilde{\varepsilon}^t_{A,i}\right\}_i, \left\{\tilde{\varepsilon}^t_{B,j}\right\}_j, \left\{\tilde{\varepsilon}^t_{AB,k}\right\}_k\}$ are mutually independent. Consumers know their own preference shocks when making purchase decisions. We have assumed that the common component is the same across all consumers. Our mechanism still works as long as there is some correlation among consumers' preference shocks. In addition, in preference specification (3), we assume that the products of both firms are perfect substitutes for AB-type consumers. This assumption is made for the sake of simplicity. The results still go through if the products of both firms are not perfect substitutes in the Y-market.

Each consumer maximizes her preference taking the product prices as given. Solving consumers' utility-maximization problems leads to the following inverse demand functions in the X-markets and Y-market, respectively:

$$p^t_{A,i} = \tilde{s}^t_{A,i} - x^t_{A,i}, i = 1, 2, ..., M; \tag{4}$$

$$p^t_{B,j} = \tilde{s}^t_{B,j} - x^t_{B,j}, j = 1, 2, ..., M; \tag{5}$$

$$p^t_y = \frac{1}{N}\left[\sum\nolimits_{k=1}^{N} \tilde{s}^t_{AB,k} - \sum\nolimits_{k=1}^{N} y^t_k\right]. \tag{6}$$

2.2 Firms and product supply

Firms A and B live for both periods. In each period, firms maximize the expected profits conditional on their information. These profit-maximization decisions lead to the product supply in the product markets. Without loss of generality, we normalize the firms' production cost to zero.

Date-0 product markets When making date-0 production decisions, firms have not received any information yet. Thus, firms choose production quantities to maximize unconditional expected profits taking as given the demand functions from consumers and the production quantities of their rivals. Given the assumption that production is costless, firm A's optimal production quantities $\left(X_{A,1}^0, ..., X_{A,M}^0, Y_A^0\right)$ are determined by

$$\max_{\left\{X_{A,i}^0\right\}_{i=1}^M, Y_A^0} \mathbb{E}\left[\underbrace{\sum\nolimits_{i=1}^M p_{A,i}^0 X_{A,i}^0}_{X\text{-market}} + \underbrace{p_y^0 Y_A^0}_{Y\text{-market}}\right], \tag{7}$$

where $p_{A,i}^0$ and p_y^0 are given respectively by demand functions (4) and (6) with $t = 0$. Firm B's decisions can be characterized similarly by changing notations.

In an X-market, the corresponding firm behaves as a monopolist and thus the two firms make decisions independently. The firms compete in the Y-market so each firm needs to take into account the other firm's production and the market-clearing condition (i.e., $Y_A^0 + Y_B^0 = \sum_{k=1}^N y_k^0$). The equilibrium computation is standard and thus omitted. We summarize the result in the following lemma.

Lemma 1 (Date-0 product market equilibrium)

In the data-0 product markets, the Nash equilibrium prices $\left(\left\{p_{A,i}^{0*}\right\}_{i=1}^M, \left\{p_{B,j}^{0*}\right\}_{j=1}^M, p_y^{0*}\right)$*, production quantities* $\left(\left\{X_{A,i}^{0*}\right\}_{i=1}^M, Y_A^{0*}, \left\{X_{B,j}^{0*}\right\}_{j=1}^M, Y_B^{0*}\right)$*, and expected profits* ($\mathbb{E}\Pi_A^{0*}$ *and* $\mathbb{E}\Pi_B^{0*}$)

are:

$$
\begin{aligned}
p_{A,i}^{0*} &= \tilde{s}_{A,i}^{0} - \frac{\bar{\theta}}{2}, X_{A,i}^{0*} = \frac{\bar{\theta}}{2}, \ \textit{for } i = 1, ..., M; \\
p_{B,j}^{0*} &= \tilde{s}_{B,j}^{0} - \frac{\bar{\theta}}{2}, X_{B,j}^{0*} = \frac{\bar{\theta}}{2}, \ \textit{for } j = 1, ..., M; \\
p_{y}^{0*} &= \frac{1}{N}\sum\nolimits_{k=1}^{N} \tilde{s}_{AB,k}^{0} - \frac{2\bar{\theta}}{3}, Y_{A}^{0*} = \frac{\bar{\theta}}{3}N, Y_{B}^{0*} = \frac{\bar{\theta}}{3}N; \\
\mathbb{E}\Pi_{A}^{0*} &= \mathbb{E}\Pi_{B}^{0*} = \left(\frac{M}{4} + \frac{N}{9}\right)\bar{\theta}^{2}.
\end{aligned}
$$

The equilibrium prices in the date-0 X-markets reveal consumers' preference shocks and hence this price data is useful for firms to make forecast about next period demand. Since the production quantities are known constants, prices and sales (i.e., prices multiplied by quantities) convey the same information. When using the wording "consumer data," we refer to the date-0 price data, $\left\{p_{A,i}^{0*}\right\}_{i=1}^{M}$ and $\left\{p_{B,j}^{0*}\right\}_{j=1}^{M}$. To ease expressions, we label these price vectors as follows: $\mathbf{P}_{A}^{0} \equiv \left\{p_{A,i}^{0*}\right\}_{i=1}^{M}$ and $\mathbf{P}_{B}^{0} \equiv \left\{p_{B,j}^{0*}\right\}_{j=1}^{M}$.

Date-1 product markets After the date-0 product markets clear, the price data is formed. Both firms observe the equilibrium Y-market price, p_{y}^{0*}. Firm A privately observes all of its X-market prices $\mathbf{P}_{A}^{0}$. Similarly, firm B privately observes its own X-market prices $\mathbf{P}_{B}^{0}$. This forms the basis of the firms' starting information structure in period 1. As we mention before, all the price data is also available to a data vendor, who in turn sells the data to firms. We will discuss the data market in the next subsection in detail. The general idea is that firm A buys price data about firm B's date-0 X-market, and vice versa.

Let $\mathcal{F}_{A} \equiv \{p_{y}^{0*}, \mathbf{I}_{A}, \mathbf{P}_{A}^{0}\}$ denote firm A's information set, where $\mathbf{I}_{A}$ indicates the vector of price data purchased by firm A. Firm A's date-1 production quantities $(\left\{X_{A,i}^{1}\right\}_{i}, Y_{A}^{1})$ are determined by

$$
\max_{\left\{X_{A,i}^{1}\right\}_{i=1}^{M}, Y_{A}^{1}} \mathbb{E}\left[\underbrace{\sum\nolimits_{i=1}^{M} p_{A,i}^{1} X_{A,i}^{1}}_{X\text{-market}} + \underbrace{p_{y}^{1} Y_{A}^{1}}_{Y\text{-market}} \ |\mathcal{F}_{A}\right], \tag{8}
$$

where the prices $p_{A,i}^{1}$ and p_{y}^{1} are given by inverse demand functions (4) and (6) with $t = 1$, respectively. We can write down a similar profit-maximization problem for Firm B.

Similar to date 0, firms behave as monopolists in their respective X-markets and make

production decisions separately. Now their optimal productions are no longer constant, but instead depend on their information sets. For instance, the optimal production policies for firm A in the ith X-markets is $X^{1*}_{A,i} = X^1_{A,i}(\mathcal{F}_A)$. In the Y-market, we need to consider the strategic interactions between the two firms, and their optimal production decisions form a Bayesian Nash equilibrium. We delegate the derivation of the date-1 product market equilibrium to Section 4, and now turn to the description of the data market which determines firms' information sets $\mathcal{F}_A$ and $\mathcal{F}_B$.

2.3 Data vendor and data market

In the data market, a data vendor sells the collected date-0 consumer data to firms who in turn use the purchased data to improve their date-1 production decisions. In the baseline model described by this section, we follow the literature on information sales in financial markets (e.g., Admati and Pfleiderer (1986)) and assume that the data vendor maximizes its own profits and behaves as a monopolist in the data market.[2] For instance, the data vendor can represent an independent transaction/settlement platform, such as Amazon.com, eBay Inc., or PayPal Holdings, Inc., who has access to various consumer data and can potentially sell data to companies. Alternatively, in our setting, the data vendor cam emerge as an equilibrium outcome, where the date-0 consumers own the data and form a data firm to monetize the value of their data (see Section 5 for more discussions).

Data transactions are completed at the beginning of date 0. We delegate the details of the transaction games to Appendix B, and the general idea is that firm A pays cost C_A to buy m_A amount of data and firm B pays cost C_B to buy m_B amount of data. As we discussed before, the consumer data is in the form of X-market prices and thus, the amount of consumer data refers to the number of X-market prices. We follow the literature (e.g., Gal-Or, 1985; Li, McKelvey, and Page, 1987; Vives, 1988; and Hwang, 1993) and assume that after firms make their data purchase decisions, the purchase amount (m_A, m_B) becomes

[2] In Section 5 where we discuss data ownership and vendor formation, we also consider a variation in which the data vendor maximizes the total profits of both firms.

common knowledge and is observable to both firms before they make their date-1 production decisions (of course, the specific values of the m prices are only observable to the firm who has purchased the data). In the terminology of Hauk and Hurkens (2001), firms do not engage in "secret information acquisition."[3]

Each firm only wants to buy the X-market prices of its rival. These prices are originally the private information of each firm who collects this information from its own local X-market transactions. Thus, with data purchase, firms effectively observe part or all of their rivals' private data. We label this resulting data exchange outcome as a "data allocation."

Definition 1 (Data allocation)

A data allocation, denoted by (m_A, m_B) with $m_A \in \{0, 1, ..., M\}$ and $m_B \in \{0, 1, ..., M\}$, refers to a situation in which, when making their date-1 production decisions, firm A observes m_A date-0 X-market prices $p^{0*}_{B,j}$ of firm B, and firm B observes m_B date-0 X-market prices $p^{0*}_{A,i}$ of firm A.

We follow Admati and Pfleiderer (1986) and assume that the data vendor can implement any data allocation through information sales. This is natural given that the data vendor is a monopolist in the data market. In Appendix B, we describe how the data vendor achieves this implementation by offering right contracts. Given data allocation (m_A, m_B), we use $\mathbb{E}\Pi^1_A(m_A, m_B)$ to denote firm A's expected profit resulting from the date-1 product market equilibrium. Specifically, we insert the optimal production policies into the objective function (8) and take unconditional expectations to compute $\mathbb{E}\Pi^1_A(m_A, m_B)$. If firm A does not buy any data from the data vendor, then its expected profit is $\mathbb{E}\Pi^1_A(0, m_B)$. Thus, firm A's willingness to pay for an amount m_A of data given that its rival has purchased an amount m_B of data is

$$C_A(m_A, m_B) = \mathbb{E}\Pi^1_A(m_A, m_B) - \mathbb{E}\Pi^1_A(0, m_B). \tag{9}$$

[3] If firms engage in "secret" information purchase (i.e., (m_A, m_B) is not observable to firms when making production decisions), then a firm will take its rival's production policies as given when considering the information value through a deviation analysis. We have shown that our results are robust under this alternative assumption.

Since the monopolist extracts all surplus, $C_A(m_A, m_B)$ constitutes its profit from selling data to firm A. We can define firm B's willingness to pay similarly and label it as $C_B(m_A, m_B)$.

A profit-maximizing data vendor's problem is to choose a data allocation to maximize its own profits as follows:

$$\max_{(m_A,m_B)\in\{0,1,\ldots,M\}^2} \left[C_A(m_A, m_B) + C_B(m_A, m_B)\right]. \tag{10}$$

Equations (9) and (10) share the same spirit as Admati and Pfleiderer (1986) who study how a monopolistic data vendor sells information in financial markets with different precision levels. In their model, the data price that the seller can charge is computed as the difference between the certainty equivalent of a trader who is equipped with the information and the certainty equivalent of a trader who is uninformed (their equation (3.1)). This corresponds to our equation (9). Similar to our equation (10), the data vendor in Admati and Pfleiderer (1986) extracts all surplus by choosing a distribution of information precision levels (see their equation (3.2)).

2.4 Equilibrium concept

Lemma 1 has already characterized the date-0 product market equilibrium. Thus, our equilibrium definition focuses only on the data vendor's profit-optimization problem on date 0 and the product-market equilibrium on date 1. We look for a Perfect Bayesian Equilibrium (PBE).

Definition 2 (Equilibrium)

A PBE consists of a date-0 data allocation (m_A^*, m_B^*) *and date-1 production policies,* $(\{X_{A,i}^1(\mathcal{F}_A)\}_{i=1}^M, Y_A^1(\mathcal{F}_A))$ *and* $(\{X_{B,j}^1(\mathcal{F}_B)\}_{j=1}^M, Y_B^1(\mathcal{F}_B))$*, such that:*

1. *Given the equilibrium amount* (m_A^*, m_B^*) *of purchased data, the date-1 policies* $X_{A,i}^1(\mathcal{F}_A)$ *and* $X_{B,j}^1(\mathcal{F}_B)$ *maximize the conditional profits in firm A's* i*th* X*-market and in firm B's* j*th* X*-market, respectively; and* $(Y_A^1(\mathcal{F}_A), Y_B^1(\mathcal{F}_B))$ *form a Bayesian Nash equilibrium in the* Y*-market.*

2. *The equilibrium amount* (m_A^*, m_B^*) *of sold data is determined by (10), and the data prices* (C_A^*, C_B^*) *are set accordingly as* $C_A^* = C_A(m_A^*, m_B^*)$ *and* $C_B^* = C_B(m_A^*, m_B^*)$.

We solve the equilibrium by backward induction. That is, we compute the date-1 product market equilibrium for any given (m_A, m_B). This allows us to determine the expression of firms' profits, $\mathbb{E}\Pi_A^1(m_A, m_B)$ and $\mathbb{E}\Pi_B^1(m_A, m_B)$. We then solve the data vendor's profit-maximization problem (10), which leads to the equilibrium data allocation (m_A^*, m_B^*). To set the stage for our analysis, in the next section we first examine a benchmark economy without a data vendor.

3 What happens without a data vendor?

In this section, we first analyze a benchmark economy without a data vendor. We then provide some background discussion on how the literature has strived to improve on the equilibrium outcome in the benchmark economy, namely, by considering information sharing among firms. However, free information sharing is not viable or costly in the case of Cournot competition and demand uncertainty. By contrast, our paper shows that information sales—both in the benchmark model of Section 2 and in the variant model of Section 5—can achieve the desired welfare improvement.

3.1 Product market equilibrium in the benchmark economy

Without a data vendor, the date-1 Y-market in our economy degenerates to the classical duopoly setting with privately informed Cournot firms (e.g., Gal-Or, 1985; Darrough, 1993). The information structure of firms is endogenously determined by the date-0 product market equilibrium. Specifically, the date-0 Y-market price p_y^{0*} serves as the public information shared by both firms. Firm A's date-0 X-market prices $\mathbf{P}_A^0 \equiv \left\{p_{A,i}^{0*}\right\}_{i=1}^M$ are firm A's private information, while firm B's date-0 X-market prices $\mathbf{P}_B^0 \equiv \left\{p_{B,j}^{0*}\right\}_{j=1}^M$ are firm B's private

information. By Lemma 1, the date-0 market prices reveal preference shocks of date-0 consumers. Formally, firm A's information set and firm B's information set are respectively:

$$\mathcal{F}_A = \{p_y^{0*}, \mathbf{P}_A^0\} = \left\{ \frac{\sum_{k=1}^N \tilde{s}_{AB,k}^0}{N} - \frac{2\bar{\theta}}{3}, \tilde{s}_{A,1}^0, ..., \tilde{s}_{A,M}^0 \right\},$$

$$\mathcal{F}_B = \{p_y^{0*}, \mathbf{P}_B^0\} = \left\{ \frac{\sum_{k=1}^N \tilde{s}_{AB,k}^0}{N} - \frac{2\bar{\theta}}{3}, \tilde{s}_{B,1}^0, ..., \tilde{s}_{B,M}^0 \right\}.$$

As standard in the literature (e.g., Gal-Or, 1985 and Darrough, 1993), we consider date-1 production policies that are linear in firms' information variables. Given that the date-0 X-market prices have the same precision level in predicting the persistent component $\tilde{\theta}$ in the future demand, it is intuitive to specify that the coefficients on these prices are the same. We therefore conjecture the following date-1 production policies for firms A and B:

$$X_{A,i}^1 = \Phi_{A_0}^X + \Phi_{A_1}^X (P_A^0 - \mu), i = 1, ..., M, \tag{11}$$

$$X_{B,j}^1 = \Phi_{B_0}^X + \Phi_{B_1}^X (P_B^0 - \mu), j = 1, ..., M, \tag{12}$$

$$Y_A^1 = \Phi_{A_0}^Y + \Phi_{A_1}^Y (P_A^0 - \mu), \tag{13}$$

$$Y_B^1 = \Phi_{B_0}^Y + \Phi_{B_1}^Y (P_B^0 - \mu), \tag{14}$$

where

$$P_A^0 \equiv \frac{1}{M} \sum_{i=1}^M p_{A,i}^{0*} = \frac{1}{M} \sum_{i=1}^M \tilde{s}_{A,i}^0, \tag{15}$$

$$P_B^0 \equiv \frac{1}{M} \sum_{j=1}^M p_{B,j}^{0*} = \frac{1}{M} \sum_{j=1}^M \tilde{s}_{B,j}^0, \tag{16}$$

are two price indices of date-0 X-market data, and

$$\mu \equiv \mathbb{E}(\tilde{\theta}|p_y^{0*}) = \bar{\theta} + \frac{N\tau_\varepsilon}{N\tau_\varepsilon + \tau_\theta} \left(p_y^{0*} - \frac{\bar{\theta}}{3} \right) \tag{17}$$

is the posterior about $\tilde{\theta}$ given the public information, the Y-market price p_y^{0*}.

Equation (11) maximizes firm A's conditional expected profit in each of its local X-markets in period 1. Note that since all the X-markets are symmetric, the optimal production policies are the same across all M local markets. Similarly, equation (12) maximizes firm B's expected profit in its date-1 X-markets. Equations (13) and (14) form a linear Bayesian Nash equilibrium in the global Y-market in period 1. The following lemma characterizes the linear date-1 product market equilibrium without a data vendor.

Lemma 2 (Date-1 product market equilibrium without data sales)

In the economy without a data vendor, there exists a unique linear PBE in which

$$X_{A,i}^{1*} = \frac{1}{2}\left[\mu + \frac{M\tau_\varepsilon}{M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta}(P_A^0 - \mu)\right], i = 1, ..., M,$$
$$X_{B,j}^{1*} = \frac{1}{2}\left[\mu + \frac{M\tau_\varepsilon}{M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta}(P_B^0 - \mu)\right], j = 1, ..., M,$$
$$Y_A^{1*} = \frac{N}{3}\mu + \frac{MN\tau_\varepsilon}{3M\tau_\varepsilon + 2(N\tau_\varepsilon + \tau_\theta)}(P_A^0 - \mu),$$
$$Y_B^{1*} = \frac{N}{3}\mu + \frac{MN\tau_\varepsilon}{3M\tau_\varepsilon + 2(N\tau_\varepsilon + \tau_\theta)}(P_B^0 - \mu).$$

The equilibrium expected profits of firm A and firm B in period 1 are

$$\mathbb{E}\Pi_A^{1*} = \mathbb{E}\Pi_B^{1*} = \frac{M}{4}\left[\bar{\theta}^2 + \frac{N\tau_\varepsilon}{(N\tau_\varepsilon + \tau_\theta)\tau_\theta} + \frac{M\tau_\varepsilon}{(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)(N\tau_\varepsilon + \tau_\theta)}\right]$$
$$+ \frac{N}{9}\left[\bar{\theta}^2 + \frac{N\tau_\varepsilon}{(N\tau_\varepsilon + \tau_\theta)\tau_\theta}\right] + \frac{MN\tau_\varepsilon(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)}{\left[3M\tau_\varepsilon + 2\left(N\tau_\varepsilon + \tau_\theta\right)\right]^2\left(N\tau_\varepsilon + \tau_\theta\right)}.$$

3.2 Data allocation, efficiency, and disclosure

3.2.1 Data allocation and welfare improvement

The equilibrium outcome characterized by Lemma 2 can be potentially improved in terms of social welfare by changing data allocations. Specifically, we consider an artificial situation in which both firms' private information $\mathbf{P}_A^0$ and $\mathbf{P}_B^0$, in addition to the public information p_y^{0*}, are commonly observed by the two firms. This corresponds to data allocation $(m_A, m_B) = (M, M)$ defined in Definition 1, and we label it with "**MM**". Under data allocation **MM**, both firms have the same information set, which is $\mathcal{F}_A^{\mathbf{MM}} = \mathcal{F}_B^{\mathbf{MM}} = \{p_y^{0*}, \mathbf{P}_A^0, \mathbf{P}_B^0\}$. Equipped with this new information set, firms still maximize their conditional expected profits in both local and global markets. The original data allocation in Lemma 2 is $(m_A, m_B) = (0, 0)$, and we label it with "$\emptyset\emptyset$" to indicate that both firms keep their private information secret.

We define the welfare variables—consumer surplus (CS) and total surplus (TS)—as fol-

lows:

$$CS \equiv \underbrace{\sum_{i=1}^{M} \frac{1}{2}\mathbb{E}(X_{A,i}^1)^2}_{A\text{-type}} + \underbrace{\sum_{i=1}^{M} \frac{1}{2}\mathbb{E}(X_{B,i}^1)^2}_{B\text{-type}} + \underbrace{\frac{1}{2}\mathbb{E}(Y_A^1 + Y_B^1)^2}_{AB\text{-type}}, \tag{18}$$

$$TS \equiv \underbrace{CS}_{\text{consumer surplus}} + \underbrace{\mathbb{E}\Pi_A^1 + \mathbb{E}\Pi_B^1}_{\text{producer surplus}}, \tag{19}$$

where $X_{A,i}^1, X_{B,i}^1, Y_A^1, Y_B^1, \mathbb{E}\Pi_A^1$, and $\mathbb{E}\Pi_B^1$ are the production policies and profits reached in the date-1 product market equilibrium when firms are equipped with their information sets.

Proposition 1 (Welfare gains)

Relative to data allocation $\emptyset\emptyset$, under data allocation **MM***, consumer surplus and total surplus are always higher, and firm profits are higher if and only if there are sufficiently many local markets. That is, $CS^{\mathbf{MM}} > CS^{\emptyset\emptyset}, TS^{\mathbf{MM}} > TS^{\emptyset\emptyset}$ for any M; and $\mathbb{E}\Pi_A^{1,\mathbf{MM}} = \mathbb{E}\Pi_B^{1,\mathbf{MM}} > \mathbb{E}\Pi_A^{1,\emptyset\emptyset} = \mathbb{E}\Pi_B^{1,\emptyset\emptyset}$ if and only if $M > \hat{M}$, where $\hat{M}$ is the unique solution of the equation below:*

$$\frac{9\hat{M}}{\hat{M}\tau_\epsilon + N\tau_\epsilon + \tau_\theta}\left[3\hat{M}\tau_\epsilon + (N\tau_\epsilon + \tau_\theta) + 2(N\tau_\epsilon + \tau_\theta) + \frac{(N\tau_\epsilon + \tau_\theta)^2}{3\hat{M}\tau_\epsilon + N\tau_\epsilon + \tau_\theta}\right] - 4N = 0. \tag{20}$$

Intuitively, when firms are equipped with better information, they collectively accommodate better consumers' needs, which improves consumer surplus and total surplus. However, because of the strategic competing behavior, firms are worse off in the Y-market when their private information becomes public. This profit loss can be compensated by their more informed production decisions in their respective local X-markets, and if the number of these local markets is sufficiently large, the overall profit effect of sharing information is positive.

3.2.2 Voluntary and mandatory disclosure

Although data allocation **MM** improves on data allocation $\emptyset\emptyset$, it is not clear how such a data allocation is achieved in the first place. The information-sharing literature has considered whether firms would like to voluntarily share their private information, for instance, by forming a trade association that discloses the signals reported by its member firms (Gal-Or, 1985; Darrough, 1993; and see Vives (2016) for a survey). However, it is shown that withholding information is always a dominant strategy for firms in oligopoly settings with

Figure 2: Efficiency and disclosure

Panel A: Welfare variables

Data allocation	Small M, $M = N/10$			Large M, $M = 10N$		
	TS	CS	$\mathbb{E}\Pi_A^1 + \mathbb{E}\Pi_B^1$	TS	CS	$\mathbb{E}\Pi_A^1 + \mathbb{E}\Pi_B^1$
$\emptyset\emptyset$	227.894	224.972	2.922	132474.67	132196.63	278.04
MM	240.379	237.463	2.916	171190.61	170836.85	353.76

Panel B: The payoff matrix for firms

$M = N/10$		Firm B	
		ND	D
Firm A	ND	$(1.461, 1.461)$	$(1.570, 1.349)$
	D	$(1.349, 1.570)$	$(1.458, 1.458)$

$M = 10N$		Firm B	
		ND	D
Firm A	ND	$(139.02, 139.02)$	$(179.02, 136.77)$
	D	$(136.77, 179.02)$	$(176.88, 176.88)$

Panel A shows the total surplus, consumer surplus, and total profits for the corresponding data allocation. Panel B is the payoff matrix for firms for the corresponding action non-disclosure, "ND", or disclosure, "D." In this numerical example, we assume $\bar{\theta} = 0, \tau_\theta = 1, \tau_\varepsilon = 0.001$ and $N = 100$. We consider two values of M: $M = N/10 = 10$ and $M = 10N = 1000$.

Cournot competition and demand uncertainty; that is, data allocation **MM** is not supported in equilibrium with voluntary disclosure.

To illustrate, let us consider the following numerical example. We set $\bar{\theta} = 0, \tau_\theta = 1, \tau_\varepsilon = 0.001$, and $N = 100$, and M can take two values: $M = \frac{N}{10} = 10$ or $M = 10N = 1000$. Consistent with Proposition 1, independent of the value of M, both consumer surplus and total surplus are higher under data allocation **MM** in Panel A of Figure 2. Also, when M is high, firms' profits are higher under **MM**, and when M is low, firms' profits are lower under **MM**.

In the context of voluntary information sharing, each firm faces a choice of disclosure (D) or nondisclosure (ND) of its own private information. This leads to the payoff matrices in Panel B of Figure 2. Each cell in this matrix is the equilibrium profits resulting from the date-1 product market equilibrium. For instance, if both firms choose not to disclose information, then the profits of each firm are given by the expression of $\mathbb{E}\Pi_A^{1*}$ and $\mathbb{E}\Pi_B^{1*}$ in Lemma 2. We can see that withholding information is a dominant strategy for each firm,

so that (ND, ND) constitutes the unique Nash equilibrium at the information-sharing stage for both values of M. In particular, when M is high, the resulting payoff matrix is the "prisoners' dilemma,"[4] which predicts that firms would have been better off if both of them could disclose, which, however, is not a viable agreement in a noncooperative setting.

Given that voluntary disclosure is not viable, the literature also suggests mandatory disclosure through regulatory agencies such as the SEC or the FASB that, in theory, can force firms to disclose the information that firms wish hidden (e.g., see Darrough, 1993). However, mandatory disclosure can be costly. The cost stems not only from the administrative cost of implementing the disclosure rules but also from some other economic costs. Firms could take strategic actions to respond to regulatory requirements, for instance, by adding noises or a large amounts of nonmaterial and raw information of little value in the public disclosure.[5] The root reason for this kind of cost is that mandatory disclosure regulations run against firms' private incentives to maximize their own profits. In the following two sections, we will show that data sales instead can incentivize firms to reach the more efficient data allocation **MM**.

4 Welfare-improving data sales

We now solve the model with data sales described in Section 2, in which the data vendor maximizes its own profits. We first solve the date-1 product market equilibrium for any

[4]Darrough (1993) also identifies a prisoners' dilemma in an information-sharing setting, although for a different reason. Specifically, in Darrough's setting, a prisoners' dilemma arises when firms' products are sufficiently different. By contrast, the firms' products are perfect substitute in our setting, and the prevalence of a prisoners' dilemma depends on the number of local markets.

[5]Evidence supporting this argument is provided by extensive studies on Regulation Fair Disclosure (Reg FD) which, promulgated by the SEC in 2000, mandates that all publicly traded companies must disclose material information to the general public at the same time. For instance, Bailey, Li, Mao, and Zhong (2003) find that the Reg FD could make the public communication become "sound bites" with "boilerplate" disclosures. A survey conducted by Security Industry Association shows that 72% of analysts interviewed during the survey mention that information communicated by issuers to the public is of lower quality after the Reg FD regulation (http://www.sia.com/testimony/html/kaswell5 -17.html). Cohen, Lou, and Malloy (2017) document that firms could "cast" their conference calls and thus control the information flow released to the public even after Reg FD. Bushee, Matsumoto, and Miller (2004) find that Reg FD had a significant negative impact on managers' decisions to continue hosting conference calls and on their decisions regarding the optimal time to hold.

given amount of data purchase, (m_A, m_B), and then solve the optimal data sales (m_A^*, m_B^*). Finally, we discuss the welfare consequences of data sales.

4.1 Product-market equilibrium

Suppose that on date 0, firms A and B have respectively purchased m_A and m_B local market prices from the data vendor. Note that the consumer data purchased by firm A is about firm B's date-0 X-markets, and vice versa. We assume that the consumer identities of the sold data are anonymous. The data vendor can achieve this goal by randomly sampling from the pool of all date-0 consumers. Nonetheless, we assume that the data vendor ensures that the data is indeed useful for firms (i.e., the data bought by firm A is drawn from firm B's X-market prices and vice versa). Let us label the randomly drawn consumers by $\{j_1, ..., j_{m_A}\}$ and $\{i_1, ..., i_{m_B}\}$ for the two sold data sets. The data sets purchased by firms A and B are, respectively,

$$\mathbf{I}_A = \{p^{0*}_{B,j_1}, ..., p^{0*}_{B,j_{m_A}}\} \text{ and } \mathbf{I}_B = \{p^{0*}_{A,i_1}, ..., p^{0*}_{A,i_{m_B}}\}.$$

As in Section 3, we still consider linear equilibria in which optimal production policies are linear in firms' information variables. Also, given the symmetry of the purchased market data, it is natural to specify that the coefficients on the purchased prices are the same. Thus, we conjecture the following date-1 production policies:

$$X^1_{A,i} = \Phi^X_{A_0} + \Phi^X_{A_1}(P^0_A - \mu) + \Phi^X_{A_2}(I_A - \mu), i = 1, ..., M, \tag{21}$$

$$X^1_{B,j} = \Phi^X_{B_0} + \Phi^X_{B_1}(P^0_B - \mu) + \Phi^X_{B_2}(I_B - \mu), j = 1, ..., M, \tag{22}$$

$$Y^1_A = \Phi^Y_{A_0} + \Phi^Y_{A_1}(P^0_A - \mu) + \Phi^Y_{A_2}(I_A - \mu), \tag{23}$$

$$Y^1_B = \Phi^Y_{B_0} + \Phi^Y_{B_1}(P^0_B - \mu) + \Phi^Y_{B_2}(I_B - \mu), \tag{24}$$

where

$$I_A \equiv \frac{1}{m_A}\sum\nolimits_{a=1}^{m_A} p^{0*}_{B,j_a} = \frac{1}{m_A}\sum\nolimits_{a=1}^{m_A} \tilde{s}^0_{B,j_a}, \tag{25}$$

$$I_B \equiv \frac{1}{m_B}\sum\nolimits_{b=1}^{m_B} p^{0*}_{A,i_b} = \frac{1}{m_B}\sum\nolimits_{b=1}^{m_B} \tilde{s}^0_{A,i_b}, \tag{26}$$

where the second equality in (25) and (26) follows from Lemma 1, and P^0_A, P^0_B, and μ are

given by equations (15), (16), and (17), respectively.

Equations (21) and (22) maximize expected profits in the local X-markets respectively for firm A and firm B. Equations (23) and (24) form a linear Bayesian Nash equilibrium in the global Y-market. The following proposition characterizes the product-market equilibrium

Proposition 2 (Product-market equilibrium)

For any given data purchase (m_A, m_B), there exists a linear product-market equilibrium characterized by equations (21)–(24), where the Φ-coefficients (equation A32) and the equilibrium expected profits $\mathbb{E}\Pi_A^{1*}$ and $\mathbb{E}\Pi_B^{1*}$ (equation A34, A35) are given in the appendix.

4.2 Equilibrium data sales

At the beginning of date 0, the data vendor designs contracts to maximize firms' willingness to pay for data, $C_A(m_A, m_B)$ and $C_B(m_A, m_B)$, given by equation (9). It turns out that the data vendor's profit is maximized when both firms purchase the maximum amount of data. These results are formalized in the following proposition.

Proposition 3 (Optimal data sales)

In equilibrium, the data vendor sells all of its data to firms, that is, $m_A^* = m_B^* = M$. The resulting data prices are

$$C_A^* = C_B^* = \left(\frac{M}{4} + \frac{N}{9}\right) \frac{M\tau_\varepsilon}{(2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)}.$$

Setting $m_A = m_B = M$ in Proposition 2, we obtain the overall sequential equilibrium for the economy with a profit-maximizing data vendor.

Proposition 4 (Overall equilibrium)

On date 0, the data vendor sells all of its data to firms. On date 1, the optimal production policies in product markets are:

$$Y_A^{1*} = Y_B^{1*} = \frac{N}{3}\left[\mu + \frac{M\tau_\varepsilon}{2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta}\left(P_A^0 - \mu + P_B^0 - \mu\right)\right],$$

$$X_{A,i}^{1*} = X_{B,j}^{1*} = \frac{1}{2}\left[\mu + \frac{M\tau_\varepsilon}{2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta}\left(P_A^0 - \mu + P_B^0 - \mu\right)\right],$$

for $i, j = 1, ..., M$. The equilibrium date-1 expected profits (gross of data price C^*) are

$$\mathbb{E}\Pi_A^{1*} = \mathbb{E}\Pi_B^{1*} = \frac{M}{4}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)(N\tau_\varepsilon + \tau_\theta)}\right] + \frac{N}{9}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)(N\tau_\varepsilon + \tau_\theta)}\right].$$

4.3 Intuitions, implementation, and welfare

We now use a numerical example in Figure 3 to illustrate better what is going on in the economy with data sales. The parameter values in Figure 3 are the same as those in Figure 2 with $M = 1000$. In Panel A of Figure 3, we plot the payoff matrix for firms. By Proposition 3, in equilibrium, the data vendor will implement data allocation **MM** and charge a price C^* for the data. Thus, firms' actions are either to reject the data vendor's contracts and not purchase data, or to accept the contracts and acquire an amount M of consumer data. Here, we allow the data vendor to be able to design "smart contracts" which allow data prices to depend on data allocations (See Appendix B for more discussions on contract implementations). Specifically, the contracts state the following: "*If data allocation is $(0, M)$, then firm A pays price t_{0M}; if data allocation is $(M, 0)$, then firm B pays price t_{0M}; and if data allocation is (M, M), then both firms pay a price t_{MM}.*" By Proposition 3, we know that the equilibrium value of t_{MM} must equal C^*, which is 41.11 in this example. We now explain why this is the case and what values t_{0M} can take.

The first observation is the following. Suppose that the data prices t_{0M} and t_{MM} are set at 0. Then, comparing the payoff matrix in Panel A of Figure 3 with that in Panel B of Figure 2, we find that the former is a transpose of the latter. Note that in Figure 2, firms' actions are disclosing or not disclosing information and thus, there, firms are considering whether to *supply* information for free. In contrast, in Figure 3, firms are considering whether to buy information at a cost, which is about the *demand* side of data. This switch between supply and demand perspectives transposes the payoff matrix, which in turn changes the equilibrium data allocations.

Recall that in equilibrium, the data vendor wants to implement data allocation (M, M),

Figure 3: Data sales and efficiency

Panel A: The payoff matrix for firms

		Firm B	
		0	M
Firm A	0	$(139.02, 139.02)$	$(136.77, 179.02 - t_{0M})$
	M	$(179.02 - t_{0M}, 136.77)$	$(176.88 - t_{MM}, 176.88 - t_{MM})$

Panel B: Welfare variables

Data allocation	TS	CS	$\mathbb{E}\Pi_A^1 + \mathbb{E}\Pi_B^1$	Data vendor's profits
$\emptyset\emptyset$	132474.67	132196.63	278.04	0
MM	171190.61	170836.85	273.54	80.22

Panel A is the payoff matrix for firms for the corresponding actions — purchasing "0" local market price or "M" local market prices. The payoff is the expected profit net cost of buying information from the data vendor. t_{0M} is the cost of data when one firm buys 0 and the other buys M, and t_{MM} is the cost when both firms purchase M local market prices. Panel B shows the total surplus, consumer surplus, total profits for firms, and total profits for the data vendor. The total surplus includes the data vendor's profits. In this numerical example, we assume $\bar{\theta} = 0, \tau_\theta = 1, \tau_\varepsilon = 0.001$ and $N = 100$. Since we focus on the "prison dilemma" problem, we consider only when M is large, i.e., $M = 10N$.

which leads to the highest profits. One way of implementation is to choose appropriate values of t_{0M} and t_{MM}, such that purchasing data is a dominant strategy for both firms. This requires the following:

$$179.02 - t_{0M} > 139.02 \Rightarrow t_{0M} < 40,$$

$$176.88 - t_{MM} > 136.77 \Rightarrow t_{MM} < 40.11.$$

Thus, by setting $t_{MM} = 40.11$ and $t_{0M} < 40$, the data vendor can sell data to both firms, collecting a total profit of $2 \times t_{MM} = 80.22$. Intuitively, the upper bounds of t_{0M} and t_{MM} are firms' willingness to pay at data allocations $(0, M)$ and (M, M), respectively. In our setting, there is strategic complementarity in firms' data purchase behavior: Firm A's willingness to pay is higher when firm B is buying data than when firm B is not (i.e., $C_A(M, M) = C_B(M, M) = 40.11 > 40 = C_A(M, 0) = C_B(M, 0)$). In consequence, the data vendor can achieve the highest profit when both firms buy data, because in this case, not only the data vendor is selling to two instead of one firm, but also each firm is willing to pay

more, relative to the case in which only one firm buys data. This complementarity result holds true in general as formalized in the following proposition.

Proposition 5 (Complementarity)

In equilibrium, firms' information purchase decisions are a strategic complement, that is, $\frac{\partial C_A(M,m_B)}{\partial m_B} > 0$ *and* $\frac{\partial C_B(m_A,M)}{\partial m_A} > 0$.

The presence of a data vendor effectively moves the equilibrium data allocation from $\emptyset\emptyset$ to **MM**. That is, in the benchmark economy without data sales, both firms keep their private information secret and thus no firm can see its rival's private information (which corresponds to data allocation $\emptyset\emptyset$). Here, with data purchase, both firms can observe the private information of their respective rivals, although at a cost. This leads to the data allocation **MM**. By Proposition 1, both consumer surplus CS and total surplus TS are improved with the introduction of a data vendor, where CS and TS are still defined by equations (18) and (19), respectively. Panel B of Figure 3 confirms these results.

Proposition 6 (Welfare-improving data sales)

The introduction of an independent for-profit data vendor improves both consumer surplus and total surplus in equilibrium.

4.4 When are the firms better off?

In Panel B of Figure 3, firms' equilibrium profits are lower with data sales than without. Thus, although introducing an independent data vendor improves total surplus, it is not a Pareto improvement. Recall that by Proposition 1, when the number M of local markets is sufficiently high, both firms can be better off under data allocation **MM** than under data allocation $\emptyset\emptyset$, if they had not paid any costs to acquire the data. The underlying reason that firms get worse off in Panel B of Figure 3 (where M is relatively high) is due to the assumption that the data vendor has all the market power in the data market. If we relax this assumption, then firms can be better off as well with the introduction of a data vendor

for a sufficiently high M, so that the introduction of a data vendor indeed leads to a Pareto improvement.

Specifically, let us consider a setting in which firms could bargain with the data vendor in the data market. This may be reasonable given that both the data vendor and the two firms are big players in the data market. Now suppose that a firm can negotiate over the data price when receiving sales contracts from the data vendor, and the data price C is set through Nash bargaining between the data vendor and the firm. We use $\beta \in (0,1)$ to denote the data vendor's bargaining power. Our baseline model corresponds to the degenerate case with $\beta = 1$.

The bargaining outcome depends on each agent's utility in the events of agreement versus no agreement. For a firm, say, firm A, the utility when agreeing on a data price C_A is $\mathbb{E}\Pi_A^1(m_A, m_B) - C_A$. If no agreement is reached, then firm A's outside option value is $\mathbb{E}\Pi_A^1(0, m_B)$. The data vendor's gain from agreement is the data price C_A. The bargaining outcome maximizes the Cobb-Douglas product of the utility gains from agreement:

$$\max_{C_A} C_A^{\beta} \left(\mathbb{E}\Pi_A^1(m_A, m_B) - C_A - \mathbb{E}\Pi_A^1(0, m_B)\right)^{1-\beta}.$$

The solution leads to the data price as follows:

$$C_A\left(m_A, m_B\right) = \beta \times \left[\mathbb{E}\Pi_A^1(m_A, m_B) - \mathbb{E}\Pi_A^1(0, m_B)\right]. \tag{27}$$

Comparing the above expression with equation (9), we find that the only difference in this generalized setting is the scaling fraction β. This fraction does not affect the data vendor's profit-maximization problem and thus, when the data vendor designs contracts to implement data allocations, it still chooses $m_A^* = m_B^* = M$. The resulting overall equilibrium is given by Proposition 4 with a smaller data price:

$$C_A^* = C_B^* = \beta \left(\frac{M}{4} + \frac{N}{9}\right) \frac{M\tau_\varepsilon}{(2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)}.$$

As a result, the net profits of firms become higher. In particular, when the data vendor has small bargaining power and when there are sufficiently many local markets, firms are better off with data sales, so that the introduction of a data vendor leads to a Pareto improvement.

Proposition 7 (Nash bargaining)

The introduction of a data vendor benefits firms and a Pareto improvement if and only if the following two conditions are simultaneously satisfied:

$$M > \hat{M} \text{ and } \beta < 1 - \left(\frac{M}{4} + \frac{N}{9}\right)^{-1} \frac{N\left[6M\tau_\varepsilon + 5(N\tau_\varepsilon + \tau_\theta)\right](2M\tau_\epsilon + N\tau_\varepsilon + \tau_\theta)}{9\left[3M\tau_\varepsilon + 2(N\tau_\varepsilon + \tau_\theta)\right]^2},$$

where $\hat{M}$ is a constant defined in equation (20).

5 Data ownership and vendor formation

In previous sections, we assume that the data vendor maximizes its own profits and we do not explore how such a profit-maximizing vendor arises. The data vendor's emergence and its resulting objective function may depend on the original data ownership. In this section, we first discuss how an independent profit-maximizing data vendor can endogenously arise in cases in which the data is originally owned by platforms (such as Amazon.com, Inc.) or by consumers. We then examine a variation setting in which the data is owned by firms and show how firms can form a data vendor to maximize their total profits. Our analysis provides useful insights for the current debates on data ownership and privacy.

5.1 Platforms, consumers, and independent data vendors

Nowadays, numerous consumers data were held by many transaction and settlement platforms, such as Amazon, PayPal, and Taobao. If the ownership of these data belongs to these platforms, then these platforms can sell the accumulated data to firms who in turn use the data to make more informed production decisions. In this case, these platforms correspond directly to the data vendor in Section 2, and their objectives are to maximize their own profits.

When consumers make purchase decisions in these platforms, they may have implicitly concurred to give up their data ownership to these platforms by signing some agreements without carefully reading the contents. Now consumers start to understand that their data

have value and that they are due some compensation. Some startups, under the concept of "data locker," have already taken this kind of initiatives to give consumers more control over their own data and the opportunity to earn compensation.[6] The recent development of blockchain technology makes such a compensation easier to implement, because this new technology is well suited for effectively defining and protecting data ownership.[7] One issue in this context is that consumers do not know how much their data are worth in terms of dollars and how to trade off this monetary benefit against the potential cost of leaking privacy.[8] Our analysis in the previous sections provides an upper bound for the potential market value of data, namely, the profits earned by the data vendor.

Formally, suppose that the transaction data is originally owned by the data-0 consumers in our setting. These consumers can seize the compensation by forming a profit-maximizing data vendor. By Proposition 3, the total profits accruing to the data vendor are $C_A^* + C_B^* = \left(\frac{M}{4} + \frac{N}{9}\right) \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)}$. If data-0 consumers' valuations toward privacy are lower than $C_A^* + C_B^*$, then it is beneficial for them to sell their transaction data.[9] When consumers' privacy concerns are heterogeneous, their decisions can be different; intuitively, in equilibrium, those consumers who care about privacy the least would like to contribute their data and become a shareholder of the data vendor. We leave a formal analysis of this kind for future research.

5.2 Firms as data owners form the data vendor

It is also natural to assume that the data-0 consumer data is owned by firms, since they are important participants in producing such data. As Section 3 shows, when firms are original data owners, they have no incentives to share their private consumer data, although

[6] "Data mining offers rich seam," February 18, 2013, *Financial Times*.

[7] One such example is called "Steem", a user-generated content platform, which "is a blockchain-based rewards platform for publishers to monetize content and grow community" (https://steem.io/).

[8] "Fuel of the future—Data is giving rise to a new economy," May 6, 2017, *Economist*. Also see Acquisti (2014) for related discussions.

[9] Existing experimental studies suggest that consumers' valuations about privacy are relatively small, ranging from 0.50 to 45.00 US dollars (see Section 5 of Acquisti (2014)).

information sharing can be better for them if there is a sufficiently large number of local markets. However, the information sharing considered by the literature is sharing "for free." Then, how about sharing "for a price"? For instance, suppose that firms can form a data vendor who purchases data from and sells data to firms. Can such a data vendor move the equilibrium data allocation from $\emptyset\emptyset$ to **MM**, as achieved in Section 4?

In relation to the information-sharing literature (e.g., Gal-Or, 1985; Vives, 2006), the data vendor corresponds to a trade association examined by the literature. In the literature, a trade association collects information from firms at no cost and distributes information to firms for free. Here, the data vendor, which is the counterpart of a trade association, pays a price to a firm that contributes data to the vendor, and charges a price from a firm that acquires data from the vendor. Given that both firms are the shareholders of the data vendor, now it is natural to assume that the data vendor maximizes the total profits of both firms (as opposed to the vendor's own profits in Section 2), and retains no profits for itself.

In this case, the data vendor has incentives to move data allocation from $\emptyset\emptyset$ to **MM** if and only if the number M of local markets is sufficiently high, since by Proposition 1, firms are better off if and only if M is high. Since the data vendor retains no profits, the data transactions are equivalent to the following transfers between firms: Firm A makes transfer t_A to firm B for firm B's private consumer data and firm B makes t_B transfer to firm A for firm A's private consumer data. Does there exist a set of transfers (t_A, t_B) that supports the data allocation **MM**, when the number M of local markets is sufficiently large? The answer to this question is positive. Now let us explain how.

Given that the data vendor now behaves like a two-sided market, we need to consider both the data supply and data demand from firms, which correspond respectively to Figures 2 and 3 in previous sections. In Figure 4, we adopt the same parameter values as those in Figure 3. Panel A of Figure 4 describes the payoff matrix when firms supply data to the data vendor. This corresponds to Panel B of Figure 2, which assumes $t_A = t_B = 0$ (i.e., public disclosure means no compensation for supplying data). From the payoff matrix, we

Figure 4: Data sales as transfers

Panel A: The payoff matrix when firms supply information

		Firm B	
		0	M
Firm A	0	$(139.02, 139.02)$	$(179.02, 136.77 + t_A)$
	M	$(136.77 + t_B, 179.02)$	$(176.88 + t_B, 176.88 + t_A)$

Panel B: The payoff matrix when firms demand information

		Firm B	
		0	M
Firm A	0	$(139.02, 139.02)$	$(136.77, 179.02 - t_B)$
	M	$(179.02 - t_A, 136.77)$	$(176.88 - t_A, 176.88 - t_B)$

Panel C: Welfare variables

Data allocation	TS	CS	$\mathbb{E}\Pi_A^1 + \mathbb{E}\Pi_B^1$
$\emptyset\emptyset$	132474.67	132196.63	278.04
MM	171190.61	170836.85	353.76

Panel A is the payoff matrix for firms when they supply information. Panel B is the payoff matrix for firms when they demand information. Panel C reports the total surplus, consumer surplus and firms' net profits, i.e., the total profits net transfers. In this numerical example, we assume $\bar{\theta} = 0, \tau_\theta = 1, \tau_\varepsilon = 0.001$ and $N = 100$, $M = 10N$.

see that when $t_A \geq 2.25$ and $t_B \geq 2.25$, supplying information is the dominant strategy for both firms. Intuitively, when data prices are sufficiently high, both firms are willing to sell their data. Panel B of Figure 4 draws the payoff matrix when firms demand data from the vendor. This corresponds to Panel A of Figure 3 (with t_{0M} and t_{MM} replaced with t_A and t_B). Apparently, when $t_A \leq 40$ and $t_B \leq 40$, the data prices are sufficiently low such that firms always want to buy data from the data vendor. Taken together, we conclude that any transfer $(t_A, t_B) \in [2.25, 40]^2$ can support data allocation **MM**.

Proposition 8

Any transfer (t_A, t_B) *in the following rectangular set can support data allocation* **MM***:*

$$(t_A, t_B) \in \left[\frac{MN\tau_\varepsilon \left[6M\tau_\varepsilon + 5(N\tau_\varepsilon + \tau_\theta)\right]}{9\left(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta\right)\left[3M\tau_\varepsilon + 2(N\tau_\varepsilon + \tau_\theta)\right]^2}, \right.$$
$$\left. \frac{M\tau_\varepsilon \Theta}{36\left(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta\right)\left(2M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta\right)\left[3M\tau_\varepsilon + 2(N\tau_\varepsilon + \tau_\theta)\right]^2} \right]^2,$$

where

$$\Theta = 81M^3\tau_\varepsilon^2 + 33M^2N\tau_\varepsilon^2 + 108M^2\tau_\varepsilon(N\tau_\varepsilon + \tau_\theta) + 44MN\tau_\varepsilon(N\tau_\varepsilon + \tau_\theta)$$
$$+ 36M(N\tau_\varepsilon + \tau_\theta)^2 + 16N(N\tau_\varepsilon + \tau_\theta)^2.$$

The above set is non-empty for a sufficiently large M*.*

Proposition 8 also illustrates why voluntary disclosure is not viable in Section 3. Specifically, as we mentioned above, voluntary disclosure essentially sets $t_A = t_B = 0$, which does not lie in the rectangular set. Intuitively, when $t_A = t_B = 0$, the data price is so low that no firms want to supply information, leading to an equilibrium data allocation $\emptyset\emptyset$.

Our discussions in Sections 5.1 and 5.2 suggest that data ownership may matter for social welfare through changing the objectives of the data vendor. Specifically, if data belongs to consumers or platforms, the data vendor is likely to maximize its own profits, and data sales always changes the equilibrium data allocation from $\emptyset\emptyset$ to **MM** independent of the number M of local markets. This change in data allocation increases total surplus. However, if firms own the data and form a data vendor that maximizes the total profits of both firms, then data sales changes data allocation and improves total surplus only for a sufficiently large M. This observation suggests that it may be better to give ownership to consumers than to firms, provided that consumers can effectively monetize the value of their transaction data.

6 Conclusion

There are many debates about the springing data economy, such as the issue of data ownership, privacy, and fairness. While many discussions are from a legal or technology per-

spective, we study the data economy from an economic perspective by considering the real consequences of selling consumer data. We cast our analysis in a classical duopoly competition setting in which duopoly firms employ past consumer data to forecast future demand and make informed production plans. Our analysis yields a few conclusions and insights:

- Without data sales, firms withhold their private consumer data to protect their respective competitive advantages. This leads to efficiency loss and in some cases, to a prisoners' dilemma for firms.

- An independent profit-maximizing data vendor can restore efficiency. In this setting, firms effectively share their consumer data in equilibrium and thus, the equilibrium data allocation maximizes total surplus. This setting is consistent with a situation in which data is originally owned by consumers or by transaction/settlement platforms such as Amazon, eBay, or Taobao. Analyzing such a setting provides a way to quantify the economic value of consumer data, which is useful for thinking about the due compensation for consumers.

- When firms own the data and form a data vendor, the data vendor may behave as a trade association to maximize the total profits of both firms (as opposed to the vendor's own profits), by buying data from and selling data to member firms. In this case, firms do not always share their consumer data in equilibrium, and they do so only to address the prisoners' dilemma. Nonetheless, once data is shared, total surplus is still improved.

Overall, our analysis provides a tractable framework to analyze economics of data. There is a welfare gain from data sharing. Without some mechanism in place to share it companies may not chose socially optimal sharing. Data sales can be such a mechanism. Data ownership and vendor governance matter for efficiency.

References

Acquisti, Alessandro, 2015, From the economics of privacy to the economics of big data, *Working Paper* pp. 1–33.

Admati, Anat R., and Paul Pfleiderer, 1986, A monopolistic market for information, *Journal of Economic Theory* 39, 400–438.

——— , 1987, Viable allocations of information in financial markets, *Journal of Economic Theory* 43, 76–115.

——— , 1988, Selling and trading on information in financial markets, *American Economic Review* 78, 96–103.

Bagnoli, Mark, and Susan G. Watts, 2015, Competitive intelligence and disclosure, *RAND Journal of Economics* 46, 709–729.

Bailey, Warren, Haitao Li, Connie X. Mao, and Rui Zhong, 2003, Regulation fair disclosure and earnings information: Market, analyst, and corporate responses, *Journal of Finance* 58, 2487–2514.

Bushee, Brian J., Dawn A. Matsumoto, and Gregory S. Miller, 2004, Managerial and investor responses to disclosure regulation: The case of Reg FD and conference calls, *Accounting Review* 79, 617–643.

Cohen, Lauren, Dong Lou, and Christopher Malloy, 2017, Playing favorites: How firms prevent the revelation of bad news, *Working Paper*.

Cong, Lin William, and Zhiguo He, 2018, Blockchain disruption and smart contracts, *Working Paper*.

Darrough, Masako N., 1993, Disclosure policy and competition: Cournot vs. Bertrand, *Accounting Review* 68, 534–561.

Gal-Or, Esther, 1985, Information sharing in oligopoly, *Econometrica* 53, 329–343.

——— , 1986, Information transmission – Cournot and Bertrand equilibria, *Review of Economic Studies* 53, 85–92.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely, 2017, Are U.S. industries becoming more concentrated?, *Working Paper*.

Hauk, Esther, and Sjaak Hurkens, 2001, Secret information acquisition in Cournot markets, *Economic Theory* 18, 661–681.

Hwang, Hae-shin, 1993, Optimal information acquisition for heterogenous duopoly firms, *Journal of Economic Theory* 59, 385–402.

Li, Jiasun, and William Mann, 2018, Initial coin offering and platform building, *Working Paper*.

Li, Lode, Richard D McKelvey, and Talbot Page, 1987, Optimal research for cournot oligopolists, *Journal of Economic Theory* 42, 140–166.

Raith, Michael, 1996, A general model of information sharing in oligopoly, *Journal of Economic Theory* 71, 260–288.

Singh, Nirvikar, and Xavier Vives, 1984, Price and quantity competition in a differentiated duopoly, *RAND Journal of Economics* 15, 546–554.

Vives, Xavier, 1984, Duopoly information equilibrium: Cournot and Bertrand, *Journal of Economic Theory* 34, 71–94.

——— , 1988, Aggregation of information in large Cournot markets, *Econometrica* 56, 851.

——— , 2016, Information sharing among firms, in *The New Palgrave Dictionary of Economics* . pp. 1–4 (Palgrave Macmillan UK: London).

Appendix A: Proofs

Proof for Proposition 1

Proof. The first order condition of the consumer utility maximization gives us the total expected consumer surplus for the A-type and B-type consumer. That is,

$$CS_A = \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left[\left(MX_{A,i}^{1*}\right)^2 \mid \mathcal{F}_A\right]\right] = \frac{M}{2}\mathbb{E}\Pi_{A,X}^{1*}. \tag{A1}$$

The last equality is implied by the market clearing condition ($x_{A,i}^{1*} = X_{A,i}^{1*}$). Similarly, the market clearing condition implies the AB-type consumer surplus is,

$$\begin{aligned} CS_{AB} &= \frac{1}{2}\mathbb{E}\left[\left(Y_A^{1*} + Y_B^{1*}\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[(Y_A^{1*})^2 + 2Y_A^{1*}Y_B^{1*} + (Y_B^{1*})^2\right] \\ &= \frac{N}{2}\left(\mathbb{E}\Pi_{A,Y}^{1*} + \mathbb{E}\Pi_{B,Y}^{1*}\right) + \mathbb{E}(Y_A^{1*}Y_B^{1*}). \end{aligned} \tag{A2}$$

⋆ The $\emptyset\emptyset$ Data Allocation

Consider the information set for the A-type and B-type consumer is $\mathcal{F}_A^{\emptyset\emptyset}, \mathcal{F}_B^{\emptyset\emptyset}$, respectively. With some computation, we have

$$\begin{aligned} CS_A^{\emptyset\emptyset} = CS_B^{\emptyset\emptyset} &= \frac{M}{2}\mathbb{E}\Pi_{A,X}^{1,\emptyset\emptyset} \\ &= \frac{M^2}{8}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta}\right) + \frac{M\tau_\epsilon}{(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)(N\tau_\epsilon + \tau_\theta)}\right]. \end{aligned} \tag{A3}$$

For the AB-type consumer, we have

$$\begin{aligned} \mathbb{E}\Pi_{A,Y}^{1,\emptyset\emptyset} &= \mathbb{E}\Pi_{B,Y}^{1,\emptyset\emptyset} \\ &= \frac{N}{9}\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta}\right) + \frac{MN\tau_\varepsilon(M\tau_\varepsilon + N\tau_\varepsilon + \tau_\theta)}{\left[3M\tau_\varepsilon + 2\left(N\tau_\varepsilon + \tau_\theta\right)\right]^2\left(N\tau_\varepsilon + \tau_\theta\right)}, \end{aligned} \tag{A4}$$

$$\begin{aligned} \mathbb{E}(Y_A^1 Y_B^1)^{\emptyset\emptyset} &= (\Phi_{A_0}^Y)^2 + (\Phi_{A_1}^Y)^2\frac{1}{N\tau_\epsilon + \tau_\theta} \\ &= \frac{N^2}{9}\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta}\right) + \left[\frac{MN\tau_\epsilon}{3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)}\right]^2\frac{1}{N\tau_\epsilon + \tau_\theta}. \end{aligned} \tag{A5}$$

Therefore,

$$CS_{AB}^{\emptyset\emptyset} = \frac{2N^2}{9}\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta}\right) + \frac{MN^2\tau_\epsilon(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)}{\left[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right]^2\left(N\tau_\epsilon + \tau_\theta\right)}. \tag{A6}$$

The total consumer surplus is

$$CS^{\emptyset\emptyset} = CS_A^{\emptyset\emptyset} + CS_B^{\emptyset\emptyset} + CS_{AB}^{\emptyset\emptyset}. \tag{A7}$$

And the total surplus is

$$TS^{\emptyset\emptyset} = \underbrace{\mathbb{E}\Pi_A^{1,\emptyset\emptyset}}_{\mathbb{E}\Pi_{A,X}^{1,\emptyset\emptyset}+\mathbb{E}\Pi_{A,Y}^{1,\emptyset\emptyset}} + \underbrace{\mathbb{E}\Pi_B^{1,\emptyset\emptyset}}_{\mathbb{E}\Pi_{B,X}^{1,\emptyset\emptyset}+\mathbb{E}\Pi_{B,Y}^{1,\emptyset\emptyset}} + CS^{\emptyset\emptyset}. \tag{A8}$$

⋆ The **MM** data allocation

For the information set $\mathcal{F}_A^{\mathbf{MM}}, \mathcal{F}_B^{\mathbf{MM}}$, we get

$$CS_A^{\mathbf{MM}} = CS_B^{\mathbf{MM}} = \frac{M^2}{8}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\epsilon}{(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}\right], \tag{A9}$$

$$\begin{aligned}
\mathbb{E}\Pi_{A,X}^{1,\mathbf{MM}} &= \mathbb{E}\Pi_{B,X}^{1,\mathbf{MM}} \\
&= \frac{M}{4}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon+N\tau_\varepsilon+\tau_\theta)(N\tau_\varepsilon+\tau_\theta)}\right], \qquad \text{(A10)} \\
\mathbb{E}\Pi_{A,Y}^{1,\mathbf{MM}} &= \mathbb{E}\Pi_{B,Y}^{1,\mathbf{MM}} \\
&= \frac{N}{9}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon+N\tau_\varepsilon+\tau_\theta)(N\tau_\varepsilon+\tau_\theta)}\right], \qquad \text{(A11)} \\
\mathbb{E}(Y_A^1 Y_B^1)^{\mathbf{MM}} &= (\Phi_{A_0}^Y)^2 + (\Phi_{A_1}^Y + \Phi_{A_2}^Y)^2 \frac{1}{N\tau_\epsilon+\tau_\theta} + 2\Phi_{A_1}^Y \Phi_{A_2}^Y \frac{1}{M\tau_\epsilon} \\
&= \frac{N^2}{9}\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{2MN^2\tau_\epsilon}{9(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}, \qquad \text{(A12)} \\
CS_{AB}^{\mathbf{MM}} &= N\mathbb{E}\Pi_{A,Y}^{1,\mathbf{MM}} + \mathbb{E}(Y_A^1 Y_B^1)^{\mathbf{MM}} \\
&= \frac{2N^2}{9}\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{4MN^2\tau_\epsilon}{9(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}, \qquad \text{(A13)} \\
TS^{\mathbf{MM}} &= \underbrace{\mathbb{E}\Pi_A^{1,\mathbf{MM}}}_{\mathbb{E}\Pi_{A,X}^{1,\mathbf{MM}}+\mathbb{E}\Pi_{A,Y}^{1,\mathbf{MM}}} + \underbrace{\mathbb{E}\Pi_B^{1,\mathbf{MM}}}_{\mathbb{E}\Pi_{B,X}^{1,\mathbf{MM}}+\mathbb{E}\Pi_{B,Y}^{1,\mathbf{MM}}} + \underbrace{CS^{\mathbf{MM}}}_{CS_A^{\mathbf{MM}}+CS_B^{\mathbf{MM}}+CS_{AB}^{\mathbf{MM}}}. \qquad \text{(A14)}
\end{aligned}$$

To avoid repetition, we will postpone the proof of above equations to **Proof of Proposition 4**.

Direct computation shows that

$$CS^{\emptyset\emptyset} < CS^{\mathbf{MM}}, TS^{\emptyset\emptyset} < TS^{\mathbf{MM}}, \tag{A15}$$

and

$$\begin{aligned}\mathbb{E}\Pi_A^{1,\mathbf{MM}} - \mathbb{E}\Pi_A^{1,\emptyset\emptyset} =& \frac{M^2\tau_\epsilon}{4(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)} \\ &- \frac{MN\tau_\epsilon(3M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)}{9(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)]^2} \\ =& \frac{M\tau_\epsilon}{36(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)} \frac{1}{[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)]^2} \\ &\times (M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)(3M\tau_\epsilon + N\tau_\epsilon + \tau_\theta) \qquad \text{(A16)} \\ &\times \underbrace{\left\{ \frac{9M}{M\tau_\epsilon + N\tau_\epsilon + \tau_\theta} \left[3M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta) + 2(N\tau_\epsilon + \tau_\theta) + \frac{(N\tau_\epsilon + \tau_\theta)^2}{3M\tau_\epsilon + N\tau_\epsilon + \tau_\theta} \right] - 4N \right\}}_{g(M)}.\end{aligned}$$

It is clear that the sign of $\mathbb{E}\Pi_A^{1,\mathbf{MM}} - \mathbb{E}\Pi_A^{1,\emptyset\emptyset}$ depends only on $g(M)$.

We find that

$$g(0) = -4N < 0, g(+\infty) > 0, \qquad \text{(A17)}$$

and

$$\begin{aligned}\frac{\partial g(M)}{\partial M} =& 9\frac{N\tau_\epsilon + \tau_\theta}{(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)^2} \left[3M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta) + 2(N\tau_\epsilon + \tau_\theta) + \frac{(N\tau_\epsilon + \tau_\theta)^2}{3M\tau_\epsilon + N\tau_\epsilon + \tau_\theta} \right] \\ &+ 9\frac{M}{M\tau_\epsilon + N\tau_\epsilon + \tau_\theta} \left[3\tau_\epsilon - \frac{3\tau_\epsilon(N\tau_\epsilon + \tau_\theta)^2}{(3M\tau_\epsilon + N\tau_\epsilon + \tau_\theta)^2} \right] > 0. \qquad \text{(A18)}\end{aligned}$$

Thus, there exists one unique solution of $g(M) = 0$.

We denote the solution as $\hat{M}$, i.e.,

$$\frac{9\hat{M}}{\hat{M}\tau_\epsilon + N\tau_\epsilon + \tau_\theta} \left[3\hat{M}\tau_\epsilon + (N\tau_\epsilon + \tau_\theta) + 2(N\tau_\epsilon + \tau_\theta) + \frac{(N\tau_\epsilon + \tau_\theta)^2}{3\hat{M}\tau_\epsilon + N\tau_\epsilon + \tau_\theta} \right] - 4N = 0. \qquad \text{(A19)}$$

When $M > \hat{M}$, $g(M) > 0$; and when $M < \hat{M}$, $g(M) < 0$. This suggest that when $M > \hat{M}$, $\mathbb{E}\Pi_A^{1,\mathbf{MM}} - \mathbb{E}\Pi_A^{1,\emptyset\emptyset} > 0$. ■

Proof for Proposition 2

Proof. Since firm A and firm B are symmetric in choosing the optimal production, we use firm A to illustrate the optimal production decision. Similar argument can be applied to firm B to yield the expression in Proposition 2.

Given any (m_A, m_B), firm A's optimal production decision is to maximize her profits in the X-market and the Y-market. Combining the market clearing condition ($x_{A,i}^1 =$

$X^1_{A,i}, \sum_{i=1}^N y^1_{AB,i} = Y^1_A + Y^1_B)$ with consumer utility maximization, we have $P^1_{A,i} = \tilde{s}^1_{A,i} - X^1_{A,i}, p^1_y = \frac{\sum_{i=1}^N \tilde{s}^1_{AB,i} - Y^1_A - Y^1_B}{N}$. Hence, firm A's production decision is

$$\max_{x^1_{A,i}} \mathbb{E}[P^1_{A,i} X^1_{A,i} \mid \mathcal{F}_A] = \max_{x^1_{A,i}} \mathbb{E}\left[\left(\tilde{s}^1_{A,i} - X^1_{A,i}\right) X^1_{A,i} \mid \mathcal{F}_A\right], \text{for } i = 1, 2, ..., M; \tag{A20}$$

$$\max_{Y^1_A} \mathbb{E}\left[\left(\frac{\sum_{i=1}^N \tilde{s}^1_{AB,i} - Y^1_A - \hat{Y}^1_B}{N}\right) Y^1_A \mid \mathcal{F}_A\right]. \tag{A21}$$

The optimal production is

$$X^{1*}_{A,i} = \frac{1}{2}\mathbb{E}\left(\tilde{\theta} \mid \mathcal{F}_A\right), \tag{A22}$$

$$Y^{1*}_A = \frac{N}{2}\mathbb{E}(\tilde{\theta} \mid \mathcal{F}_A) - \frac{1}{2}\mathbb{E}(\hat{Y}^{1*}_B \mid \mathcal{F}_A). \tag{A23}$$

Due to symmetry $\mathcal{F}_A = \left\{p^{0*}_y, \mathbf{I}_A, \mathbf{P}^0_A\right\}$ and $\mathcal{F}_B = \left\{p^{0*}_y, \mathbf{I}_B, \mathbf{P}^0_B\right\}$ are informationally equivalent to $\mathcal{F}_A = \left\{p^{0*}_y, I_A, P^0_A\right\}$ and $\mathcal{F}_B = \left\{p^{0*}_y, I_B, P^0_B\right\}$, respectively. And

$$I_A = \frac{1}{m_A}\sum_{a=1}^{m_A} p^{0*}_{B,j_a}, P^0_A = \frac{1}{M}\sum_{i=1}^{M} p^{0*}_{A,i},$$

$$I_B = \frac{1}{m_B}\sum_{b=1}^{m_B} p^{0*}_{A,i_b}, P^0_B = \frac{1}{M}\sum_{j=1}^{M} p^{0*}_{B,j}.$$

From the Bayesian updating, we have

$$\mathbb{E}\left[\tilde{\theta} - \mu \mid \mathcal{F}_A\right] = \begin{bmatrix} \frac{M\tau_\epsilon}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \\ \frac{\tau_\epsilon m_A}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \end{bmatrix} [P^0_A - \mu \quad I_A - \mu], \tag{A24}$$

$$\mathbb{E}\left[P^0_B - \mu \mid \mathcal{F}_A\right] = \begin{bmatrix} \frac{M\tau_\epsilon - \tau_\epsilon m_A}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \\ \frac{\tau_\epsilon m_A (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{M\tau_\epsilon (M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta))} \end{bmatrix} [P^0_A - \mu \quad I_A - \mu], \tag{A25}$$

$$\mathbb{E}\left[I_B - \mu \mid \mathcal{F}_A\right] = \begin{bmatrix} 1 \\ 0 \end{bmatrix} [P^0_A - \mu \quad I_A - \mu], \tag{A26}$$

where $\mu = \bar{\theta} + \frac{N\tau_\epsilon}{N\tau_\epsilon + \tau_\theta}\left(p^0_y - \frac{\bar{\theta}}{3}\right)$.

In the Y-market, when combining with the conjecture linear strategy,

$$Y^1_A = \Phi^Y_{A_0} + \Phi^Y_{A_1}(P^0_A - \mu) + \Phi^Y_{A_2}(I_A - \mu), \tag{A27}$$

$$Y^1_B = \Phi^Y_{B_0} + \Phi^Y_{B_1}(P^0_B - \mu) + \Phi^Y_{B_2}(I_B - \mu), \tag{A28}$$

we have

$$Y_A^{1*} = \begin{bmatrix} 1 & P_A^0 - \mu & I_A - \mu \end{bmatrix} \tag{A29}$$

$$\times \left\{ \frac{N}{2} \begin{bmatrix} \mu \\ \frac{M\tau_\epsilon}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \\ \frac{\tau_\epsilon m_A}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \end{bmatrix} - \frac{\Phi_{B_0}^Y}{2} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} - \frac{\Phi_{B_1}^Y}{2} \begin{bmatrix} 0 \\ \frac{M\tau_\epsilon - \tau_\epsilon m_A}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \\ \frac{\tau_\epsilon m_A (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{M\tau_\epsilon (M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta))} \end{bmatrix} - \frac{\Phi_{B_2}^Y}{2} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$$

$$= \begin{bmatrix} 1 & P_A^0 - \mu & I_A - \mu \end{bmatrix} \begin{bmatrix} \frac{N}{2}\mu - \frac{\Phi_{B_0}^Y}{2} \\ \frac{N}{2} M\tau_\epsilon \Lambda_A - \Phi_{B_1}^Y \frac{M\tau_\epsilon - \tau_\epsilon m_A}{2} \Lambda_A - \Phi_{B_2}^Y \frac{1}{2} \\ \frac{N}{2} \tau_\epsilon m_A \Lambda_A - \Phi_{B_1}^Y \frac{\tau_\epsilon m_A (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{2M\tau_\epsilon} \Lambda_A \end{bmatrix},$$

where $\Lambda_A^{-1} = M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)$.

Comparing with the conjecture strategy, we get

$$\underbrace{\begin{bmatrix} \Phi_{A_0}^Y \\ \Phi_{A_1}^Y \\ \Phi_{A_2}^Y \end{bmatrix}}_{\Phi_A^Y} = \begin{bmatrix} \frac{N}{2}\mu - \frac{\Phi_{B_0}^Y}{2} \\ \frac{N}{2} M\tau_\epsilon \Lambda_A - \Phi_{B_1}^Y \frac{M\tau_\epsilon - \tau_\epsilon m_A}{2} \Lambda_A - \Phi_{B_2}^Y \frac{1}{2} \\ \frac{N}{2} \tau_\epsilon m_A \Lambda_A - \Phi_{B_1}^Y \frac{\tau_\epsilon m_A (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{2M\tau_\epsilon} \Lambda_A \end{bmatrix} \tag{A30}$$

$$= \underbrace{\frac{N}{2} \Lambda_A \begin{bmatrix} \Lambda_A^{-1}\mu \\ M\tau_\epsilon \\ \tau_\epsilon m_A \end{bmatrix}}_{\Xi_{A_1}} - \underbrace{\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & (M\tau_\epsilon - \tau_\epsilon m_A)\Lambda_A & 1 \\ 0 & \frac{\tau_\epsilon m_A (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{M\tau_\epsilon} \Lambda_A & 0 \end{bmatrix}}_{\Xi_{A_2}} \begin{bmatrix} \Phi_{B_0}^Y \\ \Phi_{B_1}^Y \\ \Phi_{B_2}^Y \end{bmatrix}.$$

Similarly, from firm B's production decision, we can have

$$\underbrace{\begin{bmatrix} \Phi_{B_0}^Y \\ \Phi_{B_1}^Y \\ \Phi_{B_2}^Y \end{bmatrix}}_{\Phi_B^Y} = \begin{bmatrix} \frac{N}{2}\mu - \frac{\Phi_{A_0}^Y}{2} \\ \frac{N}{2} M\tau_\epsilon \Lambda_B - \Phi_{A_1}^Y \frac{M\tau_\epsilon - \tau_\epsilon m_B}{2} \Lambda_B - \Phi_{A_2}^Y \frac{1}{2} \\ \frac{N}{2} \tau_\epsilon m_B \Lambda_B - \Phi_{A_1}^Y \frac{\tau_\epsilon m_B (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{2M\tau_\epsilon} \Lambda_B \end{bmatrix} \tag{A31}$$

$$= \underbrace{\frac{N}{2} \Lambda_B \begin{bmatrix} \Lambda_B^{-1}\mu \\ M\tau_\epsilon \\ \tau_\epsilon m_B \end{bmatrix}}_{\Xi_{B_1}} - \underbrace{\frac{1}{2} \begin{bmatrix} 1 & 0 & 0 \\ 0 & (M\tau_\epsilon - \tau_\epsilon m_B)\Lambda_B & 1 \\ 0 & \frac{\tau_\epsilon m_B (2M\tau_\epsilon + (N\tau_\epsilon + \tau_\theta))}{M\tau_\epsilon} \Lambda_B & 0 \end{bmatrix}}_{\Xi_{B_2}} \begin{bmatrix} \Phi_{A_0^Y} \\ \Phi_{A_1^Y} \\ \Phi_{A_2^Y} \end{bmatrix}.$$

Thus,

$$\Phi_A^Y = \left(\underbrace{\mathbf{I}}_{\text{identity matrix}} - \Xi_{A_2}\Xi_{B_2} \right)^{-1} \left(\Xi_{A_1} - \Xi_{A_2}\Xi_{B_1} \right),$$

$$\Phi_B^Y = \Xi_{B_1} - \Xi_{B_2}\Phi_A^Y. \tag{A32}$$

In the X-market, we have

$$X_{A,i}^{1*} = \frac{1}{2}\left(\mu + \begin{bmatrix} \frac{M\tau_\epsilon}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \\ \frac{\tau_\epsilon m_A}{M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta)} \end{bmatrix} \begin{bmatrix} P_A^0 - \mu & I_A - \mu \end{bmatrix} \right). \tag{A33}$$

Replacing m_A with m_B, P_A^0 with P_B^0, I_A with I_B in the above equation, we get $X_{B,j}^1$.

The expected profit in the X-market is

$$\begin{aligned}
\mathbb{E}\Pi_{A,X}^{1*} =& M\mathbb{E}\left[\mathbb{E}\left[\left(X_{A,i}^{1*}\right)^2 \mid \mathcal{F}_A\right]\right] \\
=& \frac{M}{4}\left\{ \left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta} \right) + \frac{(M\tau_\epsilon)^2}{(M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta))^2}\mathbb{E}\left[\left(P_A^0 - \mu\right)^2\right] \right. \\
&+ \frac{\tau_\epsilon^2 m_A^2}{(M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta))^2}\mathbb{E}\left[(I_A - \mu)^2\right] \\
&\left. + \frac{2M\tau_\epsilon^2 m_A}{(M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta))^2}\mathbb{E}\left[\left(P_A^0 - \mu\right)(I_A - \mu)\right] \right\} \\
=& \frac{M}{4}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon + \tau_\theta)\tau_\theta} \right) + \frac{M\tau_\epsilon + \tau_\epsilon m_A}{(M\tau_\epsilon + \tau_\epsilon m_A + (N\tau_\epsilon + \tau_\theta))(N\tau_\epsilon + \tau_\theta)} \right].
\end{aligned} \tag{A34}$$

The expected profit in the Y-market is

$$\begin{aligned}
\mathbb{E}\Pi_{A,Y}^{1*} &= \frac{1}{N}\mathbb{E}\left[Y_A^{1*}\right]^2 \\
&= \frac{1}{N}\left\{ \mathbb{E}\left[\Phi_{A_0}^Y\right]^2 + (\Phi_{A_1}^Y)^2\mathbb{E}\left[P_A^0 - \mu\right]^2 + (\Phi_{A_2}^Y)^2\mathbb{E}\left[I_A - \mu\right]^2 + 2\Phi_{A_1}^Y\Phi_{A_2}^Y\mathbb{E}\left[\left(P_A^0 - \mu\right)(I_A - \mu)\right] \right\} \\
&= \frac{1}{N}\mathbb{E}\left[(\Phi_{A_0}^Y)^2\right] + \underbrace{\frac{1}{N(N\tau_\epsilon + \tau_\theta)}\left[(\Phi_{A_1}^Y + \Phi_{A_2}^Y)^2 + (\Phi_{A_1}^Y)^2\frac{N\tau_\epsilon + \tau_\theta}{M\tau_\epsilon} + (\Phi_{A_2}^Y)^2\frac{N\tau_\epsilon + \tau_\theta}{\tau_\epsilon m_A} \right]}_{\pi}.
\end{aligned} \tag{A35}$$

Again, replacing m_A with m_B, Φ_A^Y with Φ_B^Y in the above equation, we get $\mathbb{E}\Pi_{B,X}^1, \mathbb{E}\Pi_{B,Y}^1$.

■

Proof for Proposition 3

The following lemma is helpful for the proof.

Lemma 3

$\Delta(m_B) \equiv \mathbb{E}\Pi^1_{A,Y}(M, m_B) - \mathbb{E}\Pi^1_{A,Y}(0, m_B)$ *increases in* m_B.

Proof. Direction computation on equation (A35) shows

$$\Delta(m_B) \equiv \mathbb{E}\Pi^1_{A,Y}(M, m_B) - \mathbb{E}\Pi^1_{A,Y}(0, m_B) = \pi(M, m_B) - \pi(0, m_B). \tag{A36}$$

Differentiating $\Delta(m_B)$ with respect to m_B, we get

$$\frac{\partial \Delta(m_B)}{\partial m_B} = \frac{2M^3 N\tau_\epsilon^2}{V_2^\Delta} V_1^\Delta, \tag{A37}$$

where

$$\begin{aligned} V_1^\Delta =& 4M^6\tau_\epsilon^6(N\tau_\epsilon + \tau_\theta)\left(M - m_B\right)^2\left(M + m_B\right)\left(5M + 4m_B\right) \\ &+ 2M^5\tau_\epsilon^5(N\tau_\epsilon + \tau_\theta)^2\left(58M^4 - 29M^3 m_B + 27M^2 m_B^2 + 125Mm_B^3 + 35m_B^4\right) \\ &+ 6M^4\tau_\epsilon^4(N\tau_\epsilon + \tau_\theta)^3\left(40M^4 - 16M^3 m_B + 111M^2 m_B^2 + 80Mm_B^3 + m_B^4\right) \\ &+ 2M^3\tau_\epsilon^3(N\tau_\epsilon + \tau_\theta)^4\left(104M^4 + 90M^3 m_B + 537M^2 m_B^2 + 49Mm_B^3 - 24m_B^4\right) \\ &+ M^2\tau_\epsilon^2(N\tau_\epsilon + \tau_\theta)^5\left(4M - m_B\right)\left(16M^3 + 128M^2 m_B + 149Mm_B^2 - 5m_B^3\right) \\ &+ 3Mm_B\tau_\epsilon(N\tau_\epsilon + \tau_\theta)^6\left(4M - m_B\right)^2\left(7M + 2m_B\right) + m_B(N\tau_\epsilon + \tau_\theta)^7\left(4M - m_B\right)^3 > 0 \end{aligned} \tag{A38}$$

$$\begin{aligned} V_2^\Delta =& \left[4M^2\tau_\epsilon + 2Mm_B\tau_\epsilon + 4M(N\tau_\epsilon + \tau_\theta) - m_B(N\tau_\epsilon + \tau_\theta)\right]^3 \\ &\times \left[3M^3\tau_\epsilon^2 + 3M^2 m_B\tau_\epsilon^2 + 8M^2\tau_\epsilon(N\tau_\epsilon + \tau_\theta) + Mm_B\tau_\epsilon(N\tau_\epsilon + \tau_\theta) + 4M(N\tau_\epsilon + \tau_\theta)^2\right. \\ &\left. - m_B(N\tau_\epsilon + \tau_\theta)^2\right]^3 > 0. \end{aligned} \tag{A39}$$

So, we have

$$\frac{\partial \Delta(m_B)}{\partial m_B} = \frac{2M^3 N\tau_\epsilon^2}{V_2^\Delta} V_1^\Delta > 0. \tag{A40}$$

■

Proof. Equation (A34) represents the expected profit in the X-market. And it is obvious that $\mathbb{E}\Pi^1_{A,X}$ increases in m_A.

The key determinant of the expected profit in the Y-market is π in equation (A35). Direct computation shows that

$$\frac{\partial \pi(m_A, m_B)}{\partial m_A} = \frac{M^2 N\tau_\epsilon}{V_0^3} V_1 V_2 \geq 0, \tag{A41}$$

where

$$
\begin{aligned}
V_0 =& \tau_\epsilon^2(N\tau_\epsilon+\tau_\theta)\big(12M^4+12M^3m_A+12M^3m_B\big)\\
&+\tau_\epsilon(N\tau_\epsilon+\tau_\theta)\left[32M^3(N\tau_\epsilon+\tau_\theta)+4M^2m_A(N\tau_\epsilon+\tau_\theta)+4M^2m_B(N\tau_\epsilon+\tau_\theta)\right.\\
&\quad\left.-4Mm_Am_B(N\tau_\epsilon+\tau_\theta)\right]\\
&+(N\tau_\epsilon+\tau_\theta)\left[16M^2(N\tau_\epsilon+\tau_\theta)^2-4Mm_A(N\tau_\epsilon+\tau_\theta)^2-4Mm_B(N\tau_\epsilon+\tau_\theta)^2\right.\\
&\quad\left.+m_Am_B(N\tau_\epsilon+\tau_\theta)^2\right]>0
\end{aligned}
\tag{A42}
$$

$$
V_1 = 2M^2\tau_\epsilon+4Mm_B\tau_\epsilon+4M(N\tau_\epsilon+\tau_\theta)-m_B(N\tau_\epsilon+\tau_\theta)\geq 0 \tag{A43}
$$

$$
\begin{aligned}
V_2 =& \tau_\epsilon^3\left(88M^6+88M^5m_A+72M^5m_B+32M^4m_Am_B-16M^4m_B^2+96M^3m_Am_B^2\right)\\
&+\tau_\epsilon^2(N\tau_\epsilon+\tau_\theta)\left(272M^5+152M^4m_A-20M^4m_B+108M^3m_Am_B+36M^3m_B^2-8M^2m_Am_B^2\right)\\
&+\tau_\epsilon(N\tau_\epsilon+\tau_\theta)^2\left(224M^4+104M^3m_A-56M^3m_B-6M^2m_Am_B+12M^2m_B^2-8Mm_Am_B^2\right)\\
&+(N\tau_\epsilon+\tau_\theta)^3\left(64M^3+16M^2m_A-32M^2m_B-8Mm_Am_B+4Mm_B^2+m_Am_B^2\right)\geq 0.
\end{aligned}
\tag{A44}
$$

Hence, $\mathbb{E}\Pi^1_{A,Y}$ increases in m_A. As $\mathbb{E}\Pi^1_A=\mathbb{E}\Pi^1_{A,X}+\mathbb{E}\Pi^1_{A,Y}$, $\mathbb{E}\Pi^1_A$ increases in m_A.

Lemma 3 implies that

$$
\mathbb{E}\Pi^1_{A,Y}(M,M)-\mathbb{E}\Pi^1_{A,Y}(0,M)\geq \mathbb{E}\Pi^1_{A,Y}(M,m_B)-\mathbb{E}\Pi^1_{A,Y}(0,m_B). \tag{A45}
$$

Since $\mathbb{E}\Pi^1_{A,Y}(m_A,m_B)-\mathbb{E}\Pi^1_{A,Y}(0,m_B)$ increases in m_A, we have

$$
\mathbb{E}\Pi^1_{A,Y}(M,m_B)-\mathbb{E}\Pi^1_{A,Y}(0,m_B)\geq \mathbb{E}\Pi^1_{A,Y}(m_A,m_B)-\mathbb{E}\Pi^1_{A,Y}(0,m_B). \tag{A46}
$$

Therefore,

$$
\underbrace{\mathbb{E}\Pi^1_{A,Y}(M,M)-\mathbb{E}\Pi^1_{A,Y}(0,M)}_{C_A(M,M)}\geq \underbrace{\mathbb{E}\Pi^1_{A,Y}(m_A,m_B)-\mathbb{E}\Pi^1_{A,Y}(0,m_B)}_{C_A(m_A,m_B)}, \tag{A47}
$$

i.e., $m_A^*=m_B^*=M$ maximizes $C_A(m_A,m_B)$. By symmetry, the same applies to $C_B(m_A,m_B)$. The data vendor chooses $m_A^*=m_B^*=M$ to maximizes his revenue $C_A(m_A,m_B)+C_B(m_A,m_B)$.

Let $C_A^*=C_A(M,M), C_B^*=C_B(M,M)$. To compute C_A^*, C_B^*, we first compute

$$
\begin{aligned}
\mathbb{E}\Pi^1_A(M,M) =& \frac{M}{4}\left[\left(\bar\theta^2+\frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right)+\frac{2M\tau_\epsilon}{(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}\right]\\
&+\frac{N}{9}\left[\left(\bar\theta^2+\frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right)+\frac{2M\tau_\epsilon}{(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}\right].
\end{aligned}
\tag{A48}
$$

Then, we repeat these computation for $m_A = 0, m_B = M$ and obtain the product policy

$$\begin{bmatrix} \Phi^Y_{A_0} \\ \Phi^Y_{A_1} \\ \Phi^Y_{A_2} \end{bmatrix} = \begin{bmatrix} \frac{N\mu}{3} \\ \frac{MN\tau_\epsilon}{3(M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)} \\ 0 \end{bmatrix}. \tag{A49}$$

Plugging these to the profit computation, we obtain $\mathbb{E}\Pi^1_A(0, M)$,

$$\begin{aligned}\mathbb{E}\Pi^1_A(0, M) =& \frac{M}{4}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{M\tau_\epsilon}{(M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}\right] \\ &+ \frac{N}{9}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{M\tau_\epsilon}{(M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}\right].\end{aligned} \tag{A50}$$

At last, we obtain

$$C^*_A = C^*_B = \mathbb{E}\Pi^1_A(M, M) - \mathbb{E}\Pi^1_A(0, M) = \left(\frac{M}{4} + \frac{N}{9}\right)\frac{M\tau_\epsilon}{(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(N\tau_\epsilon+\tau_\theta)}. \tag{A51}$$

■

Proof for Proposition 4

Proof. According to Definition 2, the equilibrium with the data vendor is obtained by combining Proposition 2 and 3. More specifically, let $m_A = M, m_B = M$ in Proposition 2, we get

$$\Phi^Y_A = \Phi^Y_B = \begin{bmatrix} \frac{N\mu}{3} \\ \frac{MN\tau_\epsilon}{3(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)} \\ \frac{MN\tau_\epsilon}{3(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)} \end{bmatrix}, \tag{A52}$$

which means

$$Y^{1*}_A = Y^{1*}_B = \Phi^Y_{A_0} + \Phi^Y_{A_1}\left(P^0_A - \mu\right) + \Phi^Y_{A_2}\left(P^0_B - \mu\right). \tag{A53}$$

And from equation (A33), we have

$$X^{1*}_{A,i} = \frac{1}{2}\left(\mu + \begin{bmatrix} \frac{M\tau_\epsilon}{2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta} \\ \frac{M\tau_\epsilon}{2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta} \end{bmatrix}\begin{bmatrix} P^0_A - \mu & P^0_B - \mu \end{bmatrix}\right). \tag{A54}$$

Substituting $X^{1*}_{A,i}, Y^{1*}_A$ into the profit function (equation A34, A35), we get

$$\begin{aligned}\mathbb{E}\Pi^{1*}_A = \mathbb{E}\Pi^{1*}_B =& \frac{M}{4}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon+N\tau_\varepsilon+\tau_\theta)(N\tau_\varepsilon+\tau_\theta)}\right] \\ &+ \frac{N}{9}\left[\left(\bar{\theta}^2 + \frac{N\tau_\epsilon}{(N\tau_\epsilon+\tau_\theta)\tau_\theta}\right) + \frac{2M\tau_\varepsilon}{(2M\tau_\varepsilon+N\tau_\varepsilon+\tau_\theta)(N\tau_\varepsilon+\tau_\theta)}\right].\end{aligned} \tag{A55}$$

■

Proof for Proposition 5

Proof. See Lemma 3. ■

Proof for Proposition 6

Proof. Since the equilibrium of the economy with the data vendor reaches full information sharing (the **MM** data allocation) and the equilibrium of the economy without the data vendor is no information sharing (the $\emptyset\emptyset$ data allocation), we know from Proposition 1, $CS^{\mathbf{MM}} > CS^{\emptyset\emptyset}$ and $TS^{\mathbf{MM}} > TS^{\emptyset\emptyset}$. ■

Proof for Proposition 7

Proof. The change in the expected profit for firm A when comparing the **MM** data allocation with the $\emptyset\emptyset$ data allocation is

$$\begin{aligned}
&\mathbb{E}\Pi_A^1(M,M) - C_A(M,M) - \mathbb{E}\Pi_A^1(0,0) \qquad \text{(A56)}\\
&= (1-\beta)\left(\frac{M}{4}+\frac{N}{9}\right)\frac{M\tau_\epsilon}{(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)} + \mathbb{E}\Pi_{A,Y}^1(0,M) - \mathbb{E}\Pi_{A,Y}^1(0,0)\\
&= (1-\beta)\left(\frac{M}{4}+\frac{N}{9}\right)\frac{M\tau_\epsilon}{(2M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)(M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)}\\
&\quad - \frac{MN\tau_\epsilon(6M\tau_\epsilon+5(N\tau_\epsilon+\tau_\theta))}{9(M\tau_\epsilon+N\tau_\epsilon+\tau_\theta)[3M\tau_\epsilon+2(N\tau_\epsilon+\tau_\theta)]^2},
\end{aligned}$$

which is greater than 0, if $M > \hat{M}$, and if

$$1-\beta > \left(\frac{M}{4}+\frac{N}{9}\right)^{-1}\frac{N[6M\tau_\epsilon+5(N\tau_\epsilon+\tau_\theta)][2M\tau_\epsilon+(N\tau_\epsilon+\tau_\theta)]}{9[3M\tau_\epsilon+2(N\tau_\epsilon+\tau_\theta)]^2}. \qquad \text{(A57)}$$

$\hat{M}$ is defined in equation (20). ■

Proof for Proposition 8

Proof. By symmetry, we focus only on firm A. To ensure supplying information is a dominant strategy, we need following condition

$$\text{“Given } B \text{ does not supply, } A \text{ will supply”} \Rightarrow \mathbb{E}\Pi_A^1(0,0) \leq \mathbb{E}\Pi_A^1(0,M) + t_B, \tag{A58}$$

and

$$\text{“Given } B \text{ supplies, } A \text{ will supply”} \Rightarrow \mathbb{E}\Pi_A^1(M,0) \leq \mathbb{E}\Pi_A^1(M,M) + t_B. \tag{A59}$$

Hence, if

$$\max\left\{\mathbb{E}\Pi_A^1(0,0) - \mathbb{E}\Pi_A^1(0,M), \mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(M,M)\right\} \leq t_B, \tag{A60}$$

then both conditions are met. After computation, we get the following

$$\mathbb{E}\Pi_A^1(0,0) - \mathbb{E}\Pi_A^1(0,M) = \frac{MN\tau_\epsilon\left[6M\tau_\epsilon + 5(N\tau_\epsilon + \tau_\theta)\right]}{9\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right]^2}, \tag{A61}$$

$$\mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(M,M) = \frac{5MN\tau_\epsilon}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)}. \tag{A62}$$

We compute and obtain

$$\begin{aligned}&\left[\mathbb{E}\Pi_A^1(0,0) - \mathbb{E}\Pi_A^1(0,M)\right] - \left[\mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(M,M)\right] \\ &= \frac{M^2N\tau_\epsilon^2\left[3M\tau_\epsilon + 4(N\tau_\epsilon + \tau_\theta)\right]}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right]^2} > 0.\end{aligned} \tag{A63}$$

Therefore, condition (A60) implies

$$\frac{MN\tau_\epsilon\left(6M\tau_\epsilon + 5(N\tau_\epsilon + \tau_\theta)\right)}{9\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right)^2} \leq t_A, t_B. \tag{A64}$$

To ensure demanding information is a dominant strategy, we need following conditions,

$$\text{“Given } B \text{ does not demand, } A \text{ will demand”} \Rightarrow \mathbb{E}\Pi_A^1(0,0) \leq \mathbb{E}\Pi_A^1(M,0) - t_A, \tag{A65}$$

and

$$\text{“Given } B \text{ demands, } A \text{ will demand”} \Rightarrow \mathbb{E}\Pi_A^1(0,M) \leq \mathbb{E}\Pi_A^1(M,M) - t_A. \tag{A66}$$

Hence, if

$$t_A \leq \min\left\{\mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(0,0), \mathbb{E}\Pi_A^1(M,M) - \mathbb{E}\Pi_A^1(0,M)\right\}, \tag{A67}$$

then both conditions are met. After computation, we have

$$\mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(0,0) = \frac{M\tau_\epsilon\Theta}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right]^2}, \tag{A68}$$

$$\mathbb{E}\Pi_A^1(M,M) - \mathbb{E}\Pi_A^1(0,M) = \frac{M\tau_\epsilon\left(9M + 4N\right)}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)}. \tag{A69}$$

where

$$\begin{aligned}\Theta =& 81M^3\tau_\epsilon^2 + 33M^2N\tau_\epsilon^2 + 108M^2\tau_\epsilon(N\tau_\epsilon + \tau_\theta) + 44MN\tau_\epsilon(N\tau_\epsilon + \tau_\theta)\\ &+ 36M(N\tau_\epsilon + \tau_\theta)^2 + 16N(N\tau_\epsilon + \tau_\theta)^2.\end{aligned} \tag{A70}$$

Direct computation shows that

$$\begin{aligned}&\left[\mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(0,0)\right] - \left[\mathbb{E}\Pi_A^1(M,M) - \mathbb{E}\Pi_A^1(0,M)\right] \\ &= -\frac{M^2N\tau_\epsilon^2\left(3M\tau_\epsilon + 4(N\tau_\epsilon + \tau_\theta)\right)}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right]^2} < 0.\end{aligned} \tag{A71}$$

Therefore, condition (A67) implies

$$t_A, t_B \leq \frac{M\tau_\epsilon\Theta}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left[3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right]^2}. \tag{A72}$$

Since

$$\begin{aligned}&\overbrace{\left[\mathbb{E}\Pi_A^1(M,0) - \mathbb{E}\Pi_A^1(0,0)\right]}^{\text{upper bound}} - \overbrace{\left[\mathbb{E}\Pi_A^1(0,0) - \mathbb{E}\Pi_A^1(0,M)\right]}^{\text{lower bound}} \\ &= \frac{M\tau_\epsilon\Theta'}{36\left(M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(2M\tau_\epsilon + N\tau_\epsilon + \tau_\theta\right)\left(3M\tau_\epsilon + 2(N\tau_\epsilon + \tau_\theta)\right)^2}\end{aligned} \tag{A73}$$

where

$$\begin{aligned}\Theta' =& 81M^3\tau_\epsilon^2 - 15M^2N\tau_\epsilon^2 + 108M^2\tau_\epsilon(N\tau_\epsilon + \tau_\theta) - 20MN\tau_\epsilon(N\tau_\epsilon + \tau_\theta)\\ &+ 36M(N\tau_\epsilon + \tau_\theta)^2 - 4N(N\tau_\epsilon + \tau_\theta)^2.\end{aligned} \tag{A74}$$

Θ' is larger than 0 when $M \to \infty$. Thus, the set of transfers is not empty for large M. ■

Appendix B: Microstructure of data sales

In this appendix, we describe two mechanisms through which an independent profit-maximizing data vendor implements a particular data allocation (m_A, m_B). In the first mechanism, the vendor simultaneously offers two take-it-or-leave-it contracts to both firms, while in the

second mechanism, the vendor offers contracts sequentially. Both games feature a unique equilibrium in terms of data allocation, and the equilibrium data prices lead to equation (9) in the main text.

B.1. Simultaneous offering with contingent contracts

The data prices in the contracts offered by the data vendor are contingent on data allocations. Putting it in context, suppose that the data vendor is a platform such as Amazon or eBay that has accumulated the date-0 consumer data. For illustrative purpose, we assume that the data vendor wants to implement data allocation $(m_A, m_B) = (1000, 1000)$. Then, the data vendor may present the following two offers to firms, for example:

Contract A (on A-type consumer data): "*If you purchase 1000 data points about A-type consumers and no one else buys any data, then you pay \$30; and if you purchase 1000 data points about A-type consumers and someone else also buys some data, then you pay \$40.*"

Contract B (on B-type consumer data): "*If you purchase 1000 data points about B-type consumers and no one else buys any data, then you pay \$30; and if you purchase 1000 data points about B-type consumers and someone else also buys some data, then you pay \$40.*"

Given that only firm A is interested in contract B and only firm B is interested in contract A, the above two contracts are effectively the following: "*If the data allocation is $(m_A, 0) = (1000, 0)$, then firm A pays $t_A = 30$; and if the data allocation is $(m_A, m_B) = (1000, 1000)$, then firm A pays $t_A = 40$;*" and "*If the data allocation is $(0, m_B) = (0, 1000)$, then firm B pays $t_B = 30$; and if the data allocation is $(m_A, m_B) = (1000, 1000)$, then firm B pays $t_B = 40$.*" The contents of both contracts are observable to both firms.

These contracts correspond to the concept of "smart contracts" in the context of FinTech. Smart contracts are computer programs that execute "if this happens then do that," run and verified by many computers to ensure trustworthiness in a blockchain environment.[10] For instance, with a "Turing complete" coding system, in theory, any contingent contract

[10] See Cong and He (2018) for more discussions on smart contracts. An informal discussion on this concept can be found at: https://bitsonblocks.net/2016/02/01/a-gentle-introduction-to-smart-contracts/.

can be implemented via an Ethereum smart contract. If these contingent contracts are available, then the data vendor can use them to modify the payoff matrix of firms at the information purchase stage, such that the unique Nash equilibrium leads to data allocation (m_A, m_B) (see Section 4 for more details). There are multiple contingent contracts that implement (m_A, m_B), but for all of these contracts, the vendor's ultimate profits are given by $C_A(m_A, m_B) + C_B(m_A, m_B)$.

B.2. Sequential offering with simple contracts

In the absence of "smart contracts," we can consider a four-stage game in which the vendor offers *simple* contracts sequentially. The game's extensive form is drawn in Figure B1. In the first stage, the data vendor contacts firm A and offers a take-it-or-leave-it contract which states that "firm A can pay a cost t_A to buy an amount m_A of data." In the game, the cost t_A is the vendor's choice variable with an action space $\mathbb{R}_+$, and the data amount m_A is a fixed parameter that is exogenous to the game. Receiving the offer, firm A decides to accept or reject the offer in the second stage. If firm A accepts the offer, then it will pay t_A and purchase m_A amount of data, and if it rejects, it will not buy data.[11] In the third stage, observing firm A's choice, the data vendor then offers another contract to firm B which says that "firm B can pay a cost t_B to buy an amount m_B of data." In the last stage, firm B decides to take or reject the offer.

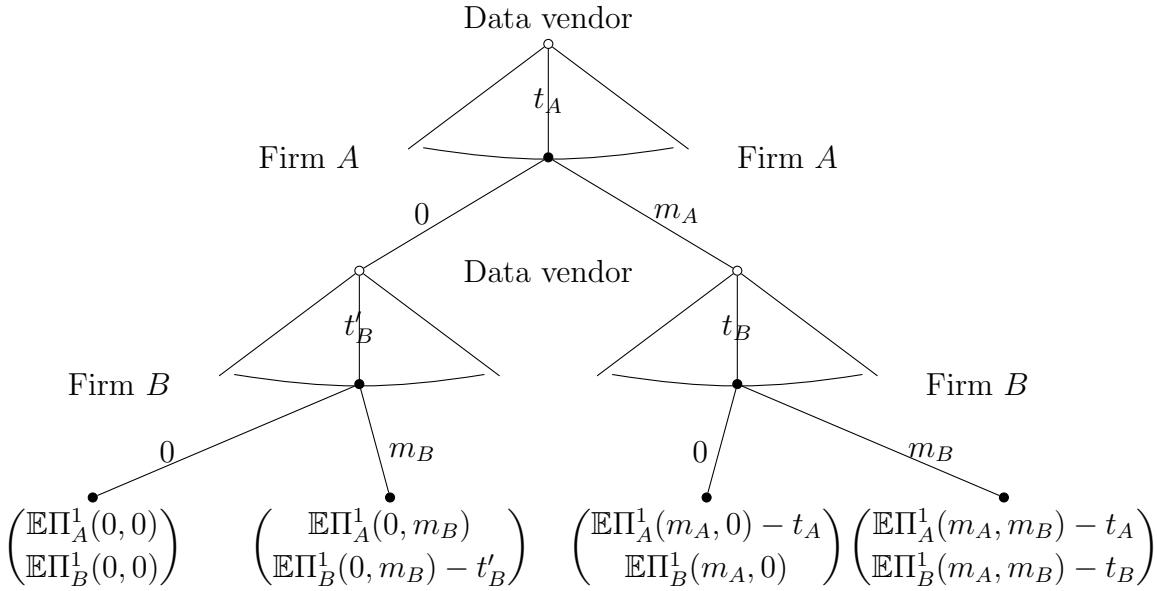
Lemma B (Sequential contract offering)

In the four-stage game described above, there is a unique sequential Nash equilibrium in which the data vendor sets data price $t_A^* = C_A(m_A, m_B)$ in the first stage, firm A accepts the offer in the second stage, then the vendor sets data price $t_B^* = C_B(m_A, m_B)$ in the third stage, and firm B also accepts the offer in the last stage.

Intuitively, the data vendor would like to sell its data to both firms and thus, it will choose

[11] When a firm is indifferent between buying and not buying data, we assume that the firm will always choose to buy data. This makes sense because if not, the data vendor can always slightly lower the data price to get a positive profit.

Figure B1: Sequential contract offering



the right data prices such that both firms would like to purchase the data. More formally, we solve the model by backward induction. In the last stage, at each node, firm B will buy data if and only if the data price t_B is sufficiently low. In the third stage, anticipating the optimal response of firm B in the last stage, the data vendor will charge the price just to the level at which firm B does not want to switch from buying, so that firm B always buys data on any possible equilibrium path. Back to stage 2, anticipating that firm B will always buy data, firm A will buy data if and only if the data price t_A is no larger than $C_A(m_A, m_B)$. To achieve the maximum profit, the data vendor sets data price at $t_A^* = C_A(m_A, m_B)$. Thus, the four-stage game implements the data allocation (m_A, m_B).

In reality, the data vendor could implement sequential sales in several ways. First, the data vendor as a monopolist has the full discretion over the timing of sales, so that it can literally do sequential data sales by contacting firms one by one. This practice is popular in many real over-the-counter (OTC) markets. Second, the data vendor can allow one firm to place a pre-order for data and then set the late-stage data price conditioning on the pre-order outcome. Pre-ordering is customary in book and video game industries. The development of Initial Coin Offerings (ICOs), in particular the utility-token sales, can facilitate pre-ordering

by directly providing token buyers with future access to the token seller's products.[12]

[12]For more discussions regarding ICOs, see Li and Mann (2018).