Privacy, Information Acquisition, and Market Competition*

Soo Jin Kim†

ShanghaiTech University

November 3, 2021

Abstract

This paper analyzes how data-driven vertical integration between a platform and one downstream seller affects market outcomes when sellers with asymmetric targeting skills target advertisements to individuals. I show that data-driven vertical integration can result in the incumbent's exclusive use of data. Therefore, a market entrant with worse targeting technology than an incumbent is disproportionately harmed by such integration. The welfare analysis shows that integration reduces welfare if consumers' privacy concerns are relatively high. Therefore, individually optimal decisions about data disclosure might not be socially optimal when aggregated.

Keywords Privacy, Information Acquisition, Data Intermediary, Targeted Advertising, Vertical Integration
JEL Codes D21; D22; D83; L15; L22; L42; L52

*I would like to thank Jay Pil Choi, Anna D'Annunzio, Jon Eguia, Hanming Fang, Thomas Jeitschko, Kyoo il Kim, Aleksandr Yankelevich, Pinar Yildirim, Hanzhe Zhang, and seminar participants at the Federal Communications Commission, Midwest Economic Theory Conference (2017), Canadian Economics Association (2017), Singapore Economic Review (2017), International Industrial Organization Conference (2018), Sauder School of Business (2018), Tilburg Law and Economics Center (2018), DIW Berlin (2018), Universitat Autónoma de Barcelona (2018), Korea Development Institute (2018), Korea Institute of Finance (2018), Bank of Korea (2018), and Yonsei Barun ICT Research Center (2018) for the valuable discussions and comments.

†sjkim@shanghaitech.edu.cn; School of Entrepreneurship and Management, ShanghaiTech University, Room 436, 393 Middle Huaxia Rd, Shanghai, China.

1 Introduction

Platforms such as internet service providers (ISPs) and social media also serve as data intermediaries that collect and sell users' personal information directly to third parties, such as downstream sellers, or use it to deliver more targeted advertising (ads) to consumers.[1] Each consumer's privacy concerns determine the amount of information available on the platforms, in turn influencing the overall effectiveness of ad targeting for each seller that obtains data from the platforms. Given that more effective targeted ads can attract a greater number of consumers, data availability for sellers affects the competitive structure of the downstream market.

If each seller has the most current information about potential customers, its targeted ads become more effective; then, consumers receive greater match benefits from this seller because they immediately obtain relevant information from these targeted ads. This potential benefit from a loss of privacy can also be asymmetric. Given that compared to entrants, incumbents have better initial targeting technology developed based on previous sales experience or existing customer data, a consumer is likely to receive a much lower match benefit from small sellers or from market entrants than from incumbents. For example, suppose that a major retailer such as Walmart.com and a new retailer buy the same dataset from a platform. Walmart.com will be better able to target consumers than the entrant because it can combine these new data with existing customer data.

An important market outcome related to this asymmetry is data-driven vertical integration. There are multiple instances of vertical integration between the platform and a downstream content provider or online retailer. In particular, vertical integration can be motivated by the seller's desire to obtain a greater collective amount of exclusive data. For example, when AT&T and WarnerMedia[2] announced their intention to merge in October 2016, they noted that the merger would benefit consumers by providing better-targeted ads based on extensive customer data. Most previous studies focus on how such data-driven integration could generate much greater privacy concerns on the consumer side (e.g., Chirita (2020); Binns and Bietti (2020)). However, none of these papers investigates the effects of such data-driven mergers, which alter data availability, on the relevant market competition, although the market consequences ultimately affect consumers. That is, if integration changes a platform's data-sharing practice for unaffiliated sellers, it changes data availability on the market, which plays a pivotal role in attracting more customers. Hence, this study adopts a different stance

[1]For example, AT&T sells advertising based on customer data via AdWorks, which is its own ad network; therefore, there is no need to sell subscribers' data to third parties so that they can sell targeted ads. However, small ISPs that do not own their own ad networks could contract with third parties and share customer data for revenue-generating purposes.

[2]Formerly Time Warner Inc.

from previous studies by focusing on how this integration affects the competitive structure in the downstream (product-selling) market. In the previous example, by maintaining exclusive use of data, WarnerMedia can target consumers much more effectively, thereby attracting more advertisers to choose WarnerMedia's content as a channel for advertising.[3] The Amazon-Whole Foods acquisition deal is another example of data-driven integration: using Amazon's extensive volume of transaction data, Whole Foods could offer better product suggestions to consumers, thereby attracting consumers away from less-informed competitors, such as local organic grocery stores. By conferring an unfair advantage upon the integrated downstream firm, mergers could lead to antitrust concerns.

Such potential anticompetitive threats are greater if a small or entrant seller is prevented from using the platform's data due to integration between the platform and the incumbent. Indeed, the platform's incentive to vertically integrate with one seller to share data depends on not only the total amount of data available for sales, which is determined by consumers' overall privacy concerns, but also the targeting asymmetry between sellers.

The latter point—that is, how asymmetry among downstream sellers affects the equilibrium outcome of data-driven vertical integration—raises several important research questions. How does initial targeting asymmetry between sellers affect the platform's incentive to engage in data-driven vertical integration with one seller? Is an entrant with weaker targeting skills, indicating that it demands considerably more personal data than an incumbent does, disproportionately affected by the market outcome? How does the asymmetry ultimately affect consumers? How do consumers' privacy concerns affect market outcomes and welfare implications?

To answer these questions, I develop a model in which two sellers—an incumbent and entrant—are asymmetric with respect to their initial targeting technologies, as described in Section 2. Without loss of generality, the incumbent is assumed to have better initial targeting skills than the entrant does. Products offered by sellers are vertically differentiated: either the incumbent or entrant provides a high-quality product. The sellers, if not integrated with the platform, decide whether to purchase consumer data from the platform to create targeted ads and subsequently engage in price competition to attract consumers who are heterogeneous in terms of the product quality valuation. A novel aspect of this model is that it permits the platform to integrate with one downstream seller, and the integrated firm then decides whether to exclusively share data with its own affiliated seller or to sell

[3] In this example, however, WarnerMedia itself is a platform, so the merger can be considered a horizontal platform-to-platform merger rather than vertical integration. Nevertheless, the primary message concerning the effect of collective and exclusive user data on better targeting remains valid. Another example that might be more relevant is the integration of WarnerMedia and HBO. As an integrated firm, HBO can more effectively target customers of premium television content than other rival content providers, such as Showtime.

data to the unaffiliated seller. Additionally, the model considers a peculiar aspect of data as nonrival goods, following Fang and Kim (2021), such that providing data, as inputs, to multiple sellers does not incur additional costs unlike in the case of rival goods. This setup allows me to investigate how such data-driven integration affects market competition through the seller's data acquisition status, which ultimately affects consumers, and how it differs from other typical vertical integration followed by input (as a rival good) foreclosure.

The main results show that the platform ends up selling data non-exclusively to both the incumbent and entrant at different prices without vertical integration due to the unobservability of a secret contracting environment, as described in Section 3.1. If a data-driven merger is permitted, as in Section 3.2, the incomplete information issue arising from the secret contracting environment is resolved because vertical integration allows both affiliated and unaffiliated sellers to have contract-related information implicitly.

In the game with vertical integration, the results depend on which seller provides a high-quality product. If the incumbent is superior in both targeting skills and product quality, the platform will vertically integrate with the incumbent and prevent the entrant from obtaining access to customer data. This outcome occurs because the incumbent's better targeting skills, combined with exclusive use of the data, allow the integrated firm to sell more affiliated products in the downstream market. These market dominance effects are sufficiently large; thus, the platform and incumbent are integrated, and the integrated firm precludes the unaffiliated entrant from gaining data access. If the entrant provides a high-quality product, the platform chooses either the incumbent or entrant, depending on the amount of data available, and the integrated firm always monopolizes data access. In particular, if sufficient data are available, the platform integrates with the incumbent because it can easily dominate the market using its better targeting technology combined with an extensive amount of exclusive data. That is, even when an entrant has an advantage in product quality, it can be foreclosed from data access. Given that the entrant has worse targeting skills, vertical integration followed by data foreclosure, which grants a considerable competitive advantage to the affiliated incumbent, can be particularly anticompetitive.

The welfare analysis results suggest that such integration can lower total social welfare, as shown in Section 4. Specifically, the market-driven integration equilibrium outcome and the resulting data foreclosure affect society negatively if consumers' privacy concerns are relatively high, meaning that less data are available for targeted ads. Intuitively, vertical integration can either enhance or reduce welfare. On the one hand, vertical integration that leads to asymmetric data acquisition for sellers harms consumers and the unaffiliated seller because the exclusive use of data allows the integrated firm to dominate the downstream market, resulting in a lack of competition. On the other hand, the exclusive use of data combined with better targeting skills (if integrated with the incumbent) or greater product quality (either one of the sellers) significantly enhances the integrated firm's profit. If overall privacy concerns are high, meaning that only a small amount of data becomes available, the positive effect of vertical integration for the integrated firm, driven by better match benefits, is negligible. Thus, vertical integration and the resulting data foreclosure are more likely to become welfare reducing as data become limited. In this respect, individually optimal decisions on data disclosure might not be socially optimal when aggregated, particularly when consumers have increased privacy concerns. Regarding the welfare implications, I show how the nonrivalry as unique feature of data plays an important role: if the nonrival feature of data is not taken into consideration, we may reach a misleading conclusion by overestimating the benefit of the exclusive use of data. This comparison between rival and nonrival inputs emphasizes the necessity of data-specific policy concerning data-driven integration followed by data (as an input) foreclosure.

In Section 5, a model extension shows that the main findings are robust to a different model specification. Based on these theoretical findings, policy implications are suggested in Section 5.2. Finally, Section 6 provides concluding remarks.

Previous Literature The stream of literature most closely related to my work includes research on the effect of privacy on a market's competitive structure. Acquisti and Varian (2005), Belleflamme and Vergote (2016), Conitzer et al. (2012), Fudenberg and Tirole (2000), Koh et al. (2017), Taylor (2004), Villas-Boas (1999), and Villas-Boas (2004) study how firms use consumer data to set prices. However, these papers either assume a monopolistic seller and thus do not examine how privacy protection or information availability affects market competition or focus on dynamic price discrimination, which I do not consider in this paper.

In models with symmetric firms, Taylor and Wagman (2014) examine how privacy enforcement produces different competitive market outcomes depending on the individual context and industries. Shy and Stenbacka (2016) suggest that there is a nonmonotonic relationship between the degree of privacy protection and equilibrium profits. Casadesus-Masanell and Hervas-Drane (2015) study firms that earn revenues from consumer data sales and from product sales to consumers who value privacy; in their model, as in that of Koh et al. (2017), consumers decide how much information to provide. Montes et al. (2018) also endogenize privacy by allowing consumers to anonymize themselves for a cost, and then they analyze how privacy concerns and the resulting information availability affect competing firms' decisions and social welfare. De Cornière and Taylor (2020) provide a general theoretical framework that could be flexibly applied to various contexts to analyze how data usage affects competition.

Asymmetries between firms are persistent for a variety of reasons arising from different previous sales experiences or the scope of product offerings. By considering such asymmetries, this paper illuminates important implications that are absent from a symmetric setting: the entrant might be disproportionately harmed by market outcomes. In this respect, Campbell et al. (2015), who demonstrate that small firms or entrants, as specialists rather than generalists, can be adversely affected by privacy regulations that impose unit costs on all firms, share one of the main implications of this paper. Similarly, Braulin and Valletti (2016) also model vertically differentiated sellers to determine how exclusive data sales affect consumer and social welfare. In contrast, my primary concern is further asymmetries—not only in sellers' product quality but also in their targeting match skills. By including such asymmetries, I offer a microfoundation for understanding how the amount of data availability for downstream sellers affects welfare.

More importantly, none of the papers from the literature places a primary focus on more diverse market outcomes related to privacy concerns, such as data-driven vertical integration, which I consider in this paper. There are several notable exceptions, such as De Cornière and Taylor (2020), who study how data-driven mergers between the monopolist platform and a symmetric downstream firm affect consumers as one of their model applications; Kim et al. (2019), who analyze how access to consumer data, which enables personalized pricing, affects the overall welfare of horizontal and vertical mergers; Gu et al. (2019), who show how the exclusive use of data affects sellers' incentives to either lead or follow price competition; and Chen et al. (2020), who study the effects of data-driven tech mergers on two linked markets—one for data collection and another for data application. My paper nevertheless differs from the aforementioned studies insofar as I focus on the relationship between privacy concerns and the downstream market's competitive structure, which affects data-driven vertical integration outcomes under an asymmetric seller setup.

2 The Model

The players in this game are as follows: a monopoly platform as a data collector, an incumbent seller, an entrant seller, and a unit mass of consumers. All of the firms (platform and sellers) have basic consumer information, such as an email address, gender, and date of birth. The amount of basic information is normalized to one.

Data Collection and Targeted Ads Depending on a platform's roles, it can collect different categories of data, such as location, browsing history, and financial data. The total aggregate amount of data provided by the platform can be sold to downstream sellers. In the paper, I interpret the amount of data as the fraction of relatively privacy insensitive consumers: given heterogeneous privacy sensitivity, only a certain proportion of consumers who have fewer privacy concerns disclose their personal information to the platform by posting daily activities on Facebook, for example.

For downstream sellers, data obtained from the platform can be used for either personalized pricing of products or better-matched targeted ads. In this paper, I focus on the targeting aspects of data since having more data about consumers' current wants or needs helps sellers recommend more suitable products to consumers, allowing them to raise product prices, and I assume that sellers do not engage in price discrimination for products.

Data-driven Merger Data-driven mergers indicate any vertical integration that immediately allows the integrated downstream seller to obtain the platform's data. In this sense, this paper is broadly related to the literature on vertical integration and input foreclosure (e.g., Chipty (2001); Rey and Tirole (2007); Reisinger and Tarantino (2015)) in that I consider data as inputs. However, data-driven integration and data foreclosure yield different implications because data are particular inputs provided by consumers: if consumers fail to consider how their data disclosure decisions, which depend on privacy concerns, ultimately affect the downstream market (in which data are used as necessary inputs), the market equilibrium outcome can be welfare reducing through the effects of integration, as shown in Section 4.

In addition, the model considers the unique features of data as nonrival goods: providing data, as inputs, to multiple sellers does not incur additional costs unlike in the case of rival goods as explained later. Given that the nonrivalry of data generates cost-reducing efficiency, I show that traditional models for rival goods may lead to a misleading welfare implication. By clarifying this hidden mechanism with data-specific aspects, my paper provides plausible policy remedies regarding data-driven integration with not only direct regulation on the integration itself but also indirect instruments concerning data collection.

Consumer There is a continuum of consumers indexed by $i \in [0,1] \times [0,1]$. Each consumer $i \in [0,1] \times [0,1]$ has a two-dimensional type τ_i and θ_i, where both types are exogenously given and independently distributed. First, τ_i denotes each consumer's privacy sensitivity, which is distributed over $[0,1]$ with distribution function F and density f. Consumer i becomes more privacy sensitive as τ_i increases. For notational convenience, let $\mathcal{D}$ denote the set of privacy insensitive consumers who disclose as much personal information as possible and $\mathcal{ND}$ denote the set of privacy sensitive consumers who do not disclose any additional personal information. The portion of each set is endogenously determined by consumers' decisions: each consumer on a continuum of τ_i compares the benefits and privacy nuisance costs from disclosing personal information to the platform and makes an optimal decision. Second, a consumer has heterogeneous valuation with respect to the product quality provided by sellers, denoted by θ_i, which is uniformly distributed over $[0,1]$. Again, each consumer on a continuum of θ_i decides from which seller (Incumbent or Entrant) to purchase a product given that each seller provides a vertically differentiated product.

First, any consumer obtains an immediate benefit from enjoying the platform's services; e.g., Facebook users enjoy the social networking service. Furthermore, as more users disclose more information to the platform, all of the other users benefit due to the network effect. In this sense, the immediate benefit increases with the total amount of detailed information available on the platform, which is an increasing function of the portion of consumers who disclose information, $\mathbb{P}(i \in \mathcal{D})$. In addition, any user $i \in \mathcal{D}$ who provides detailed personal information obtains greater benefits, such as a greater networking benefit, than $i \in \mathcal{ND}$ who provides only basic information. However, $i \in \mathcal{D}$ faces a much higher nuisance cost from privacy loss than $i \in \mathcal{ND}$, which increases in privacy sensitivity τ_i. Normalizing the net utility for $i \in \mathcal{ND}$ to zero,[4] the utility for each consumer i from the platform is given as follows:

$$v_i^{pf} = \begin{cases} v(\mathbb{P}(i \in \mathcal{D})) - \psi(\tau_i) & \text{if } i \in \mathcal{D} \\ 0 & \text{if } i \in \mathcal{ND}, \end{cases} \tag{1}$$

where the superscript pf denotes the platform; $v(\mathbb{P}(i \in \mathcal{D}))$ denotes the immediate benefit from disclosing information, with $v' > 0$ and $v'' \leq 0$; $\psi(\tau_i)$ denotes the nuisance cost, with $\psi' > 0$ and $\psi'' \geq 0$.

Because seller j's targeting effectiveness increases with the amount of consumer data, a consumer's information disclosure decision also affects the utility from the purchase of a product: any $i \in \mathcal{ND}$, the detailed information of which is not available, is likely to receive untargeted ads, whereas any $i \in \mathcal{D}$ who provides personal information receives targeted ads. As targeted ads suggest products that are better suited to consumers, I assume that $i \in \mathcal{D}$ obtains greater match benefit than $i \in \mathcal{ND}$. Normalizing the match benefit for $i \in \mathcal{ND}$ to zero, the utility specification is given as follows:[5]

$$u_{ij} = \theta_i q_j + \mathbb{1}_{\{i \in \mathcal{D}\}} \times m_j(D_j) - P_j, \tag{2}$$

where q_j is seller j's overall product quality, the indicator function $\mathbb{1}_{\{i \in \mathcal{D}\}}$ is one if consumer i discloses personal information, and $m_j(D_j)$ is the additional match benefit obtained from seller j, which is an increasing and concave function of the amount of data that seller j uses denoted by D_j.[6] Here, $i \in \mathcal{D}$

[4] This normalization is for simplicity and is not crucial for deriving the main results.

[5] As long as the match benefit for $i \in \mathcal{D}$ is greater than that from $i \in \mathcal{ND}$, the qualitative results from the current specification always hold; thus, this normalization is harmless.

[6] The assumption that having more data improves targeting effectiveness but at a decreasing rate is typical, as in

receiving targeted ads enjoys a greater match benefit as D_j increases, while $i \in \mathcal{ND}$ only receives untargeted ads. P_j denotes the price of products from seller j.

Consumer's additional match benefit As in Levin and Milgrom (2010), suppose that a male subscriber in his 20s recently had a baby and is therefore planning to buy a minivan for his family. AT&T, as a cable TV operator (platform), has privacy sensitive information about subscribers' viewing patterns, such as whether the subscriber regularly watches *SpongeBob* rather than *The Walking Dead*. Assume that Honda purchases this additional information about the subscriber, while Subaru does not. Then, Honda would produce a minivan ad, whereas Subaru would produce a sports car ad by assuming that men in their 20s prefer sports cars in general. Ultimately, when the consumer goes to Honda, he knows exactly what features to look for in the car type that he wants, whereas with Subaru, he must still look up additional details. Thus, even if he knows the price of the Subaru and that of the Honda, the transaction with Honda results in greater benefit in terms of a better match. Note that I do not consider any nuisance cost arising from targeted ads and assume that all consumers benefit from a better match. The model with a nuisance cost driven by targeted ads qualitatively yields the same results as long as the targeted match benefit outweighs the cost.

Two independent utility specifications One might question the additive separable utility specification of information disclosure and product purchasing. In this setup, consumers consider only the immediate benefits of disclosing information and do not consider any potential future benefits arising from better-targeted ads when making information disclosure decisions. This assumption is reasonable for many real examples of platforms. For example, when a consumer posts the news of the birth of his baby on Facebook, he is more likely to do so to spread good news to his friends than to see more relevant ads for baby products. Similarly, Amazon customers subscribe to items that they want to purchase on a regular basis, such as laundry detergent, to receive discounts and schedule autodeliveries through Amazon's program called *Subscribe & Save*. Although this subscription service gives Amazon additional personal information about customers' daily needs, customers are unlikely to participate in the program only to receive more targeted ads.[7]

Platform The amount of data available on the platform depends on how likely each consumer is to disclose his information to the platform, i.e., whether a consumer is privacy sensitive or privacy insensitive. The platform sells only detailed information, such as users' relationship status, through take-it-or-leave-it offers. Normalizing the total amount of detailed demographic information that the platform obtains from each consumer to one, the platform sells $\mathbb{P}(i \in \mathcal{D})$ amount of detailed

asymptotic learning theory. See Bajari et al. (2019) for a relevant discussion.

[7]For those who are interested in the case of consumers with perfect foresight, see Section 5.1.

information to any seller that wants to buy. The platform earns profits only from selling user data to any seller. Additionally, the platform is allowed to price discriminate for data: the price charged to seller j can differ from that charged to seller $-j$, where $j \neq -j$.[8] By optimally setting the per unit data price C_j, the platform solves the following profit maximization problem:

$$\max_{C_j, C_{-j}} \quad \pi_{pf}(C_j, C_{-j} | \mathbb{P}(i \in \mathcal{D})) = \begin{cases} 0 & \text{if no seller buys data} \\ \mathbb{P}(i \in \mathcal{D})(C_j - \phi) & \text{if only seller } j \text{ buys data} \\ \mathbb{P}(i \in \mathcal{D})(C_j + C_{-j} - \phi) & \text{if both sellers buy data,} \end{cases} \tag{3}$$

where the subscript pf denotes the *platform* and ϕ is the marginal cost of providing data.[9] Note that the cost structure here captures the nonrival feature of data, as similarly assumed in Fang and Kim (2021). Unlike typical rival goods, the same set of data can be provided to both sellers as nonrival goods without incurring additional marginal costs; thus, even when both sellers buy data, the platform incurs ϕ in marginal cost, not $2 \times \phi$. As described in Section 4.1, the nonrival feature yields different welfare results from the case of rival goods, which necessitates data-specific policy regarding data-driven vertical integration.

In some cases, the platform serves as more than a data broker. For instance, some platforms, such as Amazon, extract a portion of the profits from sellers by charging proportional usage fees. In this regard, Appendix A.1 analyzes how different profit models for the platform affect the results. Note that this model extension has the same qualitative implications as the main model.

Sellers Each seller $j \in \{I\text{ncumbent}, E\text{ntrant}\}$ sells a set of products to consumers. The set of products offered by each seller is vertically differentiated in terms of product quality, denoted by q_j. The match benefit is denoted by $m_j(D_j)$, where D_j is equal to $\mathbb{P}(i \in \mathcal{D}) + 1$ (i.e., detailed data provided by the platform + basic information) if seller j buys data from the platform or one otherwise. If seller j decides to buy data from the platform, the seller pays per unit data price C_j. For simplicity, I make the following assumption about seller j's match benefit $m_j(D_j)$.

Assumption 1. $m_j(D_j) = (1 + D_j)^{\alpha_j}$ *where* $\alpha \equiv \alpha_E \leq \alpha_I = 1$ *and* $\alpha \in [0, 1]$,

Assumption 1 implies that the targeting technology for seller I is better than that for seller E, which is captured by the concavity of the consumer's match benefit function; Seller E lacks sales experience and the existing customer data such that its overall targeting skills are worse than that of seller I. The model captures the worse targeting skills of seller E by considering more diminishing

[8]For instance, Google AdSense and Facebook Ads offer different quality tiers with different pricing.

[9]The choice variable C can be considered the data price if the platform sells data to third parties. If the platform is not allowed to sell data and can only use the data to create targeted ads, C can be considered a per unit advertising (intermediation) fee. Refer to https://vendorcentral.amazon.com/ and https://www.facebook.com/business/ads/pricing.

match benefits for consumers to data, i.e., $\alpha_E \leq \alpha_I$. For simplicity, I normalize α_I to one and denote α_E as $\alpha \in [0, 1]$.

Relationship with different specifications It is worth noting that the main implications from utility specification in Equation (2), combined with Assumption 1, hold under different utility specifications. For instance, it can also be the case in which seller I has more pre-existing data than seller E such that $m_I(D_I) = \delta + D_I$ and $m_E(D_E) = (1 + D_E)^{\alpha}$, where $\delta > 1$. As shown in Appendix A.2, the main findings still carry over in this extension. Additionally, as described in Cremer and Thisse (1991), Equation (2) is conceptually analogous to the horizontal differentiation model. The vertical differentiation specification in this paper, which views more data as an additional benefit for consumers by suggesting more suitable product offers, can be regarded as an effective price discount given the price-match quality ratio. In this sense, the main utility specification can be alternatively interpreted to be a lower transportation cost in the horizontal differentiation model.

Regarding product quality, either I or E can provide a high-quality product. Therefore, I examine both cases and derive the equilibrium outcomes accordingly. For simplicity, the low-quality level is normalized to one, and the high-quality level is denoted q, that is, $1 = q_{-j} < q_j \equiv q$, where $j \neq -j$. Each seller's profit maximization problem is defined as follows:

$$\max_{P_j, D_j} \quad \pi_j = P_j X_j(\mathbf{P}, \mathbf{D} | \alpha, q) - \mathbb{1}_{\{j \text{ obtains}\}} C_j \times \mathbb{P}(i \in \mathcal{D}), \tag{4}$$

where $X_j(\mathbf{P}, \mathbf{D} | \alpha, q)$ is j's aggregate market share, $\mathbf{P} = (P_1, P_2)$, and $\mathbf{D} = (D_1, D_2)$.

Contracting Environment I consider a secret contracting environment with incomplete information. As noted in previous studies (e.g., Hart and Tirole (1990); O'Brian and Shaffer (1992); McAfee and Schwartz (1994); Rey and Vergé (2004); Reisinger and Tarantino (2015)), firms facing opportunism are likely to break the commitment by renegotiating the terms of the contract. Indeed, the contracting environment is incomplete due to practical challenges in writing state-contingent contracts. To consider this limited commitment environment, reflecting possible side contracting, a secret contract in which the platform secretly and simultaneously makes a data offer to each seller j is considered.

Timing and Solution Concept All information, including the distribution of τ_i and θ_i, is common knowledge, while the true realizations of τ_i and θ_i for each i are private information. I investigate two games, namely, those with and without data-driven vertical integration. In both games, firms form beliefs about consumers' valuations given their identification status: disclosing or not disclosing for τ_i and choosing the incumbent or entrant for θ_i. In the case of no vertical integration, the timing of the game proceeds as shown in Figure 1. In the vertical integration game, I include an additional stage in which the platform decides with whom to vertically integrate at the beginning of the second stage, as in $2'$ in parentheses. Thereafter, the game proceeds as before except that, in the third stage, the affiliated seller freely obtains data from the platform, while the unaffiliated seller decides whether to buy data. After each stage, the consumer's choice of action is observed by every agent.

The solution concept that I use for this game is the perfect Bayesian Nash equilibrium (PBE) for multiperiod games with observed action, as in Fudenberg and Tirole (1991).[10]

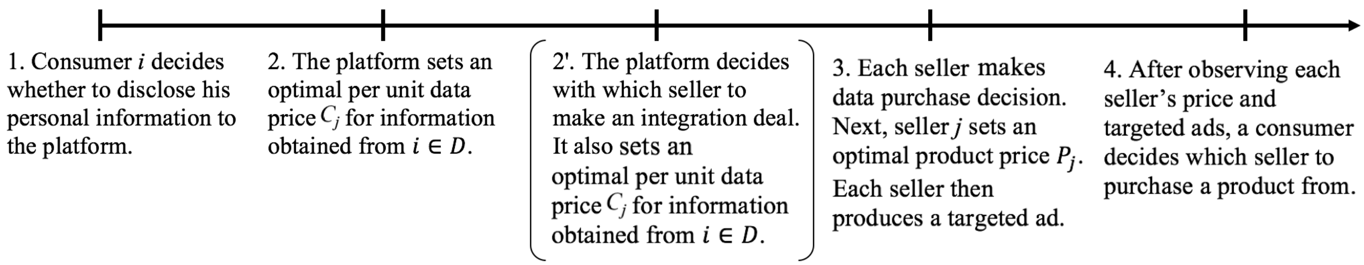

Figure 1: Timing

3 Equilibrium

In the first stage, each consumer compares the utility levels from disclosure and decides whether to disclose information. In this stage, depending on τ_i, any consumer who has $\psi(\tau_i) < v(\mathbb{P}(i \in \mathcal{D}))$ will disclose. The portions of privacy sensitive consumers (not disclosing information) and privacy insensitive consumers (disclosing) are implicitly determined by the following Proposition 1.[11]

Proposition 1. *A critical point, τ^c, satisfies the following equation.*

$$\begin{aligned} \mathbb{P}(i \in \mathcal{D}) = \mathbb{P}(\tau_i < \psi^{-1}(v(\tau^c))) = F(\psi^{-1}(v(\tau^c))) = \tau^c. \\ \mathbb{P}(i \in \mathcal{ND}) = 1 - F(\psi^{-1}(v(\tau^c))) = 1 - \tau^c. \end{aligned} \tag{5}$$

A parametric example Let $\tau_i \sim U[0,1]$, $\psi(\tau_i) = \lambda\tau_i^2$, and $v(\tau^c) = \sqrt{\tau^c}$. In this case, τ^c is the solution to $\tau^c = \psi^{-1}(\sqrt{\tau^c})$. Thus, $\tau^c = \frac{1}{\lambda^{2/3}}$. Clearly, as λ increases, i.e., as the nuisance cost increases, a greater number of consumers are reluctant to disclose information, so τ^c decreases.

[10] Note that the PBE satisfies the passive beliefs in that a seller assumes that the platform offers the equilibrium contract to its rival even when receiving an out-of-equilibrium offer (Hart and Tirole (1990); O'Brian and Shaffer (1992); McAfee and Schwartz (1994); Rey and Vergé (2004); Reisinger and Tarantino (2015)). However, as noted in McAfee and Schwartz (1994), under unobservability, the equilibrium outcome with passive beliefs is identical to that with wary beliefs.

[11] Note that when each consumer makes his or her optimal decision about information disclosure, he or she forms a rational expectation about the proportion of consumers who disclose information. In equilibrium, consumers take these probabilities as given by $\mathbb{P}(i \in \mathcal{D}) = \tau_a^c$ and $\mathbb{P}(i \in \mathcal{ND}) = 1 - \tau_a^c$, where subscript a denotes the *anticipated* proportion. In equilibrium, τ_a^c should be consistent with the true τ^c, which is aggregately determined by consumers. For notational convenience, I omit subscript a.

Given $\mathbb{P}(i \in \mathcal{D}) = \tau^c$, the amount of aggregated detailed data, I solve for the PBE using backward induction to obtain sequentially rational strategies. From the utility specification in Equation (2), the indifference condition is $\theta^c_{\mathcal{ND}} = \frac{P_j - P_{-j}}{q-1}$ for $i \in \mathcal{ND}$ and $\theta^c_{\mathcal{D}} = \frac{P_j - P_{-j} + \Delta}{q-1}$ for $i \in \mathcal{D}$, where $\Delta = (1+\tau^c)^{\alpha_{-j}} - (1+\tau^c)^{\alpha_j}$, given that $q \equiv q_j > q_{-j} = 1$. The weighted indifference condition can be rewritten in a simple way as follows:

$$X_{-j}(\mathbf{P}, \mathbf{D}) = \theta^c = \frac{P_j - P_{-j} + \Delta\tau^c}{q-1}. \tag{6}$$

Given that the market share for the high-quality seller is $X_j = 1 - \theta^c$, the solutions to profit maximization with respect to prices are given by

$$P_{-j} = \frac{q-1+\Delta\tau^c}{3}; \quad P_j = \frac{2(q-1)-\Delta\tau^c}{3}; \quad X_{-j} = \frac{q-1+\Delta\tau^c}{3(q-1)}; \quad X_j = \frac{2(q-1)-\Delta\tau^c}{3(q-1)}. \tag{7}$$

To guarantee an interior solution in any data acquisition case,[12] I assume throughout the paper that the product quality provided by two sellers is sufficiently differentiated to have positive demand for any seller. To better focus on targeting-related parameters, I often restrict my attention to a specific value of q satisfying the interior solution condition and focus on the effects of α (representing differentiation in terms of sellers' targeting skills) and τ^c (representing the amount of data). Note that restricting attention to a certain value of q, e.g., $q = 2$, is more reasonable and insightful. In the case of $q_I > q_E$, if the quality advantage given to I is overly large, it builds overly high barriers to entry: a higher q places a greater burden on E, which is likely to result in a monopolistic outcome. In the opposite case of $q_I < q_E$, it is difficult to see that E enters the market with an exceptional quality product that is incomparable with I's product quality, given E's lack of experience. Moreover, if q_E is too high, the equilibrium in the game is mostly influenced by quality differentiation, as in the standard vertical differentiation model, such that the demand-increasing effect of data becomes negligible. Thus, I limit the focus to a reasonable value of q, not simply to better focus on how the parameters of interest affect duopolistic market competition with asymmetric sellers but to consider a more reasonable setup.[13]

Given the equilibrium price and quantity in Equation (7), each seller decides whether to purchase data from the platform; thus, the equilibrium profit level is realized. By comparing profits under the two choices, I can derive thresholds of C_j that guarantee that one seller will buy data, given the rival's decision. For each seller j, there are two thresholds—offers of non-exclusive and exclusive data

[12] The sufficient condition for the interior solution on q is obtained by the solution to $X_{-j} = 0$ when seller j buys data but $-j$ does not.

[13] Note that other quantitative values of q that satisfy the interior solution condition lead to qualitatively the same results.

contracts—which are described as follows:

$$\begin{cases} \text{A non-exclusive contract is accepted if} & C_I\tau^c \leq P_I^{BB}X_I^{BB} - P_I^{NB}X_I^{NB} \equiv \bar{C}_I \text{ for } I \text{ or } C_E\tau^c \leq P_E^{BB}X_E^{BB} - P_E^{BN}X_E^{BN} \equiv \bar{C}_E \text{ for } E \\ \text{An exclusive contract is accepted if} & C_I\tau^c \leq P_I^{BN}X_I^{BN} - P_I^{NB}X_I^{NB} \equiv \bar{\bar{C}}_I \text{ for } I \text{ or } C_E\tau^c \leq P_E^{NB}X_E^{NB} - P_E^{BN}X_E^{BN} \equiv \bar{\bar{C}}_E \text{ for } E, \end{cases}$$

where the first and second superscripts denote I's and E's data acquisition decisions, respectively; e.g., BN denotes I Buying and E Not buying. The platform makes either an exclusive contract with I (at $\bar{\bar{C}}_I$) or E (at $\bar{\bar{C}}_E$) or a non-exclusive contract at $\bar{C}_I$ and $\bar{C}_E$. Clearly, $\bar{\bar{C}}_I > \bar{C}_I$ and $\bar{\bar{C}}_E > \bar{C}_E$: provided that the rival does not buy, the other firm is more likely to pay a higher price. In other words, data acquisition is a strategic substitute because $\frac{\partial^2 \pi_j}{\partial D_I \partial D_E} = -\frac{2\alpha(\tau^c)^2(D_E+1)^{\alpha-1}}{9(q-1)} < 0$.[14]

3.1 No Vertical Integration

I first analyze the case of $q \equiv q_I > q_E = 1$, in which seller I has the advantage of not only targeting skills but also product quality. Given the fees charged to sellers, the platform's profits under different data contracts can be compared. I first verify that no data selling is always dominated by either exclusive or non-exclusive selling. Additionally, given I's advantages in both dimensions, the platform would choose to make an exclusive data offer to I, instead of a non-exclusive data offer.

Next, I focus on the case in which I is better at targeting skills, but E is better at product quality; thus, $q_I < q_E$ is assumed. For instance, Walmart.com is a representative example of an incumbent firm. Because Walmart.com sells various products in many categories, it is a generalist. Any specialist retailers that sell various products in a specific category, such as apparel, can be considered high-quality entrants in that they specialize in their own business area. Campbell et al. (2015) made a similar assumption. I find that the platform would choose to make an exclusive data offer to either I or E instead of a non-exclusive data offer, depending on the relative volume of data available τ^c and E's initial targeting disadvantage α: for instance, the platform would prefer offering exclusive data access to seller I if τ^c is sufficiently large, implying that I can easily dominate the market using its better targeting skills combined with the extensive amount of data to which it has access.

However, as shown in Reisinger and Tarantino (2015), if offers are secretly proposed, this situation undermines the credibility of the platform's commitment. Thus, if the platform makes an exclusive offer to either seller I or E, the downstream sellers know that the platform's exclusive offer is not credible: each seller forms a conjecture that as soon as it accepts the exclusive data offer by paying a higher fee, the platform has an incentive to make a side offer to the rival by supplying data because the platform can extract positive rent from the seller that has not obtained exclusive data access. Then,

[14] The intuition is as follows. The data can be used to differentiate products because better-targeted ads, based on more available data, attract a greater number of consumers. Thus, the consumer information used to generate better-targeted ads increases product differentiation, which softens price competition.

the seller with exclusive data access becomes worse off because it obtains non-exclusive data access but pays more than what it would have needed to pay. By the same logic as in Lemma 1 in Reisinger and Tarantino (2015), a seller will decline the platform's exclusive data offer. Lemma 1 summarizes this finding.

Lemma 1. *If the platform makes data contracts simultaneously and secretly with both sellers, both sellers obtain non-exclusive data access at different data prices.*

Lemma 1 accords with reality. Indeed, it is easy to find real case examples in which the platform sells data to all sellers in the downstream market at different data prices. For instance, Facebook sets different prices for data based on a data buyer's willingness to pay.[15]

One might argue that a reputational concern can mitigate the platform's opportunistic incentives. However, as noted in O'Brian and Shaffer (1992), reputation does not resolve such problems due to the complex contracting environments in the real world. This fact suggests that vertical integration is an efficient way of overcoming the commitment problem in a secret contract environment, thereby achieving exclusive data contracts with more efficient sellers (Hart and Tirole (1990); O'Brian and Shaffer (1992)).

3.2 Vertical Integration

In this section, I analyze the effect of vertical integration between the platform and one of the sellers. Regarding the timing, the platform first makes an integration deal with one of the sellers.[16] After the integration is complete, the unaffiliated seller decides whether to purchase data from the platform. The affiliated seller always uses data for targeted ads. Next, the sellers simultaneously set their prices, and the consumers decide.

Note that vertical integration resolves the incomplete information issue arising from the secret contract environment, given that it allows both affiliated and unaffiliated sellers to obtain contract-related information implicitly. That is, the affiliated seller observes the data that are offered to the rival, while the unaffiliated seller credibly believes that the rival has data access at zero price. This case leads to a different data acquisition equilibrium from that in the game without integration: as I show in this section, a non-exclusive data selling equilibrium, which prevails without vertical integration per Lemma 1, no longer occurs when the platform is vertically integrated with a downstream seller.

[15] Refer to https://www.facebook.com/business/ads/pricing.

[16] If the integrated profit is less than the joint profits of the platform and affiliated seller, integration will not occur. Moreover, the platform endogenously chooses the optimal merger partner by comparing possible profit levels from integration.

3.2.1 Incumbent with High-quality Product Provision

By backward induction, each seller's price and market share are the same as earlier. Given this setting, I examine the result assuming that seller I is advantageous in both dimensions, i.e., targeting and product quality. First, it can be verified that the platform and sellers always have an incentive to vertically integrate with one another by comparing the joint profits of the platform and seller I or E under no vertical integration to the profits of the integrated firm. Suppose that the platform merges with seller I. Because seller I always uses data, seller E buys data if $C \leq \bar{C}_E^{IH}$ but does not buy data if $C > \bar{C}_E^{IH}$ where superscript IH denotes the case in which I provides a high-quality product. The profit for the integrated firm can be written as follows:

$$\pi_{VI}^{IH} = \begin{cases} P_I^{IH}(\Delta_{BB}^{IH})X_I^{IH}(\Delta_{BB}^{IH}) + \bar{C}_E^{IH} \equiv \pi_{VI,S}^{IH} & \text{if the integrated firm sells data to } E \\ P_I^{IH}(\Delta_{BN}^{IH})X_I^{IH}(\Delta_{BN}^{IH}) \equiv \pi_{VI,F}^{IH} & \text{otherwise (foreclose),} \end{cases} \tag{8}$$

where the first two letters in the subscript, *VI*, denote *Vertical integration with I*, and the last letter denotes data foreclosure (F) or sales (S) status. There is a tradeoff between selling and foreclosing data. If the integrated firm prevents the unaffiliated seller from accessing data, it can dominate the downstream market by sending better-targeted ads using exclusive data. However, data foreclosure comes at the expense of losing data sales revenue. By comparing $\pi_{VI,S}^{IH}$ to $\pi_{VI,F}^{IH}$, the integrated firm decides whether to sell data to the unaffiliated firm. The profit comparison is given as follows:

$$\pi_{VI,S}^{IH} - \pi_{VI,F}^{IH} = \frac{2\tau^c\left[(\tau^c+2)^\alpha - 2^\alpha\right]\left\{\tau^c\left[(\tau^c+2)^\alpha + 2^\alpha - 2\tau^c - 4\right] - q + 1\right\}}{9(q-1)(2\tau^c - 1)} < 0. \tag{9}$$

From Equation (9), I find that $\pi_{VI,S}^{IH} < \pi_{VI,F}^{IH}$ for $\alpha \times \tau \subset [0,1] \times [0,1]$ given q: it always wants to foreclose the unaffiliated entrant from data access because its market dominance effect from monopolizing the data outweighs the data-selling revenue effect.

Now, I examine the result when the platform merges with seller E. By the above logic, the profits for the integrated firm and the difference between the two profit levels are as follows:

$$\pi_{VE}^{IH} = \begin{cases} P_E^{IH}(\Delta_{BB}^{IH})X_E^{IH}(\Delta_{BB}^{IH}) + \bar{C}_I^{IH} \equiv \pi_{VE,S}^{IH} & \text{if the integrated firm sells data to } I \\ P_E^{IH}(\Delta_{NB}^{IH})X_E^{IH}(\Delta_{NB}^{IH}) \equiv \pi_{VE,F}^{IH} & \text{otherwise (foreclose).} \end{cases} \tag{10}$$

$$\pi_{VE,S}^{IH} - \pi_{VE,F}^{IH} = \frac{2(\tau^c)^2\left\{\tau^c\left[-2(\tau^c+2)^\alpha + \tau^c + 4\right] + q - 1\right\}}{9(q-1)(2\tau^c - 1)} > 0. \tag{11}$$

From Equation (11), I find that $\pi_{VE,S}^{IH} > \pi_{VE,F}^{IH}$ for $\alpha \times \tau \subset [0,1] \times [0,1]$ given q. To determine which seller offers sufficient incentives to induce the platform to integrate, I compare the profits from integration with seller I, resulting in foreclosing the rival, to those from integration with seller E, resulting in selling to the rival. The comparison reveals that the platform always wants to integrate with seller I and prevents unaffiliated seller E from accessing data, given that seller I is advantageous in both the targeting and product quality dimensions.

Proposition 2. *If the incumbent has an advantage in not only targeting skills but also product quality, it is always integrated with the platform, and the integrated firm forecloses the unaffiliated entrant from data access.*

Absent vertical integration, the entrant can use the platform's data at the unit price of $\bar{C}_E$. When the platform is integrated with seller I, both the platform and seller I can internalize their strategic externalities. Because this internalization incentivizes the platform to dominate the downstream market, the integrated firm is only induced to sell data to the unaffiliated E at a price sufficiently higher than $\bar{C}_E$. However, the unaffiliated E is unwilling to pay this high price and is thus foreclosed from data access.

3.2.2 Entrant with High-quality Product Provision

Next, I focus on the case in which seller E provides a high-quality product, whereas seller I has an advantage in targeting skills. The profit for the integrated firm, if integrated with I, can be written as follows:

$$\pi_{VI} = \begin{cases} P_I(\Delta_{BB})X_I(\Delta_{BB}) + \bar{C}_E \equiv \pi_{VI,S} & \text{if the integrated firm sells data to } E \\ P_I(\Delta_{BN})X_I(\Delta_{BN}) \equiv \pi_{VI,F} & \text{otherwise (foreclose).} \end{cases} \tag{12}$$

By comparing $\pi_{VI,S}$ to $\pi_{VI,F}$, the integrated firm decides whether to sell data to the unaffiliated firm. From the comparison, I find that if $\tau^c > \tau_v''$, $\pi_{VI,S} < \pi_{VI,F}$:[17] there is an incentive to sell data to the unaffiliated entrant if the available amount of data is sufficiently small. In other words, it is more likely to prevent data access as τ^c increases because if τ^c is sufficiently large, seller I can dominate the market more easily by exclusively using data for targeting consumers.

Now, I examine the result when the platform merges with seller E. The profits for the integrated firm are as follows:

$$\pi_{VE} = \begin{cases} P_E(\Delta_{BB})X_E(\Delta_{BB}) + \bar{C}_I \equiv \pi_{VE,S} & \text{if the integrated firm sells data to } I \\ P_E(\Delta_{NB})X_E(\Delta_{NB}) \equiv \pi_{VE,F} & \text{otherwise (foreclose).} \end{cases} \tag{13}$$

I find that the integrated firm wants to prevent unaffiliated seller I from accessing data if $\tau^c < \tau_v'$ because if a sufficient amount of data is available, which improves E's targeting, its advantage in

[17] τ_v' (τ_v'') is the solution to $\pi_{VE,S} = \pi_{VE,F}$ ($\pi_{VI,S} = \pi_{VI,F}$).

product quality q allows it to easily dominate the downstream market even when data are provided to the rival. To determine which seller offers a sufficient incentive to induce the platform to integrate, I compare the profits from integration with I to those from integration with E. There are three possible cases: $I - Foreclose$ or $E - Sell$ for $\tau \in [\tau_v', 1]$, $I - Foreclose$ or $E - Foreclose$ for $\tau \in [\tau_v'', \tau_v']$, and $I - Sell$ or $E - Foreclose$ for $\tau \in [0, \tau_v'']$, where $I - Foreclose$ means that the platform integrates with I and forecloses E from data access, and $E - Sell$ means that the platform and E integrate and sell data to I. Proposition 3 summarizes these results.[18]

Proposition 3. *Suppose that the incumbent has an advantage in targeting skills, but the entrant provides a high-quality product. If the amount of data is sufficient ($\tau^c > \tau_v$), then the platform has an incentive to vertically integrate with the incumbent. However, if only a limited amount of data is available ($\tau^c < \tau_v$), then the platform integrates with the entrant. In either case, the integrated firm sets the price in a way that the unaffiliated seller forgoes buying data.*

If a sufficient amount of data is available, the profit from integrating with seller I is always greater than that from integrating with seller E because seller I's better targeting skills, combined with the exclusive use of data, allow the platform integrated with seller I to sell more affiliated products in the downstream market. However, if only a limited amount of data becomes available, because of either consumers' greater privacy concerns or stricter privacy regulations, the effect of data exclusivity becomes negligible. Instead, the platform pays more attention to other aspects in the market, such as product quality. Thus, if $\tau^c < \tau_v$, then the platform has an incentive to integrate with seller E, which offers high-quality products to consumers.

This outcome suggests that the platform following integration switches its business model depending on data availability: more data availability causes the platform to switch its business model for the downstream market from being product focused to being data-driven service focused. With only a small amount of data available, a platform integrated with a seller is less likely to attract more consumers because it creates less suitable targeted ads. Then, the best strategy is to focus on other core aspects of the product, such as product quality, rather than nonproduct-focused service-related aspects, such as data-driven targeted ads.

Additionally, as summarized in Corollary 1, the comparative static result with respect to the effect of α shows that as α increases, i.e., as seller E's targeting technology improves, the platform is more likely to integrate with seller E; that is, $\frac{\partial \tau_v}{\partial \alpha} > 0$. Intuitively, if the targeting asymmetry between two sellers is sufficiently small, reflected by a sufficiently large α, then the platform places an emphasis

[18] As Montes et al. (2018) note, this exclusive data-selling strategy accords with a reality in which different firms are unlikely to obtain data on the same consumers despite doing business in the same industry.

on product quality itself, thereby making it more likely that seller E is chosen, which provides a high-quality product.

Corollary 1. *Given that the incumbent has an advantage in targeting quality and the entrant has an advantage in product quality, the platform is more likely to integrate with the entrant as the targeting technologies for the two sellers become more similar.*

Note that the data foreclosure equilibrium also encompasses the case in which the unaffiliated seller obtains data from the integrated firm but at a higher price than it used to pay without vertical integration. For instance, whereas Amazon's in-house brands, such as AmazonBasics, or its affiliated sellers, such as Whole Foods, can use the tremendous amount of consumer data that Amazon has, other unaffiliated downstream competitors are able to obtain the same data, only at a higher price. Propositions 2 and 3 imply that vertical integration increases the data price for unaffiliated sellers, providing a considerable competitive advantage to an affiliated seller. Such preferential treatment for the affiliated seller can be particularly anticompetitive if an entrant that has a greater need for consumer data due to its worse initial targeting skills is prevented from accessing data due to vertical integration, as shown in Section 3.3.

3.3 Implications

Compared to the game without vertical integration in which both sellers obtain consumer data from the platform, allowing vertical integration prevents the unaffiliated seller from obtaining consumer data, regardless of the seller with which the platform integrates. Due to this data foreclosure, the unaffiliated sellers become worse off compared to the situation with no vertical integration: the unaffiliated seller's market share is smaller when being foreclosed than when being offered data non-exclusively.

Furthermore, allowing data-driven vertical integration, followed by data monopolization, is more likely to harm E given that E is more likely to be foreclosed from data access. If seller I has advantages in both dimensions, targeting skills and product quality, as in Section 3.2.1, vertical integration always results in data foreclosure for seller E, which requires data access, so seller E is more likely to be adversely affected. Even when seller E provides a high-quality product, as in Section 3.2.2, it can be foreclosed from data access per Proposition 3, implying that harm toward its match value still exists.

This result raises an important antitrust implication regarding data-driven vertical integration. Since consumer information is vital for a small seller or a market entrant with weak targeting skills, vertical integration with a seller possessing better targeting skills is likely to have an anticompetitive effect because it prevents the unaffiliated entrant from using data to overcome its initial disadvantage.

To mitigate any harm toward the entrant, the seller must enter the market with its own distinctive advantage, such as better product quality.

The vertical integration equilibrium result is more important than ever given the recent debates on regulating exclusive data contracts. For example, Kathuria and Globocnik (2020) discuss, from an antitrust perspective, whether mandatory data sharing, i.e., requiring (B, B), is viable as a remedy for (alleged) data-related exclusionary conduct. If a regulatory authority prohibits exclusive data contracts, which always results in (B, B) in the game without vertical integration, Proposition 3 implies that the platform can always bypass the regulation by engaging in vertical integration with the incumbent. In this regard, Kim et al. (2019) also propose a remedy that ensures nondiscriminatory access because if exclusive dealing contracts are prohibited, the monopoly data broker will prefer to vertically integrate with one downstream firm to achieve the same downstream equilibrium, as under exclusive data sales. Considering the nonrival feature of data, which leads to cost-efficiency under non-exclusive data provision, this remedy is particularly welfare improving. Thus, as Section 5.2 will summarize, the policymaker must consider remedies for data-driven mergers.

4 Welfare Analysis

Based on the equilibrium results derived thus far, I examine welfare consequences in this section. First, the total social welfare function is the sum of consumer surplus and sellers' profits, including the platform's profits, as follows:

$$SW = CS + \mathbb{1}_{\{NV\}}(\pi_{pf} + \pi_I + \pi_E) + \mathbb{1}_{\{V\}}(\pi_K^I + \pi_K^E), \tag{14}$$

where the subscripts $K \in \{VI, VE\}$ and VI (VE) denote the surplus from vertical integration with seller I (E). π_{VI}^I (π_{VE}^E) is the profit for the integrated firm, and π_{VI}^E (π_{VE}^I) is that for the non-integrated firm. The indicator functions, $\mathbb{1}_{\{NV\}}$ and $\mathbb{1}_{\{V\}}$, are one if in the no vertical integration and vertical integration games, respectively. Consumer surplus can be obtained in the following manner:

$$\begin{aligned} CS = &\int_0^{\tau^c} (v(\tau^c) - \psi(x))\, dF(x) + \tau^c \left[\int_0^{\theta_{\mathcal{D}}^c} \Big(\theta q_{-j} + m_{-j}(D_{-j}) - P_{-j}\Big) d\theta + \int_{\theta_{\mathcal{D}}^c}^1 \Big(\theta q_j + m_j(D_j) - P_j\Big) d\theta\right] \\ &+ (1-\tau^c) \left[\int_0^{\theta_{\mathcal{ND}}^c} \Big(\theta q_{-j} - P_{-j}\Big) d\theta + \int_{\theta_{\mathcal{ND}}^c}^1 \Big(\theta q_j - P_j\Big) d\theta\right], \end{aligned} \tag{15}$$

where $q_j > q_{-j}$.

The comparative static results show that consumer surplus increases in τ^c. There can be two channels whereby the provision of more data benefits consumers: a greater targeting benefit and more intense market competition. If τ^c increases, a consumer receives more suitable targeted ads.

In addition, conditional that both sellers buy data, having more data implies that seller E is more likely to overcome its disadvantage in targeting, thereby fostering market competition. Proposition 4 summarizes this finding.[19]

Proposition 4. *Consumer surplus levels increase with respect to the amount of data available on the platform, which implies that having more available data makes consumers better off.*

Another insightful analysis compares welfare levels with and without vertical integration to discern whether the market equilibrium outcome, resulting in the exclusive use of data access following vertical integration, is welfare enhancing compared to non-exclusive data contracts. Indeed, social welfare under integration, which leads to data monopolization either by the affiliated incumbent, (B, N), or by the affiliated entrant, (N, B), is not always greater than that under no integration, which results in (B, B).

When both sellers symmetrically obtain data, (B, B), as in the case without vertical integration, their overall targeting becomes similarly effective, indicating less differentiation in terms of targeting. This scenario creates more intense price competition, thereby enhancing consumer surplus, which I call *intense competition effects* of no integration. In contrast, asymmetric data acquisition, (B, N) or (N, B), is likely to harm consumers through the lack of competition in the downstream market because the market is dominated by the affiliated seller, which exploits the exclusive use of data. However, asymmetric data acquisition improves the integrated firm's profits through better match quality, which I call *profit increasing effects* of data-driven integration. Thus, the asymmetric data foreclosure case might enhance total welfare if the exclusive use of data combined with better targeting skills (if integrated with I) or higher product quality (either I or E) significantly enhances the integrated firm's profit.

The relative size of two conflicting effects, namely, *intense competition* and *profit increasing effects*, depends on the amount of data available on the market. If a small amount of data becomes available due to greater privacy concerns or stricter privacy regulations, the *profit increasing effects* for the integrated firm are negligible; thus, vertical integration and the resulting data foreclosure decrease welfare, regardless of the seller with which it is integrated. However, if more data are available, data foreclosure from vertical integration is socially desirable because the firm's *profit increasing effects* from integration outweigh any harm arising from soft competition. This outcome suggests that vertical

[19] In the model, I do not consider a nuisance cost from targeted ads by focusing only on match benefits. If a consumer feels annoyed when receiving inaccurately targeted ads, this may erode the additional consumer surplus obtained from a match between suggested products and the consumer's current wants. However, as more data are collected, targeting quality is improved, such that any nuisance from inaccurately targeted ads disappears while match benefits increase. Thus, as long as nuisance costs fall as targeting is improved, Proposition 4 still holds.

integration preventing the unaffiliated seller from accessing data is particularly harmful if consumers have greater privacy concerns so that less data are available for targeting.

For example, suppose that the platform and seller I are integrated and monopolize data access. If the amount of available data is small, the effect of the affiliated seller's data monopolization on increasing the match benefit is negligible compared to the harm arising from the lack of market competition: then, (B, B) attains a greater welfare level than (B, N). However, if sufficient data become available, the effect on enhancing the match value from seller I outweighs the effect of fierce price competition in the market, so vertical integration and the resulting data foreclosure are more desirable. This finding is summarized in Proposition 5.

Proposition 5. *Data-driven vertical integration is welfare reducing compared to the absence of vertical integration if there are substantial privacy concerns.*

As the left panel in Figure 2 graphically represents the welfare comparison, if seller E provides a high-quality product, there exists a threshold on τ^c, denoted as τ_{SW}, below which a non-exclusive data contract that prevails in the game without vertical integration is welfare enhancing compared with an exclusive data contract that prevails in the game with vertical integration. That is, as data become less available, i.e., $\tau^c < \tau_{SW}$, market-driven vertical integration leaves not only the unaffiliated E worse off but also society as a whole worse off.

The welfare analysis result implies that whether the vertical integration equilibrium outcome is welfare enhancing or welfare reducing depends on the amount of information available on the platform. Specifically, consumers' privacy concerns or more stringent privacy regulations, resulting in less available data, can be welfare reducing through the exclusive use of data after integration. Thus, although each consumer makes an individually rational information disclosure decision, it might not be socially optimal after accounting for the effects of vertical integration and data foreclosure on the competitive structure of the downstream market.

4.1 Rival vs. Nonrival Inputs

As in Equation (3), the cost structure captures the nonrival feature of data, such that the marginal cost of providing data is not incurred in proportion to the amount of goods sold, unlike in the case of rival goods. If an input that is foreclosed by vertical integration is a rival good, any nondiscriminatory regulation, which forces the platform to provide the input to both downstream sellers, is too costly. However, data as nonrival goods do not incur additional marginal cost when simultaneously offered to multiple sellers, making non-exclusive data/input usage (B, B) more likely to be welfare enhancing.

The right-hand side in Figure 2 shows how the nonrivalry-driven cost-efficiency of data yields a different welfare comparison result from that in the case of rival goods: compared to the case of data, the threshold for the case of rival goods, denoted as τ_{SW}^{R}, is smaller. This implies that if nonrivalry, a unique feature of data, is not taken into consideration when designing policy remedies concerning data-driven integration followed by data (as an input) foreclosure, we may reach a misleading conclusion by overestimating the benefit of the exclusive use of data.

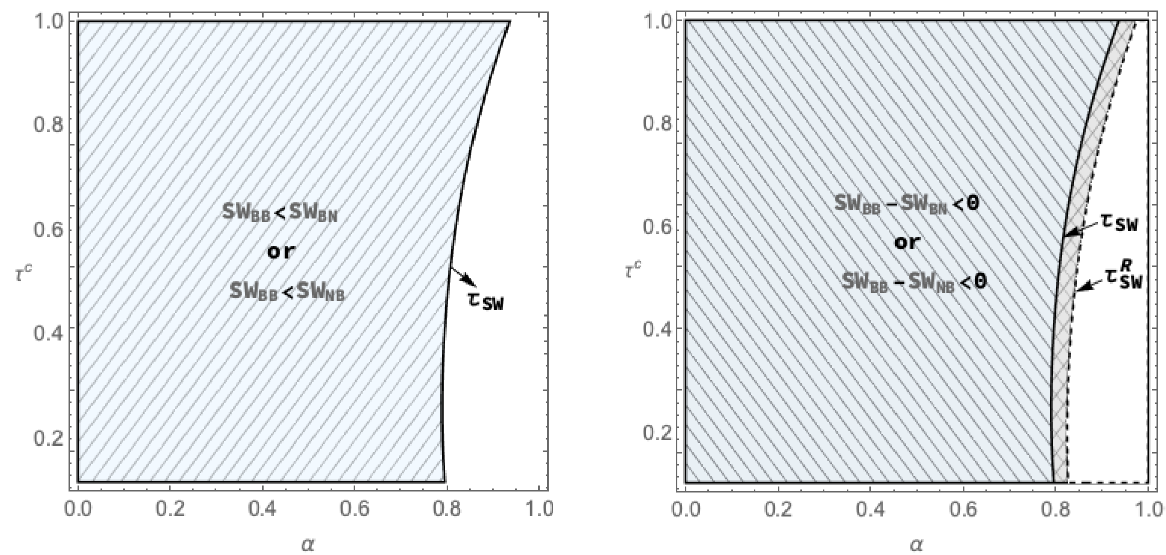

Figure 2: Welfare comparison if seller E provides a high-quality product given $q = 2$ where the right-hand side figure depicts the case of rival goods (assuming that $\phi = 0.1$)

5 Discussion and Extensions

5.1 Endogenous Amount of Data

Thus far, I have assumed that a consumer considers only the immediate benefits when making information disclosure decisions but does not consider any potential future benefits arising from better-targeted ads. Although this assumption is reasonable for many platforms, such as social media platforms, it is worth showing the result if consumers have perfect foresight when making decisions: if they are sophisticated enough to recognize that greater personal data availability on the platform will lead to more relevant personalized ads, they might take this potential effect into consideration. To capture this effect, I consider the total net utility that each consumer obtains from using the platform and from purchasing a product. For simplicity, I normalize the immediate benefit from using the platform to zero, which means that each consumer compares the privacy nuisance cost to the potential match benefits when making information disclosure decisions.[20] The aggregate utility specification is as follows.

[20]The normalization of immediate benefits to zero is harmless, because it does not change the qualitative results.

$$u_{ij}^{\text{Foresight}} = \theta_i q_j + \mathbb{1}_{\{i \in \mathcal{D}\}} \times \Big(m_j(D_j) - \tau_i^2 \times n \Big) - P_j, \tag{16}$$

where n is the number of downstream sellers with data access. That is, a consumer faces a higher privacy cost if the platform shares data with more sellers such that personal data are more exposed: in this specification, the vertical integration equilibrium and resulting data acquisition status directly affect the consumer's privacy concern, which in turn endogenously determines the amount of data.

Working backward, P_j and X_j are the same as previously. $\mathbb{P}(i \in \mathcal{D})$, which is determined in the first stage, can be implicitly derived as follows.

$$\begin{aligned} \mathbb{P}(i \in \mathcal{D}) = \tau^c &= \theta^c \mathbb{P}(i \in \mathcal{D} | i \in \mathcal{I}) + (1-\theta^c) \mathbb{P}(i \in \mathcal{D} | i \in \mathcal{E}) \\ &= \theta^c F\left(\sqrt{\frac{m_I(D_I)}{n}}\right) + (1-\theta^c) F\left(\sqrt{\frac{m_E(D_E)}{n}}\right), \end{aligned} \tag{17}$$

where θ^c, which is a function of τ^c, is given as in Equation (6) and $i \in \mathcal{I}$ ($i \in \mathcal{E}$) denotes a consumer i buying from the incumbent (entrant). Because consumers are assumed to be sufficiently sophisticated, the disclosure probability now depends on each seller j's targeting effectiveness, which implies that $\mathbb{P}(i \in \mathcal{D})$ can differ depending on the information acquisition equilibrium.

By comparing the right-hand side of Equation (17), I can rank different τ^c levels depending on each data acquisition equilibrium. Focusing on the case in which E provides a high-quality product, it can be shown that τ^c_{BB} is the lowest, whereas τ^c_{NN} is the highest, where the subscript denotes the data acquisition equilibrium. In other words, knowing that a seller acquires personal data, a consumer becomes reluctant to disclose information. The relative size of τ^c_{NB} and τ^c_{BN} depends on α, which represents the initial targeting disadvantage for E: $\tau^c_{NB} > \tau^c_{BN}$ holds if α is sufficiently large, i.e., $\alpha > \alpha'$. Intuitively, as α increases, the targeting quality between the two sellers becomes similar, which results in more consumers enjoying not only a greater match benefit but also a high-quality product from E. Thus, a consumer becomes more willing to disclose personal data under E's exclusive use of data as α increases. Proposition 6 summarizes the findings.

Proposition 6. *If a consumer has perfect foresight, the equilibrium information disclosure level is lowest when both sellers buy personal data, whereas it is highest when neither buys. The relative size of* τ^c_{NB} *and* τ^c_{BN} *depends on how much the entrant suffers from initial targeting disadvantage.*

Next, I check whether the equilibrium outcomes carry over to this extension. Obviously, the equilibrium results in the game with no vertical integration still hold if a secret contracting environment with incomplete information is maintained. Even in the game with vertical integration, the qualitative results still hold. Assuming that E provides a high-quality product, the platform integrates with either

I or E, depending on the value of α, which results in data foreclosure. Specifically, as α increases, implying that E's initial disadvantage in targeting technology gradually disappears, the platform is more likely to integrate with E and foreclose I from data access. As shown here, when the model allows a dynamic between data disclosure decisions and data-driven vertical integration outcomes, non-exclusive data provision (B, B) leads to too little data disclosure, which is more likely to incentivize the platform to monopolize data in addition to the foreclosure motive.

5.2 Policy Implications

According to the theoretical findings of this paper, data foreclosure that follows vertical integration between the platform and one seller will harm downstream market competition (e.g., by raising barriers to entry to the market in the case of integration I), which can ultimately be welfare reducing. To prevent this undesirable consequence, there are two different types of policy remedies, namely, imposing a related condition on data sharing aspects when approving a proposed merger and inducing greater data availability by reducing consumers' privacy concerns. Several remedies of these types are proposed below.

Reputation-Enhancing Program: Assuming that vertical integration is permitted, the market outcomes with vertical integration and the resulting data foreclosure are more welfare enhancing if the amount of available data is sufficiently large (i.e., $\tau^c > \tau_{SW}$). If the data collector can decrease users' privacy nuisance costs, it would help to increase the amount of information available. One way to reduce the nuisance cost is to improve the data collector's reputation. In reality, consumers are more likely to agree to app developers' data usage policies if the developers are relatively well-known rather than unknown, as shown in Kummer and Schulte (2019). In this respect, any policy that helps the platform build its reputation would be welfare enhancing.[21]

Substantial evidence demonstrates that privacy certification programs represent one such remedy. If a credible institution grants a certificate indicating that firms comply with government-enacted privacy rules, marginally privacy sensitive consumers who refuse to provide information due to possible data abuse might switch and decide to disclose personal information. Although there are a few private firms, such as *TRUSTe*, that serve a similar function, their certifications indicate self-certification at best. A credible certification program could serve as a global standard that helps participating firms improve their reputations regarding data usage. As empirical evidence has shown, privacy is a trust-based issue in that consumers care about who asks for their personal information. Because willingness

[21] This effect can be captured by slightly modifying the utility specification in Equation (1): if replacing $\psi(\tau_i)$ with $\frac{\psi(\tau_i)}{r}$, where r represents the platform's reputation, it assumes that consumers' concern for privacy is asymmetric with respect to the firm's reputation.

to disclose information depends on a firm's reputation, this remedy is likely to be effective. This policy suggestion is consistent with the policy implications in some previous literature (e.g., Campbell et al. (2015)). Additionally, the Cyber Shield Act of 2017 is in the same vein.[22]

Tax Incentives: Regulators can increase consumers' willingness to disclose their personal information by adopting a policy that encourages the platform, i.e., the data intermediary, or the integrated firm to invest more to keep consumers' data secure and, therefore, prevent an actual data breach.

It would be effective if the government were to provide tax incentives to platforms that invest a certain amount of money to obtain better data security systems. These tax incentives would encourage more investments to protect consumers' personal information from being leaked, thus reducing consumers' concerns about disclosing their data to such platforms. In other words, consumers will disclose more of their data to trustworthy data collectors, rendering more socially desirable the market-driven equilibrium outcome with vertical integration and data foreclosure between the platform and the affiliated seller.

Regulation on Mergers and Data Sharing Practice: Policymakers can implement a direct policy that regulates integrated firms' data-sharing practices. For instance, if the platform and incumbent integrate and prevent the unaffiliated entrant from accessing data, this situation will be welfare reducing in some cases. Thus, when these two entities attempt to integrate with each other, a regulatory authority could aim to either impose relevant conditions on data-sharing aspects for final merger approvals or not approve the merger at all. If the amount of data is limited due to greater privacy concerns such that vertical integration leaves consumers worse off, as in Proposition 5, the authority might need to reject a proposed data-driven merger. Additionally, it is important to design data-specific merger remedies while considering the unique features, such as nonrivalry, as emphasized in Section 4.1.

Alternatively, if a regulatory authority approves a proposed merger, it must consider not only certain traditional factors regarding the effects of the integration but also data-sharing aspects after the merger. Because consumer data have become key to sellers' business performance, data foreclosure is directly related to the competitive structure. To mitigate any detrimental effect in the case of integration, regulators might force an integrated firm to share customer data with its rivals by demanding that the price of data be set within a reasonable range to guarantee an efficient level of data availability.[23] To do so, the authorities must first create a proper nonprice metric to measure how integration creates data-related concentration, rather than relying on traditional metrics, such as typical turnover

[22] Refer to https://www.congress.gov/115/bills/s2020/BILLS-115s2020is.pdf for details.

[23] For example, U.S. media companies plan to request such a data-sharing regulation in response to the AT&T and WarnerMedia merger.

rates. By imposing such conditions on the final merger approval, even unaffiliated sellers, mostly entrants, can rest assured that the merger will not prevent them from obtaining a necessary dataset about consumers. A policy that overlooks such dynamic effects of privacy regulation could result in a more favorable market situation for the incumbent than the entrant, leading to socially undesirable outcomes.

6 Concluding Remarks

In this paper, I analyze how data-driven vertical integration between a platform and a seller affects market competition when each seller attracts potential customers by creating targeted ads based on personal information obtained from the platform. I show that the platform and the incumbent with better initial targeting technology have incentives to vertically integrate if the incumbent provides a high-quality product, or it provides a low-quality product, but more data are available. In this case, the integrated firm always wants to prevent the unaffiliated entrant from accessing data, thereby adversely affecting the entrant in the form of a smaller market share. Therefore, the entrant that requires consumer data to overcome its initial disadvantage in targeting technology is more likely to be harmed by such integration and the resulting data foreclosure.

Moreover, this process can eventually reduce total social welfare. In particular, vertical integration and data foreclosure are welfare reducing if consumers' privacy concerns are relatively high so that only a limited amount of data is available for targeting. Thus, whether vertical integration and data foreclosure are socially desirable depends on the number of privacy sensitive consumers, which suggests that consumers' privacy concerns that result in less data being available might have unintended adverse consequences on market competition through the effects of vertical integration. Additionally, by incorporating the unique feature of data as nonrival goods in that non-exclusive data provision does not incur an additional marginal cost, the welfare implications emphasize the necessity of data-specific policy concerning data-driven integration followed by data (as an input) foreclosure.

Although privacy and information availability are potentially important areas to investigate to encourage competition, regulators have not yet established concrete antitrust standards regarding this complex interaction. In this respect, the findings of my model help illustrate the relevant issues and can help in the formulation of more balanced managerial and policy recommendations regarding privacy protection and data-driven vertical integration.

References

Acquisti, A. and Varian, H. (2005). Conditioning Prices on Purchase History. *Marketing Science*, 24.

Bajari, P., Chernozhukov, V., Hortacsu, A., and Suzuki, J. (2019). The Impact of Big Data on Firm Performance: An Empirical Investigation. *AEA Papers and Proceedings*, 109:33–37.

Belleflamme, P. and Vergote, W. (2016). Monopoly Price Discrimination and Privacy: The Hidden Cost of Hiding. *Economics Letters*, 149:141–144.

Binns, R. and Bietti, E. (2020). Dissolving Privacy, One Merger at a time: Competition, Data and Third Party Tracking. *Computer Law & Security Review*, 36:105–369.

Braulin, F. and Valletti, T. (2016). Selling Customer Information to Competing Firms. *Economics Letters*, 149:10–14.

Campbell, J., Goldfarb, A., and Tucker, C. (2015). Privacy Regulation and Market Structure. *Journal of Economics and Management Strategy*, 24(1):47–73.

Casadesus-Masanell, R. and Hervas-Drane, A. (2015). Competing with Privacy. *Management Science*, 61(1):229–246.

Chen, Z., Choe, C., Cong, J., and Matsushima, N. (2020). Data-Driven Mergers and Personalization. *ISER Discussion Paper 1108*.

Chipty, A. (2001). Vertical Integration, Market Foreclosure, and Consumer Welfare in the Cable Television Industry. *American Economic Review*, 91:428–453.

Chirita, A. (2020). Data-Driven Mergers Under EU Competition Law. *The Future of Commercial Law: Ways Forward for Harmonisation*.

Conitzer, V., Taylor, C., and Wagman, L. (2012). Hide and Seek: Costly Consumer Privacy in a Market with Repeat Purchases. *Marketing Science*, 31(2):277–292.

Cremer, H. and Thisse, J. (1991). Location Models of Horizontal Differentiation: A Special Case of Vertical Differentiation Models. *Journal of Industrial Economics*, 39(4):384–390.

De Cornière, A. and Taylor, G. (2020). Data and Competition: a General Framework with Applications to Mergers, Market Structure, and Privacy Policy. *CEPR Discussion Paper*, (DP14446).

Fang, H. and Kim, S. (2021). Data Neutrality and Market Competition. *Working Paper*.

Fudenberg, D. and Tirole, J. (1991). Perfect Bayesian Equilibrium and Sequential Equilibrium. *Journal of Economic Theory*, 53(2):236–260.

Fudenberg, D. and Tirole, J. (2000). Customer Poaching and Brand Switching. *RAND Journal of Economics*, 31(4):634–657.

Gu, Y., Madio, L., and Reggiani, C. (2019). Exclusive Data, Price Manipulation and Market Leadership. *CESifo Working Paper 7853*.

Hart, O. and Tirole, J. (1990). Vertical Integration and Market Foreclosure. *Brookings Papers on Economic Activity: Microeconomics*, pages 205–276.

Kathuria, V. and Globocnik, J. (2020). Exclusionary Conduct in Data-driven Markets: Limitations of Data Sharing Remedy. *Journal of Antitrust Enforcement*, pages 1–24.

Kim, J., Wagman, L., and Wickelgren, A. (2019). The Impact of Access to Consumer Data on the Competitive Effects of Horizontal Mergers. *Journal of Economics and Management Strategy*, 28(3):373–391.

Koh, B., Raghunathan, S., and Nault, B. (2017). Is Voluntary Profiling Welfare Enhancing? *MIS Quarterly*, 41(1).

Kummer, M. and Schulte, P. (2019). When Private Information Settles the Bill: Money and Privacy in Google's Market for Smartphone Applications. *Management Science*, 65(8):3470–3494.

Levin, J. and Milgrom, P. (2010). Online Advertising: Heterogeneity and Conflation Market Design. *American Economic Review: Papers & Proceedings*, 100:603–607.

McAfee, R. and Schwartz, M. (1994). Opportunism in Multilateral Vertical Contracting: Nondiscrimination, Exclusivity, and Uniformity. *American Economic Review*, 84(1):210–230.

Montes, R., Sand-Zantman, W., and Valletti, T. (2018). The Value of Personal Information in Markets with Endogenous Privacy. *Management Science*, 65(3).

O'Brian, D. and Shaffer, G. (1992). Vertical Control with Bilateral Contracts. *RAND Journal of Economics*, 23(3):299–308.

Reisinger, M. and Tarantino, E. (2015). Vertical Integration, Foreclosure, and Productive Efficiency. *RAND Journal of Economics*, 46(3):461–479.

Rey, P. and Tirole, J. (2007). A Primer on Foreclosure. *Handbook of Industrial Organization III.*

Rey, P. and Vergé, T. (2004). Bilateral Control with Vertical Contracts. *RAND Journal of Economics*, 35(4):728–746.

Shy, O. and Stenbacka, R. (2016). Customer Privacy and Competition. *Journal of Economics and Management Strategy*, 25(3):539–562.

Taylor, C. (2004). Consumer Privacy and the Market for Customer Information. *RAND Journal of Economics*, 35(4):631–650.

Taylor, C. and Wagman, L. (2014). Customer Privacy in Oligopolistic Markets: Winners, Losers, and Welfare. *International Journal of Industrial Organization*, 34:80–84.

Villas-Boas, J. (1999). Dynamic Competition with Customer Recognition. *RAND Journal of Economics*, 30(3):604–631.

Villas-Boas, J. (2004). Price Cycles in Markets with Customer Recognition. *RAND Journal of Economics*, 35(3):486–501.

Appendix

A Further Discussions

A.1 The Role of the Platform Beyond Being a Data Broker

In the main model, I have assumed that the platform only plays the role of a data broker that derives revenue from selling data to downstream sellers. However, in some cases, platforms are more engaged in downstream market competition by internalizing some benefits from sellers. For example, transaction-based e-commerce platforms usually take a certain proportion of revenue from sellers as usage fees: an individual seller on Amazon pays \$0.99 for each sale, whereas a professional seller pays variable fees of approximately 13% of total sales revenue. To investigate such cases in which the platform is more than only a data broker, I change the model such that the platform's profit function includes not only the revenue from selling data but also the fixed portion of sellers' product sales revenue. Then, the platform's profit maximization problem is given as follows:

$$\max_{C_j^{Int},C_{-j}^{Int}} \quad \pi_p^{Int}(C_j^{Int}, C_{-j}^{Int}|\mathbb{P}(i\in\mathcal{D})) = \begin{cases} f\times \sum\limits_{j\in\{I,E\}} P_jX_j & \text{if no seller buys data} \\ \mathbb{P}(i\in\mathcal{D})(C_j^{Int}-\phi) + f\times \sum\limits_{j\in\{I,E\}} P_jX_j & \text{if only seller } j \text{ buys data} \\ \mathbb{P}(i\in\mathcal{D})(C_j^{Int}+C_{-j}^{Int}-\phi) + f\times \sum\limits_{j\in\{I,E\}} P_jX_j & \text{if both sellers buy data,} \end{cases} \tag{18}$$

where $f \in (0,1)$ is exogenously given, and the superscript Int denotes the platform that internalizes benefits from the downstream market. Accordingly, the profit function for each seller j will also be changed to $\pi_j = (1-f)P_jX_j - \mathbb{1}_{\{\text{Buy}\}}C_j^{Int} \times \mathbb{P}(i \in \mathcal{D})$. From this extension, I first find that allowing the platform to internalize some of the downstream sellers' benefits does not change the equilibrium outcomes in the game with vertical integration. Given that any unaffiliated seller's profits when using data are greater than when being foreclosed, the integrated firm now has a greater incentive to sell data. However, this incentive is offset by the lower data price charged to the unaffiliated seller due to the platform's additional rent extraction: $C_j^{Int} = (1-f)C_j$. Thus, Propositions 2 and 3 still hold in this extension.

A.2 Incumbent with Pre-existing Data

As mentioned earlier, it can also be the case in which seller I has more pre-existing data than seller E such that $m_I(D_I) = \delta + D_I$ and $m_E(D_E) = (1 + D_E)^\alpha$ where $\delta > 1$. In this section, I examine whether the main findings obtained under Assumption 1 still hold in this extension.

First, as long as a secret contracting environment with incomplete information is considered, no vertical integration game leads to non-exclusive data acquisition for both downstream sellers, as summarized in Lemma 1.

In the game with vertical integration, the results depend on which seller offers a high-quality product. Given that a pre-existing amount of data $\delta > 0$ gives additional benefits to seller I, I focus on the case in which E is advantageous in the product quality dimension. I compare the integrated firm's profit levels under different scenarios. As in Section 3.2, there are three possible cases: $I - Foreclose$ or $E - Sell$ for $\tau \in [\tau'_{v,pre}, 1]$, $I - Foreclose$ or $E - Foreclose$ for $\tau \in [\tau''_{v,pre}, \tau'_{v,pre}]$, and $I - Sell$ or $E - Foreclose$ for $\tau \in [0, \tau''_{v,pre}]$, where subscript v, pre denotes the vertical integration game with pre-existing data. By the same logic as in main model, I find a threshold on τ^c, denoted as $\tau_{v,pre}$, above (below) which the platform is integrated with (I) E, and data monopolization follows in either case; thus, the qualitative results still hold. One noticeable difference is that $\tau_{v,pre}$ is smaller than τ_v, implying that vertical integration with I and the resulting data foreclosure is more likely to emerge in equilibrium, as shown in left panel in Figure 3, when I has an additional targeting advantage with a pre-existing set of data.[24]

From the welfare analysis, I find a threshold on τ^c, denoted as τ^δ_{SW}, below which non-exclusive data acquisition, i.e., no integration equilibrium (B, B), is welfare enhancing while exclusive data acquisition, i.e., integration equilibrium (B, N), is not as shown in the right panel in Figure 3. Thus,

[24] The detailed proof, which follows a similar logic to the proof of Proposition 3 is omitted here and available upon request.

the main welfare implication, as in Proposition 4, still holds.

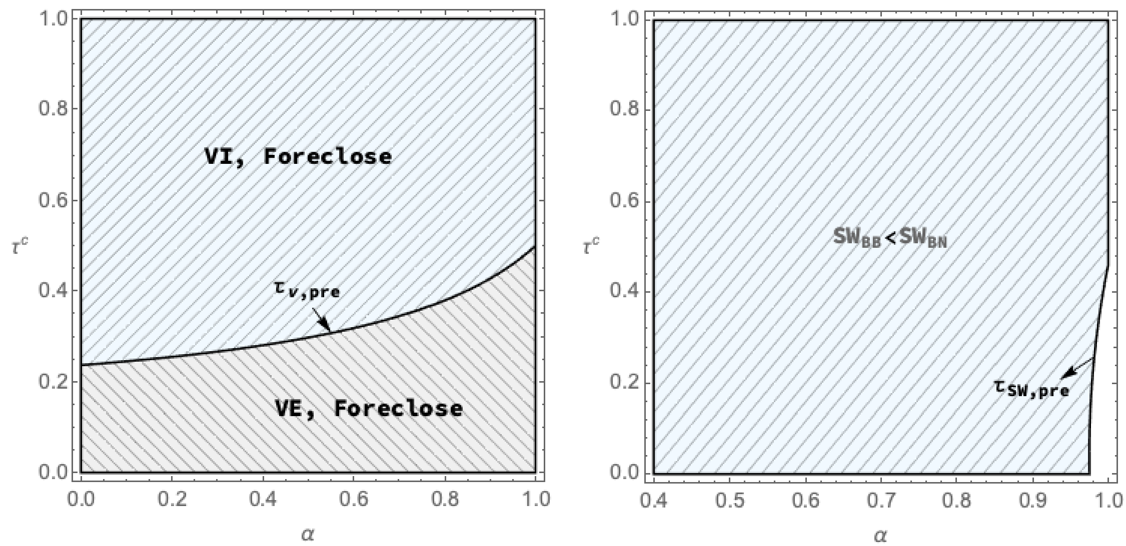

Figure 3: Welfare comparison when I has a pre-existing amount of data and E provides high-quality product, assuming that $q = 2$ and $\delta = 3/2$.

B Omitted Proofs

Proof of Proposition 1. Given that the distribution function F is continuous over the closed unit interval [0,1], there exists a fixed point, τ^c, by the fixed point theorem. □

Proof of Lemma 1. Given that the platform and one seller establish an exclusive data contract, the platform is nonetheless able to extract positive rents from the non-partner seller by providing data, which results in a non-exclusive data acquisition equilibrium. Suppose $q \equiv q_I > q_E = 1$. If the platform first offers exclusive data access to seller I, I raises its price to $P_I^{BN} = \frac{2(q-1)+\tau^c(2-2^\alpha+\tau^c)}{3(2\tau^c-1)}$. Knowing this fact, if the platform secretly offers data provision to seller E, its profit maximizing price is set at $P_E(P_I^{BN}, \Delta_{BB}) = \frac{\tau^c[3(\tau^c+2)^\alpha-2^\alpha+\tau^c-1]+2(q-1)-3\tau^c(\tau^c+1)}{6(2\tau^c-1)}$, resulting in $\pi_E = \frac{\{-\tau^c[-3(\tau^c+2)^\alpha+2^\alpha+2\tau^c+4]+2q-2\}\{\tau^c[3(\tau^c+2)^\alpha-2^\alpha+4\tau^c+2]+2q-6\tau^c(\tau^c+1)-2\}}{36(q-1)(2\tau^c-1)}$, which is decreasing in τ^c but increasing in α. Given that the minimum possible value of π_E, evaluated at $\tau^c = 1$ and $\alpha = 0$, is $\frac{(q-3)^2}{4(q-1)} > 0$, the platform maximizes its profit by renegotiating with seller E and extracting positive rents arising from positive π_E. Similarly, the platform's incentive to renegotiate with seller I when offering exclusive data access to seller E can be demonstrated. Additionally, a similar logic can be applied to the case of $q_E > q_I$. □

Proof of Proposition 2. I assume that $q \equiv q_I > q_E = 1$. By the equilibrium profit levels under different data acquisition cases derived from the solution to each seller's profit maximization problem, Equations (9) and (11) show the profit comparison in the game with vertical integration.

When the platform is integrated with seller I, it is easy to show that $\pi_{VI,S}^{IH} \leq \pi_{VI,F}^{IH}$: it is sufficient to show that the maximum possible value of $\pi^{IH}_{VI,S} - \pi^{IH}_{VI,F}$ remains negative. Given that $\pi^{IH}_{VI,S} - \pi^{IH}_{VI,F}$ is increasing in τ^c but decreasing in α, I show that $\pi^{IH}_{VI,S} - \pi^{IH}_{VI,F}$ evaluated at $\tau^c = 1$ and $\alpha = 0$ is zero, implying that $\pi^{IH}_{VI,S} \leq \pi^{IH}_{VI,F}$ for $\tau \in [0, 1]$ and $\alpha \in [0, 1]$. Similarly, when the platform is integrated with seller E, $\pi^{IH}_{VE,S} \geq \pi^{IH}_{VE,F}$ can be shown by demonstrating that the minimum possible value of $\pi^{IH}_{VE,S} - \pi^{IH}_{VE,F}$ remains positive: since it decreases in τ^c and α, its minimum is attained at $\alpha = 1$ and $\tau^c = 1$. The minimum value is $\frac{2(q-2)}{9(q-1)}$, which is always greater than zero for $q \geq 2$.

Then, I compare $\pi^{IH}_{VI,F}$ with $\pi^{IH}_{VE,S}$. First, $\pi^{IH}_{VE,S} - \pi^{IH}_{VI,F} =$
$$\frac{\tau^c\left\{\tau^c\left[2^{\alpha+1}(\tau^c+2)-4^\alpha+(\tau^c)^2+4\tau^c+2\right]-2[2\tau^c(\tau^c+1)+1](\tau^c+2)^\alpha+\tau^c(\tau^c+2)^{2\alpha}\right\}-4(2^\alpha-3)\tau^c-3q^2+q\left\{2\tau^c\left[(\tau^c+2)^\alpha+2^{\alpha+1}-6\right]-2(\tau^c)^2+6\right\}-3}{9(q-1)(2\tau^c-1)}$$
is increasing in τ^c and α. Its maximum possible value, evaluated at $\tau^c = 1$ and $\alpha = 1$, is $\frac{-5+3q(2-q)}{9(q-1)} < 0$, which is negative. Since the maximum possible value is negative, $\pi^{IH}_{VE,S} < \pi^{IH}_{VI,F}$. □

Proof of Proposition 3. Assuming that E provides a high-quality product, the profit comparison in the case with the integration with I is given as follows:

$$\pi_{VI,S} - \pi_{VI,F} = \frac{2\tau^c\left[(\tau^c+2)^\alpha - 2^\alpha\right]\left\{\tau^c\left[(\tau^c+2)^\alpha + 2^\alpha - 2\tau^c - 4\right] + q - 1\right\}}{9(q-1)}. \tag{19}$$

From Equation (19), I find that if $\tau^c > \tau''_v$, then $\pi_{VI,S} < \pi_{VI,F}$.[25] Similarly, I examine the result when the platform merges with seller E. By the above logic, the difference between the two profit levels is as follows:

$$\pi_{VE,S} - \pi_{VE,F} = \frac{2(\tau^c)^2\left\{\tau^c\left[-2(\tau^c+2)^\alpha + \tau^c + 4\right] - q + 1\right\}}{9(q-1)}. \tag{20}$$

From Equation (20), I find that if $\tau^c > \tau'_v$, then $\pi_{VE,S} > \pi_{VE,F}$.[26] To determine which seller offers a sufficient incentive to induce the platform to integrate, I compare the profits from integration with I to those from integration with E. There are three possible cases: $I - Foreclose$ or $E - Sell$ for $\tau \in [\tau'_v, 1]$, $I - Foreclose$ or $E - Foreclose$ for $\tau \in [\tau''_v, \tau'_v]$, and $I - Sell$ or $E - Foreclose$ for $\tau \in [0, \tau''_v]$. Because Equations (19) and (20) are continuous functions of τ^c and α, the domains of which are compact, their signs can be graphically mapped over the relevant set of parameters, as shown in Figure 4. First, for $\tau^c \in [0, \tau''_v]$, $\pi_{VI,S}$ is compared with $\pi_{VE,F}$ as follows:

$$\pi_{VI,S} - \pi_{VE,F} = \frac{\left\{\tau^c\left[-(\tau^c+2)^\alpha + \tau^c + 2\right] + q - 1\right\}^2 - \left\{\tau^c\left[(\tau^c+2)^\alpha - 2\right] + 2q - 2\right\}^2 + \tau^c\left[(\tau^c+2)^\alpha - 2^\alpha\right]\left\{\tau^c\left[(\tau^c+2)^\alpha + 2^\alpha - 2\tau^c - 4\right] + 4q - 4\right\}}{9(q-1)}. \tag{21}$$

Given that Equation (21) is increasing in τ^c and α, its maximum possible value, evaluated at $\tau^c = \tau''_v$ and $\alpha = 1$, can be shown to be negative, suggesting that $\pi_{VI,S} < \pi_{VE,F}$ for $\tau^c \in [0, \tau''_v]$. For $\tau^c \in [\tau''_v, \tau'_v]$, $\pi_{VI,F}$ is compared with $\pi_{VE,F}$ as follows:

[25] τ''_v is the solution to $\pi_{VI,S} - \pi_{VI,F} = 0$.
[26] τ'_v is the solution to $\pi_{VE,S} - \pi_{VE,F} = 0$.

$$\pi_{VI,F} - \pi_{VE,F} = \frac{\left[\tau^c\left(-2^\alpha + \tau^c + 2\right) + q - 1\right]^2 - \left\{\tau^c\left[(\tau^c+2)^\alpha - 2\right] + 2q - 2\right\}^2}{9(q-1)}. \tag{22}$$

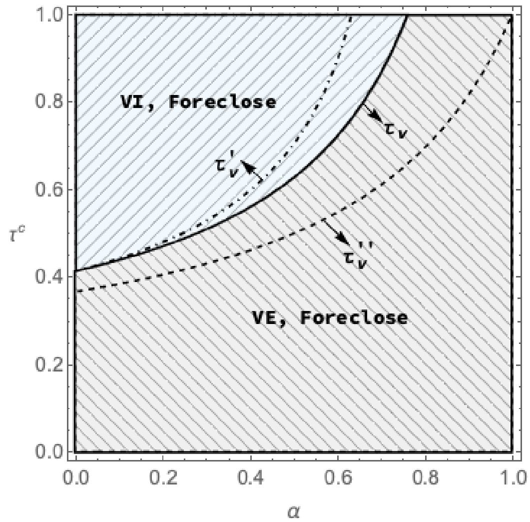

Figure 4: Vertical integration equilibrium if E provides a high-quality product.

From Equation (22), I find a threshold on τ^c, denoted as τ_v, below which $\pi_{VI,F} < \pi_{VE,F}$, where τ_v solves $\pi_{VI,F} = \pi_{VE,F}$. For $\tau^c \in [\tau_v', 1]$, $\pi_{VI,F}$ is compared with $\pi_{VE,S}$.

$$\pi_{VI,F} - \pi_{VE,S} = -\frac{(\tau^c)^2\left\{\tau^c\left[-2(\tau^c+2)^\alpha + \tau^c + 4\right] + 2q - 2\right\} - \left[\tau^c\left(-2^\alpha + \tau^c + 2\right) + q - 1\right]^2 + \left\{\tau^c\left[-(\tau+2)^\alpha + \tau^c + 2\right] - 2q + 2\right\}^2}{9(q-1)}. \tag{23}$$

Equation (23) increases in τ^c but decreases in α. Its minimum possible value, attained at $\tau^c = \tau_v'$ and $\alpha = 1$, is positive, implying that $\pi_{VI,F} > \pi_{VE,S}$. Additionally, as shown in Figure 4, I verify that $\tau_v'' < \tau_v < \tau_v'$ for $\alpha \times \tau^c \subset [0,1] \times [0,1]$.[27] □

Proof of Corollary 1. As shown above, τ_v is obtained from Equation (22). The effect of α on τ_v is obtained using the implicit function theorem. Given $G(\tau^c;\alpha) \equiv \pi_{VI,F} - \pi_{VE,F}$, I show the following.

$$\frac{d\tau_v}{d\alpha} = -\frac{2\tau^c\left\{(\tau^c+2)^\alpha \log(\tau^c+2)\left\{-\left\{\tau^c\left[(\tau^c+2)^\alpha - 2\right] + 2q - 2\right\}\right\} - 2^\alpha \log(2)\left[\tau^c\left(-2^\alpha + \tau^c + 2\right) + q - 1\right]\right\}}{2\left(-2^\alpha + 2\tau^c + 2\right)\left[\tau^c\left(-2^\alpha + \tau^c + 2\right) + q - 1\right] - 2\left[(\tau^c+2)^{\alpha-1}(\alpha\tau^c + \tau + 2) - 2\right]\left\{\tau^c\left[(\tau^c+2)^\alpha - 2\right] + 2q - 2\right\}}. \tag{24}$$

Because Equation (24) is a continuous function of τ^c and α, the domains of which are compact, their signs can be graphically mapped over the relevant set of parameters, given q, as shown in Figure 5. □

Proof of Proposition 4. First, I decompose CS into two parts: $CS = CS_{pf} + CS_K$, where CS_{pf} (CS_K) denotes the surplus from using the platform's services (from buying a product). I must show that $\frac{\partial CS}{\partial \tau^c} > 0$, where CS is defined in Equation (15). Equation (15) consists of two parts: CP_{pf} and

[27] Figures depicting the profit comparisons over parameter space $\alpha \times \tau^c \subset [0,1] \times [0,1]$ are available upon request.

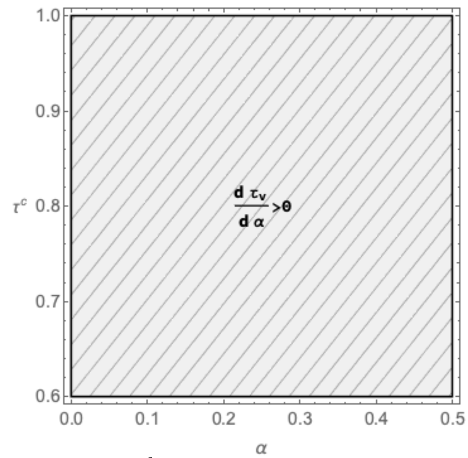

Figure 5: The shaded area represents $\frac{d\tau_v}{d\alpha} > 0$ over the $\alpha \times \tau^c$ parametric region fixing $q = 2$.

CS_K. $\frac{\partial CS_{pf}}{\partial \tau^c} > 0$ is easily shown using the Leibniz integral rule as follows:

$$\frac{\partial CS_{pf}}{\partial \tau^c} = [v(\tau^c) - \psi(\tau^c)] + \int_0^{\tau^c} v'(\tau^c) f(x) dx > 0 \quad \because v' > 0. \tag{25}$$

Next, to show that $\frac{\partial CS_K}{\partial \tau^c} > 0$, I first simplify $\frac{\partial CS_K}{\partial \tau^c}$ as follows:

$$\begin{aligned}\frac{\partial CS_K}{\partial \tau^c} = &\int_0^{\theta^c_{\mathcal{D}}} \Big(\theta + m_{-j}(D_{-j}) - P_{-j}\Big) d\theta + \int_{\theta^c_{\mathcal{D}}}^1 \Big(\theta q + m_j(D_j) - P_j\Big) d\theta - \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}} \Big(\theta - P_{-j}\Big) d\theta - \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1 \Big(\theta q - P_j\Big) d\theta \\ &+ \tau^c \left[\int_0^{\theta^c_{\mathcal{D}}} \left(\theta + \frac{\partial m_{-j}(D_{-j})}{\partial \tau^c} - \frac{\partial P_{-j}}{\partial \tau^c}\right) d\theta + \int_{\theta^c_{\mathcal{D}}}^1 \left(\theta q + \frac{\partial m_j(D_j)}{\partial \tau^c} - \frac{\partial P_j}{\partial \tau^c}\right) d\theta\right].\end{aligned} \tag{26}$$

$\frac{\partial CS_K}{\partial \tau^c}$ is positive for all possible data acquisition cases. For example, assuming that E provides a high-quality product, indicating that seller j is the entrant, i.e., $\alpha_j = \alpha$ and $\alpha_{-j} = 1$, in the case of (B, B), Equation (26) is increasing in α and τ^c. Its minimum possible value, which is attained at $\alpha = 0$ and $\tau^c = 0$, can be shown to remain positive, $\frac{4}{9}$. If seller j is the incumbent, i.e., $\alpha_j = 1$ and $\alpha_{-j} = \alpha$, Equation (26) is also increasing in α and τ^c. In the cases of (B, N) and (N, B), Equation (26) is increasing in α and τ^c. The minimum possible values of $\frac{\partial CS_K}{\partial \tau^c}$ in these two cases are all $\frac{4}{9}$, implying that CS_K is increasing in τ^c. The same logic can be applied to all other cases. □

Proof of Proposition 5. First, the difference between social welfare levels when $2 = q \equiv q_E > q_I = 1$ is obtained as follows:

$$\begin{aligned}SW_{BB} - SW_{BN} = &\frac{1}{6}\Big\{4(\tau^c+2)^\alpha + \tau^c\Big\{-8\tau^c(\tau^c+1)(\tau^c+2)^\alpha + 16(\tau^c+2)^\alpha + 4(\tau^c-1)(\tau^c+2)^{2\alpha} \\ &+ \tau^c\left[2^{\alpha+3}(\tau^c+1) - 4^{\alpha+1} - 3\right] - 5 \times 2^{\alpha+1} + 4^{\alpha+1} - 12\Big\} - 3 \times 4^\alpha + 2^{\alpha+3} - 12\Big\}. \\ SW_{BB} - SW_{NB} = &\frac{1}{6}\Big\{4\left[2(\tau^c-1)(\tau^c)^2 - 3\right](\tau^c+2)^\alpha + 3(\tau^c+2)^{2\alpha} - 2\left[2(\tau^c-1)(\tau^c+4)(\tau^c)^2 + \tau^c - 6\right]\Big\}.\end{aligned} \tag{27}$$

By depicting two equations in (27) on $\alpha \times \tau^c \subset [0,1] \times [0,1]$, it can be shown that there exists a threshold of τ^c, denoted as τ_{SW}, below which $SW_{BB} > SW_{BN}$ or $SW_{BB} > SW_{NB}$, and the reverse holds otherwise, as shown in Figure 2. Similar findings are obtained in the case in which seller I provides a high-quality product. □

Proof of Proposition 6. Assuming that E provides a high-quality product, the right-hand side of Equation (17) given each data acquisition case is as follows.

$$RHS_{BB} = \theta^c_{BB} F\Big(\sqrt{\tfrac{1+\tau^c}{2}}\Big) + (1-\theta^c_{BB})F\Big(\sqrt{\tfrac{(1+\tau^c)^\alpha}{2}}\Big); \quad RHS_{NB} = \theta^c_{NB} + (1-\theta^c_{NB})F\Big(\sqrt{(1+\tau^c)^\alpha}\Big)$$

$$RHS_{BN} = \theta^c_{BN} F\Big(\sqrt{1+\tau^c}\Big) + (1-\theta^c_{BN}); \quad RHS_{NN} = \theta^c_{NN} + (1-\theta^c_{NN}) = 1,$$

where $F\Big(\frac{1}{n}\Big) = 1$ if $n = 0$. The equilibrium τ^c in each data acquisition case is the fixed point satisfying Equation (17). I show that the $RHSs$ in the four data acquisition cases can be ranked. First, it is obvious that RHS_{NN} is the largest. Then, it remains to be shown that $RHS_{BB} < \min\{RHS_{NB}, RHS_{BN}\}$ and $RHS_{NN} > \max\{RHS_{NB}, RHS_{BN}\}$. As for RHS_{NB} and RHS_{BB}, given that $\theta^c_{NB} < \theta^c_{NN} < \theta^c_{BB} < \theta^c_{BN}$, it is enough to show that $\theta^c_{NB} - \theta^c_{BB}F\Big(\sqrt{\frac{1+\tau^c}{2}}\Big) + (1 - \theta^c_{NB})F(\sqrt{(1+\tau^c)^\alpha}) - (1-\theta^c_{BB})F\Big(\sqrt{\frac{(1+\tau^c)^\alpha}{2}}\Big) > (\theta^c_{BB} - \theta^c_{NB})F(\sqrt{\frac{(1+\tau^c)^\alpha}{2}}) + \theta^c_{NB} - \theta^c_{BB}F\Big(\sqrt{\frac{1+\tau^c}{2}}\Big) \geq 0$, implying that $RHS_{NB} \geq RHS_{BB}$. Using similar logic, $RHS_{BN} > RHS_{BB}$, $RHS_{NB} < RHS_{NN}$, and $RHS_{BN} < RHS_{NN}$ can be shown. This implies that RHS_{BB} is the lowest, while RHS_{NN} is the highest. The relative size of RHS_{BN} and RHS_{NB} depends on the size of α as shown in Figure 6.

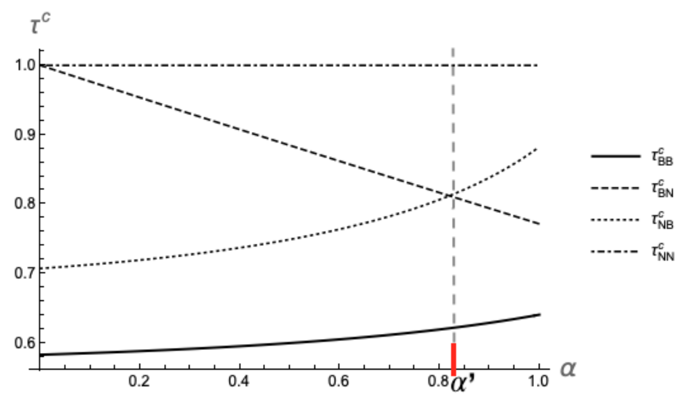

Figure 6: The equilibrium amount of data under different data acquisition cases, assuming that F is the uniform distribution and $q = 2$.

Finally, I prove that the fixed point for each case uniquely exists and that the fixed point that hits the lower RHS is smaller than another fixed point that hits the higher RHS. Let $G_k : [0,1] \to \mathbb{R}$ be defined by $G_k \equiv \tau^c - RHS_k$, where $k = \{BB, BN, NB, NN\}$. Note that $0 < RHS_k(0)$ and $1 > RHS_k(1)$, which leads to $G_k(0) < 0 < G_k(1)$. Let me first show that $\tau^c_{BB} < \tau^c_{NN}$. By the intermediate value theorem, there exists at least one $\tau^c_{BB} \in (0,1)$ such that $G_{BB}(\tau^c_{BB}) = 0$. I now prove by contradiction that there exists such a unique τ^c_{BB}. Suppose that τ^c_{BB} and τ'^c_{BB} exist such that $0 < \tau^c_{BB} < \tau'^c_{BB} < 1$ and $G_{BB}(\tau^c_{BB}) = G_{BB}(\tau'^c_{BB}) = 0$. By Rolle's theorem, there exists $\tau^c_0 \in$

$(\tau^c_{BB}, \tau'^c_{BB})$ such that $G'_{BB}(\tau^c_0) = 0$. By the mean value theorem, there exists $\tau^c_* \in (0, \tau^c_{BB})$ such that $G'_{BB}(\tau^c_*) = -G_{BB}(0)/\tau^c_{BB} > 0 = G'_{BB}(\tau^c_0)$. However, G'_{BB} is nondecreasing, hence the contradiction. Using similar logic, I can show that τ^c_{BB} and τ^c_{NN} are unique. It remains to be shown that $\tau^c_{BB} < \tau^c_{NN}$ for $RHS_{BB} < RHS_{NN}$. Given that $RHS_{BB} < RHS_{NN}$, $\tau^c_{BB} = RHS_{BB}(\tau^c_{BB}) < RHS_{NN}(\tau^c_{BB})$ and $\tau^c_{NN} = RHS_{NN}(\tau^c_{NN})$. This implies that $G_{NN}(\tau^c_{BB}) < 0 = G_{NN}(\tau^c_{NN})$. By the mean value theorem, there exists $\tau'^c_{NN} \in (0, \tau^c_{NN})$ such that $G'_{NN}(\tau'^c_{NN}) = -G_{NN}(0)/\tau'^c_{NN} > 0$. By the convexity of G_{NN}, $G'_{NN}(\tau^c) \geq G'_{NN}(\tau'^c_{NN})$ for $\forall \tau^c \in (\tau'^c_{NN}, 1)$. Thus, G_{NN} is increasing on $[\tau'^c_{NN}, 1]$. Moreover, $G_{NN}(\tau^c) \geq 0$ for $\forall \tau^c \in [\tau^c_{NN}, 1]$. Because $G_{NN}(\tau^c_{BB}) < 0$, $\tau^c_{BB} < \tau^c_{NN}$. The remaining cases can be proved in the same way. □