Privacy, Information Acquisition, and Market Competition*

Soo Jin Kim†

ShanghaiTech University

March 30, 2020

Abstract

This paper analyzes how data-driven vertical integration between a platform and one downstream seller affects market outcomes in a two-sided market where sellers with asymmetric targeting skills target advertisements to individuals who have varying privacy concerns. I show that data-driven vertical integration leads to the incumbent's exclusive use of data. Therefore, a market entrant that has worse targeting technology than an incumbent is disproportionately harmed by such integration. The welfare analysis shows that integration is welfare reducing if consumers' privacy concerns are relatively high. Therefore, individually optimal decisions on 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, 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 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

It is well known that platforms such as Internet service providers (ISPs) and social media also serve as data intermediaries that collect and sell users' personal information. These data intermediaries sell data 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, which in turn influences the overall effectiveness of ad targeting for each seller that obtains data from the platforms. Given that more effective targeted ads can attract more consumers, data availability for sellers affects the competitive structure of the downstream market.

If each seller has the most current information about potential customers, it can target its ads to better attract these customers. When a seller's targeted ads become more effective, consumers face lower mismatch costs from that seller: namely, consumers will spend less time searching for the most suitable product because they immediately obtain the relevant information from these targeted ads. This potential benefit from a loss of privacy may also be asymmetric. A consumer is likely to face a much higher mismatch cost from small sellers or from market entrants that have weaker initial targeting skills: incumbents have better initial targeting technology that has been developed based on previous sales experience or existing customer data, whereas entrants lack such experience. For example, suppose that a major retailer, such as Walmart.com (as the incumbent), and a new retailer (as the entrant) buy the same set of data 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. In a two-sided market, 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 sellers' 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

[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 who 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.

extensive customer data. Most previous studies focus on how such data-driven integration could lead to much greater privacy concerns on the consumer side (e.g., Chirita (2019); Binns and Bietti (2019)). However, none of these papers investigate the effects of such data-driven mergers, which alters data availability, on the relevant market competition, even though the market consequences ultimately affect consumers: if integration changes the platform's data-sharing practice for unaffiliated sellers, it changes the amount of information available in the market, which plays a key role in attracting more customers. Hence, this study takes a different stance 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 can be considered another example of data-driven integration: by using Amazon's extensive amount of transaction data, Whole Foods is able to 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, the merger may lead to antitrust concerns.

Such potential anti-competitive threats are greater if a small or entrant seller is foreclosed 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 for the purpose of data sharing depends not only on the total amount of data available for sales, which is determined by consumers' overall privacy concerns, but also on the targeting asymmetry between sellers.

The latter point—that is, how asymmetry in downstream sellers affects the equilibrium outcome of data-driven vertical integration—raises several interesting research questions. How does an 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, which means that it demands considerably more personal data than the incumbent, disproportionately affected by the market outcome? How does the asymmetry ultimately affect consumers

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

and social welfare? 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. Products offered by sellers are horizontally differentiated. These sellers, if not integrated with the platform, decide whether to purchase consumer data from the platform and subsequently engage in price competition. In the model, a consumer is a two-dimensional type with respect to sensitivity to privacy and the preference for one brand (seller) over the other. Depending on his privacy type, each consumer decides whether to disclose personal information to the platform by comparing the benefit arising from active participation in the platform to the nuisance cost arising from a loss of privacy. The platform aggregates all available detailed personal data and sells them to any seller who wants to use them to create targeted ads. A novel aspect of the model is that it permits the platform to integrate with one downstream seller, and the integrated firm then decides whether to exclusively use data for its own affiliated seller or to sell data to the unaffiliated seller. This analysis allows me to investigate how such data-driven integration affects market competition through the seller's data acquisition status, which ultimately affects consumers.

The main results show that the platform has an incentive to sell data to both the incumbent and entrant at different prices absent vertical integration, as described in Section 3. If a data-driven merger is permitted, as in Section 4, 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. Given that the entrant has worse targeting skills, vertical integration followed by data foreclosure that grants a considerable competitive advantage to the affiliated incumbent can be particularly anti-competitive: I show that data foreclosure makes the entrant suffer from a lower market share under vertical integration.

The welfare analysis results suggest that such integration not only harms the entrant but

also can lower consumer surplus and total social welfare, as shown in Section 5. Specifically, the market-driven integration equilibrium outcome and the resulting data foreclosure make consumers worse off if their privacy concerns are relatively high, meaning that less data are available for targeted ads. Intuitively, vertical integration can either enhance or reduce consumer welfare. On the one hand, vertical integration that leads to asymmetric data acquisition for sellers harms consumers because the exclusive use of data by an incumbent with better initial targeting skills allows the integrated firm to dominate the downstream market, leading to a lack of competition. On the other hand, the incumbent's exclusive use of data combined with its better targeting skills lowers consumers' mismatch costs; therefore, more consumers choose the incumbent in this scenario than in the scenario without vertical integration. If the incumbent's customers save sufficiently large mismatch costs, vertical integration that leads to a larger market share for the incumbent might enhance consumer welfare. If overall privacy concerns are high, meaning that only a small amount of data becomes available, the positive effect of vertical integration on lowering mismatch costs is negligible. Thus, vertical integration and the resulting data foreclosure are more likely to harm consumers as the data become limited. However, if sufficient data are available, data foreclosure following integration can benefit consumers. In this sense, individually optimal decisions on data disclosure might not be socially optimal when aggregated, especially when consumers have increased privacy concerns.

Based on these theoretical findings, policy implications with specific remedies are suggested in Section 6. First, I discuss the effectiveness of regulations on data-driven merger conditions. Additionally, other policy remedies that increase consumers' willingness to disclose personal data, such as tax subsidies and reputation-enhancing programs, are also considered, because reduced privacy concerns and a greater availability of data for targeting make the market-driven vertical integration equilibrium outcome socially desirable.

In Section 7, several model extensions show that the main findings are robust to different model specifications. In particular, Section 7.3 shows that allowing the platform to extract a portion of the sellers' profits in the form of usage fees and allowing vertical integration can be complementary in harming the entrant and ultimately reducing consumer welfare. Finally, Section 8 gives the 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), and Villas-Boas (1999, 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 they 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 leads to different competitive market outcomes depending on the individual context and industries. Shy and Stenbacka (2016) suggest that there is a non-monotonic relationship between the degree of privacy protection and equilibrium profits. Casadesus-Masanell and Hervas-Drane (2015) also study similar issues; in their model, as in 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 they then analyze how privacy concerns and the resulting information availability affect competing firms' decisions and social welfare. Recently, De Cornière and Taylor (2020) provide a general theoretical framework that can be flexibly applied in various contexts to analyze how data usage affects competition.

However, asymmetries between firms are persistent for a variety of reasons arising from different previous sales experiences or scopes of products. By taking into account such asymmetries, this paper illuminates important implications that are not present in a symmetric setting: the entrant may be disproportionately harmed by market outcomes. In this sense, Campbell et al. (2015), who demonstrate that small firms or entrants, as specialists rather than generalists, can be adversely affected by privacy regulation that imposes 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 in the horizontally differentiated sellers' market, namely, in initial targeting technology, and consumer heterogeneity with respect to privacy sensitivity. By including such asymmetries and heterogeneity, I offer a microfoundation for understanding how consumers react differently to potential privacy risk and how such endogenous privacy concerns, which determine the amount of data availability for downstream sellers, ultimately affect consumers themselves through disproportionate effects on seller market competition.

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 one of the symmetric downstream firms affect consumers as one of their model applications; Kim et al. (2018), who analyze how access to consumer data that enables personalized pricing affects the overall welfare of horizontal mergers; and Gu et al. (2019), who show how the exclusive use of data affects sellers' incentives to either lead or follow price competition. My paper still differs from the abovementioned studies insofar as I focus on the dynamic relationship between endogenous privacy concerns and the downstream market's competitive structure, which is affected by data-driven vertical integration under an asymmetric seller setup.

This paper also relates to research on privacy and online targeted advertising. De Cornière and Nijs (2016) study the platform's optimal choice of timing for disclosing users' information to advertisers prior to an ad auction. Tucker (2014) reports that users' perceptions of their control over personal information influence the effect of targeted advertising. Goldfarb (2014) emphasizes that targeted ads and information availability can be more important to small advertisers with a focus on the difference between online and offline advertising. This finding of disproportionate effects on small advertisers shares similarities with the findings of my paper. D'Annunzio and Russo (2017) find that if consumers excessively block their personal information without considering the effect of their privacy concerns on advertisers' and publishers' decisions, then the tracking in equilibrium would be too low and could harm consumers and society.[4] Again, my paper provides novel findings by showing how consumers' privacy concerns may backfire against them through the incumbent's exclusive use of data after integration.

2 Theoretical 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 consumers are registered with the platform, and all firms (platform and sellers) have basic information such as an email address, gender, and date of birth, for all consumers. The amount of basic information is normalized to

[4]Regarding the unexpected costs of privacy regulations, Goldfarb and Tucker (2011) also empirically show that privacy regulation increases the intrusiveness of advertising. Calzolari and Pavan (2006) and Kim and Wagman (2015) argue that information disclosure is not always harmful to the individual and may contribute to improving welfare.

one.

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, θ_i denotes consumer i's preference for one seller (brand) over the other, and this is uniformly distributed over the unit line. 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 horizontally differentiated product. Depending on his type (τ_i, θ_i), each consumer who has unit demand for a product makes two independent decisions: (a) whether to disclose his personal information to the platform ($\mathcal{D}$ or $\mathcal{ND}$) and (b) whether to purchase a product from the incumbent or entrant. Therefore, each consumer obtains net utility from the two sources.

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 other users benefit due to the network effect. In that sense, the immediate benefit is increasing in 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 benefit than $i \in \mathcal{ND}$ who only provides basic information: if a user shares his information with others on the platform, he will obtain a greater networking benefit than the inactive users. 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. The nuisance cost might arise from either direct economic losses (e.g., a threat of identity theft) or a negative psychological feeling about disclosing personal information. As shown in other previous studies (e.g., Kummer and Schulte (2019)), I assume that the negative effect of the nuisance cost is mitigated, as the data collector

has a stronger reputation in that privacy concerns are trust-based. In other words, if the platform has a better reputation, consumers have less concern about data breaches. Normalizing both the benefit and cost of $i \in \mathcal{ND}$ to zero,[5] the utility for each consumer i from the platform is given as follows.

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

where the superscript p denotes the platform; $v(\mathbb{P}(i \in \mathcal{D}))$ denotes the immediate benefit from disclosing information, with $v' > 0$ and $v'' \geq 0$; $\frac{\psi(\tau_i)}{r}$ denotes the nuisance cost, with $\psi' > 0$ and $\psi'' \geq 0$; and r represents the platform's reputation. I also assume that $v(x) > \frac{\psi(x)}{r}$ $\forall$ $x \in [0,1]$ and that $\psi(\tau_i)$ is continuous in τ_i.

Because seller j's targeting effectiveness increases in the amount of consumer data, a consumer's information disclosure decision also affects the utility from the purchase of a product: any $i \in \mathcal{ND}$ whose detailed information is not available is likely to suffer from a higher mismatch cost than any $i \in \mathcal{D}$ who provides personal information, as targeted ads suggest products that are better suited to consumers. Normalizing the mismatch cost for $i \in \mathcal{D}$ to zero, the utility specification is given as follows.[6]

$$u_{ij} = V - t|l_j - \theta_i| - P_j - \mathbb{1}_{\{i \in \mathcal{ND}\}} \left(\frac{1}{\gamma_j \Phi_j} \right), \tag{2}$$

where V denotes the reservation value (base utility), which is assumed to be large enough to fully cover the market, and a consumer i's preference with respect to a specific seller (i.e., i's location along the unit line) is given by $\theta_i \sim U[0,1]$. The unit transportation cost is denoted as t, l_j denotes seller j's location, P_j denotes the price of products from seller j, and $\frac{1}{\gamma_j \Phi_j}$ is the mismatch cost where γ_j denotes seller j's initial targeting technology and Φ_j denotes the amount of consumer data possessed by seller j. The indicator function $\mathbb{1}_{\{i \in \mathcal{ND}\}}$ is one if consumer i does not disclose personal information. A consumer is more likely to incur a lower mismatch cost from a seller that has better targeting skills—i.e., higher γ_j. Finally, a consumer's

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

[6] One might argue that any consumer $i \in \mathcal{D}$ should face some positive mismatch cost in the case when seller j does not purchase detailed information. However, as long as the mismatch cost for $i \in \mathcal{ND}$ is higher than that from $i \in \mathcal{D}$, the qualitative results from the current specification always hold. For instance, assuming that the mismatch cost for $i \in \mathcal{ND}$ is $\frac{1}{\gamma_j \Phi_j}$, whereas that for $i \in \mathcal{D}$ is $\frac{\alpha}{\gamma_j \Phi_j}$ where $\alpha \in [0,1)$, the qualitative results in the main model hold. Thus, this normalization is harmless.

mismatch cost is assumed to decrease as seller j obtains more consumer information to create targeted ads, and it is also assumed that the effect of data on reducing mismatch costs is nonincreasing as Φ_j increases.[7] The following is a more intuitive explanation of the mismatch cost. 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, the consumer is more likely to go to Honda than Subaru because, when a consumer goes to Honda, he knows exactly what features to look for in the car type he wants, whereas with Subaru, he still has to look up additional details. Thus, even if he knows the price of the Subaru and that of the Honda—that is, the search cost is sunk—the transaction with Honda is cheaper.

In this specification, consumers who do not provide any additional personal information to the platform also benefit as seller j obtains more aggregate information from the platform. This scenario is plausible due to *information externalities*. For example, firms can categorize consumers into subgroups based on gender and age. In each consumer category, some people provide considerable information about themselves, while others provide nothing. Such information can be transferred to other members of the peer group such that consumers who do not provide any further personal information are still likely to receive some promotional emails.[8]

Lastly, 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 to the platform and do not consider any potential future benefits arising from better-targeted ads when making an information disclosure decision. This assumption is reasonable for many real examples of platforms, such as social media or transaction-based e-commerce 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

[7]The assumption that more data improve targeting effectiveness but at a decreasing rate is typical, as in asymptotic learning theory. See Bajari et al. (2019) for a relevant discussion.

[8]See Choi et al. (2019) for a more detailed description of such information externalities.

relevant ads for baby products. Similarly, Amazon customers subscribe to items that they want to purchase on a regular basis, such as toilet paper and laundry detergent, to receive discounts and enjoy the convenience of scheduled auto-deliveries through Amazon's program called *Subscribe & Save*. Although this subscription gives Amazon additional personal information about customers' daily needs or preferred brands, customers are unlikely to participate in the program only to receive more targeted ads based on the additional pieces of information they provide.[9]

Platform The platform gathers personal information about customers while providing diverse services to them. The amount of data available 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. Although both types of consumers provide basic information to the platform to enjoy the services it offers, the platform sells only detailed information, such as users' relationship status. 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 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 $j \neq -j$.[10] By optimally setting the per unit data price C_j, the platform solves the following profit maximization problem.

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

where the subscript p denotes the *platform*.[11]

In some cases, the platform is more than just a data broker—some platforms, such as Amazon, extract a portion of the profits from the sellers by charging proportional usage fees. In this regard, Section 7.3 analyzes how different profit models for the platform affect the results. Note that this model extension not only shows qualitatively the same implications as in the main model but also serves as an aggravating factor.

[9] For those who are interested in the case of consumers with perfect foresight, see Section 7.1.

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

[11] The choice variable C can be considered to be the data price if the platform sells data to third parties. If it is not allowed to sell data but is only able to use the data to create targeted ads, C can be regarded as a per unit advertising (intermediation) fee.

Sellers Each seller j ($j \in \{I\text{ncumbent}, E\text{ntrant}\}$) sells a set of products to consumers. The set of products offered by each seller is horizontally differentiated in that sellers are located at two extremes ($l_I = 0$ and $l_E = 1$), which means maximum product differentiation. Targeting quality is denoted by $\gamma_j \Phi_j$, where Φ_j is equal to $1 + \mathbb{P}(i \in \mathcal{D})$ if seller j buys data from the platform or one otherwise. If seller j decides to buy data from the platform, he pays per unit data price C_j. Whether he does so depends on the relative magnitudes of C_j and γ_j, where γ_j captures previous sales experience and existing customer information. Without loss of generality, I assume that $\gamma_I > \gamma_E$: seller I has better targeting technology than seller E. For simplicity, I normalize γ_E to one and denote γ_I as γ where $\gamma > 1$.[12] Each seller's profit maximization problem is defined as follows.

$$\max_{P_j, \Phi_j} \quad \pi_j = P_j X_j(P_j, \Phi_j | \gamma) - \mathbb{1}_{\{\text{buy}\}} C_j \times \mathbb{P}(i \in \mathcal{D}), \tag{4}$$

where P_j is the price that seller j charges to consumers and $X_j(P_j, \Phi_j|\gamma)$ is j's aggregate market share. If j buys data from the platform, it needs to pay the price C_j set by the platform. The indicator function $\mathbb{1}_{\{\text{buy}\}}$ is one if seller j buys data from the platform.

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: 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 follows 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 the parenthesis. 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 I use for this game is the Perfect Bayesian Nash Equilibrium (PBE) for multi-period games with observed action, as in Fudenberg and Tirole (1991): PBE consists

[12] The assumption of γ is based on the fact that I has the existing customer information and thus has established stronger data analytic skills combined with previous sales experience. Hagiu and Wright (2020) and Biglaiser et al. (2019) mention such data-driven incumbency advantages. See the Appendix for a detailed discussion.

of a strategy profile for all players and a set of beliefs. These constitute a PBE if all strategies are sequentially rational given the beliefs and the beliefs are consistent given the strategies.[13]

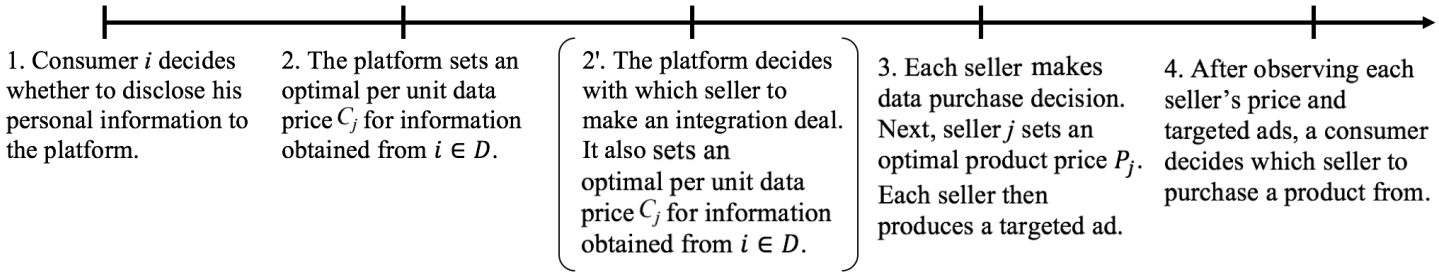

Figure 1: Timing

3 No Vertical Integration

3.1 Equilibrium

In the first stage, each consumer compares the utility levels from disclosure and decides whether to disclose information. Depending on τ_i, any consumer who has $\frac{\psi(\tau_i)}{r} < v(\mathbb{P}(i \in \mathcal{D}))$ will disclose. The portions of privacy-sensitive consumers (not disclosing information) and privacy-insensitive consumers (disclosing) are implicitly determined by the following Proposition 1.[14]

Proposition 1. *There exists a critical point, τ^c, that satisfies the following equation.*

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

A parametric example Let $\tau_i \sim U[0,1]$, $\psi(\tau_i) = \lambda\tau_i^2$, and $v(\tau^c) = 1 + \tau^c$. In this case, τ^c is the solution to $\tau^c = \psi^{-1}(r(1+\tau^c)) = \sqrt{\frac{r(1+\tau^c)}{\lambda}}$. Thus, $\tau^c = \frac{\sqrt{r(4\lambda+r)}+r}{2\lambda}$. Obviously, as λ increases, i.e., as the nuisance cost increases, more people are reluctant to disclose information, so τ^c decreases. As r increases, i.e., the platform's reputation is improved, τ^c also increases.

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

[13] Note that according to Fudenberg and Tirole (1991), if each player has only two possible types that are independent and if both types have nonzero prior probabilities, as in my model, the PBE coincides with the sequential equilibrium.

[14] Note that when each consumer makes his optimal decision on information disclosure, he 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 drop the subscript a.

in (2), the indifference condition is $\theta^c_{\mathcal{ND}} = \frac{1}{2} + \frac{P_E - P_I + (\frac{1}{\Phi_E} - \frac{1}{\gamma\Phi_I})}{2t}$ for $i \in \mathcal{ND}$ and $\theta^c_{\mathcal{D}} = \frac{1}{2} + \frac{P_E - P_I}{2t}$ for $i \in \mathcal{D}$. The weighted indifference condition can be rewritten in a simple way as follows.

$$X_I(P_I, \Phi_I | P_E, \Phi_E, \gamma) = \theta^c = \frac{1}{2} + \frac{P_E - P_I + (1-\tau^c)\Delta}{2t}, \tag{6}$$

where $\Delta = \left(\frac{1}{\Phi_E} - \frac{1}{\gamma\Phi_I}\right)$. The market share for the entrant is $X_E = 1 - \theta^c$. Given X_I and X_E, the solutions to the profit maximization problem with respect to P_j are given by

$$P_I = t + \frac{(1-\tau^c)\Delta}{3}; \quad P_E = t - \frac{(1-\tau^c)\Delta}{3}; \quad X_I = \frac{1}{2} + \frac{(1-\tau^c)\Delta}{6t}; \quad X_E = \frac{1}{2} - \frac{(1-\tau^c)\Delta}{6t}. \tag{7}$$

To guarantee an interior solution, I assume throughout the paper that $\frac{(1-\tau^c)\Delta}{3} < t$: products provided by two sellers are sufficiently differentiated to have positive demand.

Given the equilibrium price and quantity in (7), each seller decides whether to purchase data from the platform. The gap between the mismatch costs from two sellers, denoted by Δ, differs depending on each seller's choices regarding Φ_j: either $1 + \tau^c$ (if purchasing) or one (otherwise), which can be derived as follows.

$$\Delta = \begin{cases} \frac{1}{1+\tau^c}\left(1 - \frac{1}{\gamma}\right) \equiv \Delta_{BB} & \text{if both sellers buy data, (}\boldsymbol{B}\text{uy, }\boldsymbol{B}\text{uy)} \\ 1 - \frac{1}{\gamma(1+\tau^c)} \equiv \Delta_{BN} & \text{if only seller } I \text{ buys data, (}\boldsymbol{B}\text{uy, }\boldsymbol{N}\text{ot buy)} \\ \frac{1}{1+\tau^c} - \frac{1}{\gamma} \equiv \Delta_{NB} & \text{if only } E \text{ buys data, (}\boldsymbol{N}\text{ot buy, }\boldsymbol{B}\text{uy)} \\ 1 - \frac{1}{\gamma} \equiv \Delta_{NN} & \text{if both do not buy data, (}\boldsymbol{N}\text{ot buy, }\boldsymbol{N}\text{ot buy)} \end{cases} \tag{8}$$

One can rank different Δ as $\Delta_{NB} < \Delta_{BB} < \Delta_{NN} < \Delta_{BN}$,[15] where the first and second subscripts denote I's and E's decisions, respectively. Using (7) and (8), each seller's 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. The thresholds are

$$\begin{cases} \text{Given } E \text{ buys, } I \text{ buys if} \quad C_I \leq \dfrac{(1-\tau^c)\left\{(\tau^c)^2 + \gamma[6t(1+\tau^c) - 2\tau^c + 2] + \tau^c - 2\right\}}{18\gamma^2 t(1+\tau^c)^2} \equiv \bar{C}_I. \\ \text{Given } E \text{ does not buy, } I \text{ buys if} \quad C_I \leq \dfrac{(1-\tau^c)\left[(\tau^c)^2 + 2\gamma(1+\tau^c)(3t+1-\tau^c) + \tau^c - 2\right]}{18\gamma^2 t(1+\tau^c)^2} \equiv \bar{\bar{C}}_I. \\ \text{Given } I \text{ buys, } E \text{ buys if} \quad C_E \leq \dfrac{(1-\tau^c)\left\{\gamma\left[(\tau^c)^2 + 6t(1+\tau^c) + \tau^c - 2\right] + 2(1-\tau^c)\right\}}{18\gamma t(1+\tau^c)^2} \equiv \bar{C}_E. \\ \text{Given } I \text{ does not buy, } E \text{ buys if} \quad C_E \leq \dfrac{(1-\tau^c)\left\{-2(\tau^c)^2 + \gamma\left[(\tau^c)^2 + 6t(1+\tau^c) + \tau^c - 2\right] + 2\right\}}{18\gamma t(1+\tau^c)^2} \equiv \bar{\bar{C}}_E. \end{cases}$$

[15] $\Delta_{BB} = \Delta_{NN}$ and $\Delta_{NB} < 0$ if $\gamma = 1$, i.e., the two sellers' initial targeting technology is symmetric.

Unambiguously, $\bar{\bar{C}}_I > \bar{C}_I$ 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 $\frac{\partial^2 \pi_j}{\partial \Phi_I \Phi_E} = \frac{-2(1-\tau^c)^2}{9t\gamma\Phi_I^2\Phi_E^2} < 0$.[16] Nevertheless, the platform always wants to sell data to both sellers at $\bar{C}_I$ for the incumbent and $\bar{C}_E$ for the entrant to maximize its profit. That is, the profit from selling to both at different prices is greater than that from selling exclusively to one seller, either to I at $\bar{\bar{C}}_I$ or to E at $\bar{\bar{C}}_E$. Thus, the platform's profit is $\pi_p^{BB} = (\bar{C}_I + \bar{C}_E)\tau^c$, where the superscript BB denotes that both sellers buy data from the platform. The result is summarized in Proposition 2.

Proposition 2. *When there is no vertical affiliation between the platform and one seller, the platform has an incentive to charge a different data price to each seller. The data price discrimination entails that both sellers buy data.*

Proposition 2 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, any seller, regardless of its affiliation with Amazon, can obtain some of Amazon's data through a program called *Amazon Retail Analytics (ARA)*. *ARA* has two quality tiers with different pricing: *ARA Basic*, which is available for free, and *ARA Premium*, which charges an access fee. Similarly, Facebook also sets different prices for data depending on a data buyer's willingness to pay.[17]

4 Vertical Integration Case

In this section, I analyze the effect of vertical integration between the platform and one of the sellers. Regarding the timing, after each consumer decides whether to disclose information, the platform first makes a vertical integration deal with one of the sellers.[18] After the vertical integration deal is made, 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 then the consumers decide.

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

[17] Refer to https://vendorcentral.amazon.com/ and https://www.facebook.com/business/ads/pricing.

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

As I will show in this section, the data acquisition equilibrium outcome without vertical integration per Proposition 2 no longer prevails when the platform is vertically integrated with a downstream seller.

4.1 Equilibrium

By backward induction, each seller's price and market share are the same as before. Given this setting, I examine the result when the platform makes a deal with one of the sellers. First, suppose that the platform merges with seller I, which has better targeting technology. Because seller I always uses data, seller E buys data if $C \leq \bar{C}_E$ but does not if $C > \bar{C}_E$. The profit for the integrated firm can be written as follows.

$$\pi_{VI} = \begin{cases} P_I(\Delta_{BB})X_I(\Delta_{BB}) + \bar{C}_E\tau^c \equiv \pi_{VI,S} & \text{if the integrated firm sells data to } E \\ P_I(\Delta_{BN})X_I(\Delta_{BN}) \equiv \pi_{VI,F} & \text{otherwise (foreclose),} \end{cases} \tag{9}$$

where the first two letters in the subscript, *VI*, denote *Vertical integration with* I and the last letter indicates data foreclosure (F) or sales (S) status. There is a tradeoff between selling and foreclosing data. If the integrated firm forecloses the unaffiliated seller from data access, 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 $\pi_{VI,S}$ to $\pi_{VI,F}$, the integrated firm decides whether to sell data to the unaffiliated firm. The profit comparison is given as follows.

$$\pi_{VI,S} - \pi_{VI,F} = -\frac{(1-\tau^c)^2\tau^c[\gamma(2+\tau^c)-2]}{9\gamma t(1+\tau^c)^2} < 0. \tag{10}$$

Equation (10) shows that if the platform is integrated with the incumbent, it always wants to foreclose the unaffiliated entrant from accessing data 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.

$$\pi_{VE} = \begin{cases} P_E(\Delta_{BB})X_E(\Delta_{BB}) + \bar{C}_I\tau^c \equiv \pi_{VE,S} & \text{if the integrated firm sells data to } I \\ P_E(\Delta_{NB})X_E(\Delta_{NB}) \equiv \pi_{VE,F} & \text{otherwise (foreclose).} \end{cases} \tag{11}$$

$$\pi_{VE,S} - \pi_{VE,F} = \frac{(1-\tau^c)^2\tau^c(2\gamma-2-\tau^c)}{9\gamma^2 t(1+\tau^c)^2}. \tag{12}$$

From Equation (12), I find that if $\tau^c > 2(\gamma - 1) \equiv \underline{\tau}$, $\pi_{VE,S} < \pi_{VE,F}$. In contrast to the former case, if the platform is integrated with seller E, it has an incentive to sell data to the unaffiliated incumbent if the available amount of data is sufficiently small. In other words, it is more likely to foreclose data access as τ^c increases because if τ^c is sufficiently large, seller E is able to overcome his targeting disadvantage by exclusively using data for targeting and can more easily dominate the market. Therefore, the integrated firm sells data only if τ^c is small; the data-selling revenue effect is greater than the market dominance effect of data foreclosure. Note that Equation (12) decreases in γ, which means that if γ decreases, it is more profitable to foreclose data access because the combination of data monopolization and a lower γ provides a greater advantage to the integrated firm.

To determine which seller offers sufficient incentive to induce the platform to integrate, I compare the profits from integration with I to those from integration with E. There are two possible cases: $I - Foreclose$ or $E - Sell$ for $\tau^c < \underline{\tau}$ and $I - Foreclose$ or $E - Foreclose$ for $\tau^c > \underline{\tau}$, where $I - Foreclose$ means that the platform integrates with I and forecloses E from data access. Similarly, $E - Sell$ means that the platform and E integrate and sell data to I. From the comparison, the profit from integrating with I is always greater than that from integrating with E regardless of whether the platform affiliated with E sells or forecloses data: I's better targeting skills combined with the exclusive use of data allow the platform integrated with I to sell more affiliated products in the downstream market, incentivizing the platform to integrate with I. As a result, integrating with I and the resulting data foreclosure always emerge in the vertical integration game.

Furthermore, it can be verified that the platform and seller I always have an incentive to vertically integrate with one another by comparing the joint profits of the platform and seller I under no vertical integration to the profits of the integrated firm. Proposition 3 summarizes these results.[19]

Proposition 3. *If vertical integration is permitted, the platform always has an incentive to vertically integrate with the incumbent. The unaffiliated seller forgoes buying data.*

Compared to the game without vertical integration in which both sellers can obtain consumer data from the platform, allowing vertical integration keeps the unaffiliated entrant from

[19] 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.

obtaining consumer data from the platform. Absent vertical integration, the entrant is able to use the platform's data at the unit price of $\bar{C}_E$. When the platform is integrated with I, because of its incentive 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 the unaffiliated E is thus foreclosed from data access.

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. Proposition 3 implies that vertical integration raises the data price for unaffiliated sellers, which provides a considerable competitive advantage to an affiliated seller. Such preferential treatment for the affiliated seller can be particularly anti-competitive if an entrant that has a more desperate need for consumer data due to its worse initial targeting skills is foreclosed from data access due to vertical integration.

4.2 Implications

Absent vertical integration, seller E always buys data to overcome its initial disadvantage in targeting skills. Because vertical integration always leads to data foreclosure, it is more likely to adversely affect the entrant E, which always needs data access. To determine how this affects seller E, its market shares with and without vertical integration are compared. Given that the platform and I are integrated, the unaffiliated seller E always suffers from smaller market share due to the integration and resulting data foreclosure. Proposition 4 summarizes this implication.

Proposition 4. *In the game with vertical integration, in which the platform is integrated with the incumbent, the entrant suffers from a smaller market share due to data foreclosure.*

This result raises a very important antitrust implication regarding data-driven vertical integration. Because vertical integration with the incumbent seller always forecloses data access, only the entrant seller is harmed. As consumer information is vital for a small seller or a market entrant with weaker targeting skills, vertical integration with a seller possessing better target-

ing skills is likely to have an anti-competitive effect because it prevents the unaffiliated entrant from using data to overcome its initial disadvantage. Note that one might question whether such disproportionate harm toward the entrant would be substantial if the data intermediary market were competitive. As I will discuss in Section 7.4, the model with competing platforms is unlikely to undermine the main implications as long as the data are imperfect substitutes.

5 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.

$$SW = CS + \mathbb{1}_{\{NV\}}(\pi_p + \pi_I + \pi_E) + \mathbb{1}_{\{V\}}(\pi_K^I + \pi_K^E), \tag{13}$$

where the subscript $K \in \{NV, VI, VE\}$; VI (VE) denote the surplus from vertical integration with seller I (E), respectively. 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. The profits for firms under the no vertical integration and vertical integration models are all given in the paper. Consumer surplus can be obtained in the following way.

$$\begin{aligned} CS = &\int_0^{\tau^c} \left(v(\tau^c) - \frac{\psi(x)}{r} \right) dF(x) + \tau^c \left\{ \int_0^{\theta_{\mathcal{D}}^c} \Big(V - t\theta - P_I \Big) d\theta + \int_{\theta_{\mathcal{D}}^c}^1 \Big[V - (1-t)\theta - P_E \Big] d\theta \right\} \\ &+ (1-\tau^c) \left\{ \int_0^{\theta_{\mathcal{ND}}^c} \Big(V - t\theta - P_I - \frac{1}{\gamma \Phi_I} \Big) d\theta + \int_{\theta_{\mathcal{ND}}^c}^1 \Big[V - (1-t)\theta - P_E - \frac{1}{\Phi_E} \Big] d\theta \right\}. \end{aligned} \tag{14}$$

First, the comparative statics results show that consumer surplus levels increase in τ^c. There are two channels whereby more data benefit consumers: a lower mismatch cost and more intense market competition. If τ^c increases, a consumer faces a lower overall mismatch cost. In addition, having more data implies that E is more likely to overcome its disadvantage in targeting, thereby fostering market competition. Proposition 5 summarizes this finding.

Proposition 5. *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 interesting analysis is to compare welfare levels with and without vertical integration to see whether the market equilibrium outcome, leading to the integration with the

incumbent, is socially desirable or not. It can be shown that the consumer surplus under integration, which leads to data monopolization by the affiliated incumbent, (B, N), is not always greater than that under no integration, which results in (B, B). If the amount of available data is sufficiently small, no vertical integration enhances consumer surplus: when there are greater privacy concerns, which leads to a smaller amount of available data, vertical integration and the resulting data foreclosure harm consumers. However, if more data are available, data foreclosure from vertical integration benefits consumers. This suggests that vertical integration foreclosing the entrant from accessing data is particularly harmful if consumers have greater privacy concerns so that less data are available for targeting. The intuition behind this finding is as follows. When both sellers symmetrically obtain data, as in the case without vertical integration, their overall targeting becomes similarly effective, which means less differentiation in terms of targeting. In other words, the symmetric (B, B) data acquisition case leads to more intense price competition, which enhances consumer surplus. On the other hand, asymmetric data acquisition (B, N) is likely to harm consumers through the lack of competition in the downstream market because the market is dominated by the incumbent that exploits the exclusive use of data. However, (B, N) might enhance consumer welfare if the incumbent's exclusive use of data combined with its better targeting skills significantly reduces the consumers' mismatch costs. Compared to the case of no integration with (B, B), vertical integration followed by asymmetric data acquisition (B, N) leads to a greater market share for the incumbent due to a larger gap between two sellers' targeting effectiveness. If the aggregate mismatch cost reduction from a large portion of the incumbent's customers is sufficiently large, (B, N) can be welfare-enhancing. Thus, whether symmetric or asymmetric data acquisition makes consumers better off depends on the relative size between the competition-enhancing effects and mismatch cost-lowering effects. If the amount of available data is small, the effect of the incumbent's data monopolization on lowering the mismatch cost is negligible. Thus, if the amount of available data is less than a certain threshold, (B, B), which prevails under the case without vertical integration, attains greater consumer surplus than (B, N), which prevails under vertical integration. However, if a sufficient amount of data becomes available, the effect on lowering the mismatch cost outweighs the effect of fierce price competition in the downstream market, which means that vertical integration with the incumbent and the resulting data foreclosure are more desirable. The social welfare comparison shows qualitatively similar results. This finding is

summarized in Proposition 6.

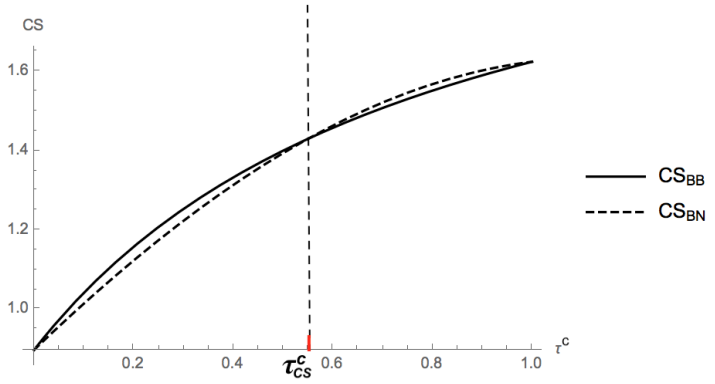

Figure 2: The equilibrium consumer surplus comparison under $V = 2$, $\gamma = 2$, and $t = 0.3$

Proposition 6. *Data-driven vertical integration with the incumbent makes the consumer and society as a whole worse off than the absence of vertical integration if there are substantial privacy concerns.*

Figure 2 offers a graphical comparison of different consumer surplus levels. As in Figure 2, there exists a threshold on τ^c, denoted as τ^c_{CS}, and if the amount of data available is smaller than τ^c_{CS}, the market-driven equilibrium that leads to vertical integration with the incumbent followed by data foreclosure reduces the consumer surplus. The threshold τ^c_{CS} is affected by external factors in the market: t representing the degree of product differentiation and γ representing the gap between initial targeting skills of the two sellers. The comparative statics results indicate that as t increases or γ decreases, τ^c_{CS} increases, which means that the market-led vertical integration with the incumbent and the resulting data foreclosure are more likely to harm consumers. If the products offered by the two sellers are more differentiated, price competition is further softened, which exacerbates the adverse effects of asymmetric data acquisition. This is why consumer surplus under no vertical integration and symmetric data usage is more likely greater than in the other case as t increases. However, as the incumbent's initial targeting skills become more effective, the benefits from better-targeted ads increase. The positive effects of a larger γ are greater under asymmetric than under symmetric data usage because more consumers choose the incumbent with better targeting skills to save more mismatch costs if the incumbent exclusively uses the data than if both sellers symmetrically use the data. Because the mismatch cost-saving effects increase as γ increases, data-driven vertical integration

is more likely to benefit consumers as the incumbent's targeting skills become more effective. Proposition 7 summarizes these findings.

Proposition 7. *Vertical integration with the incumbent and the resulting data foreclosure are more likely to harm consumers as the products offered by two sellers become more differentiated or as the incumbent's initial targeting skills deteriorate.*

The welfare analysis result implies that whether the vertical integration equilibrium outcome is welfare-enhancing or -reducing depends on the amount of information available on the platform. Specifically, it is harmful if consumers' privacy concerns, which determine the aggregate amount of information available, are very high. Thus, although each consumer makes an individually rational information disclosure decision, it may not be socially optimal once one accounts for the effects of vertical integration and data foreclosure on the competitive structure of the downstream market.

6 Policy Implications

According to the theoretical findings in this paper, the data foreclosure that follows vertical integration between the platform and incumbent will raise barriers to entry to the market, which can ultimately be welfare reducing. To prevent this undesirable consequence, there are two different types of policy remedies—imposing a related condition on data sharing aspects when approving a proposed merger and inducing more data availability by reducing consumers' privacy concerns. Several specific remedies of these types are proposed below.

6.1 Regulation on Mergers and Data Sharing Practice

Policymakers can implement a direct policy that regulates integrated firms' data-sharing practices. Again, if the platform and incumbent integrate and foreclose the unaffiliated entrant from data access, it will lead to a welfare-reducing outcome in some cases. Thus, when those two entities try to integrate with each other, a regulatory authority may want 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 makes consumers worse off, as in Proposition 6, the authority may need to reject a proposed data-driven merger.

Alternatively, if a regulatory authority approves a proposed merger, it needs to consider not only some 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 with an incumbent, 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.[20] To do so, the authority first needs to create a proper non-price metric to measure how integration leads to data-related concentration rather than relying on traditional metrics, such as typical turnover rates. By imposing such conditions on the final merger approval, even unaffiliated sellers, mostly entrants, can rest assured that the merger will not keep them from obtaining some 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, which might make consumers worse off.

6.2 Tax Subsidies

Assuming that vertical integration is permitted, the market outcome with vertical integration between the platform and the incumbent and the resulting data foreclosure are more welfare-enhancing if the amount of available data is sufficiently large ($\tau^c > \tau^c_{CS}$). Regulators can increase consumers' willingness to disclose their personal information by adopting any policy that encourages the platform, i.e., the data intermediary, or the integrated firm to invest more in keeping their data secure and thereby preventing an actual data breach.

It would be effective if the government were to provide tax subsidies to those platforms that invest a certain amount of money to obtain a better data security system. These tax subsidies would encourage more investments to protect consumers' personal information from being leaked, causing consumers to reduce concerns when disclosing their data to such platforms. In other words, consumers will disclose more of their data to trustworthy data collectors, which makes the market-driven equilibrium outcome with vertical integration and data foreclosure between the platform and incumbent more likely to be socially desirable.

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

6.3 Reputation-Enhancing Program

Similar to the suggestion above, any policy that encourages consumers' voluntary information disclosure is socially desirable. If the platform clarifies how user data are used and the potential benefit of disclosing information, more consumers will be able to understand the benefit and make a better decision: a more transparent and easy-to-read data usage policy will allow more consumers to discern any potential benefit, which will increase the immediate benefit $v(\tau^c)$ in the utility of the platform.

In addition, if the data collector can decrease users' privacy nuisance costs, it would also help to increase the amount of information available. From the utility specification in Equation (1), which assumes that consumers' concern for privacy is asymmetric with respect to firm reputation, one way to reduce the nuisance cost is to increase the data collector's reputation. In reality, a consumer has an asymmetric privacy concern with respect to the data collector's reputation: consumers are more likely to agree to app developers' data usage policy if the developers are relatively well-known rather than unknown, as shown in Kummer and Schulte (2019). Therefore, if the platform's reputation as a data collector plays a significant role in increasing information disclosure, any policy that helps the platform build its reputation would be welfare-enhancing.

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 only self-certification at best. A credible certification program could serve as a global standard that helps participating firms to increase their reputation regarding data usage. Because willingness to disclose information depends on a firm's reputation, this remedy is likely to be effective. This policy suggestion is consistent with some of the policy implications discussed in the literature (e.g., Campbell et al. (2015)). Moreover, The Cyber Shield Act of 2017 works along the same vein.[21]

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

7 Discussion and Extensions

7.1 Consumers with Foresight

Thus far, I have assumed that a consumer takes into account only 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 what happens 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 mismatch cost when making information disclosure decisions.[22] The aggregate utility specification is as follows.

$$u_{ij}^{\text{Foresight}} = V - t|l_j - \theta_i| - P_j - \mathbb{1}_{\{i \in \mathcal{ND}\}} \left(\frac{1}{\gamma_j \Phi_j} \right) - \mathbb{1}_{\{i \in \mathcal{D}\}} \left(\frac{\psi(\tau_i)}{r} \right). \tag{15}$$

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.

$$\begin{aligned} \mathbb{P}(i \in \mathcal{D}) = \tau^c &= \theta^c \mathbb{P}(i \in \mathcal{D} | i \in \mathcal{I}) + (1 - \theta^c) \mathbb{P}(i \in \mathcal{D} | i \in \mathcal{E}) \\ &= \theta^c F\left(\psi^{-1}\left(\frac{r}{\gamma \Phi_I} \right) \right) + (1 - \theta^c) F\left(\psi^{-1}\left(\frac{r}{\Phi_E} \right) \right), \end{aligned} \tag{16}$$

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 (16), I can rank different τ^c levels depending on each data acquisition equilibrium. Given that Φ_j can be either one or $1 + \tau^c$, 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

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

consumer becomes reluctant to disclose information. The relative size of τ^c_{NB} and τ^c_{BN} depends on γ and t. Put simply, $\tau^c_{NB} > \tau^c_{BN}$ is more likely to hold as γ increases given t or as t decreases given γ. That is, anticipating that E is the only data holder, a consumer becomes more willing to disclose personal information as I's targeting technology improves or the degree of product differentiation declines. Because $X_E = 1 - \theta^c$ decreases as t decreases or γ increases, the effect of E's exclusive use of data on raising P_E becomes smaller, which leads to $\tau^c_{NB} > \tau^c_{BN}$. That is, if the total demand for the seller is small, the effect of data acquisition on τ^c is negligible. Proposition 8 summarizes the findings.

Proposition 8. *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 the relative size of γ and t.*

Next, comparative statics results show that how the implicitly determined equilibrium τ^c is affected by exogenous parameters, such as γ and t. By applying the implicit function theorem to Equation (16), it is easy to see that $\frac{d\tau^c}{d\gamma} < 0$ while $\frac{d\tau^c}{dt} > 0$ for any τ^c from each data acquisition equilibrium. In other words, more consumers are willing to disclose personal data as the incumbent's initial targeting technology becomes less effective or the products provided by the two sellers become more differentiated. The intuition is that as γ becomes close to one, both sellers provide less relevant targeted ads because they lack targeting skills. Knowing that, a consumer becomes more willing to provide personal information to allow both sellers to send better-targeted ads, thereby lowering the mismatch cost. Furthermore, if t increases, the products provided by the two sellers become more differentiated, which softens price competition. In this case, one way for a consumer to save costs is to provide more personal data and incur a lower mismatch cost. Proposition 9 summarizes the finding.

Proposition 9. *Sophisticated consumers are more willing to disclose personal data as the incumbent's initial targeting technology becomes less effective or the products provided by the two sellers become more differentiated.*

7.2 Endogenous Entry

In this section, I consider a variation of the model by introducing a fixed cost of entry for the entrant. Thus, the entrant is allowed to stay out of the market if the expected profit is lower

than the entry cost. The timing of the game is modified accordingly in that the entrant decides whether to enter the market in the very first stage. Denoting the fixed entry cost as FC, the relevant thresholds of FC below which the entrant enters the market can be obtained as follows: $\overline{FC}_{BB} = \pi_E^{BB}$; $\overline{FC}_{NB} = \pi_E^{NB}$; $\overline{FC}_{BN} = \pi_E^{BN}$. Note that π_E^{BB} is equal to π_E^{BN} after paying the data price $\bar{C}_E$. If the fixed cost is higher than the equilibrium profit, the entrant stays out of the market. Given that $\overline{FC}_{BN} = \overline{FC}_{BB} < \overline{FC}_{NB}$, there are three possible cases for the entrant: (1) staying out in any case if $FC > \overline{FC}_{NB}$, (2) staying out under (B, B) or (B, N) but entering under (N, B) if $\overline{FC}_{BN} = \overline{FC}_{BB} < FC < \overline{FC}_{NB}$, and (3) entering in any case if $FC < \overline{FC}_{BN} = \overline{FC}_{BB}$. Focusing on this parametric space, I analyze how the entrant's entry decision affects the market.

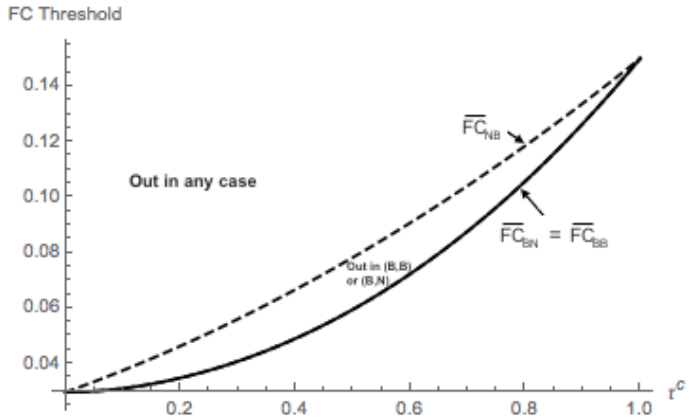

Figure 3: Entrant's entry decision depending on the fixed cost thresholds

As in Figure 3, (B, N) (or (N, B)) is the most (or least) likely to lead to entry foreclosure. That is, vertical integration with the incumbent and the resulting data foreclosure, meaning (B, N), are highly likely to lead the entrant to stay out of the market due to lower profit when entering. Obviously, when the entrant is able to use data exclusively, meaning (N, B), the entrant is the most likely to enter the market. Proposition 10 summarizes this finding.

Proposition 10. *When the entrant faces a fixed cost of entry, it is least likely to enter the market if vertical integration with the incumbent followed by data foreclosure emerges in equilibrium.*

To see how entry foreclosure affects the market outcome, I first derive the monopoly market equilibrium. Given that the entrant stays out, the incumbent can monopolize the market. Suppose that the utility specification is the same as in Equation (2). Assuming that $l_I = 0$,

the utility under monopoly implies that if consumer i's personal taste is suited to the product offered by the incumbent (i.e., θ_i is sufficiently close to I's location), he buys from the monopolist incumbent. Thus, the monopoly market share is determined by $Prob\left(\theta < \frac{V-P_I-\mathbb{1}_{\{i\in\mathcal{ND}\}}\frac{1}{\gamma\Phi_I}}{t}\right)$, which means that $X_{\mathcal{ND}}^{Mono} = \frac{V-P_I-\frac{1}{\gamma\Phi_I}}{t}$ and $X_{\mathcal{D}}^{Mono} = \frac{V-P_I}{t}$, where the superscript $Mono$ denotes the case of a monopoly. Given the weighted monopoly market share, the incumbent maximizes its profit by charging the monopoly price at $P_I^{Mono} = \frac{1}{2}\left(V - \frac{1-\tau^c}{\gamma\Phi_I}\right)$, which leads to $X_I^{Mono} = \frac{V\gamma\Phi_I-1+\tau^c}{2t\gamma\Phi_I}$. Specifically, the equilibrium profit levels under buying and not buying data from the platform can be derived as follows.

$$\pi_B^{Mono} = \frac{[V\gamma(1+\tau^c)-1+\tau^c]^2}{4t\gamma^2(1+\tau^c)^2}; \quad \pi_N^{Mono} = \frac{(V\gamma-1+\tau^c)^2}{4t\gamma^2}, \tag{17}$$

where the subscripts B and N denote Buy and Not buy, respectively. Comparing the two profit levels, the platform optimally sets the data price at $C^{Mono} = \frac{(1-\tau^c)\{2V\gamma[1+\tau^c-2+\tau^c+(\tau^c)^2]\}}{4t\gamma^2(1+\tau^c)^2}$, which is derived from $\pi_B^{Mono} - \pi_N^{Mono}$. Under monopoly, a consumer obviously becomes worse off than under duopoly as in the main model. Proposition 11 summarizes these findings.

Proposition 11. *If the entrant does not enter the market due to lower profits, the incumbent monopolizes the product market. The monopoly makes the consumer worse off.*

In other words, if the entrant needs to pay the fixed cost of entry, vertical integration with the incumbent, which is most likely to lead to monopoly in the product market, is welfare-reducing. Thus, under a dynamic model that takes the entrant's entry decision into consideration, consumers are always worse off from vertical integration with the incumbent and the resulting data foreclosure. In this case, stricter policy remedies, such as rejecting the merger between the platform and incumbent, may be needed.

7.3 The Role of the Platform Beyond 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 the downstream market competition by internalizing some of benefits from the sellers. For example, transaction-based e-commerce platforms usually take a certain portion 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 around 13% of total sales revenue. To investigate such cases in which the platform is more than just 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.

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

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 be also changed to $\pi_j = (1-f)P_j X_j - \mathbb{1}_{\{\text{Buy}\}} C_j^{Int} \times \mathbb{P}(i \in \mathcal{D})$. 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 profit when using data is greater than that when being foreclosed, the integrated firm now has a greater incentive to sell data. However, this incentive is offset by a lower data price charged to the unaffiliated seller due to the platform's additional rent extraction: $C_j^{Int} = (1-f)C_j$. Thus, Proposition 3 still holds in this extension.

However, the equilibrium outcome in the game without vertical integration is different from that of the main model. When the platform is able to extract additional rents from sellers through the usage fees, its incentive to sell data to both sellers or exclusively to one seller depends on the amount of data available on the platform. From the profit comparison, I find that there exists a threshold on τ^c, denoted as τ_{NV}^c, below which the platform sells data to the incumbent only; otherwise, it sells to both. That is, additional profits for the platform transferred from sellers incentivize the platform to engage in exclusive data sales for the incumbent even without vertical integration. Intuitively, if the available data are limited due to greater privacy concerns, the entrant is unable to overcome its targeting disadvantage, which means that the platform cannot extract higher revenue from the entrant. In this case, if the platform sells data to both sellers, the revenue extracted from the incumbent is also not sufficiently large because such symmetric data acquisition leads to more intense price competition. If the

platform instead sells data exclusively to the incumbent, it can extract higher revenue because the incumbent dominates the market. In other words, if the amount of data is so small that the effect of data on attracting consumers for the entrant is negligible, the platform maximizes its profit by allowing the incumbent to exclusively use the data, even without vertical integration. Proposition 12 summarizes these findings.

Proposition 12. *When the platform is allowed to extract the sellers' product sales revenues in the form of usage fees, it has an incentive to sell data exclusively to the incumbent, even without vertical integration, if the amount of available data is sufficiently small ($\tau^c < \tau^c_{NV}$). If the amount of data is sufficiently large ($\tau^c > \tau^c_{NV}$), it sells data to both sellers.*

Proposition 12 implies that consumers' higher privacy concerns can disproportionately harm the entrant, thereby harming consumers themselves through the lack of competition, if the platform charges usage fees to the sellers. In this sense, allowing the platform to extract revenues from the sellers and allowing vertical integration can be complementary in harming the entrant and ultimately reducing consumer welfare. Therefore, again, individually optimal decisions on data disclosure might not be socially optimal when aggregated.

7.4 Competing Platforms

Whether data-driven mergers and the resulting data foreclosure are anti-competitive or not is still debated. For instance, the Department of Justice concluded that the proposed merger between *Bazaarvoice* and *Power-Reviews*, followed by the possible exclusive use of data is likely to raise the entry barriers,[23] whereas the European Commission found that any internal use of data from the merger between *Microsoft* and *LinkedIn* could not give an anti-competitive advantage to the merged entity because alternative data sources are still available in the market.[24]

In this section, I discuss one of the main implications, the disproportionate harm toward the entrant due to the vertical integration followed by data foreclosure, which still holds even if the model allows platforms to compete in data provision. First, although there are many platforms on the Internet, most of them do not directly compete with each other, given their

[23] Refer to *USA v. Bazaarvoice Inc.m* Case No. 13-cv-00133 WHO.
Available at https://www.justice.gov/atr/case-document/file/488826/download.

[24] Refer to *Case M.8124 – Microsoft / LinkedIn.*
Available at https://ec.europa.eu/competition/mergers/cases/decisions/m8124_1349_5.pdf.

specialized functions, such as search engines or social media. When we narrow the platform market down by specialized businesses, each submarket is locally monopolized in many cases: Google is dominant in the search engine market, whereas Instagram/Facebook is dominant in the social media market.

Related to the first point, platforms in different submarkets collect different sets of data, as shown in a few previous studies (e.g., Maier (2019), Graef (2015), Stucke and Grunes (2016)). Data available on Google Maps are related mostly to consumers' location information, while those on Instagram/Facebook show consumers' social activities. In this sense, a considerable amount of data collected by different platforms is complementary, which means that data foreclosure implemented by one platform is likely to harm the unaffiliated sellers. That is, even if there are many platforms seemingly competing with each other, the integration between one platform and the incumbent and the resulting data foreclosure still lead to the competitive harms as long as the data provided by different platforms are not perfectly substitutable.

8 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 an incentive to vertically integrate. The integrated firm always wants to prevent the unaffiliated entrant from accessing the data, thereby adversely affecting the entrant in the form of a smaller market share. Therefore, the entrant that needs consumer data to overcome its initial disadvantage in targeting technology is disproportionately affected by such integration and the resulting data foreclosure. Moreover, this process can eventually lead to lower consumer surplus and lower total social welfare due to the lack of competition arising from data foreclosure. In particular, consumers become worse off from vertical integration and data foreclosure if their privacy concerns are relatively high so that only a limited amount of data is available for targeting: whether vertical integration and data foreclosure are socially desirable depends on the number of privacy-sensitive consumers. Thus, consumers' privacy concerns that result in less data being available may have unintended adverse consequences on market competition through the effects of vertical integration.

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

References

[1] Acquisti, A., Varian, H.R., "Conditioning Prices on Purchase History", *Marketing Science*, Vol. 24, No. 3, (2005), pp 367-381.

[2] Bajari, P., Chernozhukov, V., Hortaçsu, A., Suzuki, J. "The Impact of Big Data on Firm Performance: An Empirical Investigation", *AEA Papers and Proceedings*, Vol.109, (2019), pp 33-37.

[3] Belleflamme, P., Vergote, W. "Monopoly Price Discrimination and Privacy: The Hidden Cost of Hiding", *Economics Letters*, Vol.149, (2016), pp 141-144.

[4] Biglaiser, G., Calvano, E., Crémer, J., "Incumbency Advantage and Its Value", *Journal of Economics & Management Strategy*, Vol.28, (2019), pp 41-48.

[5] Binns, R., Bietti, E., "Dissolving Privacy, One Merger at a time: Competition, Data and Third Party Tracking", *Computer Law & Security Review*, (2019).

[6] Braulin, F.C., Valletti, T., "Selling Customer Information to Competing Firms", *Economics Letters*, Vol.149, (2016), pp 10-14.

[7] Calzolari, G., Pavan, A., "On the optimality of privacy in sequential contracting", *Journal of Economic Theory*, Vol.130, No.1, (2006), pp 168-204.

[8] Campbell, J., Goldfarb, A., Tucker, C., "Privacy Regulation and Market Structure", *Journal of Economics and Management Strategy*, Vol.24, No.1, (2015), pp 47-73.

[9] Casadesus-Masanell, R., Hervas-Drane, A., "Competing with Privacy", *Management Science*, Vol.61, No.1, (2015), pp 229-246.

[10] Chirita, A.D.,, "Data-Driven Mergers Under EU Competition Law", *The Future of Commercial Law: Ways Forward for Harmonisation*, (2019).

[11] Choi, J.P., Jeon, D.S., Kim, B.C., "Privacy and Personal Data Collection with Information Externalities", *Journal of Public Economics*, Vol.173, (2019), pp 113-124.

[12] Conitzer, V., Taylor, C.R., Wagman, L., "Hide and Seek: Costly Consumer Privacy in a Market with Repeat Purchases", *Marketing Science*, (2012), pp 277-292.

[13] D'Annunzio, A., Russo, A., "Ad Networks, Consumer Tracking, and Privacy", *CESifo Working Paper Series*, No. 6667, (2017).

[14] De Cornière, A., Nijs, R.d., "Online Advertising and Privacy", *RAND Journal of Economics*, Vol.47, No.1, (2016), pp 48-72.

[15] De Cornière, A., Taylor, G., "Data and Competition: a General Framework with Applications to Mergers, Market Structure, and Privacy Policy", *Working Paper*, (2020).

[16] Fudenberg, D., Tirole, J., "Customer Poaching and Brand Switching", *RAND Journal of Economics*, Vol. 31(4), (2000), pp 634-657.

[17] Fudenberg, D., Tirole, J., "Perfect Bayesian Equilibrium and Sequential Equilibrium.", *Journal of Economic Theory*, Vol.53, Issue.2, (1991), pp 236-260.

[18] Goldfarb, A., "What is Different About Online Advertising", *Review of Industrial Organization*, Vol.44, Issue.2, (2014), pp 115-229.

[19] Graef, I., "'Market Definition and Market Power in Data: The Case of Online Platforms", *World Competition*, Vol.38, No.4, (2015), pp 473-506.

[20] Gu, Y., Madio, L., Reggiani, C., "Exclusive Data, Price Manipulation and Market Leadership", *CESinfo Working Paper*, No. 7853, (2019).

[21] Goldfarb, A., Tucker, C., "Privacy Regulation and Online Advertising", *Management Science*, Vol.57, No.1, (2011), pp 57-71.

[22] Hagiu, A., Wright, J., "Data-Enabled Learning, Network Effects, and Competitive Advantage", *Working Paper*, (2020).

[23] Kim, J.H., Wagman, L., "Screening incentives and privacy protection in financial markets: a theoretical and empirical analysis", *RAND Journal of Economics*, Vol.46, No.1, (2015), pp 1-22.

[24] Kim, J.H., Wagman, L, Wickelgren, A. L., "The Impact of Access to Consumer Data on the Competitive Effects of Horizontal Mergers", *Journal of Economics and Management Strategy*, Vol.28, (2018), pp 373-391.

[25] Koh, B., Raghunathan, S., Nault, B.R., "Is Voluntary Profiling Welfare Enhancing?", *MIS Quarterly*, Vol. 41, Issue 1, (2017), pp 23-41.

[26] Kummer, M.E., Schulte, P., "When Private Information Settles the Bill: Money and Privacy in Google's Market for Smartphone Applications", *Management Science*, (2019).

[27] Levin, J., Milgrom, P., "Online Advertising: Heterogeneity and Conflation Market Design", *American Economic Review: Papers & Proceedings*, 100, (2010), pp 603-607.

[28] Maier, N., "Closeness of Substitution for 'Big Data' in Merger Control", *Journal of European Competition Law & Practice*, Vol.10, Issue 4, (2019), pp 246-252.

[29] Montes, R., Sand-Zantman, W., Valletti, T., "The Value of Personal Information in Markets with Endogenous Privacy", *Management Science*, Vol.65, No.3, (2018).

[30] Nayeem, O.A., Yankelevich, A., "Price-Cap Regulation of Firms That Supply Their Rivals", *Quello Center Working Paper*, (2017).

[31] Petrova, M., Sen, A., and Yildirim, P., "Social Media and Political Donations: New Technology and Incumbency Advantage in the United States", *CEPR Discussion Paper*, No. DP11808. (2017).

[32] Shy, O., Stenbacka, R., "Customer Privacy and Competition", *Journal of Economics and Management Strategy*, Vol.25, Issue.3, (2016), pp 539-562.

[33] Stucke, M., Grunes, A., "Big Data and Competition Policy", *Oxford University Press*, (2016).

[34] Taylor, C., "Consumer Privacy and the Market for Customer Information", *RAND Journal of Economics*, Vol. 35(4), (2004), pp 631-650.

[35] Taylor, C. and L. Wagman, "Customer Privacy in Oligopolistic Markets: Winners, Losers, and Welfare", *International Journal of Industrial Organization*, Vol. 34, (2014), pp 80-84.

[36] Tucker, C.E., "Social Networks, Personalized Advertising, and Privacy Controls", *Journal of Marketing Research*, Vol.51, No.5, (2014), pp 546-562.

[37] Villas-Boas, J., "Dynamic Competition with Customer Recognition", *RAND Journal of Economics*, Vol. 30(4), (1999), pp 604-631.

[38] Villas-Boas, J., "Price Cycles in Markets with Customer Recognition", *RAND Journal of Economics*, Vol. 35(3), (2004), pp 486-501.

Appendix. Further Discussion and Omitted Proofs.

Discussion on initial targeting technology. As mentioned earlier, the assumption of asymmetric initial targeting skill, $\gamma \equiv \gamma_I > \gamma_E \equiv 1$, can be justified by I's existing customer data

and previous experience. Given that I has established stronger data analytic skills using such previous experience, I can outperform E if both are given the same amount of data: the overall targeting effectiveness takes a multiplicative form, $\gamma_j \Phi_j$. Though this assumption is crucial to the model, the main implications are quite robust to different specifications. To check the robustness, here I modify the model in two ways: (1) additive separability of γ_j and Φ_j and (2) a symmetric targeting technology case. First, one might question whether such pre-existing data do not affect the marginal benefit of additional data acquisition but just provide a fixed amount of other data. In the latter case, in Equation (2), the overall targeting effectiveness, which is $\gamma_j \Phi_j$, should have an additive separable form, such as $\gamma_j + \Phi_j$. The main results from the main model still hold under this modification: the platfrom sells data to both sellers absent vertical integration, whereas it integrates with the incumbent and forecloses the entrant from data access when permitted. The detailed proof is omitted.

Second, if the model is changed to the symmetric setup with $\gamma = 1$, it becomes a typical horizontal differentiation model, which is not of interest in this paper.

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

Proof of Proposition 2. The platform's profit from selling data to both sellers at different prices is given by $\pi_p^{BB} = (\bar{C}_I + \bar{C}_E)\tau^c = \frac{(1-\tau^c)\tau^c\left\{(\tau^c)^2+\gamma^2\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+\gamma[6t(1+\tau^c)+4(1-\tau^c)]+\tau^c-2\right\}}{18\gamma^2 t(1+\tau^c)^2}$. Its profit from selling data to only one seller, either I or E, is given by $\pi_p^{BN} = \bar{\bar{C}}_I = \frac{(1-\tau^c)\tau^c\left[(\tau^c)^2+2\gamma(1+\tau^c)(3t-\tau^c+1)+\tau^c-2\right]}{18\gamma^2 t(1+\tau^c)^2}$ or $\pi_p^{NB} = \bar{\bar{C}}_E = \frac{(1-\tau^c)\tau^c\left\{-2(\tau^c)^2+\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+2\right\}}{18\gamma t(1+\tau^c)^2}$, respectively. When comparing the profits from different data acquisition cases, the following inequalities are obtained.

$$\begin{aligned} \pi_p^{BB} - \pi_p^{BN} &= \frac{(1-\tau^c)\tau^c\left\{\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+2(1-\tau^c)^2\right\}}{18\gamma t(1+\tau^c)^2} > 0. \\ \pi_p^{BB} - \pi_p^{NB} &= \frac{(1-\tau^c)\tau^c\left\{(\tau^c)^2+2\gamma\left[3t(1+\tau^c)+(1-\tau^c)^2\right]+\tau^c-2\right\}}{18\gamma^2 t(1+\tau^c)^2} > 0. \end{aligned} \tag{19}$$

The inequalities in (19) are guaranteed by the interior solution condition, which requires t to be sufficiently large. By Equation (19), the platform maximizes its profit by selling data to both sellers at $\bar{C}_I$ to I and $\bar{C}_E$ to E. □

Proof of Proposition 3. Suppose that the platform and I are integrated. Equation (10) shows that the integrated firm's profit from selling data to the rival is always less than that

from monopolizing the data. Suppose now that the platform and E are integrated. Equation (12) shows that the profit difference between selling and foreclosing data is less than zero if τ^c is sufficiently large. Specifically, there exists a threshold on τ^c, denoted as $\underline{\tau} \equiv 2(\gamma - 1)$, that makes Equation (12) zero. It is trivial to show that if $\tau^c > \underline{\tau}$, $\pi_{VE,S} < \pi_{VE,F}$. This suggests that if the platform is integrated with E, both selling and foreclosing data can prevail in equilibrium depending on the size of τ^c.

Knowing this, it is necessary to compare the profit from integrating with I to that from integrating with E to see which seller the platform chooses as an integration partner. The corresponding profit comparisons are given as follows.

$$\begin{aligned}\pi_{VI,F} - \pi_{VE,S} &= \frac{(1-\tau^c)\{(1-\tau^c)\tau^c\{\gamma[\gamma(2+\tau^c)-4]+\tau^c+2\}+6(\gamma-1)\gamma t(1+\tau^c)(2+\tau^c)\}}{18\gamma^2 t(1+\tau^c)^2}. \\ \pi_{VI,F} - \pi_{VE,F} &= \frac{(\gamma-1)\left[2-\tau^c-(\tau^c)^2\right]\left[(\gamma+1)(1-\tau^c)\tau^c+6\gamma t(1+\tau^c)\right]}{18\gamma^2 t(1+\tau^c)^2}.\end{aligned} \tag{20}$$

First, one only needs to show that $\gamma[\gamma(2+\tau^c)-4]+\tau^c+2$ is positive to show that $\pi_{VI,F} > \pi_{VE,S}$. The minimum value of $\gamma[\gamma(2+\tau^c)-4]+\tau^c+2$, obtained at $\tau^c = 0$ and $\gamma = 1$, is zero, which means that $\pi_{VI,F} > \pi_{VE,S}$. Similarly, it is sufficient to show that $2-\tau^c-(\tau^c)^2$ is positive to show that $\pi_{VI,F} > \pi_{VE,F}$. Given that $\tau^c \in [0,1]$, it can be shown that $2-\tau^c-(\tau^c)^2$ is always larger than zero. From Equation (20), it is evident that integrating with I yields greater profit than integrating with E in any case. □

Proof of Proposition 4. I compare E's market share under (B,B) in the no vertical integration case to that in the case of vertical integration with I—i.e., (B,N). Then, $X_E^{BB} - \underbrace{X_{VI,F}^E}_{=X_E^{BN}} = \frac{(1-\tau^c)\tau^c}{6t(1+\tau^c)}$, which can be easily shown to be positive. □

Proof of Proposition 5. First, I decompose CS into two parts: $CS = CS_p + CS_K$, where CS_p (CS_K) denotes the surplus from using the platform's services (from buying a product). I need to show that $\frac{\partial CS}{\partial \tau^c} > 0$, where CS is defined in Equation (14). Equation (14) consists of two parts: CP_p and CS_K. $\frac{\partial CS_p}{\partial \tau^c} > 0$ is easily shown using the Leibniz integral rule as follows.

$$\frac{\partial CS_p}{\partial \tau^c} = \left[v(\tau^c) - \frac{\psi(\tau^c)}{r}\right] + \int_0^{\tau^c} v'(\tau^c) f(x)dx > 0 \quad \because v' > 0. \tag{21}$$

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(V - t\theta - P_I\Big)d\theta + \int_{\theta^c_{\mathcal{D}}}^1 \Big[V - (1-t)\theta - P_E\Big]d\theta \\
&- \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}} \Big(V - t\theta - P_I - \frac{1}{\gamma\Phi_I}\Big)d\theta - \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1 \Big[V - (1-t)\theta - P_E - \frac{1}{\Phi_E}\Big]d\theta \\
&+ \tau^c\left[(V - t\theta^c_{\mathcal{D}} - P_I)\frac{\partial \theta^c_{\mathcal{D}}}{\partial \tau^c} - [V - (1-t)\theta^c_{\mathcal{D}} - P_E]\frac{\partial \theta^c_{\mathcal{D}}}{\partial \tau^c} + \int_0^{\theta^c_{\mathcal{D}}}\left(-\frac{\partial P_I}{\partial \tau^c}\right)d\theta + \int_{\theta^c_{\mathcal{D}}}^1\left(-\frac{\partial P_E}{\partial \tau^c}\right)d\theta\right] \\
&+ (1-\tau^c)\Bigg\{\left(V - t\theta^c_{\mathcal{N}\mathcal{D}} - P_I - \frac{1}{\gamma\Phi_I}\right)\frac{\partial \theta^c_{\mathcal{N}\mathcal{D}}}{\partial \tau^c} - \left[V - (1-t)\theta^c_{\mathcal{N}\mathcal{D}} - P_E - \frac{1}{\Phi_E}\right]\frac{\partial \theta^c_{\mathcal{N}\mathcal{D}}}{\partial \tau^c} \\
&+ \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}}\left(-\frac{\partial P_I}{\partial \tau^c} + \frac{\Phi_I'}{\gamma\Phi_I^2}\mathbb{1}_{\{\Phi_I>1\}}\right)d\theta + \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1\left(-\frac{\partial P_E}{\partial \tau^c} + \frac{\Phi_E'}{\Phi_E^2}\mathbb{1}_{\{\Phi_E>1\}}\right)d\theta\Bigg\} \\
&= \int_{\theta^c_{\mathcal{D}}}^{\theta^c_{\mathcal{N}\mathcal{D}}}\Big[\theta(2t-1) - P_E + P_I\Big]d\theta + \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}}\frac{1}{\gamma\Phi_I}d\theta + \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1\frac{1}{\Phi_E}d\theta \\
&+ \tau^c\left\{\frac{\partial \theta^c_{\mathcal{D}}}{\partial \tau^c}\left[(1-2t)\theta^c_{\mathcal{D}} + P_E - P_I\right] + \int_0^{\theta^c_{\mathcal{D}}}\left(-\frac{\partial P_I}{\partial \tau^c}\right)d\theta + \int_{\theta^c_{\mathcal{D}}}^1\left(-\frac{\partial P_E}{\partial \tau^c}\right)d\theta\right\} \\
&+ (1-\tau^c)\Bigg\{\frac{\partial \theta^c_{\mathcal{N}\mathcal{D}}}{\partial \tau^c}\left[(1-2t)\theta^c_{\mathcal{N}\mathcal{D}} + P_E - P_I + \Delta\right] + \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}}\left(-\frac{\partial P_I}{\partial \tau^c} + \frac{\Phi_I'}{\gamma\Phi_I^2}\mathbb{1}_{\{\Phi_I>1\}}\right)d\theta + \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1\left(-\frac{\partial P_E}{\partial \tau^c} + \frac{\Phi_E'}{\Phi_E^2}\mathbb{1}_{\{\Phi_E>1\}}\right)d\theta\Bigg\} \\
&= \frac{(1-\tau^c)[2\Delta + (1+2\tau^c)\frac{\partial \Delta}{\partial \tau^c}][\Delta(1+2\tau^c) + 3t(1-2t)]}{36t^2} + \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}}\frac{1}{\gamma\Phi_I}d\theta + \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1\frac{1}{\Phi_E}d\theta \\
&+ (1-\tau^c)\left[\int_0^{\theta^c_{\mathcal{N}\mathcal{D}}}\left(\frac{\Phi_I'}{\gamma\Phi_I^2}\mathbb{1}_{\{\Phi_I>1\}}\right)d\theta + \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1\left(\frac{\Phi_E'}{\Phi_E^2}\mathbb{1}_{\{\Phi_E>1\}}\right)d\theta - \frac{\partial P_I}{\partial \tau^c}\frac{\Delta(1-\tau^c)}{3t}\right],
\end{aligned}
\tag{22}
$$

where $\mathbb{1}_{\{\Phi_j>1\}}$ is one if seller j buys data. Given that $\Delta = \frac{1}{\Phi_E} - \frac{1}{\gamma\Phi_I}$, $\theta^c_{\mathcal{N}\mathcal{D}} = \frac{1}{2} + \frac{P_E - P_I + \left(\frac{1}{\Phi_E} - \frac{1}{\gamma\Phi_I}\right)}{2t}$, $\theta^c_{\mathcal{D}} = \frac{1}{2} + \frac{P_E - P_I}{2t}$, $\frac{\partial P_I}{\partial \tau^c} = -\frac{\partial P_E}{\partial \tau^c}$, $\Phi_j' > 0$, and $\frac{\partial P_I}{\partial \tau^c} < 0$. Following some algebraic manipulation, it can be shown that Equation (22) is positive. □

Proof of Proposition 6. First, the consumer surplus levels are obtained as follows.[25]

$$
\begin{aligned}
CS_{BB} &= CS_p + \frac{1}{36\gamma^2 t(1+\tau^c)^2}\Bigg\{\gamma^2\left\{-45t^2(1+\tau^c)^2 + 18t(1+\tau^c)[\tau^c + 2(1+\tau^c)V - 1] + \tau^c(7-8\tau^c) + 1\right\} \\
&\quad - 2\gamma(1-\tau^c)[9t(1+\tau^c) + 8\tau^c + 1] + \tau^c(7-8\tau^c) + 1\Bigg\}. \\
CS_{BN} &= CS_p + \frac{1}{36\gamma^2 t(1+\tau^c)^2}\Bigg\{\gamma^2\left\{-45t^2(1+\tau^c)^2 + 18t(1+\tau^c)[\tau^c + 2(1+\tau^c)V - 1] + \tau^c(7-8\tau^c) + 1\right\} \\
&\quad - 2\gamma\left[1 - (\tau^c)^2\right][9t(1+\tau^c) + 8\tau^c + 1] + (1-\tau^c)(1+\tau^c)^2(1+8\tau^c)\Bigg\}.
\end{aligned}
\tag{23}
$$

The difference between two consumer surplus levels is given as follows.

$$
CS_{BB} - CS_{BN} = \frac{-(1-\tau^c)\tau^c[\gamma(2+\tau^c)(1+8\tau^c) - 18\gamma t(1+\tau^c) - 2(1+8\tau^c)]}{36\gamma t(1+\tau^c)^2}. \tag{24}
$$

[25] Note that CS_p does not change across different data acquisition equilibria because the benefit from using the platform is determined in the first stage of the game.

As shown in Figure 2, I can find a threshold on τ^c below which CS_{BB} is greater than CS_{BN}. The threshold, denoted as τ^c_{CS}, solves $CS_{BB} = CS_{BN}$, which is given by $\frac{\sqrt{9\gamma^2[4t(9t-1)+25]+96\gamma(6t-5)+256}+\gamma(18t-17)+16}{16\gamma} \equiv \tau^c_{CS}$. The social welfare comparison shows qualitatively similar results.□

Proof of Proposition 7. Proposition 7 can be proved by comparative statics for τ^c_{CS} with respect to γ and t. When differentiating τ^c_{CS} with respect to γ and t, respectively, I obtain the following equations.

$$\begin{aligned}\frac{\partial \tau^c_{CS}}{\partial \gamma} &= \frac{-\sqrt{9\gamma^2[4t(9t-1)+25]+96\gamma(6t-5)+256}+3\gamma(5-6t)-16}{\gamma^2\sqrt{9\gamma^2[4t(9t-1)+25]+96\gamma(6t-5)+256}}.\\ \frac{\partial \tau^c_{CS}}{\partial t} &= \frac{9[\gamma(18t-1)+16]}{8\sqrt{9\gamma^2[4t(9t-1)+25]+96\gamma(6t-5)+256}}+\frac{9}{8}.\end{aligned} \tag{25}$$

From Equation (25), it is easy to show that $\frac{\partial \tau^c_{CS}}{\partial \gamma} < 0$, whereas $\frac{\partial \tau^c_{CS}}{\partial t} > 0$, given that t is sufficiently large (i.e., $\frac{(1+\tau^c)\Delta}{3} < t$), which is assumed throughout the paper to guarantee full market coverage. □

Proof of Proposition 8. The right-hand side of Equation (16), given each data acquisition case, is as follows.

$$RHS_{BB} = \theta^c_{BB}F\Big(\psi^{-1}\big(\tfrac{r}{\gamma(1+\tau^c)}\big)\Big)+(1-\theta^c_{BB})F\Big(\psi^{-1}\big(\tfrac{r}{(1+\tau^c)}\big)\Big);\quad RHS_{NB} = \theta^c_{NB}F\Big(\psi^{-1}\big(\tfrac{r}{\gamma}\big)\Big)+(1-\theta^c_{NB})F\Big(\psi^{-1}\big(\tfrac{r}{(1+\tau^c)}\big)\Big)$$

$$RHS_{BN} = \theta^c_{BN}F\Big(\psi^{-1}\big(\tfrac{r}{\gamma(1+\tau^c)}\big)\Big)+(1-\theta^c_{BN})F\Big(\psi^{-1}(r)\Big);\quad RHS_{NN} = \theta^c_{NN}F\Big(\psi^{-1}\big(\tfrac{r}{\gamma}\big)\Big)+(1-\theta^c_{NN})F\Big(\psi^{-1}(r)\Big).$$

The equilibrium τ^c in each data acquisition case is the fixed point satisfying Equation (16). First, I show that the $RHSs$ in the four data acquisition cases can be ranked. Given that $\gamma > 1$, $\theta^c_{NB} < \theta^c_{BB} < \theta^c_{NN} < \theta^c_{BN}$. As long as $F\Big(\psi^{-1}(r)\Big) - F\Big(\psi^{-1}\big(\frac{r}{1+\tau^c}\big)\Big)$ is sufficiently large, it is easy to show that $RHS_{BB} < RHS_{NN}$ because F is non-decreasing and ψ^{-1} is increasing. 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}, it is sufficient to show that $\theta^c_{NB}F\Big(\psi^{-1}\big(\frac{r}{\gamma}\big)\Big) > \theta^c_{BB}F\Big(\psi^{-1}\big(\frac{r}{\gamma(1+\tau^c)}\big)\Big)$. Given that γ and $F\Big(\psi^{-1}\big(\frac{r}{\gamma}\big)\Big) - F\Big(\psi^{-1}\big(\frac{r}{\gamma(1+\tau^c)}\big)\Big)$ are sufficiently large, the condition always holds. Using similar logic, $RHS_{BN} > RHS_{BB}$, $RHS_{NB} < RHS_{NN}$, and $RHS_{BN} < RHS_{NN}$ can be shown: in any case, the assumption that the effect of information acquisition on the willingness to disclose data (the difference in $F(.)$) is greater than the same effect on market share (difference in θ^c) is sufficient in this regard. This implies that RHS_{BB} is the lowest, while

RHS_{NN} is the highest, for all ranges of γ and t. The relative sizes of RHS_{BN} and RHS_{NB} depend on the sizes of γ and t, respectively.

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 by using the same logic as Nayeem and Yankelevich (2017). 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. □

Proof of Proposition 9. Because the solution to τ^c for each data acquisition case can be implicitly determined, I apply the implicit function theorem to calculate comparative statics. First, from Equation (16), let $G = \tau^c - \theta^c F\left(\psi^{-1}\left(\frac{r}{\gamma\Phi_I}\right)\right) - (1-\theta^c)F\left(\psi^{-1}\left(\frac{r}{\Phi_E}\right)\right)$. To see the effect of γ on the equilibrium τ^c, I need to show the sign of $\frac{d\tau^c}{d\gamma} = -\left(\frac{\partial G}{\partial \gamma}\right)/\left(\frac{\partial G}{\partial \tau^c}\right)$. Similarly, for the effect of t on τ^c, I need to calculate $\frac{d\tau^c}{dt} = -\left(\frac{\partial G}{\partial t}\right)/\left(\frac{\partial G}{\partial \tau^c}\right)$. Assuming that $\gamma > 1 + \tau^c$, I find the following signs: $\frac{dG}{d\gamma} = \underbrace{\frac{\partial \theta^c}{\partial \gamma}}_{(+)}\underbrace{\left[F\left(\psi^{-1}\left(\frac{r}{\Phi_E}\right)\right) - F\left(\psi^{-1}\left(\frac{r}{\gamma\Phi_I}\right)\right)\right]}_{(+)} + \underbrace{\theta^c f\left(\psi^{-1}\left(\frac{r}{\gamma\Phi_I}\right)\right)\psi'^2\frac{r}{\gamma^2\Phi_I}}_{(+)} > 0$, $\frac{dG}{dt} = \underbrace{\frac{\partial \theta^c}{\partial t}}_{(-)}\underbrace{\left[F\left(\psi^{-1}\left(\frac{r}{\Phi_E}\right)\right) - F\left(\psi^{-1}\left(\frac{r}{\gamma\Phi_I}\right)\right)\right]}_{(+)} < 0$, and $\frac{dG}{d\tau^c} = 1 + \underbrace{\frac{\partial \theta^c}{\partial \tau^c}}_{(-)}\underbrace{\left[F\left(\psi^{-1}\left(\frac{r}{\Phi_E}\right)\right) - F\left(\psi^{-1}\left(\frac{r}{\gamma\Phi_I}\right)\right)\right]}_{(+)}$

$+\underbrace{\theta^c f\Big(\psi^{-1}\big(\frac{r}{\gamma\Phi_I}\big)\Big)\psi'^2\frac{r\Phi_I'(\tau^c)}{\gamma\Phi_I^2}\mathbb{1}_{\{\Phi_I>1\}}}_{(+)}+\underbrace{(1-\theta^c)f\Big(\psi^{-1}\big(\frac{r}{\Phi_E}\big)\Big)\psi'^2\frac{r\Phi_E'(\tau^c)}{\Phi_E^2}\mathbb{1}_{\{\Phi_E>1\}}}_{(+)}>0$, which implies that $\frac{d\tau^c}{d\gamma}<0$ and $\frac{d\tau^c}{dt}>0$. □

Proof of Proposition 10. By the proof of Proposition 4, $\overline{FC}_{BN}<\min\{\overline{FC}_{BB},\overline{FC}_{NB}\}$. □

Proof of Proposition 11. At the data price under monopoly, C^{Mono}, the monopolistic incumbent always buys data. At market equilibrium, consumer surplus under a monopoly is given as follows.

$$\begin{aligned}CS^{Mono}&=\int_0^{\tau^c}\Big(v(\tau^c)-\frac{\psi(x)}{r}\Big)dF(x)+\tau^c\left[\int_0^{\theta^c_{\mathcal{D}\text{-}Mono}}\Big(V-t\theta-P_I\Big)d\theta\right]+(1-\tau^c)\left[\int_0^{\theta^c_{\mathcal{ND}\text{-}Mono}}\Big(V-t\theta-P_I-\frac{1}{\gamma\Phi_I}\Big)d\theta\right]\\&=\frac{\tau^c(2-3\tau^c)+\gamma^2(1+\tau^c)^2V^2-2\gamma\left[1-(\tau^c)^2\right]V+1}{8\gamma^2t(1+\tau^c)^2},\end{aligned}\tag{26}$$

where $\theta^c_{\mathcal{D}\text{-}Mono}=\frac{V-P_I^{Mono}}{t}$ and $\theta^c_{\mathcal{ND}\text{-}Mono}=\frac{V-P_I^{Mono}-\frac{1}{\gamma\Phi_i}}{t}$. By comparing consumer surplus levels under duopoly, the following equations are obtained.

$$\begin{aligned}CS^{Mono}-CS_{BB}&=\frac{1}{72\gamma^2t(1+\tau^c)^2}\Bigg\{\gamma^2\{90t^2(1+\tau^c)^2-36t(1+\tau^c)[\tau^c+2(1+\tau^c)V-1]\\&-2(1-\tau^c)(1+8\tau^c)+9(1+\tau^c)^2V^2\}+2\gamma(1-\tau^c)[18t(1+\tau^c)+16\tau^c-9(1+\tau^c)V+2]+\tau^c(4-11\tau^c)+7\Bigg\}.\\CS^{Mono}-CS_{BN}&=\frac{1}{72\gamma^2t(1+\tau^c)^2}\Bigg\{\gamma^2(1+\tau^c)^2\left[90t^2-36t(\tau^c+2V-1)-2(1-\tau^c)(1+8\tau^c)+9V^2\right]\\&+2\gamma\left[1-(\tau^c)^2\right](18t+16\tau^c-9V+2)+\tau^c(4-11\tau^c)+7\Bigg\}.\end{aligned}\tag{27}$$

For simplicity, I assume that $V=2$, $\gamma=2$, and $t=0.5$. In this numerical example, Equation (27) is simplified as $CS^{Mono}-CS_{BB}=-\frac{\tau^c(101\tau^c+104)+11}{144(1+\tau^c)^2}$ and $CS^{Mono}-CS_{BN}=\frac{-\tau^c\{\tau^c[64(1-\tau^c)\tau^c+165]+40\}-11}{144(1+\tau^c)^2}$. It can be shown that $CS_{BB}>CS^{Mono}$ and $CS_{BN}>CS^{Mono}$. The result still holds even under all different sets of parametric space. □

Proof of Proposition 12. Since a proportional fee f applied to the sellers' revenues does not change their profit-maximizing prices and output levels, the equilibrium price and market share for seller j are the same as in Equation (7). The unit data price charged by the platform is now a function of f as follows.

$$\begin{cases} \text{Given } E \text{ buys, } I \text{ buys if} \quad C_I^{Int} \leq \dfrac{(1-f)(1-\tau^c)\left\{(\tau^c)^2+\gamma[6t(1+\tau^c)-2\tau^c+2]+\tau^c-2\right\}}{18\gamma^2 t(1+\tau^c)^2} \equiv \bar{C}_I^{Int} = (1-f)\bar{C}_I. \\ \text{Given } E \text{ does not buy, } I \text{ buys if} \quad C_I \leq \dfrac{(1-f)(1-\tau^c)\left[(\tau^c)^2+2\gamma(1+\tau^c)(3t-\tau^c+1)+\tau^c-2\right]}{18\gamma^2 t(1+\tau^c)^2} \equiv \bar{\bar{C}}_I^{Int} = (1-f)\bar{\bar{C}}_I. \\ \text{Given } I \text{ buys, } E \text{ buys if} \quad C_E^{Int} \leq \dfrac{(1-f)(1-\tau^c)\left\{\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]-2\tau^c+2\right\}}{18\gamma t(1+\tau^c)^2} \equiv \bar{C}_E^{Int} = (1-f)\bar{C}_E. \\ \text{Given } I \text{ does not buy, } E \text{ buys if} \quad C_E^{Int} \leq \dfrac{(1-f)(1-\tau^c)\left\{-2(\tau^2)^c+\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+2\right\}}{18\gamma t(1+\tau^c)^2} \equiv \bar{\bar{C}}_E^{Int} = (1-f)\bar{\bar{C}}_E. \end{cases}$$

Given the equilibrium C_j^{Int} as above, the platform's profits under all different data-selling cases are given as follows.

$$\begin{aligned} \pi_p^{BB,Int} &= \frac{1-f}{18\gamma^2 t(1+\tau^c)^2}\Big\{\Big\{2f\left\{-2\gamma(1-\tau^c)^2+\gamma^2\left[9t^2(1+\tau^c)^2+(1-\tau^c)^2\right]+(1-\tau^c)^2\right\} \\ &\quad +(1-\tau^c)\tau^c\left\{(\tau^c)^2+\gamma^2\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+\gamma[6t(1+\tau^c)+4(1-\tau^c)]+\tau^c-2\right\}\Big\}\Big\}. \\ \pi_p^{BN,Int} &= \frac{1-f}{18\gamma^2 t(1+\tau^c)^2}\Big\{\Big\{2f\left\{\gamma^2(1+\tau^c)^2\left[9t^2+(1-\tau^c)^2\right]+[1-2\gamma(1+\tau^c)](1-\tau^c)^2\right\} \\ &\quad +(1-\tau^c)\tau^c\left[(\tau^c)^2+2\gamma(1+\tau^c)(3t-\tau^c+1)+\tau^c-2\right]\Big\}\Big\}. \\ \pi_p^{NB,Int} &= \frac{1-f}{18\gamma^2 t(1+\tau^c)^2}\Big\{\Big\{2f\left\{-2\gamma(1-\tau^c)^2(1+\tau^c)+\gamma^2\left[9t^2(1+\tau^c)^2+(1-\tau^c)^2\right]+\left[1-(\tau^c)^2\right]^2\right\} \\ &\quad +\gamma(1-\tau^c)\tau^c\left\{\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+2[1-(\tau^c)^2]\right\}\Big\}\Big\}. \end{aligned} \tag{28}$$

After some algebraic manipulation, I find that there exists a threshold on τ^c, denoted as τ_{NV}^c, below which $\pi_p^{BB,Int} < \pi_p^{BN,Int}$ and above which the reverse holds. To show whether $\pi_p^{BB,Int}$ is greater or less than $\pi_p^{BN,Int}$, it is sufficient to compare the following equation to zero.

$$\gamma(1-\tau^c)\tau^c\left\{2f(1-\tau^c)[2-\gamma(1+\tau^c)]+\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+2(1-\tau^c)^2\right\}=0. \tag{29}$$

The threshold τ_{NV}^c, which is the solution to Equation (29), is shown in Figure 4 and is also given by the following equation.

$$\tau_{NV}^c = \frac{-\gamma(1+2f)+\sqrt{[\gamma+2(\gamma-2)f+6\gamma t-4]^2+8(\gamma+2\gamma f+2)[\gamma+2(\gamma-1)f-3\gamma t-1]}+4f-6\gamma t+4}{2(\gamma+2\gamma f+2)}. \tag{30}$$

It also needs to be shown that $\pi_p^{NB,Int}$ is never greater than either $\pi_p^{BB,Int}$ or $\pi_p^{BN,Int}$. After some algebra, I can show that $\pi_p^{NB,Int} < \pi_p^{BN,Int}$ if $\tau^c < \tau_{NV}^c$ and that $\pi_p^{NB,Int} < \pi_p^{BB,Int}$ if $\tau^c > \tau_{NV}^c$. It is sufficient to compare the two profit differences to zero by checking the following equations.

$$\begin{aligned} &\pi_p^{BN,Int}-\pi_p^{NB,Int} \quad \text{from} \quad (1-\gamma)(1-\tau^c)\tau^c\left\{2(\gamma+1)f\left[(\tau^c)^2-2\right]+\gamma\left[(\tau^c)^2+6t(1+\tau^c)+\tau^c-2\right]+(\tau^c)^2+\tau^c-2\right\}. \\ &\pi_p^{BB,Int}-\pi_p^{NB,Int} \quad \text{from} \quad (1-\tau^c)\tau^c\left\{2f\left[2\gamma(1-\tau^c)+(\tau^c)^2+\tau^c-2\right]+2\gamma\left[3t(1+\tau^c)+(1-\tau^c)^2\right]+(\tau^c)^2+\tau^c-2\right\}. \end{aligned} \tag{31}$$

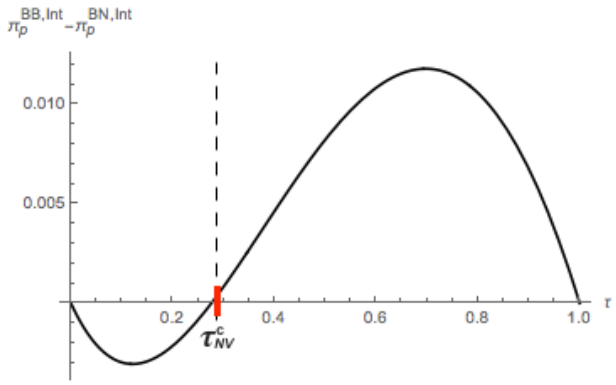

Figure 4: $\pi_p^{BB,Int} - \pi_p^{BN,Int}$ when $\gamma = 2.5$, $t = 0.3$, and $f = 0.5$

First, when τ^c is sufficiently small (i.e., $\tau^c < \tau_{NV}^c$), the curly bracket part of the first equation in (31) is always negative. Since $\gamma > 1$, this suggests that $\pi_p^{BN,Int} > \pi_p^{NB,Int}$. From the curly bracket part of the second equation in (31), it is easy to show the following: $2\gamma(1-\tau^c) + (\tau^c)^2 + \tau^c - 2$ and $2\gamma\left[3t(1+\tau^c) + (1-\tau^c)^2\right]$ are always greater than zero. Additionally, when τ^c is sufficiently large (i.e., $\tau^c > \tau_{NV}^c$), $(\tau^c)^2 + \tau^c - 2$ is negative, but its absolute value is small enough to make the curly bracket part be positive. Thus, $\pi_p^{BB,Int} > \pi_p^{NB,Int}$ is guaranteed for a relatively large range of τ^c. □