Privacy, Information Acquisition, and Market Competition*

Soo Jin Kim†

Michigan State University

January 12, 2018

Click Here for the Latest Draft

Abstract

This paper analyzes how the endogenous availability of personal information affects market outcomes in a two-sided market where sellers target advertisements to individuals who have varying privacy concerns. I focus on how a market entrant that has worse targeting technology than an incumbent is affected by a lack of information. I show that an entrant is disproportionately affected by consumers' privacy concerns. The welfare analysis shows that privacy concerns and the resulting market outcomes may lower consumer surplus and social welfare. Therefore, individually optimal decisions on data disclosure might not be socially optimal when aggregated. The empirical evidence, which is based on Google Android App Market data, corroborates the hypotheses in the model and the effectiveness of specific policy remedies that are derived from the theoretical findings.

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, 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) for valuable discussions and comments.

†kimsoo25@msu.edu; Department of Economics, Michigan State University, 486 W. Circle Dr., East Lansing, MI 48824

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 or use it to deliver more targeted advertising (ads) to consumers.[1] Each consumer's privacy concerns determine the total amount of personal data the consumer makes available to the platform, and thus, such concerns play an important role in the business of the platforms, the sellers or the advertisers, as well as in consumers' buying decisions. A platform can earn more money when it has more information because each seller or data holder who purchases data from the platform can attract more consumers through better targeting. Importantly, because the amount of information available determines the overall effectiveness of each seller's ad targeting, thereby impacting the overall consumer shopping experience, consumers themselves must also balance privacy concerns with convenience: better targeting based on more information comes at a price through the loss of privacy. Indeed, as I show in Section 6, a lack of confidence in data collection and data usage policies exacerbates consumers' privacy concerns: when more privacy-sensitive permissions are requested by a less trustworthy mobile application (app), we can observe a lower willingness to download the app.

Due to this trade-off between the benefits and the cost of privacy loss, regulators have engaged in efforts to balance firms' profit-seeking behaviors and consumers' privacy protection. However, it is always debatable where we should place more emphasis, which leads to the continuing amendment of privacy-related regulations. For example, a set of privacy rules approved by the Federal Communications Commission in October 2016, requiring ISPs to conspicuously ask for permission before collecting and selling personal data, such as browsing histories, was overturned in April 2017. The rationale for this change, which is in the industry's favor, was that privacy-related matters should be regulated on a case-by-case basis when each company violates its own privacy policies. In response to this revocation, California tried to revive the broadband privacy rules to protect consumers' privacy rights; however, this attempt ultimately failed to become law.[2] Motivated by the privacy debate, in this paper, I seek to inform researchers and

[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] Refer to Assembly Bill 375: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

policy makers how privacy sensitivity and protection can impact market outcomes and welfare.

If each seller has the most recent information about potential customers, it can target its ads to better attract them. When a seller's targeted ads become more effective, consumers face lower mismatch costs from that seller: 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 the new data with its existing customer data.

This asymmetry raises several important research questions. How does the total amount of personal information available affect market competition when an entrant with weaker targeting skills demands considerably more personal data than the incumbent? How do such privacy concerns affect market outcomes? Does an increase in privacy concerns have a greater adverse effect on market entrants? How does it ultimately affect social welfare?

To answer these questions, I develop a model in which sellers are asymmetric with respect to their initial targeting technology and their overall product quality. These sellers 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 valuation of product quality. Depending on his privacy type, each consumer decides whether to disclose personal information to the platform by comparing the benefit arising from interacting with the others on the platform and a nuisance cost arising from a loss of privacy. The platform aggregates all available detailed personal data and sells it to any seller who wants to use it to create targeted ads.

The main result shows that the entrant with weaker targeting skills always wants to buy data from the platform, whereas the incumbent buys only when privacy concerns exceed a certain threshold. In equilibrium, the entrant suffers from lower market share and revenue when the incumbent also buys data. Since the incumbent wants to buy data if less data become available

due to greater privacy concerns, consumers' privacy concerns could disproportionately harm the entrant, which can be an anti-competitive threat.

Another important market outcome related to privacy concerns is data-driven vertical integration. In a two-sided market, in particular, there are multiple instances of vertical integration between the platform and a downstream firm, such as a content provider or an online retailer. Vertical integration can also be motivated by sellers' desire to obtain a greater collective amount of exclusive data. For example, when AT&T and Time Warner Media company announced their intention to merge in October 2016, they noted that the merger would benefit consumers by providing better targeted ads based on extensive customer data. By maintaining exclusive use of these data, Time Warner can target consumers much more effectively, thereby attracting more advertisers to choose Time Warner's content as a channel for advertising. By conferring an unfair advantage upon the integrated downstream firm, the merger may lead to antitrust concerns. In particular, data-driven vertical integration is related to consumers' privacy concerns in that the amount of information available plays a key role in attracting more customers.[3]

The results of the vertical integration model considered in this paper indicate that if consumers are more likely to be privacy-sensitive such that less personal information becomes available, the platform and the incumbent will vertically integrate and prevent the entrant from obtaining access to customer data. This implies that privacy concerns have disproportionate negative effects for the entrant. The welfare analysis results show that such integration not only harms the entrant but also lowers consumer surplus and total social welfare: individually optimal decisions on data disclosure might not be socially optimal when aggregated because consumers could also be harmed by the lack of competition.

I further motivate this study by using Google Android app market data to propose more plausible policy remedies. Specifically, the evidence indicates that consumers' privacy concerns are affected by a data collector's reputation. Considering this evidence for the effect of data collectors' reputation on lowering privacy concerns in conjunction with the theoretical findings,

[3]In this example, however, Time Warner Media company itself is a platform, so the merger can be regarded as a horizontal platform-to-platform merger rather than as vertical integration. Still, 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 Time Warner Inc. and HBO. As the integrated firm, HBO can better target customers of premium television content compared to other rival content providers such as Showtime. The Amazon-Whole Foods acquisition deal can be another example of data-driven integration: by using the extensive amount of transaction data on Amazon, Whole Foods is able to suggest better products to consumers, thereby attracting consumers away from less informed competitors such as Kroger.

I suggest a specific policy, exemplified by a government-backed privacy certification program, that encourages more consumers to voluntarily disclose their personal information, which is socially optimal in this model.

Previous Literature The stream of literature most closely related to my work is research on the effect of privacy on a market's competitive structure.[4] 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. (2016) also endogenize privacy by allowing consumers to anonymize themselves for a cost and analyze how privacy concerns and the resulting information availability affect competing firms' price discrimination and data acquisition decisions, consumer surplus, and social welfare. However, in the real world, asymmetries between firms are persistent for a variety of reasons arising from different previous sales experience or scope of products. By taking into account such asymmetries, my paper leads to important implications not present in a symmetric setting: privacy concerns may disproportionately harm the entrant through data-driven vertical integration, for example. 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. Unlike Campbell et al. (2015), however, my primary concern is further asymmetries in the sellers' market in terms of initial targeting technology, product quality and heterogeneous consumer privacy sensitivity. By including such asymmetries and heterogeneity, I offer a microfoundation for understanding how consumers react differently to potential privacy risk and how sellers are disproportionately affected by privacy concerns. Braulin and Valletti (2016) also model vertically differentiated sellers to determine how exclusive data sales affect consumer and social welfare. However, they do not study the potential anti-competitive effects that arise when a seller lacks customer information. Also, none of the

[4]Taylor (2004), Acquisti and Varian (2005), Conitzer et al. (2012), Belleflamme and Vergote (2016), and Koh et al. (2017) allow consumers to actively decide how much information to disclose. However, these papers assume a monopolistic seller and thus do not examine how privacy concerns or information availability affect market competition.

papers focuses on more diverse market outcomes related to privacy concerns, such as data-driven vertical integration, which I consider in this paper. Although Kim et al. (2016) analyze how access to consumer data that enables personalized pricing affects the overall welfare of horizontal mergers, the focus is different; I focus on dynamic relationship between privacy concerns and data-driven vertical integration market outcomes under asymmetric seller setup.

This paper also relates to research on privacy and online 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 from less information availability on small advertisers shares similarities with the findings of my paper. D'Annunzio and Russo (2017) find that if consumers overly 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.[5]

Finally, my paper contributes empirical evidence that corroborates the assumptions and hypotheses made implicitly throughout the paper. In a related study that further corroborates my work, Kummer and Schulte (2017) examine a money-for-privacy trade-off in the smartphone applications market by using Google Android app market data.

The remainder of the paper is organized as follows. In Section 2, the theoretical model is provided. The no vertical integration and the vertical integration games are solved in Sections 3 and 4, respectively. In Section 5, the welfare implications are drawn. Section 6 provides the empirical evidence demonstrating the effectiveness of specific policy remedies. Section 7 considers possible model extensions and checks the robustness of the main model. Finally, Section 8 concludes by suggesting additional policy implications.

2 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 the email address,

[5]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), Kim and Choi (2010) and Kim and Wagman (2015) argue that information disclosure is not always harmful to the individual and may contribute to improving welfare.

gender, and date of birth, for all consumers. The amount of basic information is normalized to 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 horizontally 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 valuation of the overall quality of the products provided by sellers, and this is uniformly distributed over the vertical unit line. Again, each consumer on a continuum of θ_i decides which seller to purchase a product from given that each seller provides different product quality levels. As above, let $\mathcal{H}$ denote a set of high-valuation consumers and $\mathcal{L}$ denote a set of low-valuation consumers.[6] 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 a high- or a low-quality seller ($\mathcal{H}$ or $\mathcal{L}$).[7] Therefore, each consumer obtains net utility from two sources.

First, any consumer obtains 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, $P(i \in \mathcal{D})$. 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 informa-

[6] The overall quality captures not only the product quality itself, which is related to its functions, but also the service quality that sellers provide when their products are sold. I assume that targeted ads play a role in improving service quality. That said, if the seller of a low-quality product can better target its ads, it could overcome its product quality disadvantage to some extent by offering the consumer better service.

[7] Again, a consumer takes into consideration product quality and targeted ads when identifying himself as part of the *High* or *Low* valuation group.

tion with others on the platform, he will obtain 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. 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.[8] Normalizing both the benefit and cost of $i \in \mathcal{ND}$ to zero, the utility for each consumer i from the platform is given as follows.

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

where the superscript p denotes the platform, $v(P(i \in \mathcal{D}))$ 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 $\psi(\tau_i)$ is continuous in τ_i. Because it is strictly increasing and continuous in τ_i, it is invertible.[9]

Since seller j's targeting effectiveness increases in the amount of consumer data, a consumer's information disclosure decision also affects the utility from product purchase: 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.[10]

$$u_{ij} = V + \theta_i s_j - P_j - \mathbb{1}_{\{i \in \mathcal{ND}\}} \left(\frac{1}{\gamma_j D_j} \right), \tag{2}$$

where V denotes the reservation value (base utility), which is assumed to be large enough to fully

[8] In Section 6, I show that a consumer is more willing to disclose information when the data collector has a stronger market reputation. Choi et al. (2016) also discuss the role of reputation in reducing privacy nuisance costs.

[9] If a consumer discloses his information ($i \in \mathcal{D}$), he enjoys the benefit of $v(P(i \in \mathcal{D}))$, which includes immediate benefits, such as networking with friends. Simultaneously, he faces the utility loss from a nuisance cost, which might arise from either direct economic losses (e.g., a threat of identity theft) or a negative psychological feeling about disclosing personal information.

[10] 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 but generate much simpler equations.

cover the market, and the valuation of consumer i with respect to product quality is given by $\theta_i \sim U[0,1]$. The overall product quality is denoted as s_j, P_j denotes the price of products from seller j, and $\frac{1}{\gamma_j D_j}$ is the mismatch cost where γ_j denotes seller j's initial targeting technology and D_j denotes the amount of consumer data possessed by seller j. The indicator function $\mathbb{1}_{\{i \in \mathcal{ND}\}}$ is one if a 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—higher γ_j. Finally, I assume that a consumer's mismatch cost decreases as seller j obtains more consumer information for creating targeted ads and that the effect of data on reducing mismatch costs is non-increasing as D_j increases.[11]

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.[12]

Lastly, one might question the additive separable utility specification of information disclosure and product purchasing. In this setup, consumers only consider the immediate benefits from disclosing information to the platform and do not take into consideration any potential future benefits arising from better targeted ads. This assumption makes sense for many real case examples of platforms, such as social media. For example, when a consumer posts news of the birth of his baby on Facebook, he is more likely to do so to spread good news to his friends than in hopes of seeing more relevant ads on baby products.[13]

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

[11]The following is a more intuitive explanation of the mismatch cost. A consumer knows that there are two sellers (retailers) I and E that have identical product sets but that differ in overall product quality. A consumer does not have a preference for one specific brand (seller) over the other but has different needs for a specific type of product that both sellers sell. Thus, if each seller has the most recent information about potential customers, it can target its ads to be more attractive to them. Then, consumers will spend less time finding the most suitable product because they are immediately obtaining the relevant information from these targeted ads.

[12]See Choi et al. (2016) for a more detailed description of such information externalities.

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

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 $P(i \in \mathcal{D})$ amount of detailed information to any seller that wants to buy. The platform earns profits only from selling user data to any seller. By optimally setting the per unit data price C, the platform solves the following profit maximization problem.

$$\max_{C} \quad \pi_p(C|P(i \in \mathcal{D})) = \begin{cases} 0 & \text{if no seller buys data} \\ P(i \in \mathcal{D})C & \text{if one seller buys data} \\ 2P(i \in \mathcal{D})C & \text{if both sellers buy data,} \end{cases} \tag{3}$$

where the subscript p denotes the *platform*.[14]

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 vertically differentiated in terms of product quality, which is denoted by s_j, and service quality in terms of targeting quality, denoted by $\gamma_j D_j$, where D_j is equal to $1 + P(i \in \mathcal{D})$ if seller j buys data from the platform or one otherwise. The set of products for each seller overlaps, but overall quality—in terms of product and service—is different. The overall product quality increases in s_j, whereas service quality in terms of targeting increases in γ_j and in D_j. If a seller decides to buy data from the platform, he pays per unit data price C. Whether or not he does so depends on the relative magnitudes of C and γ_j which captures previous sales experience and the existing customer information. Without loss of generality, I first assume that $\gamma_I \geq \gamma_E$: seller I has better targeting technology than seller E. For simplicity, I normalize γ_E to one and denote γ_I as γ where $\gamma > 1$.[15] Regarding product quality, either I or E can provide a high-quality product. I focus on the case in which I is better at initial targeting (service quality) but E is better at product quality, and thus, $s_I < s_E$ is assumed throughout the paper. Each seller's profit maximization problem is defined

[14]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.

[15]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. See Appendix A for a detailed discussion.

as follows.

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

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

As for the other case, $s_I > s_E$, the qualitative results in the paper still hold, but it generates less embracive results than the case of $s_I < s_E$. If the incumbent has advantages in targeting as well as product quality, the room for the entrant to overcome his disadvantage by obtaining consumer data is very limited. This leads to an equilibrium in which only the incumbent benefits from data acquisition, which is less interesting because targeting has a limited effect on the competitive structure. See Appendix A for details.[16]

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: $\mathcal{D}$ or $\mathcal{ND}$ for τ_i and $\mathcal{H}$ or $\mathcal{L}$ for θ_i. In the no vertical integration case, the timing of the game follows Figure 1. In the vertical integration game, I add 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.

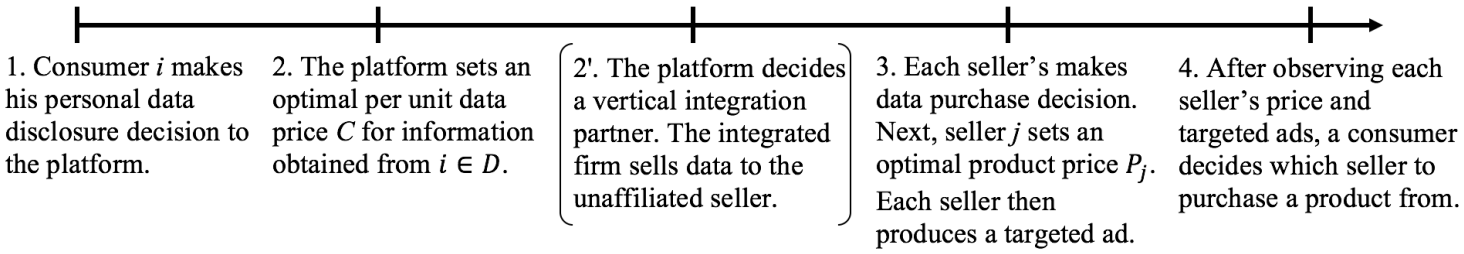

Figure 1: Timing

[16] Walmart.com is a representative example of an incumbent firm. Because Walmart.com sells various products in many categories, it is a generalist. Any specialist retailers that sell various products in a specific category, such as apparel, can be considered high-quality entrants in that they specialize in their own business area. Campbell et al. (2015) made a similar assumption.

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 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.[17]

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 $\frac{\psi(\tau_i)}{r} < v(P(i \in \mathcal{D}))$ will disclose. The portions of privacy-sensitive consumers (not disclosing information) and privacy-insensitive consumers (disclosing) are implicitly determined by the following Proposition.[18]

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

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

A parametric example Let $\tau_i \sim U[0,1]$, $\psi(\tau_i) = \lambda \tau_i^2$ where $\lambda > 2$, 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, τ^c also increases.

Given $P(i \in \mathcal{D}) = \tau^c$, the amount of aggregated detailed data, I solve for the PBE using backward induction to obtain sequentially rational strategies. From the utility specification in (2), the indifference condition is $\theta^c_{\mathcal{ND}} = \frac{P_E - P_I + (\frac{1}{D_E} - \frac{1}{\gamma D_I})}{s_E - s_I}$ for $i \in \mathcal{ND}$ and $\theta^c_{\mathcal{D}} = \frac{P_E - P_I}{s_E - s_I}$ for $i \in \mathcal{D}$. The weighted indifference condition can be rewritten in a simple way as follows.

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

[18] 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 $P(i \in \mathcal{D}) = \tau^c_a$ and $P(i \in \mathcal{ND}) = 1 - \tau^c_a$, where subscript a denotes the *anticipated* proportion. In equilibrium, τ^c_a should be consistent with the true τ^c, which is aggregately determined by consumers. For notational convenience, I drop the subscript a.

$$P(i \in \mathcal{L}) = \theta^c = \frac{P_E - P_I + (1-\tau^c)\Delta}{s}, \tag{6}$$

where $\Delta = (\frac{1}{D_E} - \frac{1}{\gamma D_I})$ and $s = s_E - s_I$.[19] The market share for each seller is given by $X_I = \theta^c = P(i \in \mathcal{L})$ and $X_E = 1 - \theta^c = P(i \in \mathcal{H})$ under $s_I < s_E$. Given X_I and X_E, the solutions to the profit maximization problem with respect to P_j are given by

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

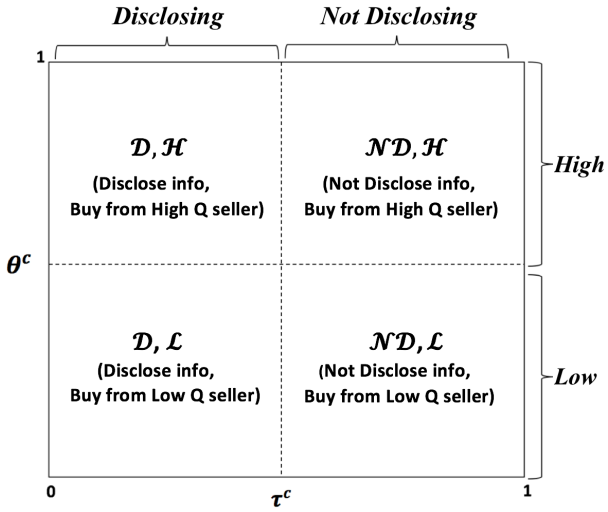

Figure 2: Endogenously derived thresholds, τ^c and θ^c, on τ_i and θ_i

To guarantee an interior solution, I assume throughout the paper that $\frac{\Delta(1-\tau^c)}{2} < s$: seller E's quality s_E is large enough to have positive demand. Note that the sufficient condition for the interior solution is $s \geq \frac{1}{2}$, which is assumed to be satisfied throughout the paper.

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 D_j: either $1 + \tau^c$ (if purchasing) or 1 (otherwise), which can be derived as follows.

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

[19] Note that Δ can be negative if τ^c is large enough and E is the only data holder, which indicates the (N, B) data acquisition case.

One can rank different Δ as $\Delta_{NB} < \Delta_{BB} < \Delta_{NN} < \Delta_{BN}$,[20] 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 that guarantee that one seller will buy data, given the rival's decision. The thresholds are

$$\begin{cases} \text{Given } E \text{ buys, } I \text{ buys if} \quad C < \dfrac{(1-\tau^c)(\Delta_{BB}-\Delta_{NB})(2s+(1-\tau^c)(\Delta_{BB}+\Delta_{NB}))}{9s\tau^c} \equiv \bar{C}_I \\ \text{Given } E \text{ does not buy, } I \text{ buys if} \quad C < \dfrac{(1-\tau^c)(\Delta_{BN}-\Delta_{NN})(2s+(1-\tau^c)(\Delta_{BN}+\Delta_{NN}))}{9s\tau^c} \equiv \bar{\bar{C}}_I \\ \text{Given } I \text{ buys, } E \text{ buys if} \quad C < \dfrac{(1-\tau^c)(\Delta_{BN}-\Delta_{BB})(4s-(1-\tau^c)(\Delta_{BN}+\Delta_{BB}))}{9s\tau^c} \equiv \bar{C}_E \\ \text{Given } I \text{ does not buy, } E \text{ buys if} \quad C < \dfrac{(1-\tau^c)(\Delta_{NN}-\Delta_{NB})(4s-(1-\tau^c)(\Delta_{NN}+\Delta_{NB}))}{9s\tau^c} \equiv \bar{\bar{C}}_E. \end{cases}$$

Unambiguously, $\bar{\bar{C}}_I > \bar{C}_I$ and $\bar{\bar{C}}_E > \bar{C}_E$: provided that the rival does not buy, it is more likely that the other firm will buy. In other words, data acquisition is a strategic substitute, since $\frac{\partial^2 \pi_j}{\partial D_I D_E} = \frac{-2(1-\tau^c)^2}{9sD_I^2 D_E^2 \gamma} < 0$. 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.

Given the interior solution assumption on s, the four thresholds are ranked as $\bar{C}_I < \bar{\bar{C}}_I < \bar{C}_E < \bar{\bar{C}}_E$. Based on the ranking on C, the platform makes a data pricing decision. Given the equilibrium decision of each seller, the platform's profits under different levels of C are

$$\begin{cases} \pi_p^{BB} = 2\bar{C}_I \tau^c & \text{if} \quad C \leq \bar{C}_I, \quad \text{(Buy, Buy)} \\ \pi_p^{NB} = \bar{\bar{C}}_E \tau^c & \text{if} \quad C \leq \bar{\bar{C}}_E, \quad \text{(Not buy, Buy)}, \end{cases}$$

because it wants to set as high a price as possible. By comparing different profit levels, the platform sets the optimal C.[21] From the profit comparison, there exists a τ_{NV} such that if $\tau^c < \tau_{NV}$, the platform sets $C^* = \bar{C}_I$, meaning that both sellers buy data, which means (B, B). Similarly, if $\tau^c > \tau_{NV}$, the platform sets $C^* = \bar{\bar{C}}_E$, meaning that only seller E buys data, whereas seller I does not, which means (N, B). The result is summarized in Proposition 2. Note that I restrict my attention to $\frac{1}{2} < s < 1$, since $s > 1$ always leads to a data acquisition equilibrium in which the entrant is the only data holder, which is not interesting.

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

[21] In Section 7, I consider the case in which the platform engages in data price discrimination.

Proposition 2. *If less data become available due to greater privacy concerns ($\tau^c < \tau_{NV}$), the platform sets a lower price, meaning that both sellers buy data, which means (B, B). However, if more data are available ($\tau^c > \tau_{NV}$), the platform sets a higher price, meaning that seller E buys data, whereas seller I does not, which means (N, B).*

Intuitively, if more data become available, E is able to overcome its disadvantage in targeting skills, thereby having higher willingness to pay for additional data. Therefore, $\tau^c > \tau_{NV}$ leads to the exclusive data selling arrangement with E.[22]

3.2 Implications

Proposition 2 states that if consumers are more privacy-sensitive (less data availability, $\tau^c < \tau_{NV}$), both sellers buy data, while only seller E buys if more data become available ($\tau^c > \tau_{NV}$). Intuitively, if only a small amount of data is available, the data are not sufficient for E to overcome its targeting disadvantage. So, if I buys the same small set of data, it can dominate the market easily. Thus, seller E always suffers from a lower market share if seller I also buys data while it can enjoy higher market revenue if it is the only data holder.[23] However, the effect on profits is ambiguous because the seller needs to pay a much higher price to the platform to secure a data monopoly. By comparing π_E^{BB} to π_E^{NB}, where the superscripts denote data acquisition status, one can find another threshold on τ_i, denoted as τ'_{NV}, that determines whether one of the profit levels is greater than the other. If $\tau^c > \frac{-\gamma-4\gamma s+\sqrt{(\gamma+(4\gamma-6)s)^2-8(\gamma-2)(-\gamma+(2\gamma-3)s+1)}+6s}{2(\gamma-2)} \equiv \tau'_{NV}$, $\pi_E^{BB} > \pi_E^{NB}$, and the reverse holds under $\tau^c < \tau'_{NV}$. As in Figure 3, τ'_{NV} can be larger than τ_{NV}, which leads to counterintuitive consequences. If $\tau^c < \tau'_{NV}$, (N, B) makes E better off, but this usually does not arise in equilibrium because I also always wants to buy data for a small range of τ^c. Since E's best response to I's buying decision is also to buy, E selects (B, B). Analogously, if $\tau^c > \tau'_{NV}$, (B, B) leads to a higher profit for E because the platform extracts too much rent given (N, B). However, for this range of τ^c, I always refuses to buy data, thereby leading to (N, B) in equilibrium. Therefore, the optimal choice for E might lead to a suboptimal result in terms of profit. Corollary 1 summarizes this finding.

[22] As Montes et al. (2016) 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.

[23] Note that seller I becomes indifferent between buying and not buying data because the platform extracts the surplus from buying data by charging the data price C.

Corollary 1. *For the entrant E, the exclusive use of data, (N, B), leads to greater market share and higher revenue than data sharing, (B, B). However, (N, B) leads to lower profit than (B, B) in most cases, except for $\tau_{NV} < \tau^c < \tau'_{NV}$.*

This result suggests that consumers' privacy concerns and the resulting decrease in available data disproportionately harm the entrant in terms of market share and revenue. Figure 3 demonstrates the results in the no vertical integration game: except for the shaded area, the optimal decision for E results in the suboptimal outcome in terms of profit.[24]

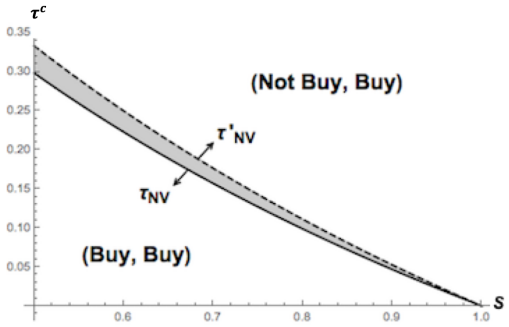

Figure 3: Equilibrium under No Vertical Integration if $\gamma = 1.8$

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. 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.

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, I assume 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.

[24] If $\tau'_{NV} < \tau_{NV}$ holds, a similar argument can be applied. Except for $\tau'_{NV} < \tau^c < \tau_{NV}$, the optimal decision for E results in the suboptimal outcome in terms of profit.

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

where the first two letters in the subscript *VI* denote *Vertical Integration with I* and the last letter indicates data foreclosure or selling status. By comparing $\pi_{VI,S}$ to $\pi_{VI,F}$, the integrated firm decides whether to sell data to the unaffiliated firm.

$$\pi_{VI,S} - \pi_{VI,F} = -\frac{2(\tau^c - 1)\tau^c\left(\gamma\left(s\tau^c + s + (\tau^c)^2 + \tau^c - 2\right) - 2\tau^c + 2\right)}{9\gamma s(\tau^c + 1)^2}. \tag{10}$$

Thus, if $\tau^c < \frac{\sqrt{\gamma^2((s-2)s+9)-4\gamma(s+3)+4}-\gamma(1+s)+2}{2\gamma} \equiv \bar{\tau}'$, $\pi_{VI,S} < \pi_{VI,F}$: as τ^c increases, it is more likely to sell data to the rival. For the incumbent, the marginal benefit of obtaining more data is much smaller due to his initial advantage in targeting technology. Thus, if the amount of data is above a certain threshold, the market dominance effect becomes smaller than the data-selling revenue effect. Accordingly, the integrated firm earns greater profits from selling data.

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.

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

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

Thus, if $\tau^c > \frac{1}{2}\left(-\sqrt{\gamma^2(s+2)^2 + 2\gamma(s-6) + 9} + \gamma(s+2) - 1\right) \equiv \underline{\tau}$, the integrated firm forecloses data access. In contrast to the former case, if the platform is integrated with seller E, 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 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 the right-hand side of the inequality decreases in s, which means that if s increases, it is more profitable to foreclose data access because the combination of data monopolization and a higher s 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. Since $\underline{\tau} < \bar{\tau}'$, there are three possible cases: I-Foreclose or E-Sell for $\tau^c < \underline{\tau}$, I-Foreclose or E-Foreclose for $\underline{\tau} < \tau^c < \bar{\tau}'$, and I-Sell or E-Foreclose for $\tau^c > \bar{\tau}'$.[25] First, if $\tau^c < \underline{\tau}$ (less data availability), the ISP integrates with I and forecloses E from data access. Second, if $\underline{\tau} < \tau^c < \bar{\tau}'$ (moderate data availability), there is another threshold on τ_i, denoted as τ_V, such that $\tau^c > \frac{\sqrt{\gamma^2((s-2)s+9)+2\gamma(s-9)+9}+\gamma(-(s+1))+1}{2(\gamma-1)} \equiv \tau_V$ leads to vertical integration with E and data foreclosure. If $\tau^c < \tau_V$, the platform and I integrate and foreclose data access. Finally, if $\tau^c > \bar{\tau}'$ (more data availability), the platform and E integrate and foreclose data access for I in equilibrium. As a result, there are two possible cases with a vertical integration equilibrium: (a) integration with I if $\tau^c < \tau_V$ and (b) integration with E if $\tau^c > \tau_V$. There is no data-selling equilibrium in any case, and data foreclosure always emerges in the vertical integration game.

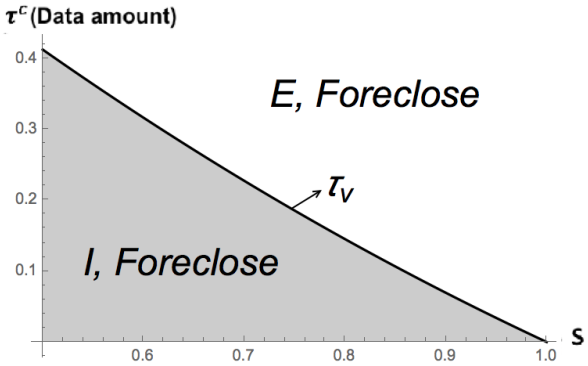

Figure 4: Vertical Integration Equilibrium on (s, τ^c) space when $\gamma = 2$

Furthermore, I must verify whether the platform and each seller have an incentive to vertically integrate with one another by comparing the joint profits of the platform and each seller under no vertical integration to the profits of the integrated firm. First, if the platform and I are integrated, they always want to be integrated if (B, B) is the no vertical integration equilibrium. However, for the case of (N, B), vertical integration is more profitable only if $\tau^c < \frac{\sqrt{4\gamma(\gamma(3\gamma-5)(3\gamma+1)+\gamma(\gamma-1)^2s^2-(\gamma(2(\gamma-3)\gamma+3)+1)s+6)+9}+2\gamma(s-\gamma(s+1))+1}{4(\gamma-1)\gamma-2} \equiv \bar{\tau}$: if τ^c is sufficiently large, the platform can extract much higher revenue from E by charging a higher price for the data. Similarly, the platform and E have an incentive to be integrated if $\tau^c > \underline{\tau}$. The intuition is similar to that above: if τ^c is small, the platform is better off from selling data to both sellers at a lower price. However, given that the integration with I only emerges if τ^c is small enough while

[25] For example, I-Foreclose means that the ISP integrates with I and forecloses E from data access. Similarly, E-Sell means that the ISP and E integrate and sell data to I.

that with E emerges in the opposite case, there is no deviation incentive for the platform.[26] As in Figure 4, vertical integration can always emerge, and the merger partner is determined by the threshold of τ_V. Proposition 3 summarizes these results.

Proposition 3. *The platform has an incentive to vertically integrate with seller I if there are more privacy-sensitive consumers and less available personal data ($\tau^c < \tau_V$), whereas integration with E emerges if there are more privacy-insensitive consumers and more available personal data ($\tau^c > \tau_V$). Regardless of which seller is involved, the unaffiliated seller forgoes buying data.*

4.2 Implications

Absent vertical integration, seller E always wants to buy data to overcome its initial disadvantage in targeting skills, while I buys data only when τ^c is small. 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 sellers, especially seller E, I compare sellers' profits with and without vertical integration. First, if the platform and seller E are integrated, the unaffiliated seller, I, becomes indifferent to the existence of vertical integration. Absent vertical integration, seller I has the same profit level whether buying or not buying data because the platform extracts all additional revenue by setting the price of data. Since the vertical integration equilibrium is (N, B), which is the same as in one case of no vertical integration, there is no difference in I's profits due to the merger between the platform and E.

If the platform and I are integrated, the unaffiliated seller E always suffers from lower profits due to the integration and resulting data foreclosure: $\min\{\pi_E^{BB}, \pi_E^{NB}\} > \pi_{VI,F}^E$ where $\pi_{VI,F}^E$ denotes E's profit under vertical integration with I. Thus, due to data foreclosure, seller E generally suffers from lower profits under vertical integration because it has a smaller market share when the platform integrates with I than it does in any case under no vertical integration, which leads to either (B, B) or (N, B). Proposition 4 summarizes this implication.

Proposition 4. *When the platform is vertically integrated with seller I, seller E suffers from a smaller market share, thereby obtaining lower profits due to data foreclosure. When the platform is vertically integrated with E, seller I faces no difference in profit, regardless of the presence of vertical integration.*

[26] Note that $\underline{\tau} < \tau_V < \bar{\tau}$.

This result raises a very important antitrust implication regarding data-driven vertical integration. Although vertical integration always forecloses data access regardless of the seller with which the deal is made, the entrant seller is more likely to be harmed. As consumer information is vital for a small seller or a market entrant with weaker targeting skills, vertical integration with a seller with better targeting skills is likely to have an anti-competitive effect because it prevents the unaffiliated entrant from using data to overcome its initial disadvantage. Moreover, given that greater privacy concerns and the resulting decrease in available personal data lead to integration with seller I, consumers' privacy concerns might disproportionately harm seller E.

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 $CS = CS_p + CS_K$, and the subscript $K \in \{VI, VE\}$; VI (VE) denote the surplus from vertical integration with seller I (E), respectively. CS_p denotes the surplus from using the platform's services. 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 under 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.

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

For simplicity, I make parametric assumptions as $\tau_i \sim U[0,1]$, $\psi(\tau_i) = \lambda \tau_i^2$ where $\lambda > 2$, $r = 1$, and $v(\tau^c) = 1 + \tau^c$; in addition, $V = 2$. The main focus here is to determine how data-driven vertical integration affects consumer surplus and total social welfare, so I focus

only on the parametric space in which vertical integration can always occur, i.e., $\underline{\tau} < \tau^c < \bar{\tau}$. First, I examine how consumer surplus from sellers is affected by vertical integration. Figure 5 shows data acquisition equilibria under a different parametric space. As shown in Section 3.1, $\tau^c = \frac{\sqrt{4\lambda+1}+1}{2\lambda}$ in this example, which means that consumer surplus is a function of λ, which is the marginal privacy nuisance cost. Given that $\tau^c > \tau_V$ leads to integration with E in equilibrium, there exists a threshold on marginal privacy nuisance cost λ, say $\bar{\lambda}$, below which $\tau^c > \tau_V$.[27] As in Figure 6, the consumer surplus under (B, N), which arises from vertical integration with I, is always lower than that under (N, B), which arises from vertical integration with E, if $s > \bar{s}$. Thus, if s is sufficiently large, the consumer surplus under integration with E is always greater than that under vertical integration with I. The comparison between different social welfare levels generates a qualitatively similar result.[28] In other words, the integration with the incumbent is welfare-reducing if s is sufficiently large.[29]

Proposition 5. *Data-driven vertical integration with the incumbent makes a consumer worse off than either no vertical integration or vertical integration with the entrant if the entrant's product quality is sufficiently higher than the incumbent's.*

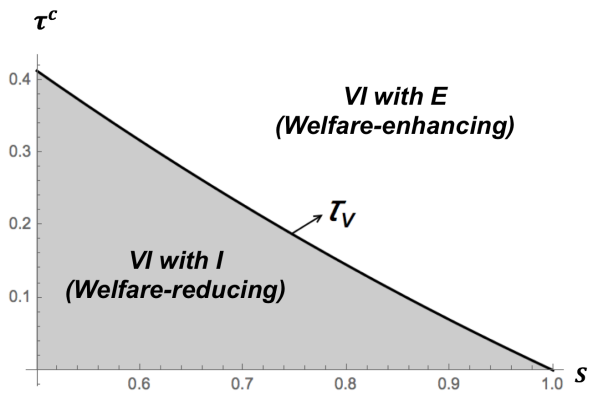

Figure 5: (s, τ^c) space when $\gamma = 2$

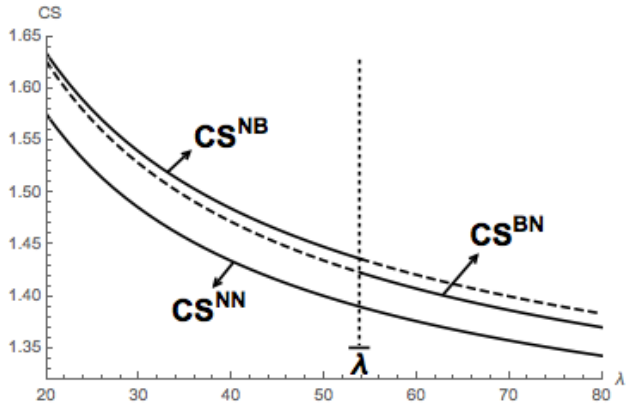

Figure 6: CS Comparison if $s > \bar{s}$ and $\gamma = 2$

The welfare analysis result implies that consumers' privacy concern, which determines the

[27] $\bar{\lambda} = \frac{(\gamma-1)\left(3\sqrt{\gamma^2((s-2)s+9)+2\gamma(s-9)+9}+\gamma(s+7)-7\right)}{2(\gamma(s-2)+2)^2}$, which can be derived from $\tau_V = \tau^c = \frac{\sqrt{4\lambda+1}+1}{2\lambda}$.

[28] Note that the threshold that leads to Proposition 5 is $s > \bar{s} \equiv \frac{\left(4\lambda+15\sqrt{4\lambda+1}+1\right)\left(\gamma^2-1\right)}{4\lambda\gamma(5\gamma-4)}$. The only difference is the threshold $\bar{s}$, which guarantees $SW^{NB} > SW^{BN}$: $s > \bar{s}^{SW} \equiv \frac{\left(20\lambda+3\sqrt{4\lambda+1}+5\right)\left(\gamma^2-1\right)}{4\lambda\gamma(7\gamma-2)}$ guarantees that $SW^{NB} > SW^{BN}$.

[29] Note that the consumer surplus level under (N, N), which is the no data selling regime that never arises in equilibrium, attains the lowest level.

aggregate amount of information availability, not only harms the entrant in a disproportionate way but may also adversely affect consumers themselves. If there is a high privacy concern, so that only limited information becomes available, the platform and the incumbent are likely to integrate and foreclose the entrant from data access. Data foreclosure obviously harms the entrant in terms of lower market share and profit, as I have shown in Proposition 4. Moreover, the welfare result shows that data foreclosure ultimately makes consumers worse off although they make individually rational information disclosure decisions.[30]

6 Empirical Evidence and Policy Implications

So far, I have shown that privacy concerns and the resulting decrease in available information not only harm the entrant but also make consumers worse off. In that sense, 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 cost, it would also help to increase the amount of information available. From the utility specification in (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, as I show in this section, 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. 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 socially desirable.

In this section, I focus on policies that decrease privacy nuisance cost, especially reputation-based remedies. To determine whether, in reality, consumers care about privacy when choosing a product and to explore how likely it is that a consumer's willingness to provide information is trust-based (related to the data collector's reputation), I analyze the mobile application

[30]Even if there is no vertical integration, a decreased amount of information results in the incumbent buying data as well, which makes the entrant worse off.

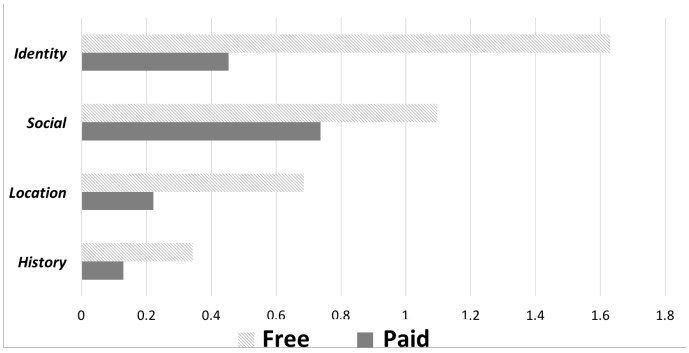

Figure 7: The most frequent permissions on average

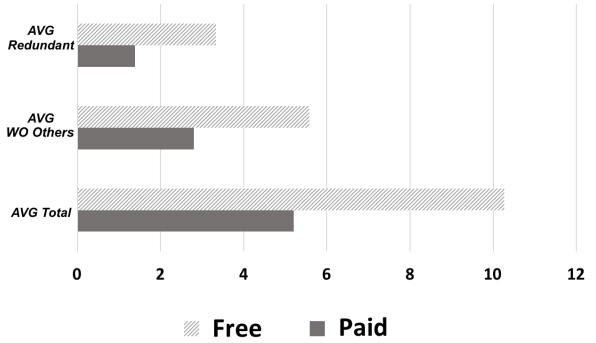

Figure 8: The number of privacy-sensitive permissions

ecosystem, using Kummer and Schulte (2017) as a key reference. Specifically, I used *Python* to scrape the necessary data from the Google Play Store, which is the official application (app) store for the Android operating system. Since users browse and download apps directly from this website, it provides all necessary app-specific information, such as price, category, and app size. Moreover, the Google Play Store provides information on which permissions each app requests, and thus, users can see those permissions before downloading apps. There are a number of different groups for each permission, such as Device and App History, Identity, Contacts, and Location. For example, *Google Photos* asks users for access to three pieces of information about Identity, three about Contacts, two about Location, and 27 different pieces of information about other groups. Thus, users can see the details regarding these permissions and decide whether to download an app.

More specifically, as privacy concerns are closely related to a user's trust in data collectors, this effect might be stronger for small app providers that lack a market reputation, while app providers that have an excellent reputation would not experience any negative effects from unnecessary permissions. Therefore, the focus here is to examine whether there is any asymmetric effect of privacy concerns with respect to a firm's size or reputation, even after controlling for app-specific characteristics and other relevant factors that affect app demand.

6.1 Data and Summary Statistics

I gathered the data from the end of October 2016 to the middle of January 2017 on a weekly basis, yielding a total sample size of 10,737.

First, a measure of app demand and the number of privacy-related permissions that each

app requests are necessary to analyze the asymmetric effect of privacy concerns on consumer demand. For app demand, I use the number of reviews for each app as a proxy demand measure because it represents at least the lower bound of demand, as some portion of customers who download each app write a review.[31] Regarding the number of permissions, Google provides 17 categories of permission groups, such as Device & App History, Identity, Contacts, and Location, from which each app developer can choose. Each app developer then might have several different permissions for each category. Among the 17 categories, "Other" represents manufacturer- or app-specific custom settings that include permissions that are relatively insignificant to privacy. Figure 7 shows that "Identity (Identity and Contact)", "Location", "Social (SMS and Phone)", and "Device & App History" are the most frequent critical permission groups requested by apps on average.[32]

Table 1: Summary statistics

Variable	**Mean**	**Std. Dev.**	**N**
ln_reviews	10.741	2.938	10,737
ln_App_Age	6.568	1.123	10,737
D_Price	0.421	0.494	10,737
Avg_rating	4.301	0.357	10,737
Top_Dev	0.481	0.5	10,737
Number_of_screenshots	13.179	6.447	10,737
Num_of_apps_per_dev	4.024	6.131	10,737
Per_Inapppurchase	0.391	0.491	10,737
Per_DeviceandApphistory	0.253	0.568	10,737
Per_Identity	0.562	0.793	10,737
Per_Contacts	0.593	0.865	10,737
Per_Calendar	0.084	0.373	10,737
Per_Location	0.495	0.820	10,737
Per_SMS	0.241	0.829	10,737
Per_Phone	0.723	1.017	10,737
Totalpermissions	15.649	10.585	10,737
Dum_Location	0.299	0.458	10,737
Dum_Social	0.570	0.495	10,737
Dum_Identity	0.509	0.5	10,737
Dum_Browsing	0.198	0.399	10,737

Furthermore, as Kummer and Schulte (2017) mention in their paper, the data show that there are some unnecessary permissions that do not affect app function. These redundant

[31] Kummer and Schulte (2017) also partly use the number of reviews as a demand measure.

[32] According to Sarma et al. (2012) and Olmstead and Atkinson (2015), those four permission groups are part of privacy-sensitive permissions or related to access to user information.

permissions might be used for monetizing purposes and therefore might increase consumer reluctance to download the app due to privacy concerns.[33] Moreover, as shown in Figure 8, free apps request more permissions than paid apps on average, which raises the concern that app developers have ulterior motives for providing free apps.

Finally, I use a set of app-specific characteristics as control variables to estimate app demand. These include price, app age since release date, average app rating, app size, a dummy for game apps, the number of screenshots on the app description page, the total number of distinct apps that each developer provides, and a dummy variable for top developer status. The top developer variable takes a value of one if the Google Play Store grants "Top Developer" status to the app. The selected set of variables is summarized in Table 1, where $Per\_$ denotes the number of permission-related variables. $Dum\_$ denotes a dummy variable for whether an app requests at least one permission related to Location, Social, Identity, or Browsing History. I use log-transformed variables for the number of reviews (as a demand measure) and for app age.

6.2 Empirical Model and Results

I estimate how the privacy-sensitive permissions affect app demand based on a pooled cross-sectional sample. The empirical model is as follows.

$$\ln\_\text{reviews}_i = \beta_0 + \beta_1 \text{Dum\_privacy}_i + \beta_2 \,\text{Top\_Dev}_i + \beta_3 \,\text{Dum\_privacy}_i \times \text{Top\_Dev}_i + \delta \ln\_\text{Price}_i + \boldsymbol{\xi} X_i + \epsilon_i, \quad (15)$$

where i denotes each app, Dum_privacy_i takes one if an app requests at least one redundant permission regarding Identity, Social, Location, or Browsing History, Top_Dev_i is a dummy variable for top developer status, β_3 is the coefficient of interest that identifies the interaction effect of the top developer and redundant privacy-sensitive permissions, and X_i is a set of app-specific characteristics used as control variables. The dependent variable, which is a proxy for app demand, is the log-transformed number of reviews for each app.

Table 2 reports the results. The first column does not include any app-specific characteristics as controls, while the remaining columns do include them. The second column does not take into account price endogeneity, whereas column (3) uses BLP type instruments as a rem-

[33] See Kummer and Schulte (2017) for a more detailed description of redundant permissions. For example, although a GPS/navigation app needs to access location information to function properly, information on a user's web browsing history would not be necessary.

edy for potential price endogeneity.[34] The problem with the cross-sectional analysis is that the treatment variable, Dum_privacy, could be endogenous and could be correlated with unobservables relegated to the error term: an app may ask for more permissions for better functionality, which might mean high quality. However, since the treatment variable Dum_privacy captures the existence of redundant privacy permissions, which do not affect apps' functionality, such concerns can be mitigated.

Table 2: Results from the Cross-Sectional Data

VARIABLES	(1) No Controls	(2) Controls	(3) IV
Dum_privacy_Top_Dev	0.670***	0.560***	0.544***
	(0.0979)	(0.0798)	(0.122)
Dum_privacy	0.0547	-0.182***	-0.177***
	(0.0636)	(0.0555)	(0.0622)
Top_Dev	0.909***	0.422***	0.460**
	(0.0863)	(0.0713)	(0.213)
Constant	11.47***	-1.248***	-1.301**
	(0.0599)	(0.433)	(0.517)
Controls	No	Yes	Yes
Time FE	Yes	Yes	Yes
Category FE	Yes	Yes	Yes
Observations	10,737	10,737	10,737
R-squared	0.495	0.692	0.690

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

As an app requests redundant privacy-sensitive permissions ($Dum\_privacy = 1$), lower demand is observed across all regressions, except for the first column, in which no controls are added. Interestingly, this effect is mitigated for apps launched by top developers, which have a stronger market reputation. For example, from the third column, having at least one privacy-sensitive permission for a non-top developer that has a lower reputation relative to top developers corresponds to a decrease in app demand of approximately 17.7%. However, for top developers, the same change is associated with a 36.7% increase in demand.[35] The estimation results show that there is an asymmetric reputation effect from privacy concerns with respect to a firm's status in the market. This empirical evidence corroborates the assumptions that are

[34] I use standard instruments including the characteristics of competing products. The instruments correct an upward bias of the price coefficient, thereby leading to more negative coefficients.

[35] See Table 3 for the full results from the cross-sectional data in Appendix B. For the robustness check, I use the pure cross-sectional data and the sample with twin apps only—as in Kummer and Schulte (2017)—and check the qualitatively same sign. The results are reported in Tables 4 and 5 in Appendix B.

imposed throughout the paper and supports a policy remedy that I suggest below.

As mentioned earlier, any remedy that makes more people willing to disclose information or be more privacy-insensitive, leading to integration with the entrant in equilibrium, is socially optimal. A privacy certification program represents one such remedy given the empirical evidence, which shows that a data collector's reputation significantly reduces consumers' privacy concerns and ultimately increases their willingness to disclose personal information. 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. As the empirical evidence has shown, privacy is a trust-based matter in that consumers care about who asks for their personal information. Because willingness to disclose information depends on a firm's reputation, this remedy is likely to be effective. This policy suggestion is consistent with the policy implications in some previous literature (e.g. Campbell et al. (2015), Kummer and Schulte (2017)).[36]

7 Extensions

7.1 Consumers with Foresight

Thus far, I have assumed that a consumer only takes into account immediate benefits when making information disclosure decisions but does not consider any potential future benefits arising from better targeted ads. Though this assumption is reasonable for the case of a social media platform, 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.[37] To capture this effect, I consider the total net utility that

[36]The Cyber Shield Act of 2017, which is a recently introduced bill, is in the same vein.

[37]Using cable TV operators or Internet service provider(ISPs) as examples of the platform, whether to disclose personal information means opting in or out of targeted ad programs. Even in this example, there can be an immediate benefit from disclosing information other than targeting benefits—e.g., AT&T used to offer a monthly discount in exchange for being allowed to collect personal information. Without such a price discount,

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.[38] The aggregate utility specification is as follows.

$$u_{ij}^{\text{Foresight}} = V + \theta_i s_j - P_j - \mathbb{1}_{\{i\in\mathcal{ND}\}}\Big(\frac{1}{\gamma_j D_j}\Big) - \mathbb{1}_{\{i\in\mathcal{D}\}}\Big(\frac{\psi(\tau_i)}{r}\Big). \tag{16}$$

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

$$\begin{aligned} P(i \in \mathcal{D}) = \tau^c &= \theta^c P(i \in \mathcal{D} | i \in \mathcal{L}) + (1-\theta^c) P(i \in \mathcal{D} | i \in \mathcal{H}) \\ &= \theta^c F\Big(\psi^{-1}\big(\frac{r}{\gamma D_I}\big)\Big) + (1-\theta^c) F\Big(\psi^{-1}\big(\frac{r}{D_E}\big)\Big), \end{aligned} \tag{17}$$

where θ^c, which is a function of τ^c, is given as in Equation (6). Because consumers are assumed to be sufficiently sophisticated, the disclosure probability now depends on each seller j's targeting effectiveness, which implies that $P(i \in \mathcal{D})$ can differ depending on the information acquisition equilibrium. By comparing the right-hand side of Equation (17), I can rank different τ^c levels depending on each data acquisition equilibrium. Given that D_j can be either one or $1+\tau^c$, it is easy to show that τ^c_{BB} is the lowest, whereas τ^c_{NN} is the highest, where the subscript denotes the data acquisition equilibrium. In other words, knowing that a seller acquires personal data, a consumer becomes reluctant to disclose information. The relative size of τ^c_{NB} and τ^c_{BN} depends on γ and s. Put simply, $\tau^c_{NB} > \tau^c_{BN}$ is more likely to hold as γ increases given s or as s 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 E's product quality advantage shrinks. Since $X_E = 1 - \theta^c$ decreases as s decreases or γ increases, the effect of E's data-buying decision 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 6. *If a consumer has perfect foresight, the equilibrium disclosure probability 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 size of γ and s.*

the immediate benefits in this example might not be obvious compared to the social media platform examples. Thus, in this case, a consumer is likely to take into account future potential targeting benefits when making an information disclosure decision.

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

Next, to check the robustness of the main findings, I compare consumer surplus levels under numerical examples. I focus on the relative size of the consumer surplus under the asymmetric data acquisition cases where only one seller buys data. I find that as data become more available, consumer surplus increases. For example, if s is sufficiently small and γ is large so that $\tau_{BN}^c < \tau_{NB}^c$ holds, the ranking of consumer surplus levels is $CS^{BN} < CS^{NB}$: I's monopoly of the data makes consumers worse off. Intuitively, if τ_{NB}^c is sufficiently large, E is able to make better targeted ads using a more extensive amount of detailed data, thereby leading to a lower mismatch cost. Though s is small, which implies that a consumer cannot enjoy more utility through product quality, a consumer is overall better off because of the sufficiently low mismatch cost stemming from greater information availability: regarding the increasing consumer surplus, the effect of the low mismatch cost dominates that of a high-quality product.

Though $CS^{BN} < CS^{NB}$ can still hold under this extension, the driving force is different. In the main model, $CS^{BN} < CS^{NB}$ holds if E's product quality advantage is sufficiently great, i.e., $s > \bar{s}$. Here is the intuition. Given that the amount of data that can be used for targeting is fixed, a higher s generates a consumer benefit, as consumers enjoy high-quality products. Also, if E buys data, thereby sending more relevant ads, a consumer can also save through a lower mismatch cost. Therefore, if high-quality product seller E buys data and is thus able to generate better targeted ads, the additional utility outweighs the high price that a consumer will be paying E, thereby leading to greater consumer surplus. However, in the model with sophisticated consumers, a higher s is likely to lead to E having a lesser amount of data available for targeting, i.e., $\tau_{BN}^c > \tau_{NB}^c$. Then, even if a consumer enjoys a high-quality product, he will face a higher mismatch cost from E due to the lack of data, which might result in $CS^{BN} > CS^{NB}$: the additional utility from a high-quality product that also incurs a high mismatch cost is comparably lower than that from a low-quality product accompanied by a lower mismatch cost. The key point here is that if s is not that high, say $s < \bar{\bar{s}}$, we can obtain $CS^{BN} < CS^{NB}$ as in the main model: if s is not that high, τ_{NB}^c will decrease only a little bit, which means that $\tau_{BN}^c < \tau_{NB}^c$ still holds. Under $\tau_{BN}^c < \tau_{NB}^c$, E's exclusive use of data leads to a greater consumer surplus level than in the other case. In other words, for $\bar{s} < s < \bar{\bar{s}}$, a consumer is better off if the platform and E integrate and use data exclusively, which has implications similar to those identified in the main model.[39]

[39]See Appendix A for detailed numerical exercise.

Lastly, comparative statics can be calculated to see how the implicitly determined equilibrium τ^c is affected by exogenous parameters, such as γ and s. By applying the implicit function theorem to Equation (17), it is easy to see that $\frac{d\tau^c}{d\gamma} < 0$ while $\frac{d\tau^c}{ds} > 0$ for any τ^c from each data acquisition equilibrium. In other words, more consumers are willing to disclose personal data as I's initial targeting technology becomes less effective or E's product quality improves. The intuition is that as γ becomes close to one, both sellers provide less relevant targeted ads because they lack personal data. 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. Also, if s increases, the products provided by the two sellers become more differentiated, which leads to soft 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 7 summarizes the finding.

Proposition 7. *Consumers are more willing to disclose personal data as I's initial targeting technology becomes less effective or E's product quality improves.*

The findings in this subsection provide an important policy implication: if a consumer has perfect foresight and thus takes into consideration any potential benefits from better targeted ads when determining his or her level of information disclosure, any policy that encourages market entrants to upgrade their product quality would encourage consumers to voluntarily provide their personal information, thereby leading to socially optimal market outcomes—integration with E and the (N, B) data acquisition equilibrium, for example.

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: $\overline{FC}_{BB} = \pi_E^{BB}$; $\overline{FC}_{NB} = \pi_E^{NB}$; $\overline{FC}_{BN} = \pi_E^{BN}$. In other words, if the fixed cost is higher than the equilibrium profit, the entrant stays out of the market. For simplicity, I focus

on a relatively small range of γ, which leads to $\overline{FC}_{BN} < \overline{FC}_{BB} < \overline{FC}_{NB}$.[40] Then, there are four 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 $\overline{FC}_{BB} < FC < \overline{FC}_{NB}$, (3) staying out under (B,N) but entering under (B,B) or (N,B) if $\overline{FC}_{BN} < FC < \overline{FC}_{BB}$, and (4) entering in any case if $FC < \overline{FC}_{BN}$. Focusing on this parametric space, I analyze how the entrant's entry decision affects the market.

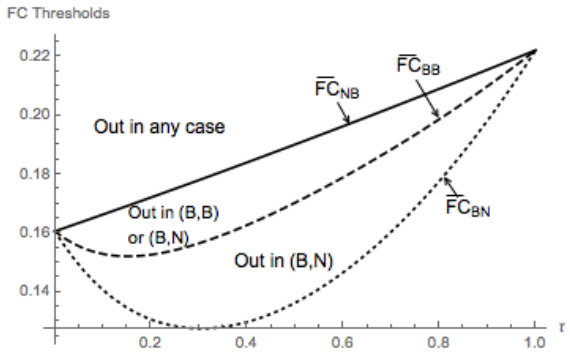

Figure 9: Entrant's entry decision depending on the fixed cost thresholds

As in Figure 9, (B,N) (or (N,B)) is the most (or least) likely to lead to entry foreclosure whereas (B,B) is somewhat in the middle. That is, vertical integration with the incumbent and the resulting data foreclosure, which means (B,N), is highly likely to lead the entrant to stay out of the market due to lower profit when entering. When both sellers buy data, which means (B,B), the entrant is more likely to enter than in the case of (B,N). Obviously, when the entrant is able to use data exclusively, which means (N,B), it is the most likely to enter the market. Proposition 8 summarizes this finding.

Proposition 8. *When the entrant faces a fixed cost of entry, it is least likely to enter the market if vertical integration with the incumbent and the consequential 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. The monopoly market share is determined by $P(\theta > \frac{\mathbb{1}_{\{i\in\mathcal{ND}\}}(\frac{1}{\gamma D_I})+P_I-V}{s_I})$, which means that $X^{Mono}_{\mathcal{ND}} = 1-\frac{\frac{1}{\gamma D_I}+P_I-V}{s_I}$ and $X^{Mono}_{\mathcal{D}} = 1-\frac{P_I-V}{s_I}$ from the utility specification as in (2). Given the weighted monopoly market share, the incumbent maximizes its profit by charging the monopoly

[40] As γ becomes larger, $\overline{FC}_{NB} < \overline{FC}_{BB}$. However, $\overline{FC}_{BN}$ is always the lowest, which implies the same messages as in Proposition 8.

price at $P_I^{Mono} = \frac{\gamma D_I(s_I+V)+\tau^c-1}{2\gamma D_I}$, which leads to $X_I^{Mono} = \frac{\gamma D_I(s_I+V)+\tau^c-1}{2\gamma D_I}$. Specifically, the equilibrium profit levels under buying and not buying data from the platform can be derived as follows.

$$\pi_B^{Mono} = \frac{(\gamma(\tau^c+1)(s_I+V)+\tau^c-1)^2}{4\gamma^2(\tau^c+1)^2} - C^{Mono} \times \tau^c; \quad \pi_N^{Mono} = \frac{(\gamma(s_I+V)+\tau^c-1)^2}{4\gamma^2}, \quad (18)$$

where the subscripts B and N denote Buy and Not buy, respectively. Comparing two profit levels, the platform optimally sets the data price at $C^{Mono} = -\frac{(\tau^c-1)\left(2\gamma\tau^c(s_I+V)+2\gamma(s_I+V)+(\tau^c)^2+\tau^c-2\right)}{4\gamma^2(\tau^c+1)^2}$, which is derived from $\pi_B^{Mono} - \pi_N^{Mono}$, meaning that $\pi_p^{Mono} = C^{Mono} \times \tau^c$. Under the monopoly, π_B^{Mono}, π_N^{Mono}, and π_p^{Mono} are all greater than the corresponding profit levels under all possible data acquisition equilibria without considering entry. However, a consumer obviously becomes worse off under the monopoly.

Proposition 9. *If the entrant does not enter the market due to lower profits, the incumbent monopolizes the product market. The monopoly leads to greater profits for the incumbent and the platform in most cases, but the consumer becomes 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.

7.3 Data Price Discrimination

So far, I have assumed that the platform charges a unit data price for all sellers. However, since the platform knows that sellers I and E have asymmetric targeting technology, it might be able to engage in data price discrimination: it might want to charge a higher data price for any seller whose willingness to pay is higher. I investigate how the platform's data price discrimination affects the market's competitive structure in this subsection.

Assuming that the platform extracts all rents from sellers in the form of data price C, data price discrimination can emerge only if both sellers buy data. Given that rivals also buy data, I and E want to buy as well if $C < \bar{C}_I$ and $C < \bar{C}_E$, respectively, where $\bar{C}_I < \bar{C}_E$. In other words, E, whose initial targeting technology is worse, has higher willingness to pay for data to overcome the disadvantage. Knowing that, the platform can charge different prices to both sellers: $C = \bar{C}_I$ to I and $C = \bar{C}_E$ to E. Then, the platform's profit is $\pi_p^{PD,BB} = \bar{C}_I + \bar{C}_E$, where

the superscript PD denotes price discrimination. By the same logic as in subsection 3.1, the platform either chooses to sell data to both sellers with different price levels or to E only by charging $\bar{\bar{C}}_E$. The difference in profits is as follows.

$$\pi_p^{PD,BB} - \pi_p^{NB} = \frac{(1-\tau^c)\tau^c\left(2\gamma\left(s(\tau^c+1)+(\tau^c-1)^2\right)+(\tau^c)^2+\tau^c-2\right)}{9\gamma^2 s(\tau^c+1)^2} \tag{19}$$

From Equation (19), if $2\gamma\left(s(\tau^c+1)+(\tau^c-1)^2\right)+(\tau^c)^2+\tau^c-2>0$, the platform wants to engage in price discrimination and sell data to both sellers. It is easy to show that the left-hand side is increasing in γ and s, respectively, and its local minimum is attained at $\tau^c = \frac{-2\gamma(s-2)-1}{4\gamma+2}$. Thus, the minimum value is $\frac{2}{3}$, which can be attained from $\tau^c = \frac{1}{3}$ with $s = \frac{1}{2}$ and $\gamma = 1$. This implies that $2\gamma\left(s(\tau^c+1)+(\tau^c-1)^2\right)+(\tau^c)^2+\tau^c-2>\frac{2}{3}$, which leads to $\pi_p^{PD,BB} > \pi_p^{NB}$. Consequently, if data price discrimination is allowed, the platform always sells data to both sellers, thereby making the entrant worse off in terms of lower market share and revenue. Moreover, the entrant now faces much lower profit than before due to a higher data price.

Proposition 10. *If the platform engages in data price discrimination, (B,B) emerges in data acquisition equilibrium and the platform charges $C = \bar{C}_I$ to I and $C = \bar{C}_E$ to E. Data price discrimination always makes the entrant worse off in terms of profit.*

8 Concluding Remarks

In this paper, I analyze how consumers' privacy concerns affect market competition when each seller attracts potential customers by creating targeted ads based on personal information obtained from a platform. In particular, I focus on the relationship between privacy sensitivity and the data-sharing aspects of vertical integration between the platform and the seller. I show that the platform and the incumbent with better initial targeting technology are more likely to vertically integrate as the number of privacy-sensitive consumers increases. The integrated firm always wants to prevent access to the data by the unaffiliated entrant, thereby adversely affecting the entrant in terms of smaller market share and lower profits. Therefore, the entrant that needs consumer data to overcome its initial disadvantage in targeting technology is disproportionately affected by a lack of access to data arising from greater privacy concerns. Moreover, this process eventually leads to lower consumer surplus and lower total social welfare

due to the lack of competition arising from data foreclosure.

The extended models also investigate aggravating factors: data price discrimination and an entrant's entry decision due to entry cost make the entrant and consumers worse off. Consequently, individually rational decisions on information disclosure, which depend on each consumer's privacy sensitivity, might not be socially optimal when aggregated.

Therefore, any policy that encourages integration with the entrant would be beneficial. In this sense, any remedy that makes more people willing to disclose information or reduces privacy concerns is socially optimal because it leads to integration with the entrant in equilibrium. One specific remedy I propose in the paper is a privacy certification program. It is worth mentioning that if consumers have perfect foresight and thus take into account future targeting benefits when making information disclosure decisions, policy makers need to consider that the willingness to disclose information then depends on targeting technology and product quality.

There is another policy implication regarding data foreclosure practices in vertical integration. Since consumer data have become key to sellers' business performance, data foreclosure is directly related to the competitive structure. This anti-competitive effect is more apparent in the case of integration with an incumbent. To mitigate this detrimental effect, regulators might force the integrated firm to share customer data with its rivals by asking that the price of data be set within a reasonable range to guarantee the efficient level of data availability.[41]

Broadly speaking, this paper emphasizes that lower privacy concerns lead to greater information availability, which, in turn, reduces barriers to entry to the marketplace. This main implication can be applied to a much broader but analogous competitive setup and supported by similar empirical research: e.g., Petrova et al. (2017) empirically show that more information channels, such as Twitter, benefit new politicians more than incumbents by reducing the gap in political donation opportunities between new and experienced politicians. Although privacy and information availability are potentially important areas to investigate in order to encourage competition, regulators have not yet established any concrete antitrust standards regarding this complex interaction. In that sense, the findings from my model help to understand the relevant issues and propose various policy implications regarding privacy protection and data-driven vertical integration.

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

References

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

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

[3] Bergemann, D., Bonatti, A., "Selling Cookies", *American Economic Journal: Microeconomics*, Vol.7, No.3, (2015), pp 259-94

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

[5] 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

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

[7] Choi, J.P., Jeon, D.S., Kim, B.C., "Privacy and Personal Data Collection with Information Externalities", *Working Paper*, (2016)

[8] 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

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

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

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

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

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

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

[15] Goldfarb, A., Tucker, C., "Privacy and Innovation", *Innovation Policy and the Economy*, Vol.12, No.1, (2012), pp 65-90

[16] Kim, J.H., Wagman, L, Wickelgren, A. L., "The Impact of Access to Consumer Data on the Competitive Effects of Horizontal Mergers", (2016) *Available at SSRN: https://ssrn.com/abstract=2728378*

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

[18] Kummer, M.E., Schulte, P., "When Private Information Settles the Bill: Money and Privacy in Google's Market for Smartphone Applications", *ZEW - Centre for European Economic Research Discussion Paper*, No.16-031, (2017)

[19] Montes, R., Sand-Zantman, W., Valletti, T., "The Value of Personal Information in Markets with Endogenous Privacy", *CEIS Working Paper*, No. 352, (2017)

[20] Norman, G., Pepall, L., Richards, D., Tan, L., "Competition and Consumer Data: The good, the bad, and the ugly", *Research in Economics*, No. 70, (2016), pp 752-765

[21] Olmstead, K., Atkinson, M., "Apps Permissions in the Google Play Store", *Pew Research Center Report*, (2015)

[22] 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)

[23] Posner, R., "The Economics of Privacy", *American Economic Review*, Vol.71(2), (1981), pp 405-409

[24] Shy, O., Stenbacka, R., "Customer Privacy and Competition", *Journal of Economics and Management Strategy*, Vol.25, Issue.3, (2016), pp 539-62

[25] Sarma, B.P., Li, N., Gates, C., Potharaju, R., Nita-Rotaru, C., Molloy, I., "Android Permissions: A Perspective Combining Risks and Benefits", *Proceedings of the 17th ACM symposium on Access Control Models and Technologies*, ACM (2012), pp. 13-22

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

[27] 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

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

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

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

Appendix A. Further Discussions 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, $\gamma_j D_j$. 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 D_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 $\gamma_j D_j$, should have an additive separable form, such as $\gamma_j + D_j$. The main results from the main model still hold under this modification. In other words, there exists a threshold on τ^c, say τ_V^{AS}, where the superscript AS denotes additive separability, such that if $\tau^c < \tau_V^{AS}$, ($\tau^c > \tau_V^{AS}$) leads to the integration with I (E) and data foreclosure. The welfare comparison still implies that the integration with I and the consequential data foreclosure arising from greater privacy concerns is welfare-reducing. The detailed proof is omitted.

Second, if the model is changed to the symmetric setup with $\gamma = 1$, it becomes a typical vertical differentiation model. By solving the model in the same way, it is easy to see that the platform always prefers vertical integration with a high-quality seller and the foreclosure of data access for a low-quality seller. Thus, the symmetric model generates results that do not depend on privacy concerns and are thus not of interest in this paper.

The Case of $\mathbf{s_I > s_E}$. The main model analyzes the case of $s_E > s_I$, which means that E as a specialist can provide a higher-quality product than I as a generalist. This specification closely resembles the setup in Campbell et al. (2015). However, it is worth checking the robustness of the main results to the opposite case.

By a similar logic to that used in the main analysis, the indifference condition is $\theta'^c_{\mathcal{ND}} = \frac{P_I - P_E + (\frac{1}{\gamma D_I} - \frac{1}{D_E})}{s_I - s_E}$ for $i \in \mathcal{ND}$ and $\theta'^c_{\mathcal{D}} = \frac{P_I - P_E}{s_I - s_E}$ for $i \in \mathcal{D}$. The weighted indifference condition can be rewritten in a simple way as follows.

$$\theta'^c = \frac{P_I - P_E + (1 - \tau^c)\Delta'}{s'}, \tag{20}$$

where $\Delta' = (\frac{1}{\gamma D_I} - \frac{1}{D_E})$ and $s' = s_I - s_E$. The market share for each seller is given by $X_I = 1 - \theta'^c$ and $X_E = \theta'^c$ under $s_I > s_E$. In this case, the ranking of data price C is $\bar{C}_E < \bar{\bar{C}}_E < \bar{C}_I < \bar{\bar{C}}_I$. This leads to the platform either setting a lower price at $C = \bar{C}_E$ to obtain both sellers or setting a higher price $C = \bar{\bar{C}}_I$ to obtain only seller I. As in Proposition 2, there exists a threshold on τ^c denoted as $\tau'_{NV} \equiv \frac{\sqrt{4(\gamma-1)\gamma(9\gamma(\gamma+1)+4(\gamma-1)\gamma s^2-4\gamma(\gamma+4)s+2s)+9}-2\gamma(\gamma+2(\gamma-1)s-2)+1}{4\gamma(\gamma+1)-2}$ below which the platform sets a higher price, which means (B, N) in equilibrium, and above which the platform sets a lower price, which leads to (B, B). In addition, the platform and I want to vertically integrate and foreclose the unaffiliated entrant from data access in equilibrium without any conditions. Thus, (B, N) is the only data acquisition equilibrium under vertical integration. Consumer surplus under no vertical integration, which involves (B, B), reaches a higher level than that under vertical integration, which means (B, N). If E stays out of the market due to data foreclosure, a monopoly makes consumers worse off, as analyzed in Section 7.2. This suggests that the main implications for the negative impact of privacy concerns on market competition and consumers still hold.

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

Proof of Proposition 2. The corresponding threshold on τ^c is derived from $2\bar{C}_I\tau^c = \pi_p^{BB} = \pi_p^{NB} = \bar{\bar{C}}_E\tau^c$, which leads to $\tau_{NV} = \frac{\sqrt{9(\gamma^2-2)^2+16(\gamma-1)^2\gamma^2s^2-8(\gamma((\gamma-9)\gamma+6)+2)\gamma s}-\gamma(\gamma+4(\gamma-1)s+4)+2}{2(\gamma-2)\gamma-4}$. The corresponding product price, P_j, and each seller's market share, X_j, can be derived by substituting the equilibrium data acquisition status into Equations (7). □

Proof of Corollary 1. The proof is in the paper.

Proof of Proposition 3. Most steps in the proof are described in the paper.

Proof of Proposition 4. First, I compare E's profit under (B, B) in the no vertical integration case to that under the vertical integration with I case—which means (B, N)—. Then, $\pi_E^{BB} - \pi_{VI,F}^E = \frac{(1-\tau^c)\tau^c\left(\left(\gamma^2-1\right)\left((\tau^c)^2+\tau^c-2\right)+2\gamma(2\gamma-1)s(\tau^c+1)\right)}{9\gamma^2s(\tau^c+1)^2}$. To show that this is always positive, all I need to show is that $((\tau^c)^2 + \tau^c - 2) + \frac{2\gamma(2\gamma-1)s(\tau^c+1)}{(\gamma^2-1)} > 0$. The second term is always positive under the assumptions of $\gamma > 1$ and $s \in [0.5, 1]$, but the first term is always negative for $\tau^c \in [0, 1]$. The second positive term attains a minimum of $4(1 + \tau^c)$ at $s = \frac{1}{2}$ and $\gamma \to \infty$, which implies that the equation attains a minimum of $(\tau^c)^2 + \tau^c - 2 + 4(1 + \tau^c) = (\tau^c)^2 + 5\tau^c + 2$, which is always positive given τ^c. Likewise, the comparison between (N, B) in the no vertical integration case and in the vertical integra-

tion case is as follows. $\pi_E^{NB} - \pi_{VI,F}^E = \frac{(1-\tau^c)\tau^c(2\gamma(\tau^c+1)(2s+\tau^c-1)+(1-\tau^c)(\tau^c+2))}{9\gamma^2 s(\tau^c+1)^2}$, which is always positive. □

Proof of Proposition 5. For simplicity, I normalize s_I to zero, which implies that $s = s_E$. The difference between the consumer surplus under (N, B) and that under (B, N) is as follows.

$$CS^{NB} - CS^{BN} = \frac{-\lambda\left(\sqrt{4\lambda+1}+12\right)\left(\gamma^2-1\right)+\left(11\sqrt{4\lambda+1}-3\right)\left(\gamma^2-1\right)+\lambda\left(\sqrt{4\lambda+1}-3\right)\gamma(5\gamma-4)s}{18\lambda^2\gamma^2 s} \tag{21}$$

Given λ and γ, it is easy to show that $CS^{NB} > CS^{BN}$ if s is larger than a certain threshold. The threshold denoted as $\bar{s}$ can be obtained from $CS^{NB} - CS^{BN} = 0$, which is $\bar{s} = \frac{\left(4\lambda+15\sqrt{4\lambda+1}+1\right)\left(\gamma^2-1\right)}{4\lambda\gamma(5\gamma-4)}$. □

Proof of Proposition 6. The right-hand side of Equation (17) given each data acquisition case is as follows.

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

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

The equilibrium τ^c under each data acquisition case is the fixed point satisfying Equation (17). First, I show that the RHS under the four data acquisition cases can be ranked. Given that $\gamma > 1$ and $s > 1/2$, $\theta_{NB}^c < \theta_{BB}^c < \theta_{NN}^c < \theta_{BN}^c$. 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 enough to show that $\theta_{NB}^c F\Big(\psi^{-1}\big(\frac{r}{\gamma}\big)\Big) > \theta_{BB}^c F\Big(\psi^{-1}\big(\frac{r}{\gamma(1+\tau^c)}\big)\Big)$. 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 is greater than the same effect on market share 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 s. The relative size of RHS_{BN} and RHS_{NB} depends on the size of γ and s, 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 used by 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 a unique such τ^c_{BB}. Suppose 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 $\forall \tau^c \in [\tau^c_{NN}, 1]$. Since $G_{NN}(\tau^c_{BB}) < 0$, $\tau^c_{BB} < \tau^c_{NN}$. The remaining cases can be proved in the same way. □

Numerical Exercise for Robustness Check in Section 7.1. From the utility specification as in Equation (16), consumer surplus can be obtained in the following way.

$$\begin{aligned} CS^{\text{Foresight}} = &\int_0^{\theta^c_{\mathcal{D}}} \int_0^{\tau^c} \left(V + \theta s_I - P_I - \frac{\psi(x)}{r}\right) dF(x) d\theta + \int_{\theta^c_{\mathcal{D}}}^1 \int_0^{\tau^c} \left(V + \theta s_E - P_E - \frac{\psi(x)}{r}\right) dF(x) d\theta \\ &+ \int_0^{\theta^c_{\mathcal{N}\mathcal{D}}} \int_{\tau^c}^1 \left(V + \theta s_I - P_I - \frac{1}{\gamma D_I}\right) dF(x) d\theta + \int_{\theta^c_{\mathcal{N}\mathcal{D}}}^1 \int_{\tau^c}^1 \left(V + \theta s_E - P_E - \frac{1}{D_E}\right) dF(x) d\theta. \end{aligned} \tag{22}$$

I make parametric and numerical assumptions: F is a uniform distribution ($\tau_i \sim U[0,1]$), $V = 2$, $\psi(\tau_i) = \lambda \tau_i$, and $r = 1$. For simplicity, I compare two cases where (1) $\tau^c_{BN} < \tau^c_{NB}$ under $s = \frac{1}{2}$, $\gamma = 2$, (2) $\tau^c_{BN} < \tau^c_{NB}$ under $s = \frac{3}{5}$, $\gamma = 2$, and (3) $\tau^c_{BN} > \tau^c_{NB}$ under $s = 1$, $\gamma = \frac{3}{2}$. The comparison is shown in the tables below.

(1) $s = \frac{1}{2}$, $\gamma = 2$ ($\tau^c_{BN} < \tau^c_{NB}$)

		(B,N)	(N,B)
$\lambda = 4$	τ^c	0.38	0.39
	CS	1.78	1.80
$\lambda = 5$	τ^c	0.33	0.35
	CS	1.75	1.78
$\lambda = 6$	τ^c	0.31	0.32
	CS	1.74	1.76
$\lambda = 10$	τ^c	0.24	0.25
	CS	1.69	1.71

(2) $s = \frac{3}{5}$, $\gamma = 2$ ($\tau^c_{BN} < \tau^c_{NB}$)

		(B,N)	(N,B)
$\lambda = 4$	τ^c	0.38	0.39
	CS	1.76	1.78
$\lambda = 5$	τ^c	0.34	0.35
	CS	1.75	1.77
$\lambda = 6$	τ^c	0.31	0.32
	CS	1.75	1.76
$\lambda = 10$	τ^c	0.24	0.25
	CS	1.70	1.72

(3) $s = 1$, $\gamma = \frac{3}{2}$ ($\tau^c_{BN} > \tau^c_{NB}$)

		(B,N)	(N,B)
$\lambda = 4$	τ^c	0.43	0.41
	CS	1.69	1.68
$\lambda = 5$	τ^c	0.39	0.37
	CS	1.67	1.65
$\lambda = 6$	τ^c	0.35	0.34
	CS	1.65	1.63
$\lambda = 10$	τ^c	0.27	0.26
	CS	1.59	1.58

Proof of Proposition 7. Since 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 (17), let $G = \tau^c - \theta^c F\Big(\psi^{-1}\big(\frac{r}{\gamma D_I}\big)\Big) + (1-\theta^c)F\Big(\psi^{-1}\big(\frac{r}{D_E}\big)\Big)$. To see the effect of γ on the equilibrium τ^c, I need to show the sign of $\frac{d\tau^c}{d\gamma} = -\big(\frac{\partial G}{\partial \gamma}\big)/\big(\frac{\partial G}{\partial \tau^c}\big)$. Similarly, for the effect of s on τ^c, I need to calculate $\frac{d\tau^c}{ds} = -\big(\frac{\partial G}{\partial s}\big)/\big(\frac{\partial G}{\partial \tau^c}\big)$. After some algebra, it can be shown that $\frac{d\tau^c}{ds} > 0$ and $\frac{d\tau^c}{d\gamma} < 0$ if $\gamma > \max\{1+\tau^c, \frac{2}{(1+\tau^c)^2}\}$. The detailed proof is omitted. □

Proof of Proposition 8. By the proof of Proposition 4, $\overline{FC}_{BN} < \min\{\overline{FC}_{BB}, \overline{FC}_{NB}\}$. □

Proof of Proposition 9. For simplicity, I assume $V = 2$. First, I compare the incumbent's monopoly profits under buying or not buying data to those under the main model in which the entrant does not stay out of the market. Since $\pi_I^{BN} > \pi_I^{BB}$, it is enough to show that $\pi_B^{Mono} = \pi_N^{Mono} > \pi_I^{BN}$: $\pi_B^{Mono} = \pi_N^{Mono}$ because the platform extracts the rents in the form of the data price. Since $\pi_I^{Mono} - \pi_I^{BN} = \frac{9s(\tau+1)^2((2+s_I)\gamma+\tau-1)^2 - 4(\gamma(\tau+1)(s-\tau+1)+\tau-1)^2}{36\gamma^2 s(\tau+1)^2}$, all I need to show is that the numerator is positive. If its minimum is positive, the proof is complete. Given that the numerator is increasing in both of γ and s, the minimum is $\frac{1}{2}(\tau^c(\tau^c(\tau^c(\tau^c+44)+44)+40)+7)$ at $\gamma = 1$ and $s = \frac{1}{2}$, which is always positive.

Next, I compare the platform's profits under the monopoly and duopoly cases. For each data acquisition case under a duopoly, $\pi_p^{Mono} - \pi_p^{BB} = -\frac{(\tau^c-1)\tau^c\big(8(\tau^c-1)(2\gamma-\tau^c-2)+2(10+9s_I)\gamma s(\tau^c+1)+9s\big((\tau^c)^2+\tau^c-2\big)\big)}{36\gamma^2 s(\tau^c+1)^2}$ and $\pi_p^{Mono} - \pi_p^{NB} = \frac{(\tau^c-1)\tau^c\big(4\gamma(\tau^c-1)(\gamma(\tau^c+2)-2(\tau^c+1))+s\big(16\gamma^2(\tau^c+1)-18(2+s_I)\gamma(\tau^c+1)-9\big((\tau^c)^2+\tau^c-2\big)\big)\big)}{36\gamma^2 s(\tau^c+1)^2}$. To show that $\pi_p^{Mono} - \pi_p^{BB} > 0$, it is enough to show that $8(\tau^c-1)(2\gamma-\tau^c-2)+2(10+9s_I)\gamma s(\tau^c+1)+9s\left((\tau^c)^2+\tau^c-2\right) > 0$. Given that the right-hand side of the equation is increasing in s, the minimum is $\gamma(9s_I(\tau^c+1)+26\tau^c-6) - \frac{7}{2}\left((\tau^c)^2+\tau^c-2\right)$ at $s = \frac{1}{2}$. The minimized value is greater than zero if τ^c is larger than approximately 0.1, which implies that $\pi_p^{Mono} - \pi_p^{BB} > 0$ holds in most cases. Using similar logic, I can also show that $\pi_p^{Mono} - \pi_p^{NB} > 0$.

Lastly, consumer surplus under a monopoly is given as follows.

$$CS^{Mono} = \int_0^{\tau^c}\Big(v(\tau^c) - \frac{\psi(x)}{r}\Big)dF(x) + \tau^c\left[\int_{\theta^c_{\mathcal{D}\_Mono}}^1 \Big(V+\theta s_I - P_I\Big)d\theta\right] + (1-\tau^c)\left[\int_{\theta^c_{\mathcal{ND}\_Mono}}^1 \Big(V+\theta s_I - P_I - \frac{1}{\gamma D_I}\Big)d\theta\right], \quad (23)$$

where $\theta^c_{\mathcal{D}\_Mono} = \frac{14\sqrt{4\lambda+1}+\lambda\big(18\lambda+29\sqrt{4\lambda+1}-9\big)+2}{24\lambda^2}$. Under the numerical examples of $V = 2, \gamma = 2, s = \frac{4}{5}$, and $s_I = \frac{1}{2}$ (which means that $s_E = 1.3$), $CS_B^{Mono} = \frac{14\sqrt{4\lambda+1}+\lambda\big(18\lambda+29\sqrt{4\lambda+1}-9\big)+2}{24\lambda^2}$ and $CS_N^{Mono} = \frac{23\big(\sqrt{4\lambda+1}+1\big)+2\lambda\big(36\lambda+34\sqrt{4\lambda+1}+57\big)}{96\lambda^2}$ where the subscripts B and N denote Buy and Not buy data. Given that CS^{BN} is always lower than CS^{NB} or CS^{BB} as in Figure 6, it

is enough to show that $CS_B^{Mono} < CS^{BN}$ since $CS_N^{Mono} < CS_B^{Mono}$. After some algebra, $CS^{BN} - CS_B^{Mono} = \frac{45\lambda\left(2\lambda-\sqrt{4\lambda+1}+5\right)-2\left(27\sqrt{4\lambda+1}+5\right)}{144\lambda^2}$, which attains the local minimum of 0.553 at $\lambda \approx 15.64$. Since the minimum value is still positive, any consumer surplus levels under entry are greater than those under a monopoly. The result still holds even under all different sets of parametric space. The detailed steps are omitted and can be provided upon request. □

Proof of Proposition 10. The proof is in the paper.

Appendix B.

Full Results from the Cross-sectional Data. Table 3 reports the full results. For a robustness check, I also run the same regression with a pure cross-sectional dataset with 2,058 observations. The result is reported in Table 4; qualitatively similar results are observed. Lastly, as in Kummer and Schulte (2017), I use the sample with twin apps—with 85 total observations—in which there are two versions of each app, one free and one paid, launched by the same developer with the same app name. The result in Table 5 shows a similar pattern. The detailed description of additional variables and results is available upon request.

Table 3: Results from the Cross-Sectional Data

	WO Controls	W Controls	IV
Dum_privacy_Top_Dev	0.670***	0.560***	0.544***
	(0.0979)	(0.0798)	(0.122)
Dum_privacy	0.0547	-0.182***	-0.177***
	(0.0636)	(0.0555)	(0.0622)
Top_Dev	0.909***	0.422***	0.460**
	(0.0863)	(0.0713)	(0.213)
D_Price	-3.828***	-3.941***	-3.634**
	(0.0611)	(0.0561)	(1.538)
ln_price	0.0688	0.0237	-0.276
	(0.0424)	(0.0372)	(1.503)
ln_App_Age		1.044***	1.042***
		(0.0549)	(0.0557)
avg_rating		1.387***	1.398***
		(0.0557)	(0.0787)
Numberofpics		0.0420***	0.0430***
		(0.00278)	(0.00580)
totalpermissions		0.0351***	0.0355***
		(0.00278)	(0.00326)
Num_of_apps_per_dev		0.0333***	0.0327***
		(0.00374)	(0.00414)
Constant	11.47***	-1.248***	-1.301**
	(0.0599)	(0.433)	(0.517)

Table 4: Results from the Pure Cross-Sectional Data

	WO Controls	W Controls	IV
Dum_privacy_Top_Dev	0.544**	0.334	-0.0197
	(0.257)	(0.211)	(0.595)
Dum_privacy	-0.377**	-0.476***	-0.559
	(0.161)	(0.137)	(0.400)
Top_Dev	1.100***	0.647***	1.415**
	(0.228)	(0.192)	(0.673)
D_Price	-3.915***	-3.899***	4.400
	(0.145)	(0.141)	(4.800)
ln_price	0.119	-0.0139	-8.868*
	(0.105)	(0.0867)	(5.121)
ln_App_Age		1.054***	0.943***
		(0.116)	(0.171)
avg_rating		1.238***	1.857***
		(0.116)	(0.455)
Numberofpics		0.0489***	0.0571***
		(0.00690)	(0.0196)
totalpermissions		0.0432***	0.0641***
		(0.00618)	(0.0182)
Num_of_apps_per_dev		0.0317***	0.0162
		(0.0108)	(0.0206)
Constant	11.01***	-1.092	-4.070*
	(0.157)	(0.877)	(2.352)

Table 5: Results from the Twin App Data

	WO Controls	W Controls	IV
Dum_privacy_Top_Dev	1.088	3.255***	3.025***
	(0.965)	(1.180)	(1.008)
Dum_privacy	0.0403	-1.472*	-1.967**
	(0.796)	(0.840)	(0.813)
Top_Dev	-0.123	-2.183*	-1.931**
	(0.741)	(1.083)	(0.892)
D_Price	-5.103***	-4.195***	-6.746***
	(0.625)	(0.623)	(1.773)
ln_price	0.267	0.189	2.730*
	(0.521)	(0.385)	(1.635)
ln_App_Age		0.801**	1.102***
		(0.381)	(0.340)
avg_rating		1.662**	1.621**
		(0.657)	(0.731)
Numberofpics		0.0445	0.0259
		(0.0526)	(0.0487)
totalpermissions		0.0852***	0.109***
		(0.0263)	(0.0265)
Num_of_apps_per_dev		0.213**	0.128
		(0.0901)	(0.0979)
Constant	13.21***	2.103	1.730
	(0.653)	(3.210)	(3.421)