Journal of Economic Theory and Econometrics, Vol. 31, No. 3, Sep. 2020, 1–22

Collection of Personal Data with Information Externality in Two-Sided Markets*

Hyungjin Kim† Jinhyuk Lee‡

Abstract We study the conditions under which an internet platform excessively collects personal information when information externality exists. A monopolistic platform runs two-sided business where users and firms form each side. Users do not pay fees to the platform, but firms pay for targeted advertisements. In an environment where marginal information externality is large, the amount of personal information collected exceeds social optimum i) when the utility of marginal user does not make up for the aggregate disutility of existing users and firms and ii) when the platform extracts large markup from firms to keep users.

Keywords Privacy, Personal Information, Private Information, Information Externality, Two-Sided Market

JEL Classification D62, L12, L86

*We thank Wonki Cho, Euncheol Shin and two anonymous reviewers for their helpful suggestions and comments. Jinhyuk Lee's work is supported by a Korea University Grant (K1710031).

†Korea University, Seoul, Korea. Email: hjk9991@korea.ac.kr

‡Korea University, Seoul, Korea. Email: jinhyuklee@korea.ac.kr. Corresponding author.

Received July 7, 2020, Revised August 24, 2020, Accepted August 26, 2020

1. INTRODUCTION

This paper studies excessive collection of personal information, which has been an important social issue. The topic has been recently studied with a one-sided business where there are direct exchanges of money and services between platforms and users (Choi *et al.*, 2019). One of examples of one-sided entrepreneurship is Netflix, where consumers pay per month for the company's streaming service. Other examples are newspaper, magazine, music streaming, and some game services in the internet. However, less is known about privacy issue in a two-sided business which features users not paying for the services of the platform. This paper contributes to the literature by studying the dynamics of distributions of firms and users on the platform with a two-sided model.

It is easy to find two-sided businesses in the internet. Many social media provide free services to users, while these internet platforms extract revenue from firms that want to post advertisements. Similarly, some free-to-use applications initially secure users and sell advertisements to firms. In addition, the successes of payment services such as PayPal in the U.S., KakaoPay in South Korea, and Alipay in China are also based on this business scheme. They provide users with convenient payment experience without charging fees. Instead, they charge firms for membership or advertisements.

Forming one side of the platform, firms are willing to pay for the platform because of the novel advantage in advertising through an internet platform. For instance, there are wedged posts for advertisements in the user interface of Facebook and Instagram. They show features, evaluations, or videos of the product to leave positive impressions to potential buyers using the platform. In some game advertisements users can even play a demo version about one minute, which is called a playable ad.[1] Through the service provided by the platform, firms could contact target consumers and attract them with highly tailored marketing strategies.

However, as one can see from Facebook privacy scandal in 2018, private information excessively collected by platforms and its management risks are becoming the heart of the discussion about privacy management. Using a Facebook personality quiz application, Cambridge Analytica collected personal information of dozens of millions of people. The information collected includes private data accessible to public, pages that victims 'liked,' timelines, news feeds, and messages, which were rich enough to infer psychographic profiles of users. The collected data were allegedly used to affect political events such as president

[1] See https://www.facebook.com/business/ads/playable-ad-format

election of the U.S. in 2016 and Brexit election. As one can see from this, more personal information is stored in the platform, the severer damage is caused in case of the leakage and misuse of personal information. In addition to data leakage, potential disutility of platform users caused by excessive data collection varies from annoyance with changing passwords regularly to less favorable contract terms when users deal with financial institutions.

Suffering of internet users can be even more serious when information externality exists; even non-users of the platform burden nuisance costs of private information collected by the platform. For instance, as the number of users grows, it becomes easier for the platform to infer a non-user's identity, preference and background information. The nuisance of non-users can be almost as serious as that of active users. Non-users can be damaged by net stalking such as persistent targeted advertising and reduced bargaining power.

This paper formally addresses privacy issues related to two-sided business. In the two-sided business model users and firms form each side across the platform. The platform earns revenue from firms by charging them advertising service fees. A firm's willingness to pay for the advertisement depends on the number of users. Therefore, the success of the platform's business depends on keeping the number of existing users and promoting the influx of new users. However, since the platform does not take the disutility of non-users caused by information externality into account, it is likely that the market equilibrium is not Pareto efficient. This paper clarifies under which circumstances the excessive private information gathering arises in such business environment. To this end, the model characterizes the monopolist's strategy with a two-sided business model charging firms only and information externality.

This paper analyzes the behavior of a monopolistic platform because the internet network effect, reputation effect, and switching cost often end up giving market power to a single platform. Recall that a usual monopoly market results in underproduction, compared to the social optimum. The monopolist in textbook keeps the quantity low and sets the price high. On the other hand, private data provision models in literature result in the opposite: The platform serves too many users or extracts too much private information, compared to the social optimum. Such suboptimal practice arises because more users mean a higher willingness to pay of firms in two-sided business. Retaining the number of users and their private data the platform can make its products more valuable to firms. When the price elasticity of users is large, the monopolist can make a higher profit by charging zero price to users and acquiring revenue from advertisements fee, rather than by charging users and having firms less willing to

advertise through the monopolist. Therefore, in this case the platform takes a profit maximizing strategy that secures a pool of users by charging very low or zero price.

The problem gets more complicated when there are interactions between firms and users. More users in the platform imply more firms paying for the advertisements through the platform. On the other hand, it is not obvious what would be the reaction of users to more firms in the platform. If advertisements are genuinely informative to users, an additional firm that enters the platform's environment would increase the number of users. However, advertisements are sometimes regarded as inappropriate or even obnoxious with no useful information. Furthermore, more advertisements often lag user experience in the platform since they are required to watch multiple commercial videos or posts. Thus, an additional firm advertising through the platform can cause some of active users to leave the platform. This paper deals with this difficulty by comparing the result of the main model with that of a simpler benchmark model.

A particular set of assumptions are adopted for simplicity, both in the benchmark and the main model. When an internet user decides to use the platform, her nuisance caused by private data that the platform holds increases. This is because to use the service, she must agree upon private data collection terms the platform offers and provide her own privacy. While more private information from active users are bad for users, it is good for the firms advertising through the platform. On the other hand, in the main model with a larger number of firms that purchase the advertisements, benefits of active users from the platform decreases. In addition to the nuisance costs from privacy provision, this type of decrease in utility from the platform worsens the social welfare.

We find that excessive amount of private information is transferred from users to the platform when marginal user's benefit from the platform cannot compensate her externality, and when the platform secures users by keeping firms with low willingness to pay out of the platform. That is, the limited entry of firms allows potential users to become active and voluntarily provide their privacy even if such provision is not socially desirable. Moreover, because the incentive of the platform is focused on securing as many users as possible, a marginal user's entry can be socially undesirable. This is because the marginal user does not consider its impact on the purchase decisions of firms and the welfare of active users in her entry decision.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature on research on collection of personal information. Section 3 contains the main body of our study on the behavior of the platform on the

collection of personal information. Section 4 concludes with a discussion and suggestions for future work.

2. LITERATURE REVIEW

Our model shares many aspects of previous studies on privacy collection. Especially, this paper aims to derive implications on social welfare based on settings of Choi *et al.* (2019). They study consequences of information externality when there are direct transactions between users and platforms. In market equilibrium excessive amount of information collection arises even when every consumer is aware of a higher nuisance cost from using the platform. They focus on the response of users to the internet platform's decision. In our model, on the other hand, users interact with another type of economic agents, firms.

This paper is also related to government regulations on privacy regulations. For instance, Athey (2014) focuses on providing general rationales for privacy regulations, pointing out that market outcomes are often inefficient due to privacy loss. She reserves the claim, however, by saying that an overly stringent regulation may lead a profitable business to shut down more than desired. This paper does not provide explicit arguments for privacy regulations. The main result in this paper suggests that the excessive information collection has a set of sufficiency conditions when it arises.

In terms of behavioral features of platforms regarding privacy, our study is also related to research on how two-sided businesses affect the information collection of the platform (Bergemann and Bonatti, 2015; Bataineh *et al.*, 2016). Previous studies explore private information selling scheme under a two-sided market environment. Under such setting, the platform can contribute to enhance matching values generated when firms and users fit one another by providing information on individual types. Bataineh *et al.* (2016) point out that there is a huge amount of unused information that can be valuable resource for firm. To solve this puzzle, they build a two-sided platform which accrues data and sells it in the secondary market. Similar to Bergemann and Bonatti (2015), they focus on creating a data trading market, rather than on information externality and its social welfare consequence. Firms and users are better off with higher matching values from using private information about users. Our discussion contributes to existing studies in the sense that we focus on the procedure where the private information stored in the platform is used to generate revenue. By doing so, we try to unveil the mechanism–often assumed as a black box function– with which the platform takes advantage of the privacy of users.

The implication of our model is related to how different market structures have different consequences in a two-sided business model. In that sense, Dimakopoulos and Sudaric (2018) suppose two platforms whose profits depend on data extracted from users. The platforms compete toward user data and form a subgame perfect Nash equilibrium. As a result of the monopolistic competition, privacy costs of users in either platform cause inefficient equilibrium. As the competition weakens, excessive information gathering becomes more likely. Although we assume a monopolistic platform, the result is similar in the sense that the platform's ignorance on the costs of users can cause inefficient amount of privacy collection.

Last but not least, this paper is related to consequence of data provision in various settings. About this particular issue, Acquisti *et al.* (2016) provide a broad theoretical background of private information issues. One of the points they make closely related to our discussion is that disclosing private information could be costly when firms, agents on the other side of the bargain, exploit the user information to increase their payoff. This reasoning justifies to assume the nuisance costs of data provision, which are adopted by many studies including ours. Meanwhile, Ali *et al.* (2019) consider the strategic formulation between platforms and users where users could opt in or out when asked to provide private information. If users have control over the provision of privacy after deciding to use the platform, they can enjoy increase in welfare from intensified competitions among firms. On the other hand, our study focuses mainly on the mechanical aspects between firms and users; One type of agents affects the other type only through the total number using the platform. Also, in our model all opt-in users must provide the same amount of private data.

We believe this paper contributes to the existing literature by providing knowledge on mechanism behind excessive privacy collection where the platform runs two-sided business. Although two-sided businesses are prevalent in reality, their implications on social welfare when users voluntarily provide their privacy are largely unknown. This paper aims to cover this gap by modifying settings of Choi *et al.* (2019), who assume one-sided business platform.

3. MODEL STRUCTURE

There is a monopolistic internet platform. The platform provides internet service to users, who values it differently from one user to another. As the platform adopts more intuitive and easier user interface by choosing lower u, it serves more users that burdens less learning cost. On the other hand, the cost of the

platform running the business is an increasing function of the number of users, due to a larger management cost. In addition, the monopolist chooses the price of advertisements, p, to be sold to firms. Once the platform chooses (u, p), users and firms make entry decisions.

Users take one of two sides across the monopolistic platform. To use the platform's service, users must put effort u given by the platform, learning how to use the service. Each user expects different utility from using the platform. In contrast to the utility of users, users are homogeneous in terms of learning ability. That is, each user burdens the same cost u as long as opting into the platform. Furthermore, to access the service any (potential) user must agree upon terms and conditions about personal information provision.

As more users opt in the data provision terms, the platform acquires ability to infer the identities and preferences of users and non-users. This inference is done through large data of users. Even with limited personal information about nonusers, the platform could match the vast data of users that have similar characteristic to those of nonusers. This means that the platform effectively stores the privacy of non-users as well, even though it is not as accurate as that of active users. This is information externality that causes nuisance to both users and non-users.

On the other side, a mass of firms observe how many consumers are using the platform and decide whether to buy the platform's advertising service. Since the platform provides the service to users for free, the sales to firms are the sole source of the platform's revenue.

It is important to shed light on how firms benefit from the increased number of users in the platform. This is because in the model the fraction of firms purchasing the advertisements depends on how many users use the platform. The influence of the number of users exists in reality. For example, there are wedged posts in the user interfaces of social network services. These advertisements often mimic usual posts and attempt to give positive impressions about the advertised product. Personal information on targeted users enables to fine-tune the contents of ads. This is because with more personal data firms can figure out i) which user group ads should target, ii) what kind of features the ad should contain, and iii) how persistent the advertisement should be when it tracks the target.[2] Hence, how effective an advertisement in the platform heavily depends on the amount of private information.

[2]See 'Are Targeted Ads Stalking You? Here's How to Make Them Stop,' Brian X. Chen, The New York Times, Aug. 15, 2018.

3.1. THE BENCHMARK MODEL: ONE-WAY INFLUENCE ENVIRONMENT

In the benchmark model, each firm's objective function depends on the number of consumers, but users' utilities from the platform are invariant to the number of firms. This setting can be found in internet businesses with a small pool of firms, whose advertisements can be kept from being offensive or annoying.

A monopolistic platform chooses u of users' learning cost. Each user has type $\hat{u} \in \mathbb{R}_+$, different one from another. $\hat{u}$ represents the user's utility on the platform's service. Thus, a consumer with type $\hat{u}$ decides to use the platform when $\hat{u}$ is no less than u. Since the platform has to lower u (lessen the learning cost) to attract more users, u is called the cutoff utility. Let a differentiable distribution function $F : \mathbb{R}_+ \to [0,1]$ and its nonzero density function $f(\cdot)$ represent the distribution of $\hat{u}$. F has monotone and bounded hazard rate, i.e. $\frac{f}{1-F}$ is a non-decreasing function and $\lim_{u\to\infty} \frac{f}{1-F} < \infty$. Let $m(u) \equiv 1 - F(u)$ denote the number of users in the platform. These settings implies that consumers do not internalize nuisance cost generated after data provision. The platform decides u and expends $C(m(u))$ of cost where $C(m) > 0$, $C' > 0$, and $C'' > 0$ for any $m \in [0,1]$. Also, the cost function satisfies an Inada condition: $\lim_{m\to 0} C'(m) = 0$.

The information from private data collected by the platform is proportional to the number of users. Assume that each user provides the same amount of data once he joins. Then, the amount of data collected by the platform has a linear relationship with the number of users using the platform. From now on, without loss of generality let m denote not only the number of users, but also the amount of data provided to the platform.

The data stored in the platform causes nuisance to both users and non-users, with different degrees. Let $\psi : [0,1] \to \mathbb{R}_+$ denote nuisance costs of active users where $\psi(m)$ represents the nuisance generated by $m \in [0,1]$ active users. ψ is increasing and differentiable positive function. Increasing property of ψ implies that more users causes larger amount of nuisance to each user; it captures the case where there is an information externality over private information. Non-users also burden $\hat{\psi} = \xi\psi$ of nuisance cost where $\xi \in [0,1)$ since their private information could also be inferred by data of m active users. ξ represents the accuracy of the platform's inference about a non-user's characteristics.

On the other side, a firm purchases the platform's advertising service if it is valuable enough. Let $a(m)$ denote the marginal benefit each firm obtains by purchasing advertisement from the platform, when m of users are in the platform. Notice that the number of consumers using the platform influences the benefit. $G(\cdot;m)$ denotes the distribution of firms' marginal benefit $a(m)$ where

m enters as a parameter. $G(\cdot,\cdot)$ is twice differentiable, and for any $m > m'$, $G(\cdot,m') \geq G(\cdot,m)$. In other words, the distribution of firm's benefit with a large number of platform users first-order stochastically dominates the one with a small number of users. Notice that this assumption does not necessarily imply every firm's benefit from the advertisement increases. Further, at each point of marginal benefit the change in the density of firms is smooth when there is a change in the number of users: The rate and acceleration of the density with respect to the number of users are bounded by some integrable function. In addition, the first-order derivative of $G(y;m)$ is strictly increasing in m, given $y \in \mathbb{R}_+$ fixed. These assumptions in marginal profit of firms are parallel to decreasing returns to scale in usual production function.

The platform's advertisement pricing determines the number of firms using the platform. Assume that the platform has uniform pricing on firm's side[3]. When the platform posts a price p, a firm with type a buys the service if and only if $a - p \geq 0$. Hence, the demand for the platform's advertisement given p is $1 - G(p;m)$. The platform decides p that simply maximizes $\pi(p) \equiv (1 - G(p;m))p$. A local solution maximizing this term is in $\mathbb{R}_+$. Let p^* denote it. G satisfies the monotone hazard rate property for any given parameter m, which ensures existence of the unique solution for the platform's price choice.

The decisions in this two-sided business model are made along with the following stages.

Stage 1 The platform decides u.

Stage 2 Each potential user decides whether to use the platform. Then m is determined accordingly.

Stage 3 Observing m, the platform decides p.

Stage 4 Each Firm decides whether to buy the platform's service or not.

This benchmark model is useful in two ways. Its first virtue is simplicity. The assumption allows an extensive form game to represent the situation, otherwise we should adopt a game with simultaneous decisions, which will be analysed in subsection 3.3. Second, unlike firms' side, it is unclear whether consumers are positively or negatively affected by the number of firms. Though many advertisements decrease the benefits of the platform users, many have amusing way of promoting the product. The game dealt in this subsection can be interpreted

[3]In this model each firm's willingness to pay is not known to the platform but its distribution is.

as a special case where users are irritated only by a nuisance generated by information externalities and indifferent about the number of firms.

3.2. SOCIAL WELFARE AND PLATFORM'S PROFIT

In this subsection we define the social welfare function and the platform's profit to derive necessary conditions for the maximization problem. Given u provided by the platform, the social welfare function is defined as

$$W(u) = \int_u^\infty x dF(x) - um - m\psi(m) - (1-m)\hat{\psi}(m) - C(m) + \int_0^\infty y dG(y,m) \quad (1)$$

where $m = 1 - F(u)$. The social welfare function consists of four parts. The first two terms are integration of utilities of consumers using the platform. The third and fourth term represent nuisance costs of users and non-users. The second last and the last term are the operation cost of the platform and the aggregate value generated by advertisements[4]. The first-order derivative of equation (1) provides a necessary condition for u to maximize W

$$\int_0^\infty y \frac{\partial}{\partial m} g(y,m) dy = \Psi(m) + C'(m) + \lambda(u) \quad (2)$$

where $\Psi(m) = \frac{d}{dm}[m\psi(m) - (1-m)\hat{\psi}(m)] = (1-\xi)\psi(m) + m\psi'(m) + (1-m)\xi\psi'(m)$ and $\lambda(u) = \frac{1-F(u)}{f(u)}$.

The interpretation of the first-order condition above is about matching marginal benefit-marginal cost. The left-hand side represents social gain changing m, which consists of the benefits of firms. The right-hand side consists of three parts: the change in nuisance costs for overall users and non-users, the increase in cost of the platform, and the ratio between the number of active users and the marginal users. Equation (2) implies that to maximize social welfare, the marginal benefit of producing additional unit of utility (LHS) should be equal to the cost of it (RHS).

Suppose the social planner decreases the cutoff utility by small amount so that more users join and provide data to the platform. The aggregate benefit of firms increases due to the increased number of users. However, the nuisance cost of users and the platform's service cost increase. The non-user at margin obtains

[4]The social welfare varies with how many firms purchase the platform's advertisement service, which depends on a. However, with positive marginal benefit of advertising it is always better to include all companies in perspective of social planner with zero price unless more firms using the platform imply decreased welfare of consumers.

u, which offsets the cost of learning. Thus, only the relative learning cost of existing users $\lambda(u)$ enters the equation.

The social planner pins down unique u^s, and thus the optimal number of users $m^s = 1 - F(u^s)$ if ψ and C are strictly convex. The proposition states the unique social optimal amount of private information gathering $m^s > 0$ exists.

Proposition 1. *Suppose $C : [0,1] \to \mathbb{R}_+$ and $\psi : [0,1] \to \mathbb{R}_+$ have positive first- and second-order derivatives and $max_{m\in[0,1]}|y\frac{\partial}{\partial m}g(y,m)| < h(y)$ for some $h(y)$ such that $\int_0^\infty h(y)dy < \infty$. Then if $\int_0^\infty y\frac{\partial}{\partial m}g(y,1)dy < (1-\xi)\psi(1) + \psi'(1) + C'(1)$, there is a unique amount of personal information gathering $m^s < 1$ that satisfies (2).*

Proof. See Appendix. □

Now we define that the platform's profit function also depends on $m(u)$. Though the price charged to firms is also a choice variable of the platform, once the number of users is given we can optimize out the platform's choice of p. Since the platform behaves as a monopolist, it effectively chooses the number of users that maximizes the profit:

$$\begin{aligned}\Pi(m) &= (1 - G(p^*(m);m))p^*(m) - C(m) \\ &= n(m)p^*(m) - C(m)\end{aligned} \tag{3}$$

where $p^*(m)$ is either the fixed point of $\phi(p;m) \equiv \frac{1-G(p;m)}{g(p;m)}$ or zero if the fixed point does not exist and $n(m) \equiv 1 - G(p^*;m)$.[5] There are two cases about the maximand $n(m)p^*(m)$: one is that it has a critical point in $(0,1]$, and the other is that it is monotone decreasing over $[0,1]$ and hence has the maximizing point at $m = 0$.

Similar to the social welfare analysis, monopolist's choice m^* satisfies the first order condition of the profit maximization problem.

$$p^*(m^*)\frac{\partial}{\partial m}G(p^*(m^*),m^*) + C'(m^*) \leq 0 \tag{4}$$

where equality holds whenever $m^* > 0$ with complementary slackness. By monotone hazard rate condition and first-order stochastic dominance assumption

[5]Given $m = F(u)$, the monopolist's rule to choose p^* is $p^* = arg\max_{p\geq 0}(1 - G(p;m))p - C(m)$. The derivative of the maximand with respect to p is $1 - G(p;m) - g(p;m)p$, and so its critical point, denoted as p^{**}, is the fixed point of the function $\phi(p) \equiv \frac{1-G(p;m)}{g(p;m)}$ over 45 degree line. This point is unique by the monotone hazard rate assumption. Thus, with complementary slackness $p^* = max\{0, p^{**}\}$.

for G, there is a unique amount of privacy collecting amount m^*. If this amount of information collection is larger than the social optimal amount, the society is said to suffer from excessive information collection. $m^* > m^s$ indicates that the platform's incentive to collect private information from a wide range of users does not align with that of the social planner.

This is because the monopolist completely ignores the nuisance costs that users bear. The social planner limits the number of active users considering firms' benefits from information on users and changes in users' benefits due to the nuisance cost generated by the amount of information collection. However, the platform as a monopolist are not obligated to take the costs of users as long as firms do not leave it. As a result, as shown in the next proposition in the one-sided business model excessive privacy is at the monopolist's hand if the benefits of firms cannot make up for the nuisance cost of users.

Proposition 2. *Given that the platform's choice m^* of privacy collection amount is greater than zero, $m^* \geq m^s$ if*

$$\int_0^\infty y \frac{\partial}{\partial m} g(y, m^*) dy + p^*(m^*) \frac{\partial}{\partial m} G(p^*(m^*), m^*) \leq \Psi(m^*) + \lambda(u^*). \quad (5)$$

where m^s is the socially efficient amount of private data collection.

This over-collection would have arisen without information externality when

$$\int_0^\infty y \frac{\partial}{\partial m} g(y, m^*) dy + p^*(m^*) \frac{\partial}{\partial m} G(p^*(m^*), m^*) \leq \lambda(u^*). \quad (6)$$

Proof. See Appendix. □

Information over-collection arises when the platform has a stronger incentive to increase the number of users compared to that of the social planner. Proposition 2 considers the social planner's marginal cost and benefit, given the data collection amount of the monopolist. If the social planner increases the number of users, the cost is the higher nuisance to users and non-users, while the platform and firms benefit from a larger number of users. Inequality (5) implies that the social planner would intend to reduce private data collection when the marginal social cost at the monopolist's choice exceeds the marginal social benefit.

In addition, Proposition 2 implies that the information externality can cause the excessive collection of privacy. Suppose inequality (6) does not hold while inequality (5) hold. Then, in the market equilibrium privacy collection exceeds socially desired amount. Should nuisance cost from information externality were

not there, however, this over-collection would not happened. Thus, the information externality plays a crucial role in excessive private information collection in this case.

This type of excessive data provision of users is likely to get severe in the future. As the technology advances and handles a larger amount of data, it is likely that the platform's inference ability about users' identity improves. This implies that larger information externality causes higher nuisance costs to users. From equation (2), if there were no nuisance costs, the social optimum amount of information gathering is higher than the optimal amount with nuisance cost. On the other hand, since equation (4) does not depend on Ψ, information externality itself does not change the platform's decision. m^* being the same, higher nuisance costs due to information externality lessens the social planner's choice of data provision from users. Therefore, more advanced data usage technologies could lead to excessive private information collection.

The result in Proposition 2 is parallel to proposition 4 in Choi *et al.* (2019), in the sense that the inequality highly depends on the magnitude of Ψ. They derived the result with an explicit assumption that the platform makes revenue with the information from users, while our study implicitly incorporates the revenue obtained by selling advertising into the inequality. Our model does not allow the platform to charge fees to users, while the revenue from users is the main source for profit in their study. In many cases, the reality is somewhere between these two extremes. Many internet platforms have two conflict incentives: to keep the users and make revenue from firms and to charge fees q to users directly for the service they provide.

3.3. TWO-WAY INFLUENCE ENVIRONMENT

In reality the distribution of payoffs on each side depends on the number of agents on the other side. The benefits that firms get form the platform service are affected by the number of users, while users' utilities are also influenced by the number of firms. As assumed in the previous section the effectiveness of a firm's advertisement through the platform can improve as there are more users and their private data. The number of firms affects the distribution of users' utility by changing the user experience in the platform. Specifically, firms are heterogeneous in what and how they advertise. As more firms advertise, it becomes more likely for users to encounter irrelevant, annoying or even harmful advertisements. The expectation for such uncomfortable experience will lower the utility of each user from the platform's service.

To see conditions that generates excessive data collection in such environ-

ment, from now on we first posit a family of F, G, ψ and C that satisfies assumptions in the previous section. Let $n \equiv 1 - G$ denote the number of firms using the platform. n enters F as a parameter since the distribution of users' utility F changes as the number of firms purchasing the advertisements changes. Assume for any $n > n'$, $F(\cdot, n) > F(\cdot, n')$ i.e. potential utility decreases as there are more firms using the platform's service. By this setting, the distributions of the utilities of users and marginal benefits of firms interact across the platform.

The distributions and entry decisions of users and firms affect those on the other side in a recursive fashion. If the number of users increases because of some exogenous shock, the distribution of firms' benefit changes in the sense that the later distribution first-order stochastic dominates the initial one. It follows that the decisions of some firms change, so does the number of firms that purchase the advertisements from the platform. This affects the distribution of users' utilities from the platform's service, and the interactions between two sides continue. Thus, the characterization should be done in a point where there is no more changes in the number of active users and firms. Suppose a simultaneous decision making process where users and firms decide whether to use the platform at the same time after they observe an offer from the platform.

Stage 1 The platform offers u and p to each side. Every agent can observe the offered price and utility.

Stage 2 Each potential user and firm decides whether to use the platform.

Suppose the platform chooses a pair $(\bar{u}, \bar{p})$. A firm with willingness to pay a chooses to purchase advertisements if its net benefit $a - \bar{p}$ is nonnegative. Similarly, a user opts in the platform's service once her utility of the service is greater than the platform's choice of $\bar{u}$. Then, the equilibrium numbers of users and firms in the platform arise from the fixed point for $m(\bar{u}, \bar{p}) = 1 - F(\bar{u}, n(\bar{u}, \bar{p}))$ and $n(\bar{u}, \bar{p}) = 1 - G(\bar{p}, m(\bar{u}, \bar{p}))$. Since F and G are continuous and strictly increasing, given the pair $(\bar{u}, \bar{p})$ its corresponding fixed point of firms and users $(\bar{m}, \bar{n})$ uniquely exists.

Therefore, it is possible to define $m : \mathbb{R}^2_+ \to [0,1]$ and $n : \mathbb{R}^2_+ \to [0,1]$. These functions imply that the platform effectively controls the number of users and firms by choosing (u, p). Assume that both functions are continuously differentiable and have the same directional derivatives in the direction of $(1, -1)$, i.e. $\frac{\partial n}{\partial u} - \frac{\partial n}{\partial p} = \frac{\partial m}{\partial u} - \frac{\partial m}{\partial p} = \mu(u, p) \neq 0$. This particular assumption says that both firm and user sides respond in the same rate to the platform's small changes in decision variable, which is represented by function μ.

When $F(u,\cdot)$ does not vary over n given any u, m is the same function as in section 3.2, so the problem degenerates to one-side influence model as well. The assumption on directional derivatives holds here if $-f(u) = g(p,m) - \frac{\partial(1-G(p,m))}{\partial m} f(u)$ where $m = 1-F(u)$. This is a special case of function $\mu(u,p)$.

Although the setting of two-sided interaction here is closer to reality, the model is barley tractable. Without full characterizations about the distribution of F, G and their relationship, it is hard to say whether the choice of the platform results in socially excessive amount of private information collection. It is usually not possible to derive algebraic formula with respect to n and m, which is essential to characterize equilibrium. This difficulty motivates a different approach for equilibrium characterizations.

We characterize excessive information collection by comparing the amount collected in two-sided business with that collected in one-sided business. Suppose that there is hypothetically excessive amount of information collection in a one-way influence environment as in 3.2: $m^* > m^s$, for which the conditions are specified in Proposition 2. Rather than characterizing how extensive the platform is collecting privacy compared to social optimum in a two-way environment, we compare the amounts of information collected in the one-way environment and the amounts collected in the two-way environment: the amounts of information gathered by the monopolist, m^* in the one-way environment and m^{*t} in the two-way environment; the social optimums, m^s in the one-way environment and m^{st} in the two-way environment. Here, superscript t denotes variables in a two-way influence environment. Suppose $m^{*t} \geq m^*$ and $m^s \geq m^{st}$. Combined with the presumption $m^* > m^s$, it immediately follows that two-sided monopolist's choice in the amount of privacy information exceeds the social optimum.

To this end, this section first defines the social welfare function and the profit function in two-sided business environment. The social welfare function is defined as

$$\begin{aligned} W(u,p) = &\int_u^\infty x dF(x,n(u,p)) - um(u,p) - m(u,p)\psi(m(u,p)) \\ &- (1-m(u,p))\hat{\psi}(m(u,p)) - C(m(u,p)) + \int_p^\infty y dG(y,m(u,p)) \end{aligned} \tag{7}$$

Whether it is always good to include every firms in the platform is the main difference between the social welfare function in two-sided environment and that in one-sided environment. Compared to equation (1), the last term in equation (7) does not integrate over the whole positive real line. Notice that it is no longer optimal for the social planner to have all firms use the platform's service. Thus, on the solution price could have nonzero value in two-sided environment.

Furthermore, it is worth to note the implication of the interactions between users and firms. More firms and more advertisements imply less total welfare for users. When the social planner chooses the platform's cutoff utility and price imposed to firms, the number of firms and users (m,n) and their distribution $F(\cdot,n)$ and $G(\cdot,m)$ are determined. As discussed in 3.2, social welfare function is the sum of user's net utility and firm's benefit from the service.

The objective function has two partial derivatives

$$\begin{aligned}
\frac{\partial W(u,p)}{\partial u} &= \int_u^\infty x\frac{\partial}{\partial n}f(x,n)\frac{\partial n(u,p)}{\partial u}dx + m(u,p) - \frac{\partial m(u,p)}{\partial u}\Psi(u,p) \\
&\quad - C'(m)\frac{\partial m(u,p)}{\partial u} + \int_p^\infty y\frac{\partial}{\partial m}g(y,m)\frac{\partial m(u,p)}{\partial u}dy = 0 \\
\frac{\partial W(u,p)}{\partial p} &= \int_u^\infty x\frac{\partial}{\partial n}f(x,n)\frac{\partial n(u,p)}{\partial p}dx - u\frac{\partial m(u,p)}{\partial p} - \frac{\partial m(u,p)}{\partial p}\Psi(u,p) \\
&\quad - C'(m)\frac{\partial m(u,p)}{\partial p} + \int_p^\infty y\frac{\partial}{\partial m}g(y,m(u,p))\frac{\partial m(u,p)}{\partial p}dy - pg(p,m) = 0.
\end{aligned} \tag{8}$$

These conditions yield a necessary condition for maximization

$$\int_u^\infty x\frac{\partial}{\partial n}f(x,n)dx + \int_p^\infty y\frac{\partial}{\partial m}g(x,m)dy + \kappa(u,p) = \Psi(m) + C'(m). \tag{9}$$

where $\kappa(u,p) \equiv \frac{u\partial m(u,p)/\partial p + m(u,p) + pg(p,m)}{\mu(u,p)} = -u + \frac{1}{\mu(u,p)}(\frac{\partial[um(u,p)]}{\partial u} + pg(p,m))$ captures the relative welfare change due to the entry of marginal agents and different learning costs where there is a small change in choice variables (u,p).

The social planner's choice here shows the principle of matching marginal benefit and marginal cost achieving welfare maximization. The distributions of benefits of both users and firms change when the social planner makes a small adjustment in (u,p). The marginal social benefit in aggregate term is in the first two terms of the left-hand side of equation (9). The change in the number of users causes privacy nuisance and service costs to change. In addition to these costs, the left-hand side of equation (9) contains the benefit that marginal agents gain, $\kappa(u,p)$. These agents start or quit to use the platform due to the change in the cutoff.

We first compare social planner's decision about the amount of private information collection. A change in business environment from one to two-sided business decreases the private data provision amount of the social planner if the following condition is satisfied.

Proposition 3. *$m^{st} \leq m^{s}$ if*

$$\kappa(u^{st}, p^{st}) \leq \lambda(u^{s}) + \int_{u^{st}}^{\infty} x \frac{\partial}{\partial n} f(x, n^{st}) dx - \int_{0}^{p^{st}} y \frac{\partial}{\partial m} g(y, m^{st}) dy \tag{10}$$

Proof. See Appendix. □

Rates of changes in the distribution of firms and users $\frac{\partial f}{\partial n}$ and $\frac{\partial g}{\partial m}$ matter when socially desirable amount of information collecting changes. When a marginal user comes into the platform, he contributes to total welfare by his own utility from the platform. However, the marginal increment in users affects the decisions of firms, which in turn changes the welfare of users in the platform. As we can see from the last term of equation (9), firms outside the platform face the opportunity costs of not purchasing the advertisements. This is because there is more information from which it can benefit. As the decrease in the social welfare from these changes offset the advantage of the entry of the marginal user, the social planner has an incentive to decrease the number of users when the business environment changes.

Now we characterize the platform maximizing the profit by choosing (u, p). The objective function of the platform consists of revenue from firms and cost of providing service to users. As in the previous section, the platform has no incentive to consider the nuisance cost to each user. Thus the platform's problem is

$$\begin{aligned} &\max_{(u,p)} \Pi(u,p) = n(u,p)p - C(m(u,p)) \\ &\quad s.t. \quad 1 - F(u, n(u,p)) = m(u,p) \text{ and } 1 - G(p, F(u,p)) = n(u,p). \end{aligned} \tag{11}$$

The first order conditions of profit maximization problem provides the solution price $p^{*t} = C'(m^{*t}) + n^{*t}/\mu^{*t}$. The second term is the markup of the monopolist, where n^{*t}/μ^{*t} is the quantity sold relative to the firms' rate of change. The following proposition compares the choices of the monopolists in one and two-sided business environment.

Proposition 4. *$m^{*} \leq m^{*t}$ if*

$$C'(m^{*}) + \frac{n^{*t}}{\mu^{*t}} \leq p^{*t} \tag{12}$$

Proof. See Appendix. □

Proposition 4 says the change in business environment from one to two-sided business results in larger amount of information collection if the price of the monopolist in two-sided environment exceeds the marginal cost of service at the one-sided platform's choice. That is, sufficiently high price optimized out of demand causes large amount of data provision in two-sided business model.

Finally, the theorem states sufficient conditions for excessive information collecting in the two-sided internet platform.

Theorem 1. $m^{st} \leq m^{*t}$ *when*

1. $\Psi(m^*) + \lambda(u^*) \geq \int_0^\infty y \frac{\partial}{\partial m} g(y, m^*) dy + p^*(m^*) \frac{\partial}{\partial m} G(p^*, m^*)$;
2. $\lambda(u^s) + \int_{u^{st}}^\infty x \frac{\partial}{\partial n} f(x, n^{st}) dx - \int_0^{p^{st}} y \frac{\partial}{\partial m} g(y, m^{st}) dy \geq \kappa(u^{st}, p^{st})$; *and*
3. $p^{*t} - C'(m^{*t}) \geq C'(m^*) - C'(m^{*t}) + \frac{n^{*t}}{\mu^{*t}}$

Proof. See Appendix. □

The interpretations of these conditions are closely related to previous propositions. In the the first condition, each side represents the marginal cost and benefit in aggregate term when the monopolist runs a hypothetical *one-way influence* business. $\Psi(m^*)$ in the left-hand side is the marginal aggregate nuisance cost that users bear in the one-way environment. $\lambda(u^*)$ is the relative learning cost of users, compared to that of marginal users. Aggregate social benefit in the right-hand side consists of two parts. The first is the aggregate benefits firms in the platform, obtained from additional information from marginal user ($\int_0^\infty y \frac{\partial}{\partial m} g(y, m^*) dy$). The second terms is the platform's additional profit from firms that just entered the platform ($p^*(m^*) \frac{\partial}{\partial m} G(p^*, m^*)$). In the first condition, costs due to learning and nuisance exceed benefits that the platform provides to firms when users are indifferent to the number of firms.

The second condition states welfare changes of firms out of the platform and users are large enough, the sum of them covering welfare change due to the decisions of marginal users about using the platform. When the marginal user enters the platform, the change in the number of users enlarges the opportunity cost firms out of the platform, as more users make advertisements more beneficial. The changes in firms' decisions due to the entries of marginal users alter the distribution of users' utility. The two-sided internet platform's privacy collection is excessive when the benefit of the marginal user cannot compensate the consequence of this additional entrance.

Lastly, the third condition says the monopolist limits firm's entry by maintaining its markup sufficiently large. The markup at the margin is socially undesirable if it is higher than the sum of ratio between the demand and its rate of change and the gap between marginal costs under effective *two-sided* environment and hypothetical *one-sided* environment. When the platform's marginal production cost gets efficient, the platform causes inefficient amount of privacy collection.

These findings indicate that if it is costly or impossible to regulate the voluntary transmission of privacy, the policymaker may want to look into some other way to resolve excessive data provision problem. When one of the inequalities in Theorem 1 does not hold, the platform's privacy collection amount does not necessarily exceed the social optimum. In addition to nuisance from information externality, the marginal agent's behaviors and the pricing of the platform causes undesirable data provision. The policy could aim to affect inequalities in Theorem 1. The results could support new policies that broaden options.

4. DISCUSSION

We explore a monopolistic two-sided business model with information externality to characterize its excessive private information collection. In the platform users do not pay for its service but firms buy advertising service based on personal information collected by the platform. We find that the excessive collection of personal information occurs when the utility of marginal user could not cover the aggregate disutility of existing users and firms and when the platform limits the number of firms to secure the number of users.

Our analysis can be extended to multiple dimensions. We do not consider the 'depth' of information that the platform acquires from each user. For simplicity, we assume that each agent provides the same amount of information to the platform to use the service. Therefore, the amount of private data from users is proportional to the number of users. This paper focuses on how extensively the platform collects privacy. In reality, however, different users can provide different amount of private information to the platform. This caveat inspires a natural extension that presumes an environment with heterogeneous data provision.

APPENDIX: PROOFS

Proof of Proposition 1

Proof. We see the right-hand side of equation 2 is a strictly decreasing function of u while the left-hand side is strictly increasing. Then, the inequality $\int_0^\infty y\frac{\partial}{\partial m}g(y,1)dy < (1-\xi)\psi(1)+\psi'(1)+C'(1)$ ensures there is a unique solution. To see the strictly decreasing RHS of 2 in u, note that

$$\frac{d}{du}[\Psi(m)+C'(m)] = -f(u)\big\{2(1-\xi)\psi'(m)+(m+(1-m)\xi))\psi''(m)+C''(m)\big\}. \tag{13}$$

Since C and ψ are both strictly increasing and strictly convex in m, and $f>0$ for each $u \geq 0$, the first derivative of $\Psi(m)+C'(m)$ with respect to u is negative.

On the other hand, $\frac{d}{du}\left\{\int_0^\infty y\frac{\partial}{\partial m}g(y,m)dy\right\} = -\int_0^\infty y\frac{\partial^2}{\partial m^2}g(y,m)dyf(u) > 0$. The second equality is from dominated convergence theorem. Now, suppose there are two different levels of socially optimal cutoff utility u^s and $u^{s'}$ satisfying equation 2. Without loss of generality, $u^s < u^{s'}$. Let m^s and $m^{s'}$ denote $1-F(u^s)$ and $1-F(u^{s'})$ each. Then, we have

$$\begin{aligned}\int_0^\infty y\frac{\partial}{\partial m}g(y,m^s)dy =& \Psi(m^s)+C'(m^s)+\lambda(u^s)\\ >& \Psi(m^{s'})+C'(m^{s'})+\lambda(u^s)\\ =& \int_0^\infty y\frac{\partial}{\partial m}g(y,m^s)dy,\end{aligned} \tag{14}$$

a contradiction. □

Proof of Proposition 2

Proof. Suppose that inequality 5 holds. By inequality 4, the inequality 5 becomes

$$\int_0^\infty y\frac{\partial}{\partial m}g(y,m^*)dy - C'(m^*) \leq \Psi(m^*)+\lambda(u^*). \tag{15}$$

Suppose $m^* < m^s$. Since $m(u) = 1-F(u)$ is strictly decreasing in u, $u^* \equiv F^{-1}(1-m^*) > u^s \equiv F^{-1}(1-m^s)$. Then, we have

$$\int_0^\infty y\frac{\partial}{\partial m}g(y,m^s)dy < C'(m^s)+\Psi(m^s)+\lambda(u^s), \tag{16}$$

which violates equation 2, a contradiction. □

Proof of Proposition 3

Proof. Let inequality 10 holds. Then, by equality 9

$$\begin{aligned}
&\kappa(u^{st},p^{st}) \leq \lambda(u^s) + \int_{u^{st}}^{\infty} x\frac{\partial}{\partial n}f(x,n^{st})dx - \int_0^{p^{st}} y\frac{\partial}{\partial m}g(y,m^{st})dy \\
\iff &\int_0^{\infty} y\frac{\partial}{\partial m}g(y,m^{st})dy + \kappa(u^{st},p^{st}) \leq \lambda(u^s) + \int_{u^{st}}^{\infty} x\frac{\partial}{\partial n}f(x,n^{st})dx + \int_{p^{st}}^{\infty} y\frac{\partial}{\partial m}g(y,m^{st})dy \\
\iff &\int_0^{\infty} y\frac{\partial}{\partial m}g(y,m^{st})dy \leq \Psi(m^{st}) + C'(m^{st}) + \lambda(u^s).
\end{aligned} \tag{17}$$

Suppose $m^{st} > m^s$. Similar to the proof of Proposition 2, we have $u^{st} < u^s$. Since the left-hand side of the last inequality is strictly increasing in u while the right hand side is strictly decreasing, we have

$$\int_0^{\infty} y\frac{\partial}{\partial m}g(y,m^s)dy < \Psi(m^s) + C'(m^s) + \lambda(u^s), \tag{18}$$

a contradiction to equation 2. □

Proof of Proposition 4

Proof. The first-order condition for the platform's problem is

$$\begin{aligned}
\frac{\partial \Pi}{\partial u} &= \frac{\partial n(u,p)}{\partial u}p - C'(m)\frac{\partial n(u,p)}{\partial u} = 0; \\
\frac{\partial \Pi}{\partial p} &= \frac{\partial n(u,p)}{\partial p}p + n(u,p) - C'(m)\frac{\partial n(u,p)}{\partial u} = 0.
\end{aligned} \tag{19}$$

These two equations yields

$$(\frac{\partial n(u,p)}{\partial u} - \frac{\partial n(u,p)}{\partial p})p = (\frac{\partial m(u,p)}{\partial u} - \frac{\partial m(u,p)}{\partial p})C'(m) + n(u,p). \tag{20}$$

Since $\frac{\partial n}{\partial u} - \frac{\partial n}{\partial p} = \frac{\partial m}{\partial u} - \frac{\partial m}{\partial p} = \mu(u,p)$, at solution p^{*t} is equal to $C'(m^{*t}) + n^{*t}/\mu^{st}$ where $\mu^{st} \equiv \frac{\partial n(u^{st},p^{st})}{\partial u} - \frac{\partial n(u^{st},p^{st})}{\partial p} = \frac{\partial m(u^{st},p^{st})}{\partial u} - \frac{\partial m(u^{st},p^{st})}{\partial p}$. The result follows immediately from the assumption that C has positive second derivative. □

Proof of Theorem 1

Proof. Let the three conditions in the statement hold. By previous positions, the result follows. Specifically, each statement implies $m^* \geq m^s$, $m^s \geq m^{st}$, and $m^{*t} \geq m^*$. □

REFERENCES

Acquisti, A., Taylor, C.R., and L. Wagman (2016). "The Economics of Privacy," *Journal of Economic Literature* 52(2), 442-492

Ali, S. N., Lewis, G., and S. Vasserman (2019). "Voluntary Disclosure and Personalized Pricing," *arXiv preprint arXiv:1912.04774.*

Athey, S. (2014). "Information, Privacy and the Internet: An Economic Perspective," *CPB Lecture*.

Bataineh, A.S., Miznouni, R., El Barachi, M., and J. Bentahar (2016). "Monetizing Personal Data: A Two-Sided Market Approach," *Procedia Computer Science* 83, 472-479.

Bergemann, D. and A. Bonatti (2015). "Selling Cookies," *American Economic Journal: Microeconomics* 7(3), 259-294.

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

Dimakopoulos, P.D. and S. Sudaric (2018). "Privacy and Platform Competition," *International Journal of Industrial Organization* 61, 686-713.