Privacy and Personal Data Collection with Information Externalities*

Jay Pil Choi† Doh-Shin Jeon‡ Byung-Cheol Kim§

October 7, 2016

Abstract

We provide a model of privacy in which data collection requires consumers' consent and consumers are fully aware of the consequences of such consent. Nonetheless, the market equilibrium is characterized by excessive collection of personal information and the loss of privacy by consumers compared to the social optimum. This result is due to firms' incentives to exploit negative privacy externalities among consumers, where disclosure of information by some consumers enables an inference of information about other consumers with the use of data analytics. We also discuss the role of data brokerage firms in the aggregation of information and provide implications for privacy policies.

Key words: big data, privacy, information externalities, coordination failure

*This version is preliminary.

†Michigan State University, 220A Marshall-Adams Hall, East Lansing, MI 48824-1038 and School of Economics, Yonsei University, Seoul, Korea. E-mail: choijay@msu.edu.

‡Toulouse School of Economics and CEPR, Manufacture de Tabacs, 21 allees de Brienne - 31000 Toulouse, France. E-mail: dohshin.jeon@gmail.com.

§School of Economics, Georgia Institute of Technology, 221 Bobby Dodd Way, Atlanta, GA 30332-0225. E-mail: byung-cheol.kim@econ.gatech.edu.

1 Introduction

The Internet is now an essential component of our daily lives, and has profoundly changed the way we work, conduct our personal lives, and interact with other people. As we rely more on the Internet, it has become more of a necessity to have constant access to it via mobile devices and computers. However, one consequence of this is that our routine online cavities such as email, search, and shopping constantly generate data about ourselves, which can be collected and used as a competitive advantage by firms. To give an indication of the scale on which "user-generated content" is created, the global Internet population is estimated to be more than 3.4 billion people (as of July 2016), with around 46% of the world population having an Internet connection.[1] According to Google CEO Eric Schmidt, they collectively create approximately 5 exabytes (10^{18}) of data every two days, which is equivalent to the amount of information created "from the dawn of civilization up until 2003."[2] This massive and unprecedented scale of personal data generation in conjunction with rapid reductions in computing costs for data storage and analytics naturally led to serious privacy concerns by the pubic and policy-makers (Schneier, 2015).

One puzzling aspect in this privacy debate is why people tend to set aside their privacy concerns and voluntarily provide their personal information to websites and content providers despite their publicly stated objections and concerns about privacy loss (Singer et al. 2001; Waldo, Lin, and Millet 2007). Certainly there are often cases where "data surveillance" is taking place without our awareness or consent, but it is also true that we frequently agree to it. For instance, we implicitly allow uninterrupted use of location tracking and camera to enjoy the sensational augmented reality game Pokémon Go,[3] and let Google have access to all the metadata we generate in exchange for the use of Gmail. In this paper we address several fundamental micro-level questions motivated by this phenomenon: Why do people tend to allow some form of data surveillance that appears to harm themselves in the end? Do we really expect that most individuals would no longer acquiesce such data surveillance

[1]Source: http://www.internetlivestats.com/internet-users/

[2]See M. G. Siegler, "Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003," *TechCrunch*, August 4, 2010 available at https://techcrunch.com/2010/08/04/schmidt-data/.

[3]The Pokemon Go raised privacy concerns by regulators. See for more details the article by Sam Biddle (9 Aug 2016), "Privacy Scandal Haunts Pokemon Go's CEO," *The Intercept*, https://theintercept.com/2016/08/09/privacy-scandal-haunts-pokemon-gos-ceo/

once they are fully aware of the deal each of them is making? Do firms collect too much or too little data? Is it enough to educate consumers the exact costs of sharing their personal information in order to protect their privacy rights?

To address these questions, we develop a model of privacy in which data collection requires consumers' consent and consumers are fully aware of the consequences of such consent. Nonetheless, the market equilibrium is characterized by excessive collection of personal information and the loss of privacy by consumers compared to the social optimum. This result is due to firms' incentives to exploit negative privacy externalities among consumers, where disclosure of information by some consumers enables an inference of information about other consumers with the use of data analytics. We also discuss the role of data brokerage firms in the aggregation of information and provide implications for privacy policies.

In particular, we consider the interactions between web users and web-based applications/content providers whose business model consists in monetizing personal digital trails by selling them to the data broker industry which is largely operating behind a veil of secrecy.[4] A primary goal of our research is to provide an economic rationale for each user's voluntary consent to the websites for their uncommitted and possibly secondary use of the collected information including sales of data to third parties. There are certainly reasons for this behavior based on non-transparency or consumers' lack of understanding about websites' data usage policy. For instance, privacy notices are "just too long for people to read through ..., making it difficult for most people to understand what they are signing up to."[5] Alternatively, it could be due to consumers' myopic and time-inconsistent preference that is responsible for such behavior.[6] We make no assumption of any bounded rationality or consumers' lack of knowledge about the website's data usage, thereby not resorting to consumers' myopia or to limited information. Instead, we assume very rational consumers who are aware of all the

[4]See the Staff Report for Chairman Rockefeller (December 18, 2013) entitled "A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes" for a detailed description of the secretive nature of the data brokerage industry.

[5]Attributed to Australian Privacy Commissioner Timothy Pilgrim. See Corinne Reichert, "Many Privacy Policies Are Long, Complex: OAIC," *ZD Net*, August 15, 2013, available at http://www.zdnet.com/article/many-privacy-policies-are-long-complex-oaic/

[6]See Dellavigna and Malmendier (2004) for such a model. In our setup, the enticing "free" services can be considered as "leisure goods" that provide immediate benefits to consumers, but imposes delayed costs of privacy loss. Another explanation put forward is that the costs of privacy loss are at best nebulous and intangible, which leads to consumers' consistent underestimation of them.

bargains they are making and optimally choose the course of actions. Even so, we show that each consumer can find it individually rational to accept the sales of their personal data and that in general there is a socially excessive monetizing on personal digital data.

Our key mechanism is negative information externalities where one person's decision to share information can adversely affect others who choose not to share. This is because the data analytics make it possible to infer useful information about people who did not offer personal data from those who disclosed information. One example illustrating such a mechanism is a study by MIT students who showed that men's sexual orientation can be predicted by an analysis of social network sites such as Facebook. This is possible because data analytics reveal that homosexual men have proportionally more gay friends than straight men, which allows to predict men's sexual orientation based solely on the sexualities of their friends.[7] One important takeaway from this example is that "you don't have control over your information"[8] even though you do not divulge of any of your personal information, if other people do.[9]

To illustrate the basic mechanism, we first consider the simple set-up of a monopolist with big data which does not need to sell data to brokerage firms. We identify a new type of distortion that drives the monopolist more likely to adopt the business model to monetize personal data than the social planner who maximizes the social welfare would do. This is primarily because the social marginal cost of serving one extra consumer is higher than the monopolist's marginal cost of doing so in the presence of negative information externalities. The intuition is as followings. When one extra consumer provides her personal information by patronizing the monopolist's web-service requiring the sharing of personal information, it makes all non-consumers' reservation utility going down in the presence of negative information externalities. While the social planner cares about this utility loss to non-consumers, the monopolist does not. Instead, the monopolist cares about the effect of an additional consumer on its ability to extract surplus from existing consumers, which is positive to the monopolist. However, this is no concern to the social planner because it is just a pure transfer. Taken together, the monopolist finds the lower private marginal cost

[7]See Johnson (2009).

[8]Hal Abelson, quoted in the Boston Globe article by Carolyn Y. Johnson.

[9]Sun Microsystems CEO Scott McNealy said plainly back in 1999: "You have zero privacy anyway. Get over it." See the article by Polly Sprenger (26 Jan 1999), "Sun on privacy: 'Get over it,' " *Wired*, http:// archive.wired.com/politics/law/news/1999/01/17538.

to serve the marginal consumer relative to the social planner. Therefore, the monopolist has socially excessive incentives to monetize personal data.

Then, we consider how this negative information externalities work at the level of websites where we consider a continuum of heterogeneous websites who decide the entry and subsequently their business model. Note first that once personal data are sold to the brokerage firms, consumers lose control of their usage. As consumers mostly do not directly deal with data brokerage firms, the process of the usage of the gathered data is completely opaque. Therefore, the brokerage firms (and the advertising firms who buy data from them) have no incentive to internalize the nuisance from its usage to consumers. As a consequence, they will induce excessive usage of personal data. In our model, consumers are fully aware of such excessive usage and hence each website takes it into account when inducing consumer participation. However, we find an equilibrium in which too many websites enter to collect and sell personal data and all consumers consent to data collection and sales of their personal data. As a result, each costumer is worse off than in the benchmark without the data brokerage industry.

This is because once a large amount of data on many consumers are gathered and sold to data brokerage, it generates negative externalities to those consumers who refuse sales of personal data. These negative information externalities relax the individual rationality constraint of each consumer by worsening his reservation utility, with respect to the benchmark of no data brokerage industry. Therefore, even if each consumer finds it individually rational to accept the sales of their personal data, each of them is worse off. Furthermore, websites generate positive externalities among themselves as the higher is the number of websites monetizing personal data, the worse is each consumer's reservation utility, which explains excessive entry. In sum, what drives our result is a coordination failure among consumers generated by negative information externalities. Therefore, consumers collectively end up being hurt by the presence of the brokerage industry. This calls for a policy response that addresses a collective choice problem as in the case of the provision of public goods.

Most academic work to date assumes that the data is already available and possessed by a third party. Bergemann and Bonatti (2015) considers a model of data provision and data pricing in which a single data provider controls a large database about the match value information between individual consumers and individual firms. They analyze the equilibrium data acquisition and pricing policies when such information allows targeted advertising.

Their focus is on the data provider's optimal pricing policy and how the price of data influences the composition of the targeted set, but do not address the issue of privacy and how such database is acquired. Montes, Sand-Zantman, and Valletti (2015) investigate the value of personal information in a Hotelling-type duopoly model. They consider a game in which competing firms can acquire information about consumers' characteristics, which enables them to practice personalize pricing while consumers at the same time can pay a privacy cost to avoid such price discrimination. They analyze the effects of such price discrimination and privacy costs on competition and social welfare. Once again, they just assume that such consumer information is already available from an upstream data supplier and do not consider the mechanism in which such data is collected and aggregated. In our model, we show how the emergence of data brokerage firms can facilitate the collection of such data through websites whose business model is to acquire consumer data.

Bataineha *et al.* (2016) adopt a two-sided market approach to analyze how a data monetizing platform can be used to generate higher profit for both data providers and data consumers compared to other market mechanisms. However, they do not consider privacy issues and information externalities in the provision of personal data.

As in our paper, several legal scholars pointed out the public good nature of privacy and the ineffectiveness of the "notice and choice" approach as a solution to the unwanted privacy invasion. In essence, the notice and choice approach is based on the promise of individual control of disclosure and dissemination of information and thus inadequate to address a collective choice problem. MacCarthy (2011), for instance, emphasizes the concept of privacy externalities that disclosure of personal information by some people reveals information about others. He argues that in the presence of negative information externalities with information leakage by some potentially harming others, reliance on individual consent to determine the collection and usage of personal information will be ineffective. In a similar vein, Fairfield and Engel (2015) characterize privacy as a public good because "[a]n individual who is careless with data exposes not only extensive information about herself, but also others as well." They thus calls for a collective choice approach to address the privacy issue. Our paper formalizes these ideas. We show how the equilibrium with excessive privacy invasion can emerge in the framework of rational consumers who are fully aware of the consequences of their consent due to privacy externalities and how the existence of data brokerage firms can facilitate monetization of collected data when each website is too small to do so independently.

Our research thus has important implications for the recent policy debate regarding data brokerage and privacy. European Commission, for instance, introduced new data protection policies which require website to get consumer approval for transferring personal data to third parties (such as data brokerage firms). Such policy would have some effect on naive consumers by alerting them to be aware of such data transfer. The basic premise of the notice-and-choice approach is to rely on fully informed individuals to make rational decisions. However, the effects of such an individualistic approach may be limited in addressing the negative information externalities problem even if consumers are well-informed and make fully rational decisions. The reason is that privacy in the presence of information externalities is akin to public goods and an effective protection may require a collective choice approach.[10]

The rest of the paper is organized as follows. In section 2, we develop a model of monopolist to illustrate the basic mechanism of negative information externalities. Section 3 extends the analysis to a model with a continuum of small websites. We consider a scenario in which each individual firm's scale of operation is too small to justify the independent use of "data sales" as a business model, yet show how the existence of data brokerage firms can serve as a channel of information aggregation and enables the emergence of data sales as a collective business model. Section 4 discusses implications for optimal privacy policies and section 5 concludes.

2 Monopoly Model

We first consider the simple set-up of a monopolist to illustrate the basic mechanism. There is a mass one of consumers who consume digital products delivered online, simply referred to as 'content' hereafter. Consumers' valuation for the service provided by the monopolist is given by u, which is assumed to be distributed over $[\underline{u}, \overline{u}]$ with distribution function F and density f. The monopolistic content provider can collect consumers' personal data in the process of providing its service, which can be potentially utilized for other purposes. For instance, it can be used for targeted advertising or promotion of other ancillary services.

We assume that consumers incurs a nuisance cost of "privacy loss" when personal information is used for such purposes. There can be many sources of such utility loss. For

[10]Fairfield and Engel (2015) and Schneier (2015) suggest to adopt a similar approach to privacy as the one addressing pollution problems to protect environment. See Hirsch (2006).

instance, there could be direct economic losses due to personalize pricing enabled by the detailed knowledge of personal preferences. We can also think of a variety of psychological reasons for negative feelings about privacy loss. A newly released smartphone app called Google Trips, promoted to provide a "personalized tour guide in your pocket," is a case in point. The modus operandi of this app developed by Google is to "use what it already knows about you, based on data it has collected from your Gmail account, and combines it with established features from its other offerings, like Destinations, and its large database of crowd-sourced reviews," which led a New York Times reviewer for this app to comment that "It's Kind of Creepy."[11] We assume that the extent of knowledge about a consumer depends not only on the information directly collected from the consumer but also on the amount of information revealed by other consumers because the monopolist can do data mining to indirectly infer information about the consumer. In particular, we assume that the monopolist may infer some information about the consumer even if he does not patronize the content provider. For instance, we can imagine that the firm may have from the beginning some information about consumer i, which they obtained from data available from off-line or on-line using public or private sources. Then, they can always find some consumer i' whose personal data matches consumer i (up to the information they have about consumer i). Hence, even if consumer i does not use the service, the fact that there are several consumers similar to i who use the service may allow some inference about consumer i.

We thus represent a consumer's nuisance costs as $\psi_1(m)$ and $\psi_0(m)$ respectively, depending on whether the consumer uses the service or not, where m is the mass of consumers who use the service. The nuisance costs are increasing in m, that is, $\psi_k^{'}(m) > 0$, where $k = 0, 1$, and $\psi_1(m) > \psi_0(m)$ with $\psi_0(0) = 0$. When the personal information collected is utilized, the monopolist can generate additional revenue of $R(m)$, with $R^{'}(m) > 0$. To focus on the consumers' coordination failure and negative externalities in the nuisance costs, we assume that $R(m)$ also represents social benefits. As we are concerned with a digital product/service, the marginal cost of the content is assumed to be zero.

The timing of the game is as follows. At the beginning of the game, the monopolist

[11]To quote, "Before you create your first trip, you'll see some of your previous trips that you didn't even share. That's because it has already pulled in information from your Gmail account, so it knows which hotels you stayed in and where you rented a car from and stores this information under Reservations." See Justin Sablich, "How to Use Google to Plan Your Trip," *New York Times*, September 21, 2016.

commits to a privacy regime with the adoption of a business model of "pure content pricing" or "personal data usage". In the pure content pricing regime, the monopolist commits not to collect personal information or not to use it for advertising or other purpose of ancillary revenue generation including data sales to a third party. In this case, privacy of all consumers is protected and there is no nuisance cost for consumers. In the personal data usage regime, consumers make rational usage decisions with the understanding that his personal information can be used by the monopolist and subsequently he is subject to a nuisance cost.

2.1 Social Optimum

We first analyze a socially optimal outcome as a benchmark in which a social planner chooses a price under incomplete information: the planner is not allowed to make perfect price discrimination.[12] Let u denote the cutoff type of consumers such that all consumers whose valuation exceeds or equal to u use the service. The mass of consumers who use the service is given by $m = 1 - F(u)$. Welfare given a cutoff type u is given by

$$W(u) = \int_u^{\overline{u}} x dF(x) + R(1 - F(u)) - (1 - F(u))\,\psi_1(1 - F(u)) - F(u)\psi_0(1 - F(u)).$$

The welfare-maximizing cutoff type u can be derived by the following first order condition.

$$-uf(u) - R'f(u) + (1 - F(u))\,\psi_1' f(u) + \psi_1 f(u) - f(u)\phi_0 + F(u)\psi_0' f(u) = 0,$$

which is equivalent to

$$u + R' = (\psi_1 - \psi_0) + (1 - F(u))\,\psi_1' + F(u)\psi_0'. \tag{1}$$

The RHS of (1) represents the social marginal cost (SMC) of nuisance when additional user joins the customer base. There are three channels through which SMC is affected when an additional user starts to use the content service. First, the marginal consumer's status change from a non-consumer to a consumer directly affects his nuisance cost by $(\psi_1 - \psi_0)$. Additionally, a new consumer inflicts externalities not only on the consumer group he joins, but also on the non-consumer group he leaves behind. The nuisance cost of an existing

[12] However, the main insight can be obtained even if we assume perfect price discrimination both for the social planner and the monopolist.

consumer changes by ψ_1' as a new consumer joins; thereby the aggregate change for the consumer group is equal to $(1 - F(u))\psi_1'$. In addition, the nuisance cost of an existing non-consumer also changes by ψ_0' with the aggregate effect being $F(u)\psi_0'$.

2.2 Monopoly

We now derive the monopolist's optimal regime choice and price.

2.2.1 Personal Data Usage Regime

Given the monopolist's price p, let u be the cutoff type of consumers who are indifferent between using the service or not. For the cut-off type u, the IR can be written as

$$(IR : u) \quad u - p - \psi_1(1 - F(u)) \geq -\psi_0(1 - F(u)),$$

where $-\psi_0(1-F(u))$ is the reservation utility of type u consumer. Therefore, the monopolist will solve the following problem:

$$\underset{u}{Max}\ \Pi(u) = (1 - F(u))\left\{u - [\psi_1(1 - F(u)) - \psi_0(1 - F(u))]\right\} + R(1 - F(u)).$$

The first order condition for profit maximization is

$$(1 - F(u))[1 + (\psi_1' - \psi_0')f(u)] - f(u)[u - (\psi_1 - \psi_0)] - R'f(u) = 0.$$

This is equivalent to

$$\left[u - \frac{(1 - F(u))}{f(u)}\right] + R' = (\psi_1 - \psi_0) + (1 - F(u))(\psi_1' - \psi_0'). \tag{2}$$

Note that if consider the standard monopoly model without additional source of revenue from personal data usage and nuisance costs, that is, $\psi_1(m) = \psi_0(m) = R(m) = 0$,condition (2) reduces to the standard monopoly condition:

$$u - \frac{(1 - F(u))}{f(u)}(\equiv u^v) = 0,$$

where the LHS represents the virtual type u^v, which is assumed to be increasing with u.

For the monopolist in our model, the LHS of (2) is changed into $R' + u^v$ to reflect the additional revenue R'. The RHS of (2) represents the private marginal cost (PMC) of nuisance.

A comparison of PMC in (2) with SMC in (1) indicates that there is another source of distortion in our model. More specifically, the difference between these two is given by $F(u)\psi_0' + [1 - F(u)]\,\psi_0' = \psi_0'$. That is, we have

$$SMC - PMC = \psi_0' > 0.$$

This new type of distortion we identify can be explained in the following way. When one extra consumer is served and his data adds to the monopolist's database, it inflicts additional negative externality to $F(u)$ measure of non-consumers even though they do not use the monopolist's content. This effect on non-consumers' reservation utility is $F(u)\psi_0'$. While the social planner cares about this negative externality, the monopolist does not because they are non-consumers. Instead, the monopolist cares about the effect of an additional consumer on its ability to extract surplus from existing consumers. However, this is no concern to the social planner because it is just a pure transfer. More specifically, in order to induce one more additional consumer the monopolist's price needs to be adjusted below by $(\psi_1' - \psi_0')$ to compensate the differences in the nuisance cost change. Note that as additional consumer also negatively affects non-consumers and reduces the reservation value of the marginal consumer, the price compensation needs to be only $(\psi_1' - \psi_0')$, not ψ_1'. As a result, the negative profit impact via a reduced price to the inframarginal consumers is given by $(1 - F(u))(\psi_1' - \psi_0')$ whereas the social planner only cares about the real impact on the inframarginal consumers which is $(1 - F(u))\psi_0'$. This creates an additional difference of $(1 - F(u))\psi_0'$; the extent to which the reservation utility of the marginal consumer is reduced with an additional consumer (ψ_0') represents more ability to extract surplus from consumers for the monopolist in the amount of $(1 - F(u))\psi_0'$, but this is a transfer which does not figure in the social planner's calculus. In other words, the social planner cares about the real nuisance effect of the addition of a marginal consumer on *non-consumers* (with a measure of $F(u)$) while the monopolist cares cares only about its ability to extract surplus from *consumers* (with a measure of $1 - F(u)$) through its effect on the marginal consumer's willingness to pay by reducing the reservation utility. Taken together, the total difference between SMC

and PMC becomes $F(u)\psi_0' + (1 - F(u))\psi_0' = \psi_0'$. Thus, this type of distortion leads to the monopolist to serve too many consumers and the extent to which the monopolist's decision departs from the social planner's depends on the additional consumer's impact on the reservation utility. The effect of this distortion is in the opposite direction of the standard monopoly result that the monopolist serves too few consumers. The overall effect thus depends on the relative magnitudes of these two opposing effects. If the negative externality effect of making the monopolist to serve more than socially optimal number of consumers is greater than the standard monopoly distortion, too many consumers can be served by the monopolist compared to the socially efficient level.

Let u^* and u^{FB} be the monopoly cutoff and the first-best cutoff types, respectively, and $m^* \equiv 1 - F(u^*)$ and $m^{FB} \equiv 1 - F(u^{FB})$ be the respective ly corresponding measures of consumers served. Then, we have the following proposition that summarize thus far analysis.

Proposition 1 (Monopolist vs. Social Planner)

(i) The monopolist serve more consumers than the social planner, i.e., $u^* < u^{FB}$ (or, $m^* > m^{FB}$) if and only if

$$\frac{1 - F(u^{FB})}{f(u^{FB})} < \psi_0'(1 - F(u^{FB})).$$

(ii) The monopolist serves all consumers while the social planner does not if

$$[\psi_1(1) - \psi_0(1)] + [\psi_1'(1) - \psi_0'(1)] + \frac{1}{f(\underline{u})} < R'(1) + \underline{u} < [\psi_1(1) - \psi_0(1)] + \psi_1'(1)$$

Proposition 1 (ii) requires a necessary condition of $\psi_0'(1) > \frac{1}{f(\underline{u})}$. This means that the stated result will be obtained under a sufficiently low reservation utility due to the negative marginal externality.

To illustrate the result, consider a simple parametric example in which u is uniformly distributed on $[0, 1]$ and $R(m) = rm$ with $\psi_1(m) = \kappa m$ and $\psi_0(m) = \xi\kappa m$,where $\xi \in (0, 1)$. Later we will provide more micro-foundation for these nuisance costs from a CES type functional form. In this case, the SMC and PMC are respectively given by

$$\begin{aligned} SMC &= \kappa[2(1 - \xi)m + \xi] \\ PMC &= \kappa[2(1 - \xi)m], \end{aligned}$$

with $SMC - PMC = \psi_0' = \kappa\xi$.

Then, the socially optimal level of consumption and the equilibrium level of consumption are characterized by

$$\begin{aligned} m^{FB} &= \frac{1+r-\xi\kappa}{1+2\kappa(1-\xi)}; \\ m^{*} &= \frac{1+r}{2[1+\kappa(1-\xi)]} \end{aligned}$$

We have $m^* > m^{FB}$ if and only if $2\xi\kappa[1+\kappa(1-\xi)] > 1+r$. As is clearly seen from the explicit expressions of m^* and m^{FB}, if $\xi = 0$, we have $m^* < m^{FB}$ due to the standard monopoly distortion. However, for a sufficiently high $\kappa\xi$, the opposite result can be obtained, which is also confirmed by the fact that the LHS, $2\xi\kappa[1+\kappa(1-\xi)]$, is increasing in κ and ξ for all $\xi \in (0,1)$.

2.2.2 Pure Content Pricing Regime

In the pure content pricing regime in which privacy is protected and no personal data is utilized, we have the standard result, where the virtual type is equalized to marginal cost, which is zero.

$$u - \frac{(1-F(u))}{f(u)} (\equiv u^v) = 0$$

Let the solution to the above problem be denoted as $\widetilde{u}^*$ and the corresponding number of consumers as $\widetilde{m}^* = 1 - F(\widetilde{u}^*)$. Then, the monopolist's maximized profit with data collection is given by

$$\widetilde{\Pi}^* = \widetilde{m}^* F^{-1}(1-\widetilde{m}^*)$$

2.3 Monopolist's Choice of Business Model

Suppose that the monopolist can choose between pure content pricing model with no data collection and data collection model (we may extend to analyze how much private data to utilize in an agreement with users, as the case of intensity of usage analysis, but for now, let us it fixed).

As we already analyzed in Subsection 1.2.1, the profit maximizing choice of u, denoted by u^*, is determined by (2). Recall that the corresponding number of consumers as $m^* =$

$1 - F(u^*)$ and the monopolist's maximized profit with data collection is given by

$$\Pi^* = m^*\{F^{-1}(1-m^*) - [\psi_1(m^*) - \psi_0(m^*)]\} + R(m^*)$$

Now suppose that the monopolist adopts the pure content pricing business model without collecting any information or by committing not to utilize private information for any other purposes. As we analyzed in Subsection 2.2.2, the monopolist's optimal choice $\widetilde{u}^*$ is determined by $u^v = 0$ the monopolist's maximized profit with data collection is given by

$$\widetilde{\Pi}^* = \widetilde{m}^* F^{-1}(1-\widetilde{m}^*).$$

Lemma 1 *A sufficient condition for the monopolist to choose the business model of data collection over pure content pricing is that* $R(m) - m\left[\psi_1(m) - \psi_0(m)\right] \geq 0$ *when it is evaluated at* $m = \widetilde{m}^*$.

Proof. *By the revealed preference argument, we have*

$$\begin{aligned}
\Pi^* &= m^* F^{-1}(1-m^*) + \{R(m^*) - m^*\left[\psi_1(m^*) - \psi_0(m^*)\right]\} \\
&\geq \widetilde{m}^* F^{-1}(1-\widetilde{m}^*) + \{R(\widetilde{m}^*) - \widetilde{m}^*\left[\psi_1(\widetilde{m}^*) - \psi_0(\widetilde{m}^*)\right]\} \\
&= \widetilde{\Pi}^* + R(\widetilde{m}^*) - \widetilde{m}^*\left[\psi_1(\widetilde{m}^*) - \psi_0(\widetilde{m}^*)\right]
\end{aligned}$$

Therefore, if $R(\widetilde{m}^*) - \widetilde{m}^*\left[\psi_1(\widetilde{m}^*) - \psi_0(\widetilde{m}^*)\right] \geq 0$, *we have* $\Pi^* \geq \widetilde{\Pi}^*$. ■

For now, let us solve a parametric model. Assume our canonical parametric example in which $\psi_1(m) = \kappa m$, $\psi_0(m) = \xi\kappa m$, $R(m) = rm$, and u is uniformly distributed on $[0, 1]$. Then, we can easily derive that

$$\begin{aligned}
m^* &= \frac{1+r}{2[1+\kappa(1-\xi)]} \text{ and } \widetilde{m}^* = \frac{1}{2} \\
\Pi^* &= \left(\frac{1+r}{2[1+\kappa(1-\xi)]}\right)^2 \text{ and } \widetilde{\Pi}^* = \frac{1}{4}
\end{aligned}$$

From the comparison between Π^* and $\widetilde{\Pi}^*$, the pure content pricing will be used if and only if $\kappa(1-\xi) > r$. This is when the consumer's marginal loss in privacy cost (and thus the compensation needed to make up for the loss) exceeds that consumer's marginal revenue to the monopolist. The social planner's optimal use of data condition is given by m^{FB}. We

already know that $m^* > m^{FB}$ if $2\xi\kappa[1+\kappa(1-\xi)] > 1+r$. Thus, if the following condition holds

$$\kappa(1-\xi) < r < 2\xi\kappa[1+\kappa(1-\xi)] - 1,$$

the monopolist adopts a data collection model and there will be too much data collection and loss of privacy. Here we can interpret κ as a scale parameter. We normalize the number of consumers at 1, but κ can be reinterpreted as the size of market. There will be such r that satisfies the above condition if $2\xi\kappa > 1$. This means that we will have too much data collection and loss of privacy as long as the marginal externality intensity ξ and/or the size of market κ is large enough. Even if this condition as of today may not be satisfied, as the data mining advances further (so that ξ increases enough) our society will suffer more from the data brokerage firms.

Below we illustrate when the data collection by the monopolist is adopted over the pure content pricing model and when such a choice is socially inefficient. In the space of (ξ, κ), we have total four sets I, II, III and IV depending on the business model choice and on the socially excessive data collection or not. Set I denotes the parameter constellation of (ξ, κ) where the monopolist adopts the data collection model but such choice is not leading to socially excessive privacy loss. However, Set II captures the situation in which such data collection is socially harmful. This occurs when ξ is large enough while k is not that high. This is because the higher k means the greater Spence distortion so that the monopolist ends up not serving enough consumers from the welfare perspective. Set III and Set IV denote the parameter set where the monopolist uses only the content price, with commitment if possible to no data collection. Set IV represents the set of excessive data collection if the monopolist had adopted the data monetizing business model.

Many websites that adopt the business model of data sales offer their services for free. This may be explained if we introduce some marginal cost $c > 0$ of billing and maintaining accounts. If the monopolist charges a zero price for its content, the marginal type of consumer is defined by $u^z = [\psi_1(m^z) - \psi_0(m^z)]$ and the monopolist's profit is simply $\Pi^z = R(m^z)$, where $m^z = 1 - F(u^z)$. If $\Pi^z \geq \max[\Pi^*, \widetilde{\Pi}^*]$, the monopolist's business model is to provide free content in exchange for personal data and derive all revenues from targeted advertising, i.e., $R(.)$. This can explain the prevalence of websites providing free services.[13] In equilibrium,

[13]In our parametric example, this condition can be written as $\frac{r}{1+(1-\xi)\kappa} \geq [\left(\frac{1+r-c}{2[1+\kappa(1-\xi)]}\right)^2, \frac{1}{4}]$.

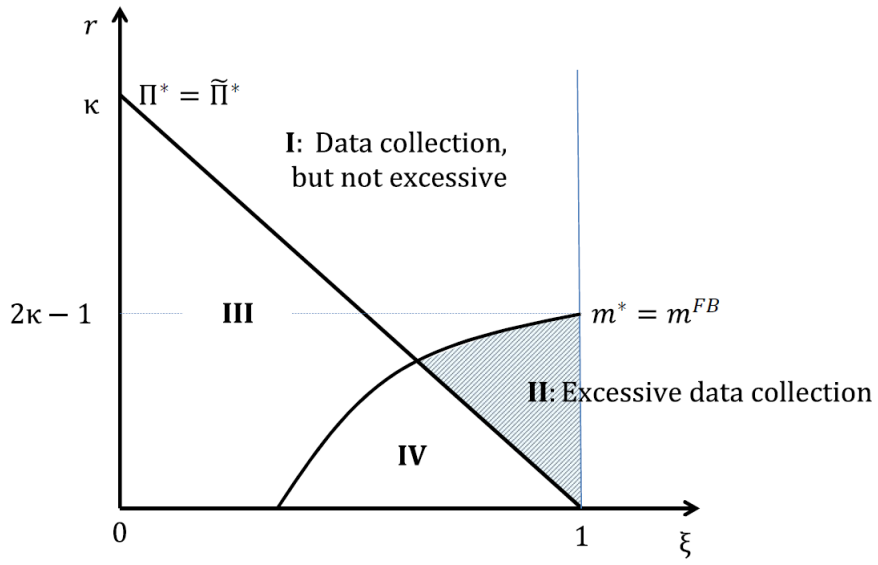

Figure 1: Monopolist's choice of business model and social (in)efficiency collection

too many consumers are served, which is equivalent to too many consumers giving up their privacy: too much personal data are collected and used.

3 Data Brokerage Firms and Big Data

In the previous section, we considered a monopoly website and its incentives to collect persona data as a business model. We showed the monopolist website may have excessive incentives to collect personal data with the resulting loss of privacy for consumers compared to the social optimum. The main mechanism responsible for this outcome was the gap between the monopolist's private marginal cost of serving one more consumer and the social marginal cost of doing so, which is due to negative privacy externalities.

In this section, we consider an alternative mechanism at the level of web sites. As a building block, we consider monopolistic websites that has no incentives to collect data alone due to its small customer base. We show that the emergence of data brokerage firms that purchases and aggregate data from websites can restore incentives to collect data.

Consider there is a mass one of monopolistic websites in their own market niche and a mass one of homogeneous consumers. Consumers multihome. Websites are heterogeneous in terms of the value that their content generates to consumers. Let v denote the value of a website's content to each homogeneous consumer. v follows a distribution function G with density g over the interval $[\underline{v}, \overline{v}]$.with $\underline{v} > 0$. We assume that all websites have the same fixed

cost of entry $K > 0$.

We assume that in the absence of brokerage industry, each website has no incentive to collect data since the scale of their data is too small and thus adopts the pure content pricing model.

Now consider the presence of data brokerage firms who can aggregate data from individual websites and use big data to better utilize collected information. Let $R(n)$ denote the aggregate revenue of the data brokers where n is the measure of websites who feed the personal information about their users to data brokers when all consumers in each monopolistic market use the corresponding website. Let $\psi(n)$ denote the aggregate nuisance to the consumers in such a case. We assume that both the revenue and the nuisance increase with n, $R^{'}(n) > 0$.and $\psi'(n) > 0$.

We also assume some scale economy in data brokerage such that

$$R(\varepsilon) < \psi(\varepsilon) \text{ for } \varepsilon(> 0) \text{ small enough.} \tag{3}$$

Hence, if no other website sells data, a single website has no incentive to collect and sell personal data instead of adopting a pure content pricing model. For expositional clarity, we will often consider a simple case in which both $R(n)$ and $\psi(n)$ are linear with $R(0) = \psi(0) = 0$. Then, the above assumption implies that $R(n) < \psi(n)$ for any n. However, we do not need such strong assumption and our results are also obtained when $R(n) > \psi(n)$ for some interval of n.

We consider a three-stage game with the following timing.

- Stage 1: each website simultaneously decides whether to incur the fixed cost of entry $K(> 0)$.
- Stage 2: The websites which entered simultaneously decide their business model (including a price) and make offers.
- Stage 3: Consumers decide which websites to use among those already entered and made an offer.

3.1 Competition when all websites monetize data

Consider stage 2 and suppose that all websites of measure n use the business model of data sales (supplemented with a content pricing). We below characterize the equilibrium in which every consumer uses all the websites.

Let j and k represent two different websites with v^j and v^k. Let p^j and p^k the content price they charge. Since all websites are identical in terms of the nuisance they generate from data sales, the following equation must be satisfied for any pair of (j,k) in equilibrium:

$$v^j - p^j = v^k - p^k (\equiv v(n)),$$

Hence, a consumer's payoff in equilibrium is $nv(n) - \psi(n)$.

Suppose that all consumers consume all websites and consider the deviation of consumer i. Let $\psi(n', n)$ represent the nuisance of consumer i depending on the measure $n' \leq n$ of websites he consumes given that all the other consumers consume all the websites of measure n. We assume that $\psi(n', n)$ is increasing in each element. For instance, $\psi(0, n)$ represent the nuisance he experiences even if he does not consume any website. We have $\psi(n) = \psi(n, n)$ and $\psi(0, 0) = 0$.

Then, a necessary condition for $v(n)$ to constitute an equilibrium is that the following incentive constraint is satisfied for any $n' < n$:

$$(IC : n', n) \quad nv(n) - \psi(n, n) \geq n'v(n) - \psi(n', n) \quad \text{for any } n' \leq n.$$

The RHS of the inequality represents the deviation payoff and its first term is linear in n'. If $\psi(n', n)$ is concave, then RHS is convex in n' and its maximum is attained either at $n' = 0$ or at $n' = n$. Therefore, the incentive constraint is satisfied for any $n' \leq n$ if $(IC : 0, n)$ is satisfied:

$$nv(n) - \psi(n, n) \geq -\psi(0, n).$$

Of course, in equilibrium, the above inequality must be satisfied with equality: otherwise, each website has an incentive to raise a bit its price. Therefore, the consumer will have no unilateral incentives to deviate if

$$v(n) = \frac{\psi(n, n) - \psi(0, n)}{n}.$$

In summary, conditional on that all websites of measure n adopt the business model of data sales, the equilibrium prices are such that for any pair of firms (j,k)

$$v^j - p^j = v^k - p^k = \frac{\psi(n,n) - \psi(0,n)}{n}.$$

In the equilibrium, every consumer consumes all the websites.

In addition, we show in Appendix A that the price that each website will receive from selling its data to the brokerage market is equal to $R'(n)$. Namely, we provide a microfoundation for this result by assuming competition among any given number of symmetric brokerage firms.

Then, each website j's equilibrium payoff is given by

$$R'(n) + v^j - \frac{\psi(n,n) - \psi(0,n)}{n}.$$

Now, consider the deviation of website j at stage 2 by adopting a pure content pricing business model. Upon such deviation, the payoff of website j is independent of other websites and is given by v^j. Thus, all websites adopting the data sales business model is an equilibrium if

$$R'(n) \geq \frac{\psi(n,n) - \psi(0,n)}{n}.$$

We thus have the following proposition.

Proposition 2 *Suppose that n measure of websites entered at stage 1. If $\psi(n',n)$ is concave in n' and $R'(n) \geq \frac{\psi(n,n)-\psi(0,n)}{n}$, there is an equilibrium in which all websites adopt the business model of selling data to the data brokerage firm. In the equilibrium,*

(i) each website j with the content value v^j charges

$$p^j = v^j - \frac{\psi(n,n) - \psi(0,n)}{n};$$

(ii) each website j's profit is $R'(n) + v^j - \frac{\psi(n,n)-\psi(0,n)}{n}$.

3.2 Free Entry

We have conducted an analysis the game that starts at stage 2 when a fixed measure of websites n entered the market at stage 1. We now study stage 1 by making the entry

endogenous. Let n^* be the equilibrium number of websites. Then, the marginal website's value to consumers v^* is given by $1 - G(v^*) = n^*$. This implies that in the first stage, the extent of entry is determined by the following conditions.

$$G^{-1}(1 - n^*) - \frac{\psi(n^*, n^*) - \psi(0, n^*)}{n^*} + R'(n^*) = K \tag{4}$$

Let us consider the optimal number of entrants from the social planner's viewpoint. Given the marginal cutoff type of entrant v, social welfare can be written as

$$W(v) = \int_v^{\overline{v}} x dG(x) + R(1 - G(v)) - \psi(1 - G(v)) - ((1 - G(v))K$$

The welfare-maximizing cutoff type v can be derived by the following first order condition.

$$-vg(v) - R'g(v) + \psi'(1 - G(v))g(v) + Kg(v) = 0,$$

which is equivalent to

$$v + R' - \psi'(1 - G(v)) = K$$

Let v^{FB} be the cut-off value in the first best outcome and let $n^{FB} = 1 - G(v^{FB})$. Then, the condition for social optimum can be characterized by

$$G^{-1}(1 - n^{FB}) + R'(n^{FB}) - \psi'(n^{FB}) = K \tag{5}$$

The comparison of (4) and (5) reveals that the comparison of socially optimal number of websites and the market equilibrium boils down to the relative magnitudes of $\psi'(n)$ and $\frac{\psi(n,n)-\psi(0,n)}{n}$. For instance, if ψ is convex, ψ' is increasing, which implies that $\psi'(n) \geq \psi(n, n)/n$. Therefore, as long as $\psi(0, n^{FB}) > 0$, we have socially excessive entry.

Summarizing, we have:

Proposition 3 *There is an excessive entry of websites (i.e., $n^* > n^{FB}$) if the following condition holds*

$$\frac{\psi(n^{FB}, n^{FB}) - \psi(0, n^{FB})}{n^{FB}} < \psi'(n^{FB}). \tag{6}$$

If $\psi(n)$ is convex, there is excessive entry of websites.

To focus on the main driving force, consider the case in which $\psi(n)$ is linear. Then,

we have $\psi(n^{FB}, n^{FB})/n^{FB} = \psi'(n^{FB})$ and therefore, the inequality (6) is always satisfied. $-\psi(0, n)$ represents the reservation utility of a consumer who does not use any website when all other consumers use all websites. Suppose that initially n^{FB} measure of websites entered. This reduces the reservation utility of a non-user from $-\psi(0, 0) = 0$ to $-\psi(0, n^{FB})$. Therefore, each marginal website can extract more than its social contribution by $\psi(0, n^{FB})/n^{FB}$. This implies $n^* > n^{FB}$. In other words, the entry of some websites generate positive externalities to other websites who are contemplating their entry by worsening consumers' reservation utility, which generates socially excessive entry.

Note that the mechanism for our excessive entry result is very different from the standard business-stealing effect of Mankiw and Whinston (1986). In our setup, there is no room for business stealing because we assumed that each website market is segmented and each website enjoyed complete monopoly power in its niche market. The excessive result in our model is coming from the negative information externalities, namely, each entrant's effect on consumers' reservation utility through the privacy channel.

To illustrate our results, consider the following parametric example in which we assume a CES nuisance cost of $\psi(x, n) = \kappa[\alpha x^\rho + (1-\alpha)n^\rho]^{\frac{1}{\rho}}$, where $0 < \alpha < 1$, $\rho < 1$, and κ is a scale parameter. It can be easily verified that $\psi(x, n)$ is concave in x.[14] Note that this functional form implies that $\psi(n, n) = \psi(n) = \kappa n$ and $\psi(0, n) = \xi\kappa n$, where $\xi = (1-\alpha)^{\frac{1}{\rho}}$ and $0 < \xi < 1$. As usual, with this CES nuisance cost function, α is the share parameter and ρ determines the degree of substitutability of one's own personal data and other people's data where the elasticity of substitution is given by $\sigma = \frac{1}{1-\rho}$. In the extreme, if $\rho = 1$,we have a perfect substitute case in terms of the nuisance cost as a consumer's own data can be perfectly substituted by other people's data. If $\rho = -\infty$, they are perfect complements.

With this parametric example, there is an equilibrium in which all websites adopt the data sales model if

$$r \geq \kappa(1-\xi)$$

[14]The CES nuisance cost function means the equal percentage response of the relative marginal nuisance costs of x and n to a percentage change in the ratio of their quantities. For example, consider two different consumers who respectively use 10% and 20% smaller number of websites relative to all other consumers. Suppose that the marginal nuisance cost saved by using 10% smaller websites is 5%. Then, the marginal nuisance cost saved by using 20% smaller websites must be 10%.

In the free-entry equilibrium, we also need to have

$$r + \underbrace{u^* - \kappa(1-\xi)}_{p(n^*)} = K$$

for the marginal entrant type. Using $u^* = 1 - n^*$ under the uniform distribution over $[0, 1]$, we have

$$n^* = 1 - \kappa(1-\xi) + r - K$$

In contrast, the socially optimal number of entrants is given by

$$r + u^{FB} - \kappa = K$$

which is equivalent to

$$n^{FB} = 1 - \kappa + r - K$$

Note that if $r < \kappa$, the revenue is smaller than the nuisance such that monetizing personal data is socially undesirable. However, if $r > \kappa(1-\xi)$, all websites adopt the business model of monetizing personal data and there is an excessive entry of such websites. The excessive entry region with excessive data usage is represented in Figure 2.

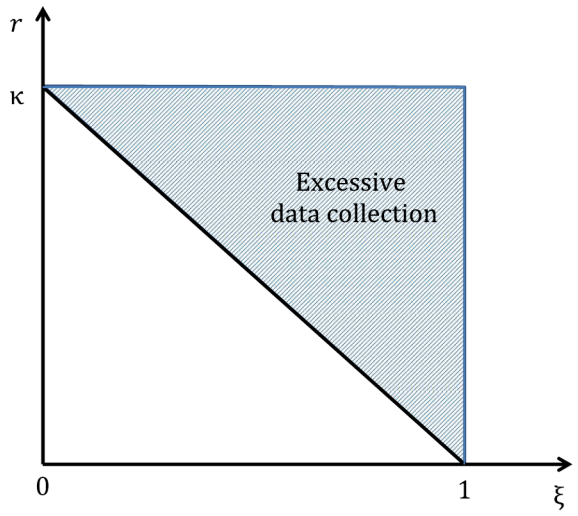

Figure 2: Free entry model and excessive data collection

4 Policy Implications

Our model of negative privacy externalities has implications for evaluating various policy proposals and the design of optimal policies concerning privacy. In particular, it suggests ineffectiveness of the current policy framework of the "informed consent model." (MacCarthy, 2011). The informed consent model is based on the premise that an individual's informed consent provides legitimacy for any information collection and use practice. This model is intuitively appealing because it allows data subjects to control information about themselves and to make decisions according to their own preferences. Yet, there have been wide criticisms against the informed consent model as the privacy notices are rarely read, and even if read, not easy to understand. Instead of these traditional criticisms, we provide a theoretical foundation for why the informed consent model cannot be effective under negative privacy externalities. In our model, even costless reading of all privacy policies would not change any behaviors of data subjects and data collectors.

Another policy implication drawn from our model is that we need to examine the dynamics of negative privacy externalities when constructing a policy remedy. As a case in point, let us consider the recent event in Germany where a regulator ordered Facebook to stop collecting and storing data on its subsidiary messaging service WhatsApp users. The regulator even asked Facebook to delete all information already forwarded (about 35 million German users) from WhatsApp.[15] According to the article, the German regulator asserted that 'neither the internet messenger nor Facebook had received individuals' permission to share the information and had potentially misled people over how their data would be used in the future.' Now suppose that a substantial fraction of users agree to the proposed data merger because they expect to receive a better service from combined data. Then, even other users who initially had not given their consent may well find it better to give their consent as well in the presence of negative privacy externalities. While we provide a static model in this paper, the interpretation can be extended into this dynamic setting. Our model shows that how the informed consent model can be vulnerable to dynamic acceleration process though negative privacy externalities.

A slightly different angle to view the case of Facebook and WhatsApp is when consumers

[15]Mark Scott, "Facebook Ordered to Stop Collecting Data on WhatsApp Users in Germany" *New York Times* Sept. 28, 2016. page B6.

would consider the data integration of multiple services welfare-enhancing instrument and when they would find it detrimental. We note that our model implicitly assumes that the data collectors cannot commit to the use of personal information as a welfare-enhancing tool. This was innocuous assumption for our primary research goal. Now let us relax this assumption to consider a possibility that a monopoly data collector with good reputation has the lower nuisance cost compared to a data collector with bad reputation. Then consider the following remedy: the regulator forces the monopoly to provide an option to consent on the usage of personal information *without integrating information from both services at the same price* as the option for the usage of integrated personal information. In Appendix B, we show that we can construct an equilibrium where the only good reputation firm would obtain the consent to integrated information whereas the one with bad reputation cannot obtain the consent to merge the two different information sources. Therefore, the proposed policy remedy may induce consumers to coordinate on the better option according to the level of data collectors' reputation. This kind of remedy can be applied to data sales with many websites. The regulator should force a website to offer an option to not to agree on data sales at the same price as the option to agree on data sales. Then, whenever consumers are concerned about data sales to some dubious source, it is a strictly dominant strategy to choose the option of no data sales. Therefore, the remedy allows consumers to coordinate on the best option. As a response, any website who wants to sell data will have incentives to find a data broker with reputation not to abuse the delivered data.

5 Concluding Remarks

At this information age, our life-style is heavily dependent on all sorts of computerized devices such as computers, laptops, tablets and mobile phones of which the use is constantly producing data. Such data become so valuable that so-called data broker industry has been fast-growing. Numerous websites and applications provide their content for free or at a highly subsidized price in exchange for the users' data agreement. And, many data brokers are willing to purchase such personal data and sell them to many advertising and marketing companies who need the data for advertising advantages. As such, the harms and costs to individuals and society become a main concern in the context of privacy loss. However, one puzzling aspect behind all the debate is that many consumers are voluntarily giving con-

sent to almost arbitrary use of their personal information by websites and content providers despite their concerns about privacy loss.

To address this problem, in this paper we provide a model of privacy based on the idea of negative privacy externalities. Even if data collection requires consumers' consent and consumers are fully aware of the consequences of such consent, we show that the market equilibrium is characterized by excessive collection of personal information and the loss of privacy by consumers compared to the social optimum. Therefore, we find that the current main privacy regulatory framework of the informed consent model may be ineffective to address the privacy concerns associated with the data broker industry.

To quote Schneier (p.238), "[d]ata is the pollution problem of the information age, and protecting privacy is the environmental challenge." As the pollution problem of the industrial age challenges us economists to come up with various policies—either market-oriented mechanisms or direct regulations—we now need to take a similar approach to the data surveillance problem. As pollutants have negative externalities and any preventive efforts such as abatement have the public good problem, the privacy protection in this big data world also requires solid understanding of information externalities and the privacy as the nature of public good. We hope that our research provides a first-step in this direction of economics research.

References

[1] Ahmed Saleh Bataineh, Rabeb Mizouni, May El Barachi, and Jamal Bentahar. Monetizing personal data: A two-sided market approach. *Procedia Computer Science*, 83:472–479, 2016.

[2] Dirk Bergemann and Alessandro Bonatti. Selling cookies. *American Economic Journal: Microeconomics*, 7(3):259–294, 2015.

[3] US Senate Commerce Committee. A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes. December 2013.

[4] Joshua AT Fairfield and Christoph Engel. Privacy as a public good. *Duke LJ*, 65:385, 2015.

[5] Dennis D Hirsch. Protecting the inner environment: What privacy regulation can learn from environmental law. *Georgia Law Review*, 41(1), 2006.

[6] Carolyn Y. Johnson. Project 'gaydar': At mit, an experiment identifies which stu- dents are gay, raising new questions about online privacy. Boston Globe, September 20 2009. available at http://archive.boston.com/bostonglobe/ideas/articles/2009/09/20/project gaydar an mit experiment.

[7] Mark MacCarthy. New directions in privacy: Disclosure, unfairness and externalities. *ISJLP*, 6:425, 2010.

[8] N Gregory Mankiw and Michael D Whinston. Free entry and social inefficiency. *The RAND Journal of Economics*, pages 48–58, 1986.

[9] Rodrigo Montes, Wilfried Sand-Zantman, and Tommaso M Valletti. The value of personal information in markets with endogenous privacy. 2015.

[10] Bruce Schneier. *Data and Goliath: The hidden battles to collect your data and control your world*. WW Norton & Company, 2015.

[11] Eleanor Singer, John Van Hoewyk, Roger Tourangeau, Darby M Steiger, Margrethe Montgomery, and Robert Montgomery. Final report on the 1999-2000 surveys of privacy attitudes. 2001.

[12] James Waldo, Herbert Lin, and Lynette I Millett. *Engaging privacy and information technology in a digital age*. National Academies Press, 2007.

A Appendix: A microfoundation of R(m)

Since the data broker market is extremely hidden from public knowledge in terms of their market structure, revenue and cost structure, and business practices, we adopted a reduce form approach by considering $R(m)$ without further description of how it is determined. In this appendix, let us provide one particular micro-foundation to determine how much each website would obtain the revenue from the data sales to the data broker market. For this purpose, let us consider a simplest setting that there are n symmetric data brokers. Each broker has a revenue function $B(m)$ where m is the measure of websites. Assume $B' > 0$ and $B'' \leq 0$. Then, we can establish the following lemma:

Lemma 2 *There is an equilibrium in which each brokerage firm proposes a price per website equal to* $B'(m/n)$.

Proof. *If every firm proposes the same price, then each firm gets a profit of* $B(m/n) - B'(m/n)m/n > 0$. *Now consider a firm's deviation. It has no incentive to propose a lower price as it is not going to obtain any data. It has no incentive to propose a higher price. Upon the deviation, it attracts all website and the upper bound of its profit is*

$$\begin{aligned} & B(m) - B'(m/n)m \\ = \; & B(m/n) - B'(m/n)m/n + m(n-1)/n \left\{ \frac{B(m) - B(m/n)}{m(n-1)/n} - B'(m/n) \right\} \\ < \; & B(m/n) - B'(m/n)m/n \end{aligned}$$

where the inequality is from the fact that the bracket term is negative if B *is strictly concave and zero if* B *is linear.* ■

There are several remarks following from the lemma. First, even if one allows for a deviation in which the deviating broker proposes a higher price but limits the offer to a certain first–arrived websites, then there will be no profitable deviation. This is because, by charging a lower price, that broker cannot attract any website and by charging a higher price (say as close as $B'(m/n)$), attracting more than m/n websites leads to a lower profit.

Second, this implies that there is no other symmetric NE in which all firms charge a price lower than $B'(m/n)$). Third, there may exist another symmetric NE with prices higher than $B'(m/n)$). However, the equilibrium in lemma will be Pareto-superior to any other symmetric equilibrium from brokerage firms' point of view.

We have $R(m) = nB(m/n)$. And $R'(m) = B'(m/n)$. Therefore, each website gets a profit of $R'(m)$. Let us first provide a condition such that if no other website sells data, a single website has no incentive to sell data. Suppose that the website is the only who sells data. Let ε the amount of its data; its profit from data sales is approximately $R'(\varepsilon)\varepsilon$. An individual consumer's IR constraint requires

$$u - p - \psi_1(\varepsilon) \geq \psi_0(\varepsilon)$$

Therefore the website has the overall profit of

$$R'(\varepsilon)\varepsilon + u - [\psi_1(\varepsilon) - \psi_0(\varepsilon)] .$$

Hence, we need to assume

$$R'(\varepsilon)\varepsilon + \psi_0(\varepsilon) < \psi_1(\varepsilon) \text{ for } \varepsilon \text{ small enough.}$$

Basically, we expect $\psi_0(\varepsilon)$ is zero: a single website has no impact on the outside option. Hence, a sufficient condition is

$$R(\varepsilon) < \psi_1(\varepsilon) \text{ for } \varepsilon \text{ small enough.}$$

B Appendix: Reputation-Based Remedy Proposal with the Same Price Constraint

In this Appendix let us provide a potential policy remedy that expands consumer choices and induces the monopoly data collector to build its reputation not to abuse its big data against consumers. The key idea underlying this proposal is as follows. When websites can offer different prices, bad websites can induce an equilibrium (offering a lower price) in which consumers allow integration of information and end up being worse off. With one price,

consumers have no incentives to allow integration of websites for bad sites, which may lead to a Pareto-superior outcome. We will illustrate our policy recommendation with a monopoly situation and then discuss how it can be used for the case of many small websites.

Consider a monopoly which offers two different services and there is a mass one of homogeneous consumers. Each consumer obtains a utility of $u > 0$ per service. The monopoly can have either a good reputation or a bad reputation and consumers know its reputation. In the absence of the proposed policy intervention, the monopoly asks each consumer to consent on the integrated use of personal information obtained from both services. Let $\psi_2^I(m;\theta)$ (respectively, $\psi_1^I(m;\theta)$, $\psi_0^I(m;\theta)$) denote the total nuisance when a consumer uses both services (respectively, one service and no service) when m measure of consumers use both services given the reputation of the monopoly $\theta = g, b$.

We assume that $\psi_i^I(m;\theta)$ strictly increases with m for $i = 0, 1, 2$ and for $\theta = g, b$. In addition, we assume

$$\begin{aligned} \psi_2^I(m;\theta) &> \psi_1^I(m;\theta) > \psi_0^I(m;\theta) \text{ for } \theta = g, b; \\ \psi_2^I(m;\theta) - \psi_1^I(m;\theta) &> \psi_1^I(m;\theta) - \psi_0^I(m;\theta) \text{ for } \theta = g, b, \\ \psi_i^I(m;b) &> \psi_i^I(m;g) \text{ for } i = 1, 2, 3 \end{aligned}$$

where the second line is assumed for expositional sinplicity. We assume that $R^I(m;\theta)$ increases with m.

In the absence of any policy intervention, the monopoly firm will obtain the consent to use integratd personal data by charging p per service such that the following IR is satisfied with equality:

$$2u - 2p - \psi_2^I(1;\theta) = \max\left\{u - p - \psi_1^I(m;\theta),\ -\psi_0^I(m;\theta)\right\}.$$

Under the above assumptions, this condition becomes equivalent to

$$u - p = \frac{\psi_2^I(m;\theta) - \psi_0^I(m;\theta)}{2}.$$

So regardless of the reputation status, consumers end up consenting to integrated use of personal information.

Consider now the following remedy: the regulator forces the monopoly to provide an

option to consent on the usage of personal info *without integrating information from both services at the same price* as the option for the usage of integrated personal information. We introduce $\psi_i^S(m;\theta)$ and $R^S(m;\theta)$ where S means separation of data and assume that they satisfy the same properties that $\psi_i^I(m;\theta)$ and $R^I(m;\theta)$ do. In addition, let us assume

$$\begin{aligned} R^I(m;\theta) &> R^S(m;\theta); \\ \psi_i^I(m;g) &< \psi_i^S(m;g); \\ \psi_i^I(m;g) &> \psi_i^S(m;g). \end{aligned}$$

Basically, good reputation means that the monopoly does not abuse the power from the integrated big data such that both the monopoly and the consumers benefit from the integrated data. Forcing the firm to offer such an option generates the following nice properties:

- For any given p, if the firm has good reputation, for each consumer it is a strictly dominant strategy to choose the integrated usage of personal data between the two options, regardless of m;
- For any given p, if the firm has bad reputation, for each consumer, it is a strictly dominant strategy to choose the usage of personal data without integration between the two options, regardless of m.

The policy remedy allows consumers to coordinate on the best option between the two by following the dominant strategy. Hence, the firm with good reputation charges the above price and obtain the consent to use integrated data. The firm with bad reputation charges the following price and obtain consent to use data without integration;

$$u - p = \frac{\psi_2^S(m;b) - \psi_0^S(m;b)}{2}.$$

This kind of remedy can be applied to data sales. The regulator should force a website to offer an option to not to agree on data sales at the same price as the option to agree on data sales. Then, whenever consumers are concerned about data sales to some untrustworthy source, it is a strictly dominant strategy to choose the option of no data sales. Therefore, the remedy allows consumers to coordinate on the best option. As a response, any website which wants to sell data, it has an interest to find a buyer (or a data broker) with reputation

not to abuse the power of the big data.

There are two caveats. In order to create a permanent incentive to maintain good reputation, we should make the firm to obtain the consent on a regular basis. That is, consumers should be allowed to change the option they chose whenever they are worried about data abuse. This is because the firm may have incentive to abuse data after obtaining consent if the consent is permanent. If the proposed remedy provides incentives to build good reputation, consumers will consent on more gathering and usage of big data, which in turn will be Pareto improving. Second, $\psi_2(m)$ can be decreasing for firms with good reputation. For instance, in the case of Google, $2u - \psi_2(m)$ is net benefit from using search service and Gmail when Google uses big data for consumer benefit, which can be increasing with m.