Privacy and Personal Data Collection with Information Externalities

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

January 31, 2018

Abstract

We provide a theoretical model of privacy in which data collection requires consumers' consent and consumers are fully aware of the consequences of such consent. Nonetheless, excessive collection of personal information arises in the monopoly market equilibrium which results in excessive loss of privacy compared to the social optimum. In a fragmented market with a continuum of firms, no individual website has incentives to collect and monetize users' personal data in the presence of scale economies in data analytics. However, the emergence of data brokerage industry can restore these incentives. Our results have important policy implications for the ongoing debate regarding online privacy protection: excessive loss of privacy emerges even with costless reading and perfect understanding of all privacy policies. We support the view that privacy is a public good and propose alternative policy remedies beyond the current informed-consent approach.

Key words: privacy, personal data, information externalities, big data analytics

[*] We thank Alexandre de Cornière, Saara Hämäläinen, Christian Peukert, David Laband, Liad Wagman and seminar participants at the 2017 International Industrial Organization Conference, 2017 IDEI-TSE-IAST conference on "The Economics of Intellectual Property, Software and the Internet," and University of Alabama for helpful comments. Jay Pil Choi's research was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2016S1A5A2A01022389).

[†] 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.

[§] Department of Economics, Finance, & Legal Studies, Culverhouse College of Commerce, University of Alabama. Tuscaloosa, AL, 35487, USA. E-mail: byung-cheol.kim@ua.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 development is that our routine online activities, such as email, search, and on-line 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 the former Google CEO Eric Schmidt, they collectively create approximately five exabytes 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 public and policy-makers (Schneier, 2015).

One puzzling aspect of this privacy debate is why people 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 with neither our awareness nor consent, but it is also true that we frequently agree to it. For instance, we let Google have access to all the metadata we generate in exchange for the use of Google apps such as Gmail, YouTube, Google Maps, etc.[3] We also implicitly allow uninterrupted use of location tracking and camera to enjoy the sensational augmented reality game Pokémon Go.[4]

In this paper we address the following fundamental questions motivated by these phenomena: Do firms collect too much personal data from a social planner's perspec-

[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] There have been simply too many articles on privacy concerns published in the news media. For just one instance, see "The End of Privacy" by Andrew Burt and Dan Geer (Oct. 5, 2017) *The New York Times.*

[4] 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.*

tive, or too little? If too much, why do people tend to allow some form of personal data collection which appears to harm themselves in the end? Do we really expect that most individuals would no longer voluntarily agree to such data collection as soon as they become fully aware of the 'deals' each of them is making? Put differently, would it be enough to educate consumers about the exact costs of sharing their personal data for a socially desirable privacy protection? What are the appropriate policies to address the privacy concerns?

To address these questions of interest, we first develop a simple model of privacy with a monopoly website, often referred to as 'firm' selling content. The firm decides its user privacy policy whether to commit to privacy protection with no data collection and transfer to a third party or to ask for the user's agreement to the data collection and unrestricted use. Under the latter regime, each user knows the possibility that her personal data can be used for advertising or other purposes for ancillary revenue generation for the firm. In this environment we show that the market equilibrium is easily characterized by excessive consent and thus collection of personal data, which results in excessive loss of consumer privacy compared to the social optimum. This equilibrium arises even if consumers are fully aware of the consequences of their consent. This suggests that the provision of information or education would not fix the problem.

The main mechanism for the result is information externalities: some people's decision to share their personal information may allow the parties accessing to the information to know more or better about others even if they choose not to share their information. Information externalities have been more powerful due to significant advances in big data analytics which have made it possible to draw more accurate inference about those consumers who had not shared their personal data based on the data gleaned from those who had shared. In this environment, even if each user supposedly knows the potential harm of personal data release to herself, she may not take into account any spillover effects of her data release, either positive or negative, on other users. As a result, individually optimal decisions by even fully informed agents may not lead to a socially efficient outcome. If positive externalities exceed negative ones such as privacy infringement, the equilibrium would feature the situation where individual's privacy concerns deter socially desirable construction of greater data-networks. If the opposite situation arises, however, the data collection reaches socially harmful levels.

As an extension of this basic framework, we also consider an alternative market structure with a continuum of small websites to explore the role of data brokerage firms in the aggregation of information. In this set-up, we focus on how these information externalities operate at the level of websites. We show that even if each website alone has no incentives to collect personal data due to its small scale of operation, the emergence of data brokerage markets that purchase and aggregate data from multiple websites can restore incentives to collect personal data. The intuition behind this result is as follows. Once a large amount of data on many consumers is gathered and sold to data brokers, it can generate externalities to even those consumers who refused the sale of personal data. These information externalities influence the individual rationality constraint by affecting the individual's reservation utility evaluated when her data is not provided. From the perspective of potential entrants, the incumbent websites may lower their entry costs as the presence of negative externalities lowers compensation to be made for consumer nuisance. As a result, even without any business stealing effects as in Mankiw and Whinston (1986), we can find an equilibrium in which too many websites enter to collect and sell personal data to data brokers. Though each consumer may find it individually rational to accept the sale of her personal data, it is possible for every consumer to get worse off in equilibrium. In this sense, our model captures a coordination failure among consumers.

Our research thus has important implications for the recent policy debate regarding data brokerage and privacy. The European Commission, for instance, introduced data protection policies which require websites to receive consumer approval for transferring personal data to third parties such as data brokerage firms. The U.S. Federal Communications Commission passed a similar rule. The new rule requires Internet providers such as AT&T, Verizon, and Comcast to obtain their customers' explicit consent before using or sharing sensitive data with third parties such as marketing firms, which was one of their main sources of revenue generated by turning customers' behavioral data into better information basis for targeted advertising.[5] These policies may have some effects on naive consumers by alerting them to be aware of such data transfers. Nonetheless, the overall effects of such a consent-based approach may be limited in addressing the negative information externalities problem since well-informed, fully rational consumers may not change their behaviors because opting-out may not be individually rational in the presence of information externalities.

As in our paper, several legal scholars pointed out the public good nature of privacy and warned of the ineffectiveness of the 'informed consent model' as a solution

[5] For the EU policy, see Directive 95/46/EC (the data protection Directive). For the FCC's new privacy ruling, see Brian Fung and Craig Timberg, " The FCC just passed sweeping new rules to protect your online privacy." *The Washington Post* Oct. 27, 2016.

to protect against invasion of privacy.[6] In essence, this notice-and-consent approach is based on the premise that each individual should have control for disclosure and dissemination of his own personal information. However, this individualistic choice approach is inadequate in addressing the invasion of privacy problem due to information externalities. MacCarthy (2011), for instance, argues that the reliance on individual consent to determine the collection and use of personal information will be ineffective in the presence of negative information externalities and potential risks of information leakage. In a similar vein, Fairfield and Engel (2015) propose to label privacy as a public good and thus call for a collective choice approach to address the privacy issue. Earlier, Hirsch (2006) also pointed out similarities between privacy regulation and environmental laws. Our paper formalizes these ideas.

The rest of the paper is organized as follows. In Section 2 we discuss several strands of literature to which our paper is closely related. Then we briefly discuss the information externalities in Section 3 before we introduce the model of monopolist in Section 4. We extend the model to the analysis for the entry game among a continuum of small websites in Section 5. In Section 6 we discuss implications of our analysis for effective privacy policies and propose two different policy remedies: taxing revenue from data monetization and mandatory opt-out option. Section 7 contains concluding remarks.

2 Related Literature

The literature on privacy is vast and extensive. We thus do not intend to provide an exhaustive review of the literature. Instead, we limit our discussion to selective strands of literature more directly related to this article. For more comprehensive reviews of economic perspectives on privacy and the Internet, we refer to Athey (2014) and Acquisti, Taylor, and Wagman (2016).[7]

One branch of literature to which we make a meaningful contribution is the recent debate on how we should address privacy concerns against prevailing data broker industry. According to the U.S. Senate (2013) and President's Council (2014),

[6] It appears that there is no universally accepted terminology: 'notice-and-choice' and 'notice-and-consent' approach are other terms often used interchangeably. Essentially, they all describe the same approach that data operators should inform individuals of the data policy and each individual decides to agree to it or not.

[7] For reviews with broader perspectives, we find Lane et al. (2014) and Smith, Dinev, and Xu (2011) very helpful. For behavioral approaches to privacy issues, see Acquisti (2009) and Acquisti and Grossklasgs (2004, 2007). For legal perspectives, Tene and Polonestky (2011, 2012) and Solove (2007, 2013) are good resources.

"data brokers" have vibrantly collected, packaged, and traded sensitive consumer data mostly behind a veil of secrecy and thus pose substantial privacy concerns for consumers. The supply chain appears to start from the interactions between web users and web-based applications/content providers of which business model consists in monetizing personal digital trails. Those websites feed the collected data to data brokers who then sell the data after some processing to interested third parties such as advertisers and marketers. Apparently, the use of those data is not expected to be limited to designated purposes only and there could be further transfers to others. In this article we focus on the very early stage when each user voluntarily agrees to uncommitted data use; we aim to provide an economic rationale for the users' consent to the websites.[8]

Certainly, there should be some convincing behavioral reasons for why users give away their personal data. Some have mentioned consumers' lack of understanding about websites' data use policy. For instance, the Australian Information Commissioner and Privacy Commissioner Timothy Pilgrim remarked that privacy notices are just too long for people to read through and most people find it difficult to understand what they are signing up to.[9] Alternatively, it could be due to consumers' myopic and time-inconsistent preference. For discussion's sake let us envision a user's data sharing decision from the perspective of 'contract design and self-control' à la Dellavigna and Malmendier (2004). Then, users would view the enticing free (or highly subsidized) content services as 'leisure goods' that provide immediate benefits but impose delayed costs of privacy loss. Any naive time-inconsistent users will easily opt for enjoying the free content services now by agreeing to the data use policy even if they are well aware of the future costs. Another explanation put forward is that the costs of privacy loss are at best nebulous and intangible so that the users end up underestimating them substantially. Admitting the persuasiveness of those behavioral exposition, for the purpose of this paper we rather make no assumption of any bounded rationality or consumers' lack of knowledge about the website's data use. In other words, we assume the very rational consumers who are fully aware of all consequences from their choices so that there is no way to resort to consumers' myopia or to limited information. Even so, we show that each consumer can find it individually rational to accept the third party use of their personal data and that

[8] Voluntary over-disclosure in web-forms was also found in a field experiment (Preibusch, Krol, Beresford, 2013).

[9] "Many privacy policies are long, complex: OAIC. ZDNET. (Aug. 15, 2013) by Corinne Reichert. http://www.zdnet.com/article/many-privacy-policies-are-long-complex-oaic/

in general there is socially excessive monetizing of personal digital data in the presence of negative information externalities. Thus, we provide a theoretical foundation calling for a different policy beyond the current notice-and-choice approach, which takes the same stance as Hirsch (2006), MacCarthy (2011) and Fairfield and Engel (2015).[10]

Our research is also associated with the literature on data acquisition and pricing which mostly adopts two-sided market configurations. For instance, Bergemann and Bonatti (2015) consider 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. Since we focus on information externalities on the consumer side and how the database can be aggregated through the data brokerage markets, our work is complementary to theirs, for the two works combined provide a more comprehensive perspective on the use of consumer data. Our work also reminds of Bataineh *et al.* (2016) in that they propose a data monetization platform intermediating individuals (personal data sellers) with merchants (data users). We explore the adverse effects of data monetization on individual privacy whereas they view active data trading as a potential market mechanism for higher profits for both data sellers and buyers.[11]

In addition, our work is related to recent studies where firms benefit from better targeting but consumers try to avoid the privacy costs. For examples, Goh, Hui, and Png (2016) empirically examine the effects of privacy and marketing externalities from U.S. Do Not Call registry. Johnson (2013) studies targeted advertising by merchants and advertising avoidance by consumers for their privacy or reducing ads annoyance by installing ad-blockers. Montes, Sand-Zantman, and Valletti (2015) study a Hotelling-type duopoly model where competing firms can acquire information about consumers' characteristics for a better personalized pricing while consumers can pay a 'privacy cost' to avoid such price discrimination. As these studies have in common,

[10] Campbell, Goldfarb, Tucker (2015) suggest a new distortion by the commonly used consent-based approach that may disproportionately benefit big firms but adversely affect small and new firms.

[11] In our model, consumers can get subsidized for their web-content when they agree to the uncommitted data use policy, but they do not actively seek for monetary compensation by selling their personal data as a valuable economic good via the intermediating platform enabling the data aggregation for higher valuation.

consumers may take various actions to avoid the costs of privacy loss. In our model, in the absence of information externalities, each consumer can avoid all privacy costs by choosing no consumption but with heterogeneous valuation on consumption the high valuation consumers decide to sacrifice the privacy for entertaining the content service. In our model the consumers are rather passive in the sense that their choice is limited and do not take active actions considered in these papers. It seems an interesting future research to study interplay between information externalities, ads avoidance, and targeted marketing.

We also notice that many economists studied various privacy issues in the Internet such as the effects of competition on privacy (Casadesus-Masanell and Hervas-Drane 2015), impact of taxation on data collection (Bloch and Demange 2018; Bourreau, Caillaud, and De Nijs 2018), effects of a privacy regulation on firm's investment in quality (Lefouili and Toh 2017). Generally our research adds to this burgeoning literature on information, privacy, and the Internet.

3 Information Externalities

The key driving force in our paper is information externalities. More specifically, we consider a scenario in which some people's decisions to share their personal information may allow the parties accessing the information to know more or better about others who choose not to share their information.[12] 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 one to predict men's sexual orientation based solely on the sexuality of their friends (Johnson, 2009).

We recognize that information externalities can be either positive or negative. As an example of positive externalities, one's data may help those who have access to them improve others' user experiences because more data can lead to better matching between agents or better online search results. Furthermore, there may exist some

[12] Information externalities have been featured by prior works in various contexts. To name a few, we can recall the informational cascades by earlier adopters in IT adoption (Li, 2004), negative payoff externalities by the first-come, first-served rule among depositors in bank runs (Chen, 1999), informational externalities by a firm's hiring conditional on the duration of unemployment on other firm's inference toward a worker's productivity remaining in the labor market (Lockwood, 1991), information externalities in common pools by one firm's investment decision such as drilling exploratory wells in oil tracts on neighbors' investment decisions (e.g., Hendricks and Kovenock 1989, Hendricks and Porter 1996, Lin 2013).

scale of economies in data inference process: the more data points, the more accurate the inference will be. While we keep in mind those plausible positive externalities, as we focus on the privacy concerns in this information age, we pay more attention to the *negative privacy externalities* (MacCarthy, 2011). In fact, one takeaway by Hal Abelson, a computer science professor at MIT, from the above Facebook study is that "you don't have control over your information" even though you do not divulge your personal information, if other people do. This interpretation exactly meant the negative information externalities in our study.[13]

We assume in the model that consumers incur a nuisance cost of privacy loss when personal information is collected and used. There can be many sources of such utility loss. For instance, there could be direct economic losses due to personalized pricing enabled by the detailed knowledge of personal preferences. This kind of loss reminds of the classic argument by the Chicago School scholars (e.g., Posner (1978, 1981) and Stigler (1980)) that one main reason for demanding privacy is to avoid exploitation by potential trading partners who might take advantage of the released information against the revealing individuals.[14] 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."[15]

The information externalities suggest that data operators may infer some information about consumers even if the consumers have not patronized the services of the data operators. For instance, we can consider the situation that firms may have

[13] Once Scott McNealy, a co-founder of Sun Microsystems, even uttered plainly back in 1999 that "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.

[14] Privacy protection in this context is likened to protection of fraudulent claims by Posner as there is no social benefits from privacy protection (except a taste for privacy itself). In contrast to this disclosure literature, there is a vast literature on costly screening and distortionary signaling about private types. Daugherty and Reingaum (2010) provide a new model of the economics of privacy related to both strands of literature.

[15] 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.

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. This process is often referred to as a 'doppelganger search' which is feasible by a zooming-in with big data (Stephens-Davidowitz, 2017).

Following this spirit, in the next monopoly model, we represent a consumer's nuisance costs as $\psi_1(m)$ and $\psi_0(m)$ respectively, depending on whether the consumer uses the service (1) or not (0), where m is the mass of consumers who use the service. Note that the monopoly platform can provide a higher quality of service to users by making use of data. Then $\psi_1(m)$ should be interpreted as the total nuisance minus this benefit from data.[16] We assume that 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 $\psi_0(0) = 0$.

4 The Monopoly Model

We first consider a simple set-up of a monopolistic content provider 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 monopolist's service is given by u, which is assumed to be distributed over $[\underline{u}, \overline{u}]$ with distribution function F and density f. We assume that F satisfies the standard monotone hazard rate condition (MHRC), that is, $f/(1-F)$ is non-decreasing. 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.

Recall that a consumer's nuisance costs are measured as $\psi_1(m)$ for a user and $\psi_0(m)$ for a non-user, where m is the mass of the service users. We assume that when the personal information collected is utilized, the monopolist can generate additional revenue of $R(m)$, with $R^{'}(m) > 0$. For simplicity, we further assume that $R(m)$ represents social benefits as well. Alternatively, we could assume that there exist other channels through which social benefits are generated from the collected data.

[16] We are well aware of the fact that Bloch and Demange (2018) and Bourreau et al. (2018) both explicitly distinguish consumer nuisance from consumer benefit from data use. In our model, what really matters is the nuisance net of the benefit.

By assuming away such consideration, we can focus on the consumers' coordination failure and negative externalities in the nuisance costs.[17] As we are mainly 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 commits to a particular privacy regime. In particular, we consider two possible privacy regimes to choose. One choice is to adopt the business model with pure content pricing where there is either no or little data collection (with commitment to no harms from the collected data).[18] The other option is to inform of potentially uncommitted use of the personal data. In the pure content pricing regime, the privacy of all consumers is under protection and consumers are free of any nuisance costs. In the personal data usage regime, each consumer makes a rational decision with the understanding that her personal information can be used by the monopolist and subsequently she is subject to nuisance costs.

4.1 Social Optimum: The First Best Benchmark

We first analyze the socially optimal outcome as a benchmark in which a social planner chooses the allocation (i.e., the set of consumers who use the service). The social planner chooses the cutoff type of consumers, u^o, such that all consumers whose valuation exceeds or is equal to u^o use the service. The mass of consumers who use the service is given by $m^o = 1 - F(u^o)$. Social 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^o can be derived by the following first order condition.

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

[17] With additional benefits generated from the collected data, the socially desirable level of data collection would increase compared to the currently assumed situation. This consideration would not affect qualitative results of this paper.

[18] For instance, in Section 6.2 we discuss the cases of Facebook/WhatsApp and Google/Nest labs mergers where readers can find that WhatsApp and Nest labs charged users a nominal fee with neither advertising nor data sharing with third parties.

which is equivalent to

$$u + R' = \underbrace{(\psi_1 - \psi_0) + (1 - F(u))\, \psi_1' + F(u)\psi_0'}_{\text{social marginal cost (SMC) of nuisance}}. \tag{1}$$

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 joins to use the content service. First, the marginal consumer's status change from a non-user to a user directly affects his nuisance cost by $(\psi_1 - \psi_0)$. In addition, a new user inflicts externalities not only on the user group he joins, but also on the non-user group he leaves behind. The nuisance cost of an existing user changes by ψ_1' as a new user joins, with the aggregate change for the user group being equal to $(1 - F(u))\psi_1'$. The nuisance cost of an existing non-user also changes by ψ_0' with the aggregate effect being $F(u)\psi_0'$. The last two terms represent negative information externalities.

4.2 Monopoly

We now derive the monopolist's optimal regime choice and price. We analyze the personal data usage model first and then the pure content pricing model as the latter is relatively simple.

Under the personal data usage regime

Given the monopolist's price p, let u be the cutoff type of consumer who is indifferent between using the service or not. For the cutoff type u, the individual rationality (IR) constraint 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. As the IR constraint is binding, the monopolist solves the following problem:

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

The first order condition for profit maximization is

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

Define $u - \frac{1-F(u)}{f(u)} \equiv u^v(u)$ to be the "virtual valuation" of a consumer with value u.

Then, the first order condition above can be rewritten as

$$\underbrace{u^v(u)}_{\text{virtual valuation}} + R' = \underbrace{(\psi_1 - \psi_0) + (1 - F(u))(\psi_1' - \psi_0')}_{\text{private marginal cost (PMC) of nuisance}}. \tag{2}$$

Note that if we consider the standard monopoly model without additional source of revenue from personal data use and nuisance costs, that is, $\psi_1(m) = \psi_0(m) = R(m) = 0$, condition (2) reduces to the standard monopoly condition, $u^v(u) = 0$.[19] In contrast, for the monopolist in our model, the LHS of (2) becomes $u^v(u) + R'$ to reflect the additional revenue R' from data monetization and the RHS represents the private marginal cost (PMC) of nuisance. The comparison of (1) and (2) shows a new type of distortion we identify that the private marginal cost differs from the social marginal cost.

$$SMC - PMC = F(u)\psi_0' + [1 - F(u)]\,\psi_0' = \psi_0' > 0.$$

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-users even though they do not use the monopolist's content. This effect on non-users' reservation utility is $F(u)\psi_0'$. While the social planner cares about this negative externality, the monopolist does not because they are non-users. Instead, the monopolist cares about the effect of an additional user on its ability to extract surplus from existing users. 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 to consume the content, 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 the additional user also negatively affects non-users 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 user group is given by $(1 - F(u))(\psi_1' - \psi_0')$ whereas the social planner only cares about the real impact on the user group which is $(1 - F(u))\psi_1'$. This creates an additional difference of $(1 - F(u))\psi_0'$.

In summary, the social planner cares about the real nuisance effect of the addition of a marginal consumer on *non-users* (with a measure of $F(u)$) while the monopolist cares only about its ability to extract surplus from *users* (with a measure of $1 - F(u)$) through its effect on the marginal consumer's willingness to pay by reducing the

[19] Because the virtual valuation $\phi(u)$ is non-decreasing with u under MHRC, there is a unique solution to the monopoly problem.

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 user'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^o be the monopoly cutoff and the socially optimal cutoff type, respectively, and $m^* \equiv 1 - F(u^*)$ and $m^o \equiv 1 - F(u^o)$ be the respectively corresponding measures of consumers served. Then, we have the following proposition that summarize the analysis up to this point.

Proposition 1 (Monopolist vs. Social Planner) *Suppose that the monopolist uses the business model of data collection.*

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

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

(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 following simple parametric example:

Example U: $u \sim \mathcal{U}[0,1]$, $R(m) = rm$, $\psi_1(m) = \kappa m$, $\psi_0(m) = \xi\kappa m$, where $\xi \in (0,1)$.

In this example, the SMC and PMC are respectively derived as

$$SMC = \kappa[2(1-\xi)m + \xi]$$
$$PMC = \kappa[2(1-\xi)m],$$

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

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

$$m^o = \frac{1 + r - \xi\kappa}{1 + 2\kappa(1 - \xi)}; \tag{3}$$

$$m^* = \frac{1 + r}{2[1 + \kappa(1 - \xi)]} \tag{4}$$

We have $m^* > m^o$ if and only if

$$2\xi\kappa[1 + \kappa(1 - \xi)] > 1 + r. \tag{5}$$

As is clearly seen from the explicit expressions of m^* and m^o, if $\xi = 0$, we have $m^* < m^o$ 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 $\xi \in (0, 1)$.

Under the pure content pricing regime

In the pure content pricing regime in which privacy is protected and no personal data is utilized, we have $\psi_1(m) = \psi_0(m) = 0$ for any number of users m. Therefore, we obtain the standard result, where the virtual type is equalized to marginal cost, which is zero:

$$u - \frac{(1 - F(u))}{f(u)} = 0.$$

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 without data collection is given by

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

4.3 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.

Under the data collection model, as we already analyzed in the preceding subsection, the profit maximizing choice of u, denoted by u^*, is determined by (2). Recall that the corresponding number of users is $m^* = 1 - F(u^*)$ and the monopolist's maximized profit with data collection is given by

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

Then, we find the following sufficient condition under which the monopolist will choose the data collection model instead of pure content pricing model.

Lemma 1 *A sufficient condition for the monopolist to choose the business model of data collection over pure content pricing is that $R(m) - m\psi_1(m) \geq -m\psi_0(m)$ 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 $\Pi^* \geq \widetilde{\Pi}^*$. ■

The proof basically shows that given m, the monopolist prefers the data collection model if $R(m) - m\left[\psi_1(m) - \psi_0(m)\right] \geq 0$. However, social welfare maximization requires the pure content pricing model if $R(m) - m\psi_1(m) - (1-m)\psi_0(m) < 0$ at a given m. Therefore, if $(1-m)\psi_0(m) > R(m) - m\psi_1(m) > -m\psi_0(m)$, the monopolist chooses the data collection model even if this leads to a lower welfare than the pure content pricing model. Again, the social total nuisance $m\psi_1(m) + (1-m)\psi_0(m)$ is larger than the private total nuisance internalized by the monopolist $m\left[\psi_1(m) - \psi_0(m)\right]$ by $\psi_0(m)$, which can lead to excessive choice of the data collection model by the monopolist.

Back to Example U:

For the simple example that we described in Example U, we can derive the mass of active users

$$m^* = \frac{1+r}{2[1+\kappa(1-\xi)]} \quad \text{and} \quad \widetilde{m}^* = \frac{1}{2}$$

and the profits

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

From the comparison between Π^* and $\widetilde{\Pi}^*$, the data monetization model will be adopted if and only if

$$\kappa(1-\xi) < r. \tag{6}$$

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^o. From (5) we know that $m^* > m^o$ if $2\xi\kappa[1+\kappa(1-\xi)] > 1+r$. Thus, combining (5) and (6), the monopolist adopts the data collection model under which there will be too much data collection and loss of privacy if

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

Here we note that κ can be interpreted as a scale parameter for the size of market while we normalize the number of consumers to one. There will be such r that satisfies the above condition whenever $2\xi\kappa > 1$ holds. This means that we will have too much privacy loss if the marginal externality intensity ξ and/or the size of market κ is large enough. This result has an important implication: If the data mining advances further so that ξ is large enough, our society will suffer more from excessive data collection and loss of privacy.

Figure 1 illustrates the business model choices and whether the data collection, if adopted, is socially excessive or not in the space of (ξ, r). Regions I and II denote the set of parameters in which the monopolist adopts the data collection model; in Region III, pure content pricing model is adopted. Particularly, Region II represents where the data collection is socially harmful. This occurs when the externality intensity ξ is large enough while the revenue generated by data collection r is not that high.

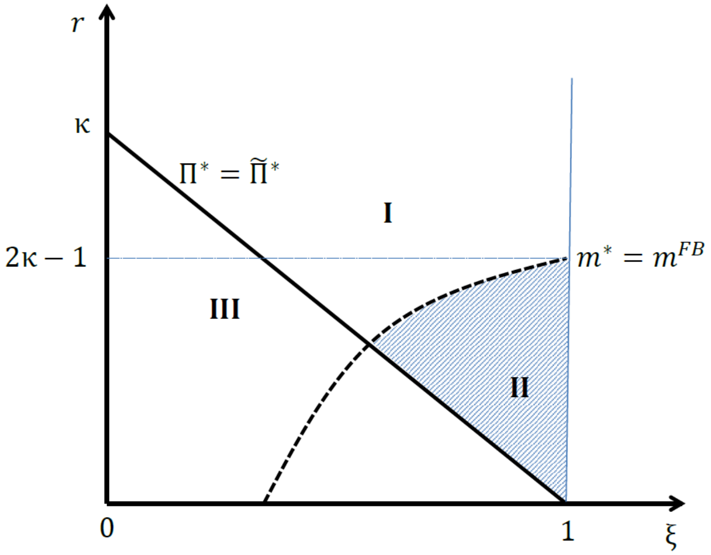

Figure 1: Monopolist's choices of business models and excessive data collection

Many websites that adopt the business model of data sale 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 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., $R(\cdot)$. This can explain the prevalence of websites providing free services. In Example U, this condition can be written as

$$\frac{r}{1+(1-\xi)\kappa} \geq \left[\left(\frac{1+r-c}{2[1+\kappa(1-\xi)]}\right)^2, \frac{1}{4}\right].$$

Again, in equilibrium, too many consumers can give up their privacy.

5 Data Brokerage Firms and Big Data

In the previous section, we considered a monopoly website and its incentives to collect personal data as a business model. We showed that a monopoly 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 wedge between the monopolist's private marginal cost and the social marginal cost of serving marginal consumers due to negative information externalities.

In this section, we consider an alternative market structure with a large number (i.e. a continuum) of small websites to explore the role of data brokerage firms in the aggregation of information. In this set-up, we focus on how these information externalities operate at the level of websites. To analyze this alternative channel of information externalities, we consider a setting in which each website alone has no incentives to collect personal data due to its small scale of operation even though each firm is assumed to be monopolistic in its own niche market. Nonetheless, we show that the emergence of data brokerage firms that purchase and aggregate data from websites can restore incentives to collect data.

5.1 The Model

Consider a mass one of monopolistic websites in their own niche market and a mass one of consumers. All consumers are homogeneous and they may patronize multiple websites. The websites are heterogeneous in terms of the value their content generates for consumers: let v denote a consumer's valuation of content, which is assumed to be distributed according to 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 envision a situation in which the types of information collected across websites are different. Suppose that all websites use the business model of data monetization. Then, 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.[20] Let $\phi(n', n)$ represent the nuisance cost inflicted on consumer i when the consumer uses a measure $n'(\leq n)$ of websites while all the other consumers patronize websites of measure n. Though each consumer ends up using the same number of websites in equilibrium in our setup of homogeneous consumers, the off-the-equilibrium nuisance costs $\phi(n', n)$ needs to be specified in deriving deviation payoffs in order to analyze each consumer's consumption incentives. For instance, $\phi(0, n)$ represent the nuisance she experiences even if she does not consume any website. We assume:

A1: $\phi(n', n)$ is strictly increasing in each element and concave in n' with $\phi(0, 0) = 0$.

As an example satisfying A1, we can consider the following CES nuisance cost of

$$\phi(x, n) = \kappa[\alpha x^{\rho} + (1-\alpha)n^{\rho}]^{\frac{1}{\rho}},$$

where $0 < \alpha < 1$ and $\rho < 1$. When all consumers use the same websites of measure n, the equilibrium nuisance cost is simply denoted by $\Phi(n) = \phi(n, n)$. Both the revenue and the nuisance costs increase with n, i.e. $R^{'}(n) > 0$ and $\Phi'(n) > 0$. In the case of the CES nuisance cost, $\Phi(n) = \kappa n$.

We assume a scale economy in data brokerage:[21]

A2: $R'(0) < \phi_1(0, 0)$ where ϕ_1 is the partial derivative with respect to the first element.

This assumption simply states that in the absence of the brokerage industry, no website (of measure zero) has incentives to collect data on its own (alternatively, in the presence of the brokerage industry if no other website sells data for aggregation), because the size of the data that can be collected and monetized by each firm is too

[20] Since we consider homogeneous consumers, the size of each website's consumer base is 1. $R(n)$ is thus measured with a mass one of consumers for each monopolistic website.

[21] See the Appendix for a justification of this assumption.

small to justify the nuisance costs. This implies that each website adopts the pure content pricing model. This starting point is purposefully set this way because we aim to show the role of data brokerage firms to revive each firm's incentive to data monetization.

We consider a three-stage game with the following timing. In Stage 1, each website simultaneously decides whether to be active or not: to be active, a website must incur the fixed cost of entry $K > 0$. In Stage 2, each active website simultaneously chooses its business model and the content price. In Stage 3, each consumer decides which websites to patronize among the active websites given each firm's privacy policy and content price offers.

5.2 Competitive Personal Data Monetization

Let us first analyze the case that all active websites of measure n adopted the business model of data monetization at Stage 2. We focus on characterizing the equilibrium in which every consumer patronizes all these active websites. Let j and k represent two different websites with v^j and v^k. Let p^j and p^k denote the respective content prices they charge. Since all websites are identical in terms of the nuisance cost they generate from data sales, for any pair of websites (j, k) in equilibrium we have

$$v(n) := v^j - p^j = v^k - p^k. \tag{7}$$

Hence, each consumer's payoff in equilibrium will be given by $nv(n) - \Phi(n)$: each website yields the surplus $v(n)$ and hence the consumer obtains $nv(n)$ as the total surplus, but pays the total nuisance cost of $\Phi(n)$.

Then, at Stage 3, $v(n)$ induces all consumers to patronize all n websites if the following incentive constraint (IC) is satisfied for any $n' \leq n$:

$$[IC : (n', n)] \quad nv(n) - \phi(n, n) \geq n'v(n) - \phi(n', n) \quad \text{for any } n' \leq n. \tag{8}$$

The RHS of the inequality represents a possible deviation payoff. Note that $n'v(n)$ is linearly increasing with n' and thus the RHS is convex in n', which implies that its maximum is attained either at $n' = 0$ or at $n' = n$. Hence, the IC will be satisfied for any $n' \leq n$ once the $[IC : (0, n)]$ is satisfied, which is given by

$$nv(n) - \phi(n, n) \geq -\phi(0, n). \tag{9}$$

Clearly, in equilibrium, inequality (9) must hold with equality because otherwise each website finds an incentive to raise its price at Stage 2. As the IC condition is binding, the necessary condition (8) implies

$$v(n) = \frac{\phi(n,n) - \phi(0,n)}{n}. \tag{10}$$

In summary, we derived the condition that determines the surplus $v(n)$ from each website in the proposed equilibrium as (10). For any active website j given its value of v^j, the equilibrium price for the content service is derived as

$$\hat{p}^j = v^j - \frac{\phi(n,n) - \phi(0,n)}{n} \quad \text{for } \forall j \tag{11}$$

In addition, the rent that each website will receive from selling its data to the brokerage market is equal to $R'(n)$, the marginal contribution of its data to the data aggregation for a competitive brokerage market: See Appendix for a micro-foundation. Then, each website j's equilibrium payoff is equal to $R'(n) + \hat{p}^j$.

As the last step, let us check the deviation incentive by website j at Stage 2 to a pure content pricing model. Then, the individual rationality constraint of each consumer is given by:

$$[IR : (n,n)] \quad v^j - p^j - \phi(n,n) \geq \phi(n,n),$$

where we set $n' = n$ as we consider the deviation of one among a continuum of websites. Hence, the deviation payoff of website j is v^j independent of other websites. Thus, all websites adopting the data monetization can be established as an equilibrium if

$$R'(n) + \hat{p}^j \geq v^j \Longleftrightarrow R'(n) \geq \frac{\phi(n,n) - \phi(0,n)}{n}.$$

Proposition 2 *Suppose that n measure of websites entered at Stage 1. Under A1 if $R'(n) \geq \frac{\phi(n,n)-\phi(0,n)}{n}$, there is an equilibrium in which all websites end up adopting the business model of selling data to a competitive data brokerage market. In the equilibrium, each website j with the content value v^j charges $\hat{p}^j$ defined in (11) and receives a profit of $R'(n) + \hat{p}^j$.*

5.3 Excessive Entry of Websites

Our analysis so far has been confined to the ex post entry stage (from Stage 2) when a fixed measure of websites n entered the market at Stage 1. Let us move backwards and study Stage 1 by making the entry decision 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 free-entry condition:

$$G^{-1}(1 - n^*) - \frac{\phi(n^*, n^*) - \phi(0, n^*)}{n^*} + R'(n^*) = K \tag{12}$$

because $v^* = G^{-1}(1 - n^*)$ and the fixed cost of entry is K.

Given the marginal cutoff type of entrant v, the optimal number of entrants from the social planner's viewpoint is a solution to maximize the following social welfare function,

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

where $1-G(v)$ measures the total number of active websites. The welfare-maximizing cutoff type v can be derived by the following first-order condition:

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

which is simplified as

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

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

$$G^{-1}(1 - n^o) + R'(n^o) - \Phi'(n^o) = K \tag{13}$$

The comparison of (12) and (13) reveals that the comparison of socially optimal number of websites and the market equilibrium boils down to the relative magnitudes of the two terms, $\Phi'(n)$ and $\frac{\phi(n,n)-\phi(0,n)}{n}$. For instance, if Φ is weakly convex, the marginal nuisance exceeds its average, i.e., $\Phi'(n) \geq \phi(n, n)/n$. This implies that we have socially excessive entry as long as $\phi(0, n^o) > 0$. Hence, in the case of CES nuisance cost, there is an excessive entry of websites. Alternatively, if $\phi(n', n)$ is linear in the first element, we have $\frac{\phi(n,n)-\phi(0,n)}{n} = \phi_1(n, n)$ and hence $\Phi'(n) = \phi_1(n, n) + \phi_2(n, n)$ is larger than $\frac{\phi(n,n)-\phi(0,n)}{n}$.

Proposition 3 *Under A1, there is an excessive entry of websites (i.e., $n^* > n^o$) if* *the following condition holds*

$$\frac{\phi(n^o, n^o) - \phi(0, n^o)}{n^o} < \Phi'(n^o), \tag{14}$$

which will be the case if $\Phi(n)$ *is weakly convex with* $\phi(0, n^o) > 0$ *or if* $\phi(n', n)$ *is linear in the first element.*

To provide a more intuitive exposition for the key driving force here, consider a linear function $\Phi(n)$ under which we have $\phi(n^o, n^o)/n^o = \Phi'(n^o)$ and inequality (14) is always satisfied for $-\phi(0, n) > 0$. Recall that $\phi(0, n)$ represents the reservation utility of a consumer who does not use any website when all other consumers use all websites. Suppose that initially n^o measure of websites entered. This reduces the reservation utility of a non-user from $-\phi(0, 0) = 0$ to $-\phi(0, n^o)$. Therefore, each marginal website can extract more than its social contribution by $\phi(0, n^o)/n^o$. Put differently, the entry of some websites generate positive externalities to other websites who are contemplating their entry by worsening consumers' reservation utility.

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.

5.4 A Simple Example: CES Nuisance Costs

Consider the CES nuisance cost of

$$\phi(x, n) = \kappa[\alpha x^\rho + (1 - \alpha)n^\rho]^{\frac{1}{\rho}},$$

where $0 < \alpha < 1$ and $\rho < 1$. 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.[22] Note that this functional form implies that $\Phi(n) = \kappa n$ and $\phi(0, n) = \xi\kappa n$, where $\xi = (1 - \alpha)^{\frac{1}{\rho}}$ and $0 < \xi < 1$. The elasticity of substitution is given by $\sigma = \frac{1}{1-\rho}$. If $\rho = 1$,we have a perfect substitute case in

[22] 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%.

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 sale model if

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

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

$$r + u^o - \kappa = K$$

which is equivalent to

$$n^o = 1 - \kappa + r - K.$$

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.

6 Discussion: Policy Implications and Remedies

Our model has important policy implications for the ongoing policy debate regarding privacy protection from the collection and potential sales of personal data to third parties such as data brokerage firms.

Our model formally shows the limitations of the informed consent model, which is based on the premise that an individual's informed consent provides legitimacy for any information collection and its use. Despite its intuitive appeal, there has been wide criticism against such approach. One argument is that privacy notices are rarely read, and even if read, not easy to fully understand (The White House, 2014).[23] This criticism has naturally led to the discussion of how we can improve transparency about firms' data practices (Federal Trade Commission, 2012). No one will dispute the importance of improving readability and transparency of data policies. However, we assert that such approach may not warrant effective enhancement of privacy protection.[24] In our model, we show that even *costless* reading and *perfect* understanding lead to an equilibrium with an excessive privacy loss.

For discussion's sake, consider the European Union personal data protection policy that entitles the data subject to be informed of any data processing such as a transfer to a third party at every occasion even after the initial consent. Our model implies that such policy may have some effect of alerting naive or uninformed consumers to be aware of potential data transfers, but may not change their behaviors and resolve the excessive loss of privacy due to the information externalities we identify. We below propose two alternative remedies for potential policy reforms.

6.1 Taxing Platforms

It is natural to use tax instruments to reduce excessive loss of privacy resulting from negative information externalities.[25] For instance, Bloch and Demange (2018) study how different taxes affect the degree of exploitation of personal data even though they do not consider information externalities. They examine three different tax regimes: (i) tax on the platform's revenue, (ii) tax on the platform per user, and (iii) a differentiated revenues tax system whereby a different tax is levied depending on the way the revenue is generated. They find that simply taxing the revenue or taxing the platform per user would not reduce the amount of privacy lost, but that the approach (iii) can be effective.

Their result can be applied to our model in which a platform may earn from both subscription and data exploitation through targeted advertising and/or data sales. We give some caution to the idea of taxing the revenue from subscription because it may have the effect of penalizing the pure content pricing model over the data monetization business model. And, taxing the platform per user basis is not desirable if it ends up penalizing the content pricing model particularly when many

[23] McDonald and Cranor (2008) estimated the total time opportunity cost being worth of $781 billion per year if all web visitors had read all privacy policies.

[24] Solove (2013) points out the ineffectiveness in addressing the current privacy concerns by means of providing consumers with more transparency on their personal data collection and use.

[25] Regarding taxation in two-sided markets, see Kind, Koethenbuerger, and Schjelderup (2008) and the special issue on taxation in digital economy in Journal of Public Economic Theory (February 2018).

consumers are to patronize the platform under the content pricing in the absence of the tax regulation. In contrast, taxing the revenue from data monetization will help reduce excessive loss of privacy.

6.2 Mandatory opt-out option to build up reputation

We here propose a mandatory opt-out option to induce platforms to build up good reputation in the usage of data. After explaining it in the context of post-merger data integration, we show how the remedy can be applied more broadly.

Consider a data integration issue that arises when a platform in data-intensive industries acquires another company as in the cases of Facebook/WhatsApp and Google/Nest labs (and Google/Dropcam). In all of these acquisition cases, a major platform such as Facebook or Google acquired another company which had a different business model in terms of privacy. For instance, WhatsApp neither sold advertising space nor collected a lot of personal data on its users; instead it charged users a nominal fee. Similarly, Nest labs had a paid-for-business model without advertising and did not share its data with anyone. Both WhatsApp and Nest labs pledged their intent to protect their customer data from their parent company, which means that without a customer's content, her/his data will not be transferred from WhatsApp (Nest Labs) to Facebook (Google) (Stucke and Grunes, 2016, p.83). However, Facebook and Google have strong incentives to transfer the data from acquired company to merge them with their own data to improve targeted advertising.

Recall that our model implicitly assumes that platforms cannot commit to use personal information only to enhance users' surplus. We worked with this assumption to focus on excessive loss of privacy. Now let us relax this assumption and think of a policy remedy that induces platforms to build reputation not to abuse their big data against users. For this purpose, we define a platform with good (bad) reputation as the one which provides a higher (a lower) surplus net of nuisance to its users the more information they share with the platform. Suppose that a platform acquired a service provider and that both companies share some common user base as in the case of Facebook/WhatsApp. We propose the following remedy. The regulator mandates the platform to provide users with the following two options *at the same price*: (i) OPT-IN: an user allows for data transfer and integration set as default and (ii) OPT-OUT: an user sets no approval for data transfer and integration as default.

The idea behind our proposal is as follows. When the platform is allowed to offer these two options at different prices, the bad reputation platform can induce the undesirable equilibrium in which consumers agree to the data integration and end up being worse off because of the negative information externalities. With the same price regulation, however, consumers may authorize only the good reputation platform to obtain the consent to the data integration. This is because each consumer has an incentive to set the opt-in at default only if the data operator has kept good reputation, *regardless of the choices made by other consumers.* Therefore, the proposed policy remedy aims to facilitate consumer coordination on the better option depending on the platform's reputation.

To be more specific about this suggestion, consider the following simple environment whereby each and every consumer of mass one patronizes both services provided by the two firms of which reputation is indexed by $\theta \in \{G, H\}$. Let $R(m;\theta)$ represent the total revenue of the merged entity from data exploitation when m mass of consumers agreed to data integration. As before, assume that $R(m;\theta)$ strictly increases with m for $\theta \in \{G, H\}$. Let $\psi^I(m;\theta)$ ($\psi^S(m;\theta)$) represent the total nuisance for a consumer whose data from both firms are integrated (separated). We assume that for any given m

$$\begin{aligned} \psi^I(m;G) &< \psi^S(m;G); & \psi^I(m;B) &> \psi^S(m;B): \\ \frac{d\psi^I}{dm}(m;G) &\leq 0; & \frac{d\psi^S}{dm}(m;G) &\leq 0; \\ \frac{d\psi^I}{dm}(m;B) &> 0; & \frac{d\psi^S}{dm}(m;B) &> 0. \end{aligned}$$

The first line defines the role of reputation: the good reputation firm does not abuse the power of the integrated big data such that the consumers benefit from the integrated data (as well as the firm); the bad reputation firm does and harm the consumers. The second and third lines imply that the marginal harm is not increasing with more data for the good firm but the opposite for the bad. In addition, we assume that welfare is maximized with the full data integration for the good, but with non-integration for the bad. Formally,

$$\begin{aligned} 1 &= \arg\max_{m\in[0,1]} R(m;G) - m\psi^I(m;G) - (1-m)\psi^S(m;G) \\ 0 &= \arg\max_{m\in[0,1]} R(m;B) - m\psi^I(m;B) - (1-m)\psi^S(m;B). \end{aligned}$$

As long as each consumer can choose the two options at an equal price, the social optimum is achieved as then we have:

> If $\theta = G$ (B), for each consumer it is a strictly dominant strategy to choose OPT-IN over OPT-OUT regardless of m (resp. the reverse for B).

For discussion's sake, suppose that the firm was allowed to propose the two alternatives at unequal prices and now possibly it provides some bonus $b > 0$ for a consumer's consent to the data integration (though each consumer can maintain data separation as her default choice without any compensation). Then, by providing a bonus $b = \psi^I(1; B) - \psi^S(1; B)$, the firm can induce all consumers to approve data integration and realize a profit of $R(1; B) - (\psi^I(1; B) - \psi^S(1; B))$, which can be strictly larger than $R(0; B) - \psi^S(0; B)$. Indeed, in the presence of strong negative information externalities, the compensation required $\psi^I(1; B) - \psi^S(1; B)$ can be quite small relative to the full nuisance from data integration $\psi^I(1; B)$.

In fact, the same idea can be applied more broadly. The regulator may want to force the firm to offer the opt-out option at the same price as the opt-in option: of course, the website is not restricted for its pricing choice but it must treat equally all consumers regardless of whether a consumer exercises the opt-out option. Such a policy will play the role of guiding consumers to approve data collection only when the website's reputation is good enough. Moreover, in order to provide some incentive to maintain good reputation, the regulator should make the firm obtain the consent on a regular basis. This is because the firm may find it better to abuse the collected data if the consent is permanent and thus there is no need to keep the reputation. If the proposed remedy provides incentives to build good reputation, consumers will consent to more data gathering and good use of big data, which in turn will be Pareto improving.

As a real-world case in point, let us revisit the Facebook/WhatsApp case. Under the parent company Facebook's control, WhatsApp's reputation for privacy has significantly diminished (Ibid. p.83). According to media reports,[26] WhatsApp never informed its users that it was collecting data for its business intelligence and there was no choice of opt out without uninstalling the app. The France privacy watchdog CNIL said that this practice violates "the fundamental freedoms of users". Facebook/WhatApp merger and their plausible data sharing have gained attention from European regulators. Earlier Germany and then the UK ordered Facebook to stop collecting data from WhatsApp users. Recently the CNIL issued the same order after in August 2016 WhatsApp announced updates to its privacy policy terms, including the possibility of linking WhatsApp users' phone numbers with Facebook users'

[26] For instance, see https://www.theverge.com/2017/12/18/16792448/whatsapp-facebook-data-sharing-no-user-consent

identities.[27] Our policy proposal is consistent with the regulators' concern that even the users who opt out should be able to consume platform services without being excluded from the market.

7 Concluding Remarks

As our life-style becomes increasingly reliant on the Internet, our daily activities through all kinds of computer and mobile devices leave digital trails, constantly producing up-to-date information about our personal activities. Such data becomes so valuable that now many websites and content providers offer their content for free or at a highly subsidized price in exchange for users' agreement to more or less uncommitted use of personal data, and the collected data is handed over to the data broker markets. This has raised critical privacy concerns about potential harms and costs to individuals and society.

In this paper we provide a model of privacy based on the concept of information 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 resulting loss of privacy 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]data 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 personal data. 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 generates information externalities and the privacy protection may be viewed as a public good. We hope that our research provides a step conducive to more research in this direction.

[27] On May 18, 2017, the European Commission fined Facebook €110 million for providing incorrect information during the Commission's 2014 investigation of Facebook/WhatsApp merger because Facebook alleged that it would be unable to establish reliable automated matching between the Facebook users' accounts and WhatsApp users' accounts, which turned out incorrect.

References

Acquisti, A. (2004). Privacy and security of personal information. *Economics of Information Security*, 179–186.

Acquisti, A. (2009). Nudging privacy: The behavioral economics of personal information. *IEEE security & privacy 7*(6).

Acquisti, A. and J. Grossklags (2004). Privacy attitudes and privacy behavior. In *Economics of information security*, pp. 165–178. Springer.

Acquisti, A. and J. Grossklags (2007). What can behavioral economics teach us about privacy. *Digital Privacy: Theory, Technologies and Practices 18*, 363–377.

Acquisti, A., C. R. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 52*(2), 442–92.

Athey, S. (2014). Information, privacy and the internet: an economic perspective. *CPB Lecture*.

Bataineh, A. S., R. Mizouni, M. El Barachi, 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.

Bloch, F. and G. Demange (2018). Taxation and privacy protection on internet platforms. *Journal of Public Economic Theory 20*(1), 52–66.

Bourreau, M., B. Caillaud, and R. De Nijs (2017). Taxation of a digital monopoly platform. *Journal of Public Economic Theory 20*(1), 40–51.

Campbell, J., A. Goldfarb, and C. Tucker (2015). Privacy regulation and market structure. *Journal of Economics & Management Strategy 24*(1), 47–73.

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

Chen, Y. (1999). Banking panics: The role of the first-come, first-served rule and information externalities. *Journal of Political Economy 107*(5), 946–968.

Crémer, J. (2015). Taxing network externalities. *Taxation and the digital economy: A survey of theoretical models*.

Daughety, A. F. and J. F. Reinganum (2010). Public goods, social pressure, and the choice between privacy and publicity. *American Economic Journal: Microeconomics 2*(2), 191–221.

DellaVigna, S. and U. Malmendier (2004). Contract design and self-control: Theory and evidence. *The Quarterly Journal of Economics 119*(2), 353–402.

Fairfield, J. A. and C. Engel (2015). Privacy as a public good. *Duke LJ 65*, 385.

Goh, K.-Y., K.-L. Hui, and I. P. Png (2015). Privacy and marketing externalities: evidence from do not call. *Management Science 61*(12), 2982–3000.

Hendricks, K. and R. H. Porter (1996). The timing and incidence of exploratory drilling on offshore wildcat tracts. *The American Economic Review*, 388–407.

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

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

Johnson, J. P. (2013). Targeted advertising and advertising avoidance. *The RAND Journal of Economics 44*(1), 128–144.

Kind, H. J., M. Koethenbuerger, and G. Schjelderup (2008). Efficiency enhancing taxation in two-sided markets. *Journal of Public Economics 92*(5-6), 1531–1539.

Lane, J., V. Stodden, S. Bender, and H. Nissenbaum (2014). *Privacy, big data, and the public good: Frameworks for engagement.* Cambridge University Press.

Lefouili, Y. and Y. L. Toh (2017). Privacy and quality. Technical report, Toulouse School of Economics.

Lin, C.-Y. C. (2013). Strategic decision-making with information and extraction externalities: A structural model of the multistage investment timing game in offshore petroleum production. *Review of Economics and Statistics 95*(5), 1601–1621.

Lockwood, B. (1991). Information externalities in the labour market and the duration of unemployment. *The Review of Economic Studies 58*(4), 733–753.

MacCarthy, M. (2010). New directions in privacy: Disclosure, unfairness and externalities. *ISJLP 6*, 425.

Mankiw, N. G. and M. D. Whinston (1986). Free entry and social inefficiency. *The RAND Journal of Economics*, 48–58.

Montes, R., W. Sand-Zantman, and T. M. Valletti (2015). The value of personal information in markets with endogenous privacy. mimeo.

Posner, R. A. (1978). The right of privacy. *Georgia Law Review 12*(3), 393–422.

Posner, R. A. (1981). The economics of privacy. *American Economic Review 81*(2), 405–409.

Preibusch, S., K. Krol, and A. R. Beresford (2013). The privacy economics of voluntary over-disclosure in web forms. In *The Economics of Information Security and Privacy*, pp. 183–209. Springer.

Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world.* WW Norton & Company.

Singer, E., J. Van Hoewyk, R. Tourangeau, D. M. Steiger, M. Montgomery, and R. Montgomery (2001, December). Final report on the 1999-2000 surveys of privacy attitudes. Technical report, Washington, DC, US Bureau of the Census, Planning, Research and Evaluation Division.

Smith, H. J., T. Dinev, and H. Xu (2011). Information privacy research: an interdisciplinary review. *MIS quarterly 35*(4), 989–1016.

Solove, D. J. (2007). *The future of reputation: Gossip, rumor, and privacy on the Internet.* Yale University Press.

Solove, D. J. (2013). Privacy self-management and the consent dilemma. *Harv. L. Rev. 126*, 1879–2479.

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are.* HarperCollins. May.

Stigler, G. J. (1980). An introduction to privacy in economics and politics. *Journal of Legal Studies 9*(4), 623–644.

Stucke, M. E. and A. P. Grunes (2016). *Big data and competition policy.* Oxford University Press.

Tene, O. and J. Polonetsky (2011). Privacy in the age of big data: a time for big decisions. *Stan. L. Rev. Online 64*, 63.

Tene, O. and J. Polonetsky (2012). Big data for all: Privacy and user control in the age of analytics. *Nw. J. Tech. & Intell. Prop. 11*, xxvii.

The White House (2014). Big data and privacy: a technological perspective. *Washington, DC: Executive Office of the President, President's Council of Advisors on Science and Technology*.

US Senate (2013, December). A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes. Technical report, Washington, DC: Committee on Commerce, Science, and Transportation, US Senate.

Waldo, J., H. Lin, and L. I. Millett (2007). *Engaging privacy and information technology in a digital age.* National Academies Press.

Appendix: A microfoundation for the revenue from data aggregation

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 revenue each website would obtain from the data brokers. For this purpose, let us consider a simple environment of n symmetric data brokers. Let $B(m)$ denote each broker's revenue function where m is the measure of websites. We assume that the revenue increases but at a decreasing rate, i.e., $B' > 0$ and $B'' \leq 0$. Then, we can establish the following lemma:

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

Proof. *If every data broker proposes the same price $B'(m/n)$, then each broker will get the profit of $B(m/n) - B'(m/n)m/n > 0$. Now consider a brokerage 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. This is because, upon the deviation, it would attract all websites and the upper bound of its profit will be*

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

where the inequality is from the fact that the bracketed term in the second line takes on a negative value if B is strictly concave and becomes zero if B is linear. ■

There are several remarks associated with the above lemma. First, suppose that one allows a deviation in which the deviating broker proposes a higher price but can limit the offer to only a certain number of first-arrived websites. Even so, 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 will still lead to a lower profit. Second, the lemma implies that there is no other symmetric Nash equilibrium in which all brokers 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 the lemma will be Pareto-superior to any other symmetric equilibrium from the viewpoint of each brokerage firm.

We can set $R(m) = nB(m/n)$ so that $R'(m) = B'(m/n)$. In this case, each website gets the profit of $R'(m)$.

We below provide the condition in which a single website has no incentive to sell data if no other website sells data. Suppose that the website is the only who sells data. Let ε be the data amount of this website. Then, its profit from the data sale is approximately $R'(0)\varepsilon$. An individual consumer's IR constraint requires

$$u - p - \Phi(\varepsilon) \geq \phi(0, \varepsilon).$$

Therefore the website has the overall profit of

$$R'(0)\varepsilon + u - [\Phi(\varepsilon) - \phi(0, \varepsilon)] .$$

Hence, we need to assume

$$R'(0)\varepsilon + \phi(0, \varepsilon) < \Phi(\varepsilon) \text{ for } \varepsilon \text{ small enough,}$$

which is equivalent to

$$R'(0) < \phi_1(0, 0),$$

which is the assumption **A2**.