MONASH University

MONASH BUSINESS SCHOOL

Department of Economics
ISSN number 1441-5429

Privacy Costs and Consumer Data Acquisition: An Economic Analysis of Data Privacy Regulation

Discussion Paper no. 2022-07

Zhijun Chen

Abstract:

General Data Protection Regulation (GDPR) aims to protect consumer data privacy, however, its adverse effects have been widely documented. We present a new model for the analysis of consumer data acquisition under privacy regulation. We treat both data and analytics as separate strategic variables and consider the heterogeneity of privacy costs across consumers. Using this model to examine the impact of GDPR, we identify a market failure before GDPR and find that GDPR activates a market for data acquisition by imposing consent requirements on data acquisition. We further study the optimal design of the mechanism for consumer data acquisition and deliver important policy implications for implementing the social optimum.

Keywords: Data acquisition, Privacy Costs, and Data Analytics

JEL Classification: D47, L11, L40, K21

Zhijun Chen: Monash University, Department of Economics (email: chenzj1219@gmail.com).

monash.edu/business/economics

ABN 12 377 614 012 CRICOS Provider Number: 00008C

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Privacy Costs and Consumer Data Acquisition: An Economic Analysis of Data Privacy Regulation[*]

Zhijun Chen[†]

April 17, 2022

Abstract

General Data Protection Regulation (GDPR) aims to protect consumer data privacy, however, its adverse effects have been widely documented. We present a new model for the analysis of consumer data acquisition under privacy regulation. We treat both data and analytics as separate strategic variables and consider the heterogeneity of privacy costs across consumers. Using this model to examine the impact of GDPR, we identify a market failure before GDPR and find that GDPR activates a market for data acquisition by imposing consent requirements on data acquisition. We further study the optimal design of the mechanism for consumer data acquisition and deliver important policy implications for implementing the social optimum.

Key Words: Data acquisition, Privacy Costs, and Data Analytics.

JEL Codes: D47, L11, L40, K21

[*] I gratefully acknowledge financial support from the Australian Research Council (grant number DP210102015). I am grateful to Chongwoo Choe for his effort and contribution to the research project of data privacy. His important comments helped improve the presentation of the paper. I thank my research assistant Ratul Das Chaudhury for providing supplementary reading materials "Cookies" and "Data Analytics" in the Online Appendix G and H.

[†] Monash University, contact email: chenzj1219@gmail.com

1 Introduction

Consumer data has become a fundamental resource for the modern digital economy.[1] Recent technological progress in data science has facilitated enormous growth in the scale and precision of consumer data. These advances have led firms to explore new products and services and convert to new business models, which has generated a new source of revenue for firms and extra benefits to consumers.

However, the unprecedented scale of consumer data generation through invasive and opaque acquisition practices by digital platforms has raised privacy concerns to the forefront of the policy debate. Digital platforms typically offer “free” content or services for consumers, and digital businesses collect and process consumer data generated from the use of these “free” services, which they then monetize. Such a raw data set contains heterogeneous and complex attributes and dimensions of personal information related to a consumer’s online activities, which might include private and sensitive information. In addition, consumer data can be harvested across different devices and processed through different parties other than the digital platforms, making it impossible to track the dispersion of such sensitive information. An increasing number of data breach cases have been exposed,[2] but the true scale of privacy breaches is difficult to estimate.

Government regulators have taken action to protect consumer privacy in the digital era. The European Union (EU) has endeavoured to enact such legislation by introducing the General Data Protection Regulation (GDPR) in 2016; this is the toughest privacy and security law in the world, which aims to harmonize data privacy laws across of its member countries as well as provide greater protection and rights to individuals. GDPR has become a blueprint for privacy regulation in many other countries and states, including Australia, New Zealand, Brazil, and India, as well as California and Vermont in the United States.

Two years after its entry into effect in May 2018, the EU claims that "the GDPR enhances transparency and gives individuals enforceable rights, such as the right of access, rectification, erasure, the right to object and the right to data portability" and that "the GDPR has been

[1] Consumer data refers to the behavioural, demographic and personal information trail that consumers leave behind as a result of their Internet use. The terminologies of “consumer data” and “personal data” are often used interchangeably. Here, we prefer using “consumer data” to emphasize the commercial purpose of such data.

[2] According to the Identity Theft Resource Center’s 2021 Data Breach Report, there were 1862 reported cases of data breaches and the case number increased by 68 percent from the previous year. See https://www.cnet.com/news/privacy/record-number-of-data-breaches-reported-in-2021-new-report-says/.

an overall success, meeting many of the expectations".[3] However, the negative economic consequences of GDPR have also been widely seen. GDPR's rollout causes an immediate negative impact on digital platforms' data acquisition and their profitability. One of the most common ways consumer data is collected is through the use of cookies.[4] An empirical study by Aridor et al. (2020) indicates a 12.5% drop in total cookies, while Johnson et al. (2020) find that the significant reduction of cookies reduces the platforms' revenue by 52% from opt-out consumers. The decrease of revenue also leads to a significant reduction of investments in digital infrastructure,[5] and the strict regulation on data acquisition will have significant impacts on the development of emerging technologies.[6] Moreover, GDPR compliance costs a significant level of resources for platforms, and it is expected that such high costs will eventually be passed on to consumers.[7]

The above controversy raises several fundamental economic questions on data privacy regulation. First, how do we reconcile the conflict between consumer data acquisition and privacy protection? Second, on what economic basis do we evaluate the overall impact of regulation? Third, what is the optimal regulation of data acquisition that maximizes social welfare?

Privacy protection often takes a fundamental rights perspective, although there is no consensus among scholars on the concept and the value of privacy. A distinction between privacy rights and data rights is also commonly acknowledged, in which the former is formally protected and cannot be traded- whereas the latter can be traded under the consent of the data subject.[8] Such distinction makes it possible to reconcile the trade-off between data rights and privacy protection. Moreover, an essential step for such goal is to establish a proper market mechanism for consumer data acquisition under privacy protection.

[3] See https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1166

[4] Cookies are small text files that are planted into the user's device by a web browser when the user visits a particular website. By varying the number and types of cookies, digital platforms are able to change the scale of data that it will collect from each consumer. We provide a detailed discussion of cookies in the Online Appendix.

[5] Jia et al. (2021) find a 26% reduction of venture investment in digital sectors by EU ventures compared to their US counterparts.

[6] See the discussion on the impact of GDPR on global technology development by Li et al. (2019).

[7] According to the PricewaterhouseCoopers survey, 68% of American companies are expected to spend between US$1 million and US$10 million to meet the GDPR requirements, and 9% are expected to spend more than US$10 million (PwC, 2017, https://www.pwc.com/us/en/services/consulting/library/gdpr-readiness.html.)

[8] Most data protection regulations including GDPR do not recognize property rights over consumer data. Digital platforms can harvest data and possess data without data subjects retaining any property rights over their data, provided the data subjects have consented to data collection.

To understand key features of such a market and examine the impact of GDPR, we develop a new theoretical model of consumer data acquisition. Our first step is to introduce the concept of privacy costs which captures two key features of the economics of privacy.[9] First, harvesting consumer data with sensitive personal information causes a privacy concern, which results in a loss of utility or a cost to consumers. Moreover, the privacy cost increases with the amount of data collected by the platform. Second, privacy sensitivities and attitudes are subjective and idiosyncratic, because what constitutes sensitive information differs across consumers. Thus, it is essential to incorporate the heterogeneity in consumers' privacy sensitivities.

Second, we treat "data" and "process" as two separate inputs. Raw consumer date does not have much value per se. It needs to be processed and analyzed to create value, and its value depends on the scale of data and the firm's capability in extracting valuable information from the data. We call this capability "data analytics". Broadly speaking, a digital platform's data analytics is the aggregation of its data analysis technologies, computation infrastructures, and most importantly, the data science team.[10] Data analytics is not a software development cycle such as machine learning. Instead, it is an exploratory undertaking closer to research and development than it is to software engineering. As stated by Provost and Fawcett (2013), the investment in data analytics is the most important asset for a digital platform.

Hence, we consider both raw data and data analytics as two essential inputs, and the combination of these inputs generates a revenue to platforms and a benefit for consumers respectively. We take a reduced-form for the revenue and benefit from the data with the assumption of complementarity between the two inputs. This approach allows us to establish a general social welfare function by taking into account both the benefits and the costs of data acquisition, which can be used as a benchmark for policy evaluation. The social benefits from consumer data rely on contributions from both platforms and consumers, by which platforms incur a cost for data analytics while consumers bear a privacy cost. However, the nature of these costs is somewhat different. Digital platforms invest a large amount of physical costs in the capacity-building of data analytics, and these physical costs are observable. By contrast, data provision incurs a psychological cost to consumers, with such privacy costs being subjective and heterogeneous across consumers, which remains a piece of private information to consumers.

Equipped with these novel features, our model provides a new lens through which to examine the impact of GDPR and a new approach to analyze the optimal mechanism design for consumer

[9] See Acquisti et. al (2016) for detailed discussions on the economics of privacy.

[10] We provide a brief summary of data analytics in the Online Appendix.

data acquisition. Before GDPR, digital platforms bundle the "free" digital service with the requirement of data provision; that is, by offering their digital services for "free", the digital platforms harvests consumer data at zero marginal cost. Without taking consumers' privacy costs into account, digital platforms collect more consumer data than needed for the social optimum to maximize total social welfare, which further promotes additional investment in data analytics than the social optimum due to the complementarity of both inputs. Such market failure before GDPR harms consumers and distorts resource allocation.

A principal rule for data acquisition under GDPR is the requirement of consent from the data subject, which is defined as "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her" (article 4). When data acquisition is inevitably accompanied by privacy concerns, such consent requirement entitles consumers to trade their personal data for benefits at the cost caused by privacy concerns, with consumers only willing to give consent if the benefits from providing data exceed the cost of privacy. GDPR allows consumers to use the free service of a platform without accepting non-essential cookies (i.e., GDPR opt-out). If consumers also accept non-essential cookies (i.e., GDPR opt-in), they can enjoy extra benefits from sharing data but will incur a privacy cost. Consumers will choose opt-in if the extra benefit from providing data exceeds their privacy cost. To incentivize participation, digital platforms need to compensate opt-in consumers with additional benefits (or transfers) for their data provision.

Thus, GDPR activates a market for consumer data acquisition in which consumers are entitled to trade their personal data for extra benefits while platforms need to pay for data acquisition. Fixing the market failure benefits consumers by reducing the scale of data acquisition and improves the efficiency of investment in data analytics. We first analyze the data acquisition mechanism when the digital platform is committed to a uniform data collection policy for all consumers (by offering a single option for all non-essential cookies). Comparing the equilibrium before and after GDPR allows us to deliver some useful policy implications. In particular, we find GDPR improves consumer welfare when the mean of consumer privacy costs is sufficiently high.

The uniform policy is not an optimal mechanism with heterogeneous types of consumers. Consumers who are more privacy-sensitive will opt out, resulting in a welfare loss. To counter such welfare loss, digital platforms are gradually adopting more sophisticated incentive schemes with a menu of options for different types of cookies, through which consumers are able to choose how much and what type of data to share according to their privacy sensitivity. We then study the optimal design of the data acquisition mechanism that maximizes total social welfare under the private information of privacy sensitivity. This optimal mechanism requires digital platforms to provide an incentive compatible policy with a type-contingent data scale and compensation for all types of consumers.

Hence, by fixing the market failure, GDPR opens a door to maximizing social welfare under privacy protection. Of course, achieving such a goal requires well-designed guidelines for the regulation. The key obstacle of implementation is the asymmetric knowledge on data acquisition whereby consumers often do not understand the implications and the real value of given options. We propose a guideline for the categorization and standardization of cookie specifications, through which the accurate and specific information of cookies can be provided in a standardized and plain language.[11]

GDPR's rollout leads to a significant reduction of third-party cookies.[12] In particular, several dominant digital platforms, including Amazon, Facebook, and Google, are currently moving to phase out third-party cookies, which causes serious antitrust concerns. Using a variant of the baseline model, we analyze a digital platform's benefit and cost for replacing third-party cookies. This analysis incorporates two key features: first, digital platforms typically share the revenue from the third-party advertiser without incurring the cost for data analytics; and second, third-party cookies cause more serious privacy concerns (and higher privacy cost) than first-party cookies.

Prior to GDPR, digital platforms did not differentiate third-party cookies from first-party cookies according to their resulting privacy costs. After GDPR, digital platforms might favour first-party over third-party cookies, since they need to compensate opt-in consumers for their privacy cost. For dominant digital platforms who have their own technology and capacity for online advertising, they will phase out the third-party's advertisement when the privacy concern becomes serious. We characterize the equilibrium condition for such replacement and find that phasing-out third-party cookies increases consumer surplus.

Digital platforms also use consumer data to offer personalized products and charge personalized prices. We extend the baseline model to analyze data acquisition with personalization.

[11] Apple's movement of Privacy Labels in 2020 is an endeavour to achieve such standardization.

[12] Third-party cookies are placed by external domains that differ from the site the user is browsing, mainly for the purpose of online advertisements. Libert et al. (2018) find that the number of third-party cookies has gone down by more than 30% in the EU's news websites after GDPR.

Before GDPR, digital platforms can fully extract consumer benefits from data provision through personalized pricing, leaving consumers with a negative extra surplus from providing their data. By contrast, GDPR opt-out secures consumers a non-negative extra surplus from data sharing, and digital platforms must leave a positive surplus to opt-in consumers as compensation for their privacy costs. Hence, GDPR improves consumer welfare unambiguously.

The rest of the paper is organized as follows. Section 2 sets up the model and provides a benchmark for social optimum. Section 3 studies the equilibrium before and after GDPR and the mechanism design to achieve the second-best outcome. We discuss policy implications in Section 4 and analyze the equilibrium with personalization in Section 5. Finally, Section 6 reviews the literature and concludes the paper.

2 The Model and Benchmark

2.1 The Baseline Model

A digital platform provides free digital services to consumers. The consumer data is then collected, from which the platform earns revenue by selling targeted advertisements, providing various data-driven services to other businesses, or selling products/services directly to consumers. The platform's investment in the digital service is a sunk cost and is not relevant here. There is a continuum of consumers with the total population normalized to 1. Consumers gain a value u from using the digital platform's free service.

A consumer's online activities through the digital platform generate a data set that the platform can harvest using cookies. Let $s \in [0, \bar{s}]$ be the measure of the data set harvested by the platform and call it the data scale, with $\bar{s}$ being the maximum.[13] The digital platform can choose the data scale by varying the number and types of cookies. The raw data set contains unstructured and complex attributes and dimensions of information and it must be processed to generate valuable information. Let $\phi \in [0, +\infty)$ denote the capacity of data analytics. The platform's choice of ϕ is a strategic investment in data analytics, including the infrastructure for data collection and storage, data analysis, and data interpretation, and the cost of investment

[13] Defining and measuring a big data set is quite challenging because of its complexity. Big data has commonly been characterized by four "V"s: Volume, Velocity, Variety, and Value of the data. Here we propose the terminology of "data scale" as a measurement of a data set for the purpose of economic analysis. It can be viewed as a real-valued function that maps different attributes of a data set into the real number set $\Re_+$.

in data analytics is $I(\phi)$ with $I' > 0$ and $I'' > 0$.[14] We focus only on $I(\phi)$ as the cost that is relevant to this study.

The platform earns a revenue $r(s,\phi)$ from each consumer's data set, while a consumer derives an additional benefit $b(s,\phi)$ from providing data, which includes a personalized service provided by the platform or attractive targeted offers from third-party sellers on the platform, for example. Both $r(s,\phi)$ and $b(s,\phi)$ increase in the scale of data s as well as the capacity of data analytics ϕ.

Consumer data contains sensitive personal information and data acquisition causes a privacy concern. Such privacy concerns transform into a loss of utility or a cost to a consumer, which we call the privacy cost. Privacy sensitivities and attitudes are subjective and idiosyncratic, because what constitutes sensitive information differs across consumers. A consumer's privacy sensitivity is characterized by parameter θ, which is distributed between $\underline{\theta}$ and $\bar{\theta}$ according to cumulative distribution function F and strictly positive density f, and the mean denoted by μ. We assume $h = F/f$ is increasing. It is natural to think that the privacy cost increases in the scale of data collected by the platform. Thus, we assume the privacy cost for consumer θ is given by $s\theta$.[15]

Remark 1: Privacy Costs. We treat consumers' privacy concerns as a cost increasing with the scale of data being collected and moreover consider the heterogeneity of privacy sensitivity across consumers. This modelling approach is motivated by some experimental and empirical evidence from a growing literature of preferences for privacy.[16] Lin (2021) provides experimental evidence on how consumers' privacy preferences (concerns) can be divided into intrinsic and instrumental preferences, the distinction first made by Becker (1980). The intrinsic preference of privacy is related to the intrinsic moral value. Privacy is considered as an aspect of human dignity, because it provides personal autonomy and independence. Privacy laws also justify privacy protection on moral grounds. We consider a consumer's privacy sensitivity as an intrinsic

[14] Large digital platforms such as Google and Amazon build their own capacities of data analytics. Small digital platforms outsource the service of data analytics to several leading specialist companies such as Adobe Analytics. Data processing is not a "standard" product or service. Data transformation, data mining, and data evaluation have to be customized according to a customer's specific business model and unique data features. Hence, digital platforms need to pay for the specific investment cost of data analytics when they outsource the task of data processing.

[15] A consumer's privacy cost can be expressed by a general function form $C(s,\theta)$. For the tractability of the analysis we assume a simple function form $C(s,\theta) = s\theta$.

[16] We refer to Acquisti et al. (2016) for a more comprehensive review of the literature.

privacy preference. Experimental evidence from Lin (2021) shows the strong heterogeneity of intrinsic privacy preferences across consumers. The instrumental preference of privacy is endogenously determined by how the private information is used in transactions. For instance, consumers may be concerned that firms can use their data to charge personalized prices according to their willingness-to-pay. Such instrumental privacy concerns can be expressed as a cost. We incorporate the instrumental cost into a consumer's benefit function such that $b(s, \phi)$ is treated as a consumer's net benefit from data provision.

Before GDPR, digital platforms set "accepting all cookies" as a default option for using the digital platform's service. Even when consents are required in principle, most consumers are not aware of their rights under such practice. Essentially, this is equivalent to a requirement of data provision. Thus, we consider that digital platforms bundle their digital services to the requirement of data provision before GDPR. Then consumers using the digital service obtain utility $U(s, \phi, \theta) = u + b(s, \phi) - s\theta$. After GDPR's rollout, however, consumers have two options for sharing their data with the platform. First, in order to use the platform's service, consumers have to allow essential cookies from the platform, in which case they obtain utility u. For simplicity we assume away the privacy cost of allowing essential cookies. Second, if they also allow non-essential cookies, then they derive an additional benefit but incur the privacy cost, and their total utility is expressed in the same way as before GDPR: $U(s, \phi, \theta) = u + b(s, \phi) - s\theta$. For simplicity, we use the term "opt-in" when consumers accepts non-essential cookies and "opt-out" when they do not.

Assumptions

We restrict the analysis to a reasonably large range of θ. Roughly speaking, the lower bound $\underline{\theta}$ is not too small and the upper bound $\bar{\theta}$ is not too large. Specifically, we assume $u \geq \bar{s}\bar{\theta}$, so that all consumers are willing to provide data if this is a requirement to obtain utility u. The following regular assumption is needed to guarantee that optimal policies are well-defined:

Assumption A: The per-consumer revenue function $r(s, \phi)$ and consumer benefit $b(s, \phi)$ are concave.[17]

It is well-understood that data scale and data analytics are *complementary.* The scale of data relies on the infrastructures and technology of data gathering, storage, and cleaning, whereas the capacity of data analytics can be improved through learning-by-doing in processing the large

[17] Formally, this assumption requires the second-order derivatives $r_{ss}(s, \phi)$ and $r_{\phi\phi}(s, \phi)$ are negative and $H \equiv r_{ss}r_{\phi\phi} - r_{s\phi}r_{\phi s} \geq 0$. The same applies to $b(s, \phi)$.

scale of data. This feature is captured by the following assumption:

Assumption B: The cross-derivatives of $r(s,\phi)$ and $b(s,\phi)$ are positive: $r_{s\phi}(s,\phi) > 0$ and $b_{s\phi}(s,\phi) > 0$.

We assume that the function forms $r(s,\phi)$, $b(s,\phi)$, and $I(\phi)$ are common knowledge. The digital platform announces its data scale s and data analytics ϕ publicly. However, the privacy sensitivity θ is a consumer's private information that the digital platform cannot observe.

In the baseline model, we focus on the type of platforms that offer free service/products to consumers in order to collect consumer data (which is the business model for many digital platforms), and moreover use the reduced forms for $r(s,\phi)$ and $b(s,\phi)$. An illustrative example of microfoundation for $r(s,\phi)$ and $b(s,\phi)$ from online advertising is provided below. In Section 5, we analyze the digital platform's pricing decisions when it supplies personalized products in Section 5.

Microfoundation: Targeted Advertising/Recommendation

Advertising markets are typically two-sided markets in which advertisers aim to match with users who are most interested in their products. Consider a continuum of consumers with heterogeneous preferences over some product/service, where their taste x is uniformly distributed along the line $[0,1]$. There is also a continuum of sellers (advertisers) located on the same unit line with its location y and each seller's product is listed at a competitive price a.[18] A consumer x is matched with a seller y with probability $1-(y-x)^2$, in which case this consumer derives a matching value v and zero otherwise (here the distance $(y-x)^2$ measures the utility loss due to mismatch).

Consumer data is collected through cookies. Each internet user is associated a unique cookie ID. The digital platform can monetize the display of matching advertisements through auctions, usually via first-price auctions. In these auctions, advertisers do not bid directly, but rather via "Demand Side Platforms" (DSPs), which select from among the millions of advertising opportunities available on the internet on behalf of their advertiser clients. Most large platforms including Amazon, Facebook, and Google, however, run their own DSPs. When the DSPs are operated by parties other than the digital platform, the competing DSPs will receive an internet user's third-party cookie identifier prior to bidding. The cookie allows the DSP to track the user's data and impute this user's taste. The DSP then uses this information to determine which of its advertiser clients would be the best match for this user, and then submit a bid on behalf of this advertiser in the auction.

[18] The production cost is normalized to zero.

When a consumer (user) visits the platform, its tracking technology identifies this user's ID and a DSP generates a public signal among advertisers about this consumers' taste, $\delta \sim N\left(x, 1/m\left(s,\phi\right)\right)$, where $m\left(s,\phi\right)$ measures the precision of targeted advertising. The most relevant seller/advertiser is located at $y = \delta$. If the DSP chooses to display y's product (advertisements), this seller/advertiser's revenue is $r\left(s,\phi\right) = a\left(1 - E\left[\left(y-x\right)^2 | y = \delta\right]\right) = a\left(1 - 1/m\left(s,\phi\right)^2\right)$. The advertiser's revenue $r\left(s,\phi\right)$ will be shared by advertising intermediaries (DSPs) and the digital platform. The total social benefit generated by consumer data is the extra consumer surplus through improved matching: $B\left(s,\phi\right) = v\left(1 - 1/m\left(s,\phi\right)^2\right)$,[19] from which the consumer receives its share $b\left(s,\phi\right) = \left(v-a\right)\left(1 - 1/m\left(s,\phi\right)^2\right)$.

2.2 The Benchmark

As a comparison benchmark, we first consider a hypothetical economy in which a social planner runs the digital platform, providing a free service and processing consumer data. A consumer's data generates a social benefit $B\left(s,\phi\right) = r(s,\phi) + b\left(s,\phi\right)$, whereas this consumer bears a privacy cost $s\theta$. In addition, processing consumer data incurs an investment cost of data analytics $I\left(\phi\right)$. We further assume that the social planner possesses full information on each consumer' privacy sensitivity θ, and can design a type-contingent offer $\left\{s\left(\theta\right),\ t\left(\theta\right)\right\}$ for each consumer, where $s\left(\theta\right)$ is the type-contingent data scale while $t\left(\theta\right)$ is the associated transfer to the consumer. The social planner's offer satisfies each consumer's participation constraint in data provision, such that each consumer's extra surplus from providing data is non-negative, i.e., $V\left(\theta\right) \equiv b\left(s\left(\theta\right),\phi\right) + t\left(\theta\right) - s\left(\theta\right)\theta \geq 0$.

The social planner chooses $\left\{s\left(\theta\right),\ t\left(\theta\right)\right\}$ and ϕ to maximize the following social welfare

$$SW = \int_{\underline{\theta}}^{\bar{\theta}} \left[B\left(s\left(\theta\right),\phi\right) - s\left(\theta\right)\theta\right] dF\left(\theta\right) - I\left(\phi\right),$$

subject to $V\left(\theta\right) \geq 0$. The maximization of SW with respect to $s\left(\theta\right)$ requires that the term under the integral $B\left(s\left(\theta\right),\phi\right) - s\left(\theta\right)\theta$ be maximized with respect to $s\left(\theta\right)$ for all θ. That is, the optimal data scale $s^*\left(\theta\right)$ balances the marginal benefit of data acquisition $B_s\left(s\left(\theta\right),\phi\right)$ and a consumer's marginal cost, the privacy sensitivity of type θ:

$$B_s\left(s\left(\theta\right),\phi\right) = \theta. \tag{1}$$

The above FOC determines the equilibrium path $s^*\left(\theta,\phi\right) = B_s^{-1}\left(\theta\right)$, which increases in ϕ under Assumption B and decreases in θ under Assumption A. Throughout the paper, we make the

[19] Assume consumers receive zero payoff without using the third-party cookies.

following assumption on the lower bound of θ to ensure an interior optimum for the lowest type: $s^*(\underline{\theta}) < \bar{s}$.[20]

Assumption C: $s^*(\underline{\theta}) < \bar{s}$.

The socially optimal data analytics is determined by equating the marginal cost for data analytics to the expected marginal social benefit:

$$\int_{\underline{\theta}}^{\bar{\theta}} B_\phi\left(s\left(\theta\right),\phi\right) dF\left(\theta\right) = I'\left(\phi\right). \tag{2}$$

Substituting $s^*(\theta,\phi) = B_s^{-1}(\theta)$ into the above FOC, the first-best data analytics ϕ^* is the solution of

$$\int_{\underline{\theta}}^{\bar{\theta}} B_\phi\left(s^*\left(\theta,\phi\right),\phi\right) dF\left(\theta\right) = I'\left(\phi\right). \tag{3}$$

Solving the FOCs (2) and (1) determines the first-best outcomes ϕ^* and $s^*(\theta)$. The transfer must satisfy $t^*(\theta) \geq s^*(\theta)\theta - b(s^*(\theta),\phi^*)$.

The first-best outcome is summarized in the following lemma:

Lemma 1 *Suppose a social planner runs the digital platform and has complete information on a consumer's privacy sensitivity. Under Assumptions A, B, and C, the first-best data scale and data analytics are characterized by* (1) *and* (2)*. The optimal data scale decreases in θ.*

Leading Example

For further illustration, we provide a leading example. The functions $r(s,\phi)$ and $b(s,\phi)$ take the form of Cobb-Douglas with $0 < \rho < 1$:

$$r\left(s,\phi\right) = \alpha s^\rho \phi^{1-\rho},\ b\left(s,\phi\right) = \left(1-\alpha\right) s^\rho \phi^{1-\rho}.$$

In addition, $I(\phi) = \phi^2/2$. The detailed calculation of all equilibrium outcomes is provided in Online Appendix B.

Using this example, the first-best data analytics and data scale are given respectively by

$$\phi^* = \left(1-\rho\right)\int_{\underline{\theta}}^{\bar{\theta}} \left(\frac{\rho}{\theta}\right)^{1/(1-\rho)} dF\left(\theta\right),\ s^*\left(\theta\right) = \left(\frac{\rho}{\theta}\right)^{1/(1-\rho)} \phi^*.$$

[20] This is guaranteed when the marginal benefit of data decreases to zero for s approaching to infinity: $\lim_{s\to\infty} B_s(s,\phi) = 0$. See detailed characterization in the Online Appendix A.

3 The Role of GDPR

There is a market failure in data acquisition before GDPR. We first analyze such a market failure and its underlying harm to consumers and society, and then study the role of GDPR in fixing the market failure.

3.1 Market Failure before GDPR

Before GDPR, the digital platform bundles its digital service with a default option of accepting all cookies. Then consumers using the digital service obtain utility $U(s, \phi, \theta) = u + b(s, \phi) - s\theta$, and the assumption $u \geq \bar{s}\bar{\theta}$ implies that all consumers use the platform. With full participation, the platform's total revenue is $r(s, \phi)$. The platform chooses s and ϕ to maximize its profit $\Pi = r(s, \phi) - I(\phi)$.

The digital platform's revenue increases in data scale s while the marginal cost of data acquisition is zero. Hence, the platform will collect the maximum scale of consumer data: $s = \bar{s}$. Meanwhile, the digital platform builds the optimal capacity of data analytics such that the marginal cost of investment in data analytics $I'(\phi)$ is equal to the marginal revenue $r_\phi(s, \phi)$, as given by

$$r_\phi(s, \phi) = I'(\phi). \tag{4}$$

The optimal data analytics as a function of data scale, as denoted by $\phi^b(s)$ (the superscript b stands for "Pre-" GDPR) increases in s under Assumption B, and the equilibrium data analytics is $\phi^b = \phi^b(\bar{s})$.

The market failure causes a negative impact on consumer surplus. The digital platform acquires excessive consumer data compared to the first-best. Over-collection of consumer data makes consumers worse-off. A consumer's extra surplus from data provision is $V(\theta) = b(\bar{s}, \phi^b) - \bar{s}\theta$. Then, consumers relatively high privacy sensitivity (i.e., $\theta > \hat{\theta} \equiv b(\bar{s}, \phi^b)/\bar{s}$) receive negative extra surplus from data provision,[21] whereas they could be better off if they were allowed to use the digital service without providing their data. The aggregate extra consumer surplus from data provision can be expressed as

$$CS^b = \int_{\underline{\theta}}^{\bar{\theta}} \left(b\left(\bar{s}, \phi^b\right) - \bar{s}\theta\right) dF(\theta) = b\left(\bar{s}, \phi^b\right) - \bar{s}\mu,$$

which turns to be negative when the mean of privacy cost μ is high such that $\mu > \hat{\theta}$.

[21] Nevertheless, consumers still receive a positive utility from free service as $U(\theta) = u + b(\bar{s}, \phi^b) - \bar{s}\theta > 0$ under the assumption $u > \bar{s}\bar{\theta}$.

By contrast, the market failure leads to two countervailing effects on data analytics. First, the platform has less incentives in data analytics because it does not internalize consumer benefits. Before GDPR, the platform offers a free service to consumers in exchange for their data, which is independent of its data analytics. Thus, the platform does not internalize consumer benefits $b(s,\phi)$ in its investment in data analytics. In addition, full consumer participation is guaranteed before GDPR, implying that the platform needs not to improve its data analytics to induce consumers' participation decision. Second, the platform chooses the maximum data scale, i.e., $s = \bar{s}$. The over-collection of consumer data, however, contributes to increasing the investment in data analytics since $\phi^b(s)$ increases in s, which is a positive effect on data analytics due to complementarity with data scale.

In some scenarios the digital platform can capture all social benefits through personalized offers (see the discussion in Section 5), i.e., $b(s,\phi) = 0$, in which the negative effect on data analytics vanishes. Then, ϕ^b is determined by $B_\phi(\bar{s},\phi) = I'(\phi)$. Compared to the first-best data analytics ϕ^* as given by (2) and noting that $s^*(\theta) < \bar{s}$ for all θ while $B_{\phi s}(s,\phi) > 0$, it appears that $\phi^b > \phi^*$. That is, over-collection of consumer data leads to excessive investment in data analytics, at a cost to the society.

Summarizing the above analysis leads to:

Proposition 1 *There is a market failure in data acquisition before GDPR. As a result, the digital platform acquires the maximum scale of consumer data (i.e., $s = \bar{s}$), and consumers with high privacy sensitivity (i.e., $\theta > \hat{\theta}$) receive a negative extra surplus from data provision. When the digital platform can capture all social benefits, the over-collection of consumer data leads to excessive investment in data analytics.*

3.2 Fixing Market Failure after GDPR

GDPR compliance requires that digital platforms must have "specific, unambiguous consent" from data subjects for data provision and must "allow users to access your service even if they refuse to allow the use of certain cookies". Under GDPR, the digital platform is required to unbundle its digital service from the default consumer consent for data collection and allow consumers to use its service and obtain utility u by accepting only essential cookies (i.e., GDPR opt-out). If consumers also allow non-essential cookies, i.e., GDPR opt-in, then they can enjoy additional benefits $b(s,\phi)$ but incur the privacy cost $s\theta$.

GDPR opens a door to fix the market failure. The digital platform now must compensate consumers for data acquisition through non-essential cookies. In addition to the consumer benefit $b(s, \phi)$ from data provision, the platform might offer additional benefits to opt-in consumers. These may include enhanced services (personalized services) or even a monetary payment (vouchers) for opt-in. We collectively call this a transfer from the platform to opt-in consumers, denoted by t. Consumers will then choose opt-in only if the overall benefits from data collection more than offset their privacy cost, i.e., if $b(s, \phi) + t \geq s\theta$.

Digital platforms respond to GDPR compliance by providing explicitly consent forms for cookies (cookies policies) on their websites. Many digital platforms only provide one option for all non-essential cookies, which we refer to as the uniform policy for data collection. Others, however, offer a menu of options consisting of several categories of cookies for consumers to choose to accept. We analyze the uniform data policy first and then study the optimal design of data policy with a menu of options in the next subsection.

When the digital platform is committed to the uniform data policy, it only provides one option for all non-essential cookies whereby consumers cannot choose how much and what kind of data is being collected through non-essential cookies. Since we restrict analysis to the uniform data policy, we only need to consider a constant transfer.[22] The timing of the game is given as follows. First, the digital platform announces the policy $\{s, \phi, t\}$. Second, observing the policy, consumers choose to opt in or opt out.

Under the uniform policy $\{s, \phi, t\}$, a consumer will opt in if $\theta \leq \tau \equiv (b(s, \phi) + t)/s$, where τ is the cut-off value of θ for the marginal consumer indifferent between opt-in or opt-out. Then the platform's revenue from each opt-in consumer is $r(s, \phi) - t$ and the population of opt-in consumers is reduced to $F(\tau)$. Insofar as $\tau < \bar{\theta}$, GDPR reduces the opt-in population, which decreases the platform's profit given the same data scale and data analytics. Substituting $t = s\tau - b(s, \phi)$ into the platform's profit, we have

$$\Pi^u = F(\tau)\left(r(s, \phi) - t\right) - I(\phi) = F(\tau)\left(B(s, \phi) - s\tau\right) - I(\phi).$$

That is, the platform's revenue from each opt-in consumer is equal to the social benefit less the privacy cost of the marginal consumer τ. It follows that, given s and ϕ, choosing the optimal t is equivalent to choosing the optimal threshold τ. Hence, the platform chooses the optimal policy

[22] When the transfer t is made in terms of enhanced services or personalized services, it may cost $c(t)$ for the digital platform. We assume $c(t) = t$ for simplicity of analysis. The analysis for a general form of $c(t)$ is a bit complicated. However, as long as the cost $c(t)$ is independent of s and ϕ, the main results and insights of the equilibrium analysis do not change qualitatively. When digital platforms sell products as well, the transfer can be made through a price discount, in which case the cost $c(t)$ is equal to the benefit t; see the discussion in Section 5.

$\{s^u, \phi^u, \tau^u\}$ to maximize the above profit, where the superscript 'u' indicates 'uniform policy'.

Before GDPR, the platform sets the maximum data scale because its marginal cost for harvesting consumer data is zero. After GDPR, the platform has to compensate opt-in consumers for data acquisition. Consumers choose opt-in or opt-out by comparing total benefits $b(s, \phi) + t$ with their privacy cost $s\theta$. GDPR opt-out reduces the population of opt-in consumers to $F(\tau)$. The platform can counter this by making a transfer to opt-in consumers, however, this reduces the platform's revenue per opt-in consumer. An important implication is that, through the transfer and opt-in decisions, the platform internalizes the consumer benefit $b(s, \phi)$ but also compensates consumers for their privacy cost, which is not the case before GDPR. For each opt-in consumer, the platform captures social benefits $B(s, \phi)$ but pays a price equal to the privacy cost of the marginal consumer. This leaves each opt-in consumer with an extra surplus $V(\theta) = b(s, \phi) + t - s\theta = s(\tau - \theta)$, equal to the difference between her actual privacy cost and the privacy cost of the marginal type.

GDPR reduces the data scale as the digital platform now has to take into account the privacy cost of opt-in consumers. The platform's optimal choice of data scale s^u, which balances the marginal social benefit with the marginal cost of the consumer with type $\theta = \tau$, is an interior solution of the following FOC:

$$B_s(s, \phi) = \tau. \tag{5}$$

By contrast, GDPR generates two opposite effects on data analytics. First, with the transfer, the digital platform is able to internalize its externality of data analytics on opt-in consumers, and its net per-consumer revenue becomes $B(s, \phi) - s\tau$, which is a positive impact on data analytics. Second, since only $F(\tau)$ consumers opt in, the digital platform can only recoup its investment cost from a reduced population of consumers, which generates a negative impact on data analytics. More specifically, an increase in data analytics generates a marginal social benefit $F(\tau) B_\phi(s, \phi)$ at the marginal cost $I'(\phi)$, and the optimal data analytics ϕ^u satisfies the following FOC:

$$F(\tau) B_\phi(s, \phi) = I'(\phi). \tag{6}$$

Meanwhile, the platform chooses the optimal threshold τ^u such that the per consumer social benefit offsets the privacy cost for the marginal consumer $s\tau$ plus a rent $sh(\tau)$ from all opt-in consumers

$$B(s, \phi) = s\tau + sh(\tau). \tag{7}$$

This leaves the digital platform a per-consumer revenue: $B(s,\phi) - s\tau = sh(\tau)$. Combining the above equation with (5), we can derive the equilibrium threshold τ^u as the solution of

$$\frac{\tau^u}{\tau^u + h(\tau^u)} = \varepsilon_s \equiv \frac{B_s(s,\phi)\, s}{B(s,\phi)}.$$

The equilibrium threshold τ^u depends on the elasticity of social benefit on data scale, ε_s; it does not rely on the cost function $I(\phi)$ directly. When the social benefit function has a constant elasticity as in the case of the leading example, τ^u is independent of s and ϕ. This nice property simplifies analysis. Finally, the equilibrium transfer is $t^a = s^u\tau^u - b(s^u, \phi^u) = r(s^u, \phi^u) - s^u h(\tau^u)$, and the equilibrium profit is $\Pi^u = s^u h(\tau^u) F(\tau^u) - I(\phi^u)$.

We show in Online Appendix C how to characterize the equilibrium s^u, ϕ^u, and τ^u through the FOCs (6), (5), and (7).[23] GDPR opt-out reduces consumer participation in data provision, causing an efficiency loss. Compared to the first-best data scale $s^*(\theta,\phi)$, it follows that $s^*(\theta,\phi) > s^u(\phi)$ for any $\theta < \tau^u$, since $B_s(s,\phi)$ decreases in s. That is, the digital platform collects less consumer data than the first-best level for each opt-in consumers under the uniform data policy. Hence, it collects much less total amount of consumer data under the uniform policy. This further implies a lower capacity of data analytics under the uniform policy since ϕ is complementary to data scale.

Summarizing the above analysis leads to:

Proposition 2 *GDPR activates a market for data acquisition. Suppose the digital platform is committed to a uniform data policy after GDPR. The digital platform's optimal data scale s^u, optimal data analytics ϕ^u, and equilibrium threshold τ^u are characterized by* (5), (6), *and* (7) *respectively. Opt-in consumers receive an extra surplus from data provision $V(\theta) = s^u(\tau^u - \theta)$. Compared to the social optimum, the digital platform collects less data and invests less in data analytics.*

Using the leading example, we can solve for the optimal policy, as given by

$$s^u = (1-\rho)\, F(\tau^u) \left(\frac{\rho}{\tau^u}\right)^{(1+\rho)/(1-\rho)}, \ \phi^u = (1-\rho)\, F(\tau^u) \left(\frac{\rho}{\tau^u}\right)^{\rho/(1-\rho)},$$

where the equilibrium threshold τ^u is given by $\rho h(\tau^u) = (1-\rho)\,\tau^u$.

Impact of GDPR

[23] The assumption that $B(s,\phi)$ is concave and $I(\phi)$ is convex ensures that these FOCs are also sufficient for the optimum.

Fixing the market failure benefits consumers from two aspects. First, GDPR changes consumers' default choice and allows consumers to opt-out but still use the digital platform's free service when they have a relatively high privacy cost. In other words, GDPR entitles consumers to trade their personal data for extra benefits, under which a consumer will do so if and only if the extra gain exceeds her privacy cost. Thus, no consumers will receive negative extra surplus under GDPR. In contrast, consumers were not granted with (or not aware of) such rights before GDPR, and some of them with relatively high privacy sensitivity actually receive negative extra surplus from providing data. Second, the digital platform takes into account consumers' privacy costs under GDPR and collects less consumer data, which can further benefit opt-in consumers by reducing their privacy cost. After GDPR, opt-in consumers receive extra surplus $V^u(\theta) = s^u(\tau^u - \theta)$, whereas opt-out consumers get zero. So, the aggregate extra consumer surplus CS^u is always positive after GDPR, whereas consumer surplus before GDPR CS^b becomes negative when $\mu > r\left(\bar{s}, \phi^b\right)/\bar{s}$.

GDPR has significantly reduced the total amount of data collected by digital platforms.[24] This could generate negative effects on data analytics. First, GDPR opt-out reduces the digital platform's incentives in its investment on data analytics. Fixing the data scale and differentiating both sides of $F(\tau) B_\phi(s, \phi) = I'(\phi)$ with respect to τ, it is straightforward to show that the equilibrium data analytics increases in the opt-in population:

$$\frac{d\phi}{d\tau} = \frac{f(\tau) B_\phi(s, \phi)}{I''(\phi) - F(\tau) B_{\phi\phi}(s, \phi)} > 0.$$

Second, the digital platform collects less consumer data after GDPR, resulting further in a lower ϕ as $\phi^u(s)$ increases in s.

For further comparison, recall that the optimal data analytics before GDPR $\phi^b(\bar{s})$ satisfies $r_\phi(\bar{s}, \phi) = I'(\phi)$, whereas the optimum after GDPR $\phi^u(s)$ solves $F(\tau) B_\phi(s, \phi) = I'(\phi)$. When the digital platform captures all social benefits from data mining, i.e., $r(s, \phi) = B(s, \phi)$, $\phi^b(s) > \phi^u(s)$ for any given s. In such a scenario, GDPR results in a lower level of data analytics. $\phi^b = \phi^b(\bar{s}) > \phi^b(s^u) > \phi^u(s^u) = \phi^u$.

The negative effect in digital platforms' investments is evidenced by the recent empirical studies including Jia et al. (2021), in which they find that, shortly after GDPR's rollout, the venture investment in technology by digital firms in the EU drops by more than 30% relative to both their US counterparts and counterparts in the rest of the world. They also find that

[24] The empirical study by Aridor et al. (2020) finds that GDPR resulted in approximately 12% reduction in total number of cookies.

the negative effect of GDPR on technology investment appears particularly pervasive for firms relying heavily on consumer data, including those in the healthcare and finance categories. Economists are also concerned that the significant reduction of investments in data analytics might cause a long-run negative effect on social welfare, since the innovation in data science becomes the main driving force of economic growth. We do not investigate such long-run effect on data analytics in this paper.

3.3 Mechanism Design for Consumer Data Acquisition

A uniform data collection policy cannot maximize total social welfare when consumers have heterogeneous privacy costs. Digital platforms are learning to adopt sophisticated mechanisms for consumer data acquisition. Many platforms now offer a menu of options for consumers to select different types of cookies according to their privacy preferences, instead of a single option for all non-essential cookies. The provision of a menu of options (contracts) for consumers has become common practice in many industries, including the telecom and electricity sectors among others. Not surprisingly, more digital platforms are expected to use such mechanisms to maximize their profits. In this section, we use the mechanism design approach to characterize the second-best policy. We briefly sketch the analysis here while leaving the detailed computation in Online Appendix D.

Suppose a digital platform is committed to a type-contingent data scale $s(\theta)$ and moreover offers a type-contingent transfer $t(\theta)$ for each type-θ consumer. Without loss of generality, we focus on the direct mechanism in which the digital platform requires a consumer to report her true type θ and then recommends the policy $\{s(\theta), t(\theta)\}$ accordingly. By reporting her true type, a type-θ consumer receives net utility $U(\theta) = u + V(\theta)$, where $V(\theta) = b(s(\theta), \phi) + t(\theta) - s(\theta)\theta$ is the consumer's extra benefit from opt-in. This consumer prefers opt-in to opt-out if $V(\theta) \geq 0$, which defines a consumer's participation constraint.

Since a consumers' marginal privacy cost θ is private information, the optimal policy must satisfy a consumer's incentive compatibility constraint to prevent this consumer from mis-reporting. A consumer of type θ, who mis-reports his type as $\hat{\theta}$, will receive

$$V\left(\tilde{\theta}, \theta\right) = t\left(\tilde{\theta}\right) + b\left(s\left(\tilde{\theta}\right), \phi\right) - s\left(\tilde{\theta}\right)\theta.$$

The mechanism is incentive compatible if $V(\theta) \geq V\left(\tilde{\theta}, \theta\right)$ for any $\tilde{\theta} \neq \theta$. In the Online Appendix, we check that $V(\theta)$ satisfies the single-crossing condition. When $s(\theta)$ is monotonic, the Incentive Compatibility (IC) constraint can be transformed into the following first-order condition, according to the standard mechanism design approach:

$$t'(\theta) = -\frac{\partial V}{\partial s} s'(\theta). \tag{8}$$

Moreover, using the Envelope Theorem, a consumer's net value from opt-in can be expressed as

$$V(\theta) = \int_{\theta}^{\bar{\theta}} s(x)\, dx.$$

The digital platform's profit under full participation is

$$\Pi = \int_{\underline{\theta}}^{\bar{\theta}} \left(r\left(s(\theta), \phi\right) - t(\theta)\right) dF(\theta) - I(\phi).$$

Substituting $t(\theta) = V(\theta) - b\left(s(\theta), \phi\right) + s(\theta)\theta$ and using integration by parts, we obtain

$$\Pi = \int_{\underline{\theta}}^{\bar{\theta}} W(\theta, \phi)\, dF(\theta) - I(\phi),$$

where $W(\theta, \phi) \equiv B\left(s(\theta), \phi\right) - s(\theta)\left(\theta + h(\theta)\right)$ is the net social benefit from type θ consumer and the term $s(\theta) h(\theta)$ is the information rent for type θ consumer to meet her IC constraint.

GDPR enables a market for consumers to trade their data for extra benefits. The optimal data scale $s(\theta)$ must equate a consumer's marginal social benefit from data provision, $B_s\left(s(\theta), \phi\right)$, to her privacy sensitivity plus the information rent, $\theta + h(\theta)$, as given by

$$B_s\left(s(\theta), \phi\right) = \theta + h(\theta). \tag{9}$$

The above FOC determines the equilibrium path $s^{**}(\theta, \phi) = B_s^{-1}\left(\theta + h(\theta)\right)$, which is decreasing in θ and increasing in ϕ. It follows that the digital platform collects less consumer data than the first-best level, given the same ϕ: $s^{**}(\theta, \phi) = B_s^{-1}\left(\theta + h(\theta)\right) < B_s^{-1}(\theta) = s^{*}(\theta, \phi)$.

The digital platform chooses the optimal data analytics such that its marginal cost is equal to the expected marginal social benefit, which is given by

$$\int_{\underline{\theta}}^{\bar{\theta}} B_{\phi}\left(s(\theta), \phi\right) dF(\theta) = I'(\phi). \tag{10}$$

Substituting $s^{**}(\theta, \phi)$ into the above FOC, the second-best ϕ^{**} is determined by

$$\int_{\underline{\theta}}^{\bar{\theta}} B_{\phi}\left(s^{**}(\theta, \phi), \phi\right) dF(\theta) = I'(\phi).$$

Compared to the first-best data analytics ϕ^{*} (given by (3)), then $s^{**}(\theta, \phi) < s^{*}(\theta, \phi)$ implies $B_{\phi}\left(s^{**}(\theta, \phi), \phi\right) < B_{\phi}\left(s^{*}(\theta, \phi), \phi\right)$ since $B_{\phi s}(s, \phi) > 0$. It follows that $\phi^{**} < \phi^{*}$ as $I'(\phi)$ increases in ϕ. This further implies $s^{**}(\theta) = s^{**}(\theta, \phi^{**}) < s^{*}(\theta, \phi^{**}) < s^{*}(\theta, \phi^{*}) = s^{*}(\theta)$, since

$s^*(\theta, \phi)$ increases in ϕ. Hence, the digital platform collects less data for each consumer and invests less in data analytics than the first-best outcome.

Finally, substituting $s^{**}(\theta)$ and ϕ^{**} into the optimal transfer gives

$$t^{**}(\theta) = \int_{\theta}^{\bar{\theta}} s^{**}(x)\,dx + s^{**}(\theta)\,\theta - b\left(s^{**}(\theta), \phi^{**}\right), \tag{11}$$

which leaves each consumer with a positive extra surplus $V(\theta) = \displaystyle\int_{\theta}^{\bar{\theta}} s(x)\,dx$. The second-best transfer $t^{**}(\theta)$ is not monotonic in θ, since it depends on the shape of consumer benefit as well.

The optimal mechanism $\{s^{**}(\theta), t^{**}(\theta), \phi^{**}\}$ is incentive compatible, enables full consumer participation, and implements the second-best outcome that maximizes total social welfare under asymmetric information, as summarized below:

Proposition 3 *The second-best policy for data acquisition* $\{s^{**}(\theta), t^{**}(\theta), \phi^{**}\}$ *is characterized by* (9)*,* (11)*, and* (10)*. Compared to the first-best outcome, the digital platform collects less consumer data and invests less in data analytics.*

Leading Example:

Using the leading example, the second best outcomes are given by

$$\begin{aligned} \phi^{**} &= (1-\rho)\int_{\underline{\theta}}^{\bar{\theta}} \left(\frac{\rho}{\theta + h(\theta)}\right)^{\rho/(1-\rho)} dF(\theta), \\ s^{**}(\theta) &= \left(\frac{\rho}{\theta + h(\theta)}\right)^{1/(1-\rho)} \phi^{**}. \end{aligned}$$

Compared to the first-best outcomes, we have $\phi^{**} < \phi^*$ and $s^{**}(\theta) < s^*(\theta)$.

4 Policy Implications

4.1 Implementation of Optimal Mechanism: A Guideline

Our key finding in Proposition 3 is that the second-best social optimum can be implemented through a type-dependent policy.[25] In practice, how to implement such contract remains an important policy issue. GDPR establishes a set of principles for consumer data collection, including

[25] The second-best outcome we have characterized is based on the assumption of continuous distribution of type θ. Of course, implementing the menu of options in the real world requires the segmentation of consumers into n different groups according to their privacy preferences and accordingly offer n different options (s_i, t_i), $i = 1, ..., n$.

lawfulness, fairness and transparency, purpose limitation, and data minimization, among others. However, it does not provide detailed guidelines. As a result, GDPR-compliance cookie policies can take different forms, as long as they meet the requirement for opt-in consent.[26] While digital platforms can control the data scale through the number and types of cookies, most of them do not provide detailed specifications on the different types of cookies. For the most part, consumers do not understand how much and what kinds of personal data will be collected through different types of cookies. Choosing an option of cookies is not as straightforward as choosing a mobile phone plan. Thus, the main obstacle to implementing the optimal policy is the asymmetric knowledge between the digital platforms and consumers regarding data collection.

The GDPR website has classified cookies into four categories according to their purposes: strictly necessary cookies, preference cookies, statistics cookies, and marketing cookies.[27] However, a typical website contains hundreds of different types of cookies, and a coarse classification such as this does not help consumers identify the features and attributes of their data being collected through these different cookies. GDPR compliance of cookie policies requires that digital platforms "must provide accurate and specific information about the data each cookie tracks and its purpose in plain language before consent is received." However, there does not exist a commonly acceptable interpretation for the "accurate and specific information in plain language" without any standards as references.

Thus, the most important guideline is the categorization and standardization of cookies, and this guideline must specify key features and attributes of consumer data being collected by a particular cookie, which include:

- 1. Variety of data: What types of personal information will be collected and in what kinds of formats?
- 2. Volume of data: How much personal data will be collected within a given time period, say one hour?
- 3. Purpose of data: Which parties are going to use these consumer data and for what purpose?
- 4. Analysis of data: Which party is going to process these data and what kind of data analytics tools will be used?

[26] See detailed discussion in the Online Appendix "Cookies".

[27] See https://gdpr.eu/cookies/.

- 5. Value of data: What is the potential or estimated value that the collected data can generate? What are the potential benefits to consumers?

- 6. Risk of data breach: What is the potential risk of data breach? How does the digital platform prevent this? If data breach occurs, how will consumers be informed and compensated?

Items 1 to 4 provide essential information about the scale of data being collected through a particular cookie, while items 5 and 6 help consumers assess the benefit and the cost of data sharing when they accept this particular cookie. Regulators should provide standardized templates for cookie specifications, as they did for consent forms in the GDPR website.

Apple's Privacy Labels

Apple's privacy labels introduced in December 2020 are a case in point. They require all Apple application developers to disclose their data collection practices by filling out privacy "nutrition" labels. The official form provided by Apple to app developers defines 14 data types, 32 specific data items, and six data usages. The six data usages include third-party advertising, developer's advertising or marketing, analytics, product personalization, other purposes, and app functionality. Moreover, Apple has classified three categories of purposes according to how widely data is shared with other parties: *Data Used to Track You, Data Linked to You, and Data Not Linked to You.*

Apple's movement is an important endeavour to standardize specifications of features on consumer data, a similar idea to the Nutrition Facts label on food packaging. Privacy labels disclose several key properties of data collection: variety of data, purpose of data, and analysis of data. The information is presented in a standardized format, which is designed to be easy to read for users. However, Apple's design of privacy labels does not allow app developers to provide consumers with a menu of options in the way our second-best policy describes. Hence, Apple's practice contributes to implementing the optimal uniform policy but not the second-best outcome.

4.2 Impact on Third-Party Cookies

GDPR categorizes cookies into two types according to its attribute of provenance: first-party cookies and third-party cookies. First-party cookies are placed directly in the user's device by the website (domain) the user is browsing, while third-party cookies are generated by external domains that differ from the site the user is browsing. Typically, third-party cookies are placed in the user's device by advertisers or web analytics providers.[28]

Since GDPR's rollout, several large digital platforms have blocked third-party cookies by default, including Apple's Safari and Mozilla's Firefox web browsers. In a widely-expected move, Google has announced its plan to block by default all third-party cookies for its Google Chrome browser in 2023. Online advertisers rely on third-party cookies. Google's move to phase out third-party's cookies will have an enormous negative impact on the online advertising market. Empirical studies by Alcobendas et al. (2021) find that such a ban would reduce publishers' revenue by 51%, and advertisers' surplus by 41%. However, they do not examine the welfare effect on consumers, which is the key purpose of antitrust intervention.

In this subsection, we use a variant of the baseline model to analyze the GDPR's impact on third-party cookies. The variant captures two key features related to the third-party cookies.

First, the platform runs the business of online advertising through the third-party, from which it receives a share from the third-party's revenue, however, the platform does not bear the cost of data analytics for such business. For the simplicity of analysis, we assume the digital platform obtains a share β from the third-party's revenue.[29]

Second, consumers are more concerned about privacy issues caused by third-party cookies than first-party. Unlike first-party cookies, third-party cookies can track a user across websites. The considerable use of such cookies creates an environment where cookies are continuously sent between browser and server. This behaviour magnifies the diffusion of user information and unnecessarily escalates potential interception by an adversary. Hence data collection through third-party cookies causes higher privacy sensitivity $\gamma\theta$ with the multiplier $\gamma > 1$.

Denote by s_t the scale of data collected by the third-party for advertising and by ϕ_t the related data analytics, where the subscript t stands for third-party. Online advertising generates a per-consumer social benefit $\hat{B}\left(s_t, \phi_t\right) = \hat{r}(s_t, \phi_t) + \hat{b}(s_t, \phi_t)$, where the third-party receives a revenue $\hat{r}(s_t, \phi_t)$ and consumers gain $\hat{b}(s_t, \phi_t)$. The third-party specializes in online advertising and its cost of data analytics is $\hat{I}\left(\phi_t\right)$. The digital platform can control the data scales s_t by monitoring and adjusting the number and categories of cookies. We focus the analysis on the uniform data policy here. Applying the methodology of the optimal mechanism design to this

[28] See the discussion "Cookies" in the Online Appendix.

[29] The digital platform shares the third-party's revenue from online advertisements. Such revenue-sharing is achieved through intermediaries such as the "Supply Side Platforms" and "Demand Side Platforms" (DSPs), using online auctions as the main revenue-extraction mechanism. See the discussion of microfoundation in Section 2.

case is quite straightforward.

Before GDPR, both the digital platform and the third-party incur zero marginal cost for data acquisition. The digital platform can share the third-party's revenue without incurring the cost of data analytics. Since $\hat{r}(s_t, \phi_t)$ increases in s_t, the third-party is allowed to collect the maximum scale of consumer data $s_t^b = \bar{s}_t$ before GDPR.

After GDPR, however, the digital platform has to compensate opt-in consumers for their privacy cost, caused not only by the first-party cookies but also by the third-party cookies. Since data collection through the third-party cookies causes a higher (marginal) privacy cost than that by the first-party cookies, it is not surprising that the third-party's online advertising business will be affected more severely under GDPR. For simplicity of analysis, we isolate data acquisition through third-party cookies from that through first-party cookies. Then, a consumer ticking the box of third-party cookies (opt-in) receives an extra surplus $V_t(\theta) = \hat{b}(s_t, \phi_t) - \gamma s_t \theta$, and they will opt in if $\theta \leq \tau_t = \left(\hat{b}(s_t, \phi_t) + t\right) / (\gamma s_t)$.

Most digital platforms do not have their own DSPs and only display the third-party's advertisements before GDPR. After GDPR, these digital platforms still rely on third-party's DSP but the third-party's data scale s_t decreases as a response of consumer opt-out. Such reduction of the data scale s_t can be quite significant due to the high privacy cost: the equilibrium data scale s_t^u decreases in γ. This further decreases the equilibrium data analytics, since ϕ_t and s_t are complementary, and reduces the third-party's profits.

The most profound impact on third-party cookies comes from several dominant digital platforms, including Google, Facebook, and Amazon, which run their own online advertising but also display advertisements from independent DSPs. That is, these digital platforms use both the first-party and third-party cookies for online advertising, although these two types of businesses are operated through separate Demand Side Platforms.

Suppose the digital platform still uses the third-party cookies after GDPR, its profit becomes

$$\begin{aligned}\Pi_t^u &= F(\tau)\left(\beta \hat{r}(s_t, \phi_t) + \hat{b}(s_t, \phi_t) - \gamma s_t \tau_t\right) \\ &= F(\tau)\left(\hat{B}(s_t, \phi_t) - (1-\beta)\,\hat{r}(s_t, \phi_t) - \gamma s_t \tau_t\right).\end{aligned}$$

When the digital platform can extract full surplus from the third-party such that $(1-\beta)\,\hat{r}(s_t, \phi_t) = \hat{I}(\phi_t)$, its maximum profit from accommodating the third-party is

$$\Pi_t^u = F(\tau)\left(\hat{B}(s_t, \phi_t) - \hat{I}(\phi_t) - \gamma s_t \tau\right).$$

If, instead, these platforms replace the third-party's DSP by its first-party DSP after GDPR, they could earn a profit $\Pi^u = F(\tau)\left(\hat{B}(s, \phi) - s\tau\right) - I(\phi)$. Since Π_t^u decreases in γ whereas Π^u

does not depend on γ, it follows that the digital platform will phase-out the third-party's online advertising when the multiplier γ is sufficiently large.

To examine the impact on third-party cookies, we first analyze the equilibrium in which the digital platform still uses the third-party's DSP, and then compare it with the counterfactural equilibrium in which the digital platform phases out the third-party cookies. The characterization of equilibrium is similar to the baseline model and is provided in Online Appendix E. For further comparison, we use a variant of the leading example as follows:

Example V: $\hat{B}(s_t, \phi_t) = s_t^{\sigma} \phi_t^{1-\sigma}$, $I(\phi_t) = \phi_t^2/2$, and $\hat{I}(\phi_t) = j\phi_t^2/2$ with $j \leq 1$.

We show in the Online Appendix that when $\gamma > \bar{\gamma}$, where $\bar{\gamma}$ is a cut-off threshold as defined in Online Appendix E, the digital platform finds it profitable to phase out the third-party. Comparing the equilibrium with and without third-party cookies, we find the following:

First, the equilibrium data scale s_t is determined by $\hat{B}_s(s_t, \phi_t) - (1-\beta)\,\hat{r}_s(s_t, \phi_t) = \gamma\tau_t$. Compared with the FOC for s, $\hat{B}_s(s, \phi) = \tau$, it follows that consumer data collected through third-party cookies generates less marginal social benefits but causes higher marginal privacy costs, given $\phi_t = \phi$ and $\tau_t = \tau$. Hence, the optimal data scale by the third-party is lower than the scale of data that would have been collected by the first-party, all other things equal. Moreover, the optimal data scale s_t^u decreases in γ.

This result indicates that GDPR's rollout has a more severe impact on the third-party's data collection, which is supported by the evidence that the number of third-party cookies has gone down by more than 30% in the EU's news websites after GDPR according to the study by Libert et al. (2018). However, we find that the phase-out increases the allocative efficiency of consumer data, which further allows the digital platform to collect more consumer data: $s^u > \gamma s_t^u > s_t^u$.

Second, the third-party's equilibrium data analytics ϕ_t^u satisfies $(1-\beta)\,\hat{r}_\phi(s_t, \phi_t) = \hat{I}'(\phi_t)$. Its investment in data analytics only partially captures the marginal social benefit, but it does not need to discount the reduction of opt-in consumers. By contrast, the first-party's optimal data analytics ϕ^u, which is determined by $F(\tau)\hat{B}_\phi(s, \phi) = I'(\phi)$, fully captures the marginal social benefit but is also discounted by the reduction of opt-in consumers ($F(\tau) < 1$). The net effect depends on the parameters α, β, and γ, as well as the difference between cost functions $I(\phi_t)$ and $\hat{I}(\phi_t)$. Using Example V, we show that the replacement increases the capacity of data analytics, i.e., $\phi^u > \phi_t^u$ if $\gamma > \tilde{\gamma}$, where $\tilde{\gamma}$ is a cut-off threshold defined in Online Appendix E.

Third, when the social benefit takes the function form of Cobb-Douglas, phasing out third-party's cookies does not change the cut-off threshold: $\tau^u = \tau^u$. This result simplifies the comparison of consumer surplus. The net surplus for opt-in consumers is $V_t(\theta) = \gamma s_t^a (\tau^u - \theta)$

with third-party cookies and becomes $V(\theta) = s^u(\tau^u - \theta)$ without them. Thus, the replacement benefits consumers if $s^u > \gamma s_t^a$. We show in the Online Appendix that the replacement increases net consumer surplus (when $\gamma > \bar{\gamma}$).

The above analysis is summarized as follows:

Proposition 4 *Suppose digital platforms can replace the third-party's DSP by its own DSP. Such a phase-out happens when $\gamma > \bar{\gamma}$ in Example V, in which the replacement increases data scale and total consumer surplus, and moreover improves data analytics if $\gamma > \tilde{\gamma}$.*

Google's movement of phasing out third-party cookies was originally expected to launch in early 2022 and was postponed to 2023 due to criticism and concern voiced by industry and government. We find that, while the replacement reduces the third-party's profitability, it reduces consumer privacy costs, increases the allocative efficiency of consumer data, and increases consumer surplus. Hence, prohibiting Google's move to phase-out third-party cookies can harm consumers. However, Google could potentially take advantage of such a move to further expand its market share in online advertising, and such monopolization could dampen competition in online advertising. Hence, competition authorities need to balance the short-run consumer gain and the long-run competition harm.

5 Data Acquisition with Personalization

The analysis in the baseline model is focused on intermediary digital platforms that do not sell their own products/services. When digital platforms also sells their own products, consumer data collected through its cookies can be used for personalization. Through data mining, the digital platform can recommend a best-matching product (personalized product) to a consumer according to her taste, while charging a personalized price based on her willingness to pay. In this section, we consider an extension of the baseline model and apply the same methodology in order to analyze the equilibrium data policy with personalization.

Suppose the digital platform sells a product in addition to its free service. A standard version of the product is supplied in a competitive market at a price normalized to zero. Each consumer demands one unit of the product and derives utility $v - x$ from consuming the standard version, where x represents the consumer's general taste and it is uniformly distributed in $[0, 1]$. The digital platform knows a consumer's taste x through data mining from her past activities on the website.

Consumer data will be collected through the digital platform's first-party cookies. If a consumer does not consent to data collection, she can only get the standard version. The digital platform cannot charge this consumer a personalized price based on her taste x because the standard version is supplied in a competitive market at a price zero.

If a consumer consents to data collection (or by default before GDPR), the platform can offer her a personalized version of the product, from which this consumer obtains a higher level of utility $v - (1 - m(s,\phi))\,x$, where $v > 1$. Here, $m(s,\phi)\,x$ is the extra value from improved matching. The provision of a personalized product relies on the prediction of the consumer preferences on some specific attributes of the product, not limited to her general taste x, thus, data collection is essential for personalization. The precision of such prediction, $m(s,\phi)$, depends on the data scale s and data analytics ϕ, with $m(s,\phi) < 1$, $m'(s,\phi) > 0$ and $m''(s,\phi) \leq 0$. We assume away the additional cost of supplying personalized version. Meanwhile, the digital platform can charge the consumer a personalized price $p(x)$ for the personalized version. Each personalized version (with a personalized price) is offered personally and privately, which is not comparable across consumers.

Hence, consumer data generates a social benefit $m(s,\phi)\,x$ through personalization, from which the digital platform earns a revenue $p(x)$ and leaves this consumer with a benefit $m(s,\phi)\,x - p(x)$. For illustration, we focus on the uniform policy only, while the analysis of the type-dependent policy is provided in Online Appendix F.

Before GDPR

Before GDPR, the digital platform acquires consumer data at zero marginal cost and a consumer's privacy cost $s\theta$ is her "sunk" cost when she uses the platform's free service. This consumer will purchase the personalized version rather than the standard version if the extra benefit from personalization $m(s,\phi)\,x$ exceeds the personalized price $p(x)$. This pins down the digital platform's personalized price to $p(x;s,\phi) = m(s,\phi)\,x$, in which each consumer receives zero net surplus from providing data. The digital platform's profit is

$$\Pi^b = \int_0^1 m(s,\phi)\,x dx - I(\phi) = \frac{m(s,\phi)}{2} - I(\phi).$$

The platform acquires the maximum scale of data: $s^b = \bar{s}$, which exceeds the social optimum. Its optimal data analytics equates the expected marginal benefit from the improved matching value to the marginal cost, as given by $m_\phi\left(\bar{s},\phi^b\right)/2 = I'\left(\phi^b\right)$. Since data analytics is complementary to data scale, the digital platform builds excessive capacities of data analytics compared to the social optimum. Consumers actually receive a negative extra surplus from providing data:

$V(\theta,x) = m(s,\phi)x - p(x) - s\theta = -s\theta$.

After GDPR

Consumers can choose to opt-out under GDPR. The timing of the game is given as follows. In stage one, the digital platform announces the data policy $\{s,\phi\}$. Observing the policy, a consumer with taste x and type θ decides whether or not to opt out. Opt-out consumers will purchase the standard version and receive $v - x$. In stage 2, the digital platform offers each opt-in consumer a personalized version and charges a personalized price. Consumers then decides whether or not to accept the personalized offer. Moreover, a consumer rejecting the personalized offer can choose to opt-out to avoid the privacy cost, because such an option is always open. This ensures the consumer a reservation payoff $v - x$ by choosing to opt-out at any time.

A consumer's privacy costs $s\theta$ is not a sunk cost after GDPR. An opt-in consumer accepting the personalized offer receives $v - (1 - m(s,\phi))x - p(x) - s\theta$, whereas this consumer can always secure a payoff $v - x$ by choosing opt-out. Thus, a consumer with taste x and type θ will accept the personalized offer if the net surplus $m(s,\phi)x - p(x)$ exceeds the privacy cost $s\theta$. That is, if

$$\theta \leq \tau(x) \equiv \frac{m(s,\phi)x - p(x)}{s},$$

where $\tau(x)$ is the cut-off value of θ for the marginal consumer. The digital platform's profit is

$$\Pi^u = \int_0^1 p(x) F(\tau(x)) dx - I(\phi).$$

The digital platform sets personalized prices $p(x) = m(s,\phi)x - s\tau(x)$, which leaves each opt-in consumer a positive surplus $V(x,\theta) = m(s,\phi)x - p(x) - s\theta = s(\tau(x) - \theta) > 0$.

We first solve for the digital platform's optimal personalized prices given $\{s,\phi\}$. Substituting $p(x)$ into the above profit function, we obtain

$$\Pi^u = \int_0^1 (m(s,\phi)x - s\tau(x)) F(\tau(x)) dx - I(\phi).$$

Thus, given $\{s,\phi\}$, choosing $p(x)$ is equivalent to choosing $\tau(x)$ in the maximization. The maximization of Π^u with respect to $\tau(x)$ requires that the term under the integrand $(m(s,\phi)x - s\tau(x)) F(\tau(x))$ be maximized with respect to $\tau(x)$ for all x. Differentiating the integrand with respect to $\tau(x)$ and solving for the FOC leads to

$$m(s,\phi)x = s(\tau(x) + h(\tau(x))). \tag{12}$$

Similar to the result in the baseline model, the equilibrium cut-off threshold $\tau(x)$ is given by equating the per consumer social benefit to the privacy cost for the marginal consumer $\theta = \tau(x)$ plus a rent $sh(\tau(x))$ from all opt-in consumers.

The optimal data scale and data analytics can be determined using the similar approach as in the baseline model, which is left in the Online Appendix. Since the digital platform must take into account the opt-in consumers' privacy cost, it will collect less consumer data than before GDPR, which results in a lower level of data analytics. Hence, the main results and insights in the baseline model hold here. Furthermore, GDPR improves consumer welfare since opt-in consumers get a positive extra surplus whereas opt-out consumers get zero, compared to the negative extra surplus before GDPR.

Summarizing the above analysis leads to:

Proposition 5 *Suppose the digital platform sells personalized products and charges personalized prices to opt-in consumers. Before GDPR, the platform can extract full consumer surplus from data provision through personalized pricing, leaving consumers a negative extra surplus from data-sharing. After GDPR, the digital platform collects less consumer data and leaves positive extra surplus to opt-in consumers. GDPR improves consumer welfare.*

6 Literature and Conclusion

Related Literature

There is a growing literature of theoretical research focusing on privacy rights and data security, but only a few of them examine the impact of GDPR by modelling explicitly consumers' opt-in or opt-out choices.

Choi et al. (2019) provide a model of privacy and personal data collection with information externalities. They consider the heterogeneity of privacy sensitivity on different types of personal data but assume that such sensitivity does not vary across consumers. Moreover, they assume that privacy costs increase as more users share personal information, which reflects data externalities. By contrast, we emphasize the heterogeneity of privacy sensitivity across consumers while using a measurement of data scale to incorporate different types of information. We also consider a consumer's privacy cost increasing with the scale of data collected from that consumer but assume away the cross-consumer externality. They focus on the digital platform's data collection policy but do not consider data analytics, whereas we treat both data scale and data analytics as two strategic inputs, while also considering the different features of these input costs.

These different modelling features lead to different theoretical findings and policy implications. They identify excessive data collection by a monopoly digital platform before GDPR, but the main mechanism for this result is information externalities across users and users' coordination failure in data sharing. They show that such information externalities can make GDPR ineffective and argue that monetary inducement for opt-in should not be allowed. By contrast, we find that the excessive data collection is the result of a market failure in which the digital platform bears zero marginal cost in data acquisition due to the bundling of the digital services with data collection. We show that monetary inducement can play an important role in fixing such market failure and, furthermore, characterize the optimal mechanism with type-contingent compensation for consumers.

Ke and Sudhir (2020) analyze a model of behavioural-based pricing that endogenizes a consumer's decision to exercise the rights to opt-in, erases personal data and data portability, and in which firms can offer personalized products to opt-in consumers. They assume exogenous privacy costs that are determined by the probability of data breach and the expected loss from the breach. They emphasize the role of GDPR in promoting data security and find that by reducing expected privacy breach costs, data security mandates increase opt-in, consumer surplus and a firm's profit. By contrast, we investigate the impact of GDPR through its consent requirement for data acquisition and shed light on its fundamental role in fixing the market failure.

Another closely related paper is that of Fainmesser et al. (2021), who develop a model of digital privacy where a digital platform chooses the data level (scale) and the data protection strategies, in which a data breach by third parties can impose privacy costs on users. A consumer's privacy costs increase with her online activities (data scale) but are assumed homogeneous among all consumers. They do not explicitly model consumers' opt-in or opt-out choices and thus do not examine the impact of GDPR. Instead, they focus on the optimal data protection strategies against adversaries and show that the social optimal policy combines a minimum data protection requirement with a tax proportional to the amount of data collected.

We consider consumer heterogeneity in their privacy sensitiviy. Consumers' opt-in decision is endogeneously characterized by a cut-off threshold of privacy sensitivity balancing the benefits and costs of data sharing. This allows us to examine the impact of GDPR on social welfare. We characterize the social optimal data acquisition mechanism, which combines a type-contingent data scale with a type-contingent compensation for consumers.

Our paper is also related to a growing literature of data collection and data intermediation, which treats consumer data as an informative signal for the prediction of a consumer's willingness to pay and/or for the improvement of product recommendation. The precision of such prediction or recommendation relies on the scale and quality of data and can be also affected by information externalities. Acemoglu et al. (2021) find that digital platforms over-collect data due to information externalities and moreover characterize conditions to shut down data markets for welfare improvement. Bergemann et al. (2021) propose a model of data intermediation to analyze the incentives for sharing individual data in the presence of information externalities. They show that the intermediary enables firms to offer personalized product recommendations but not personalized prices. Ichihashi (2021a) considers a single data intermediary and asks how complements or substitutes to consumer signals affect the equilibrium price of the individual data under information externalities.

In the absence of information externalities, Ichihashi (2020) studies both personalized pricing and product recommendations, and shows that a seller benefits from committing not to use the consumer's information to set prices. Ichihashi (2021b) considers a dynamic model of consumer privacy and platform data collection. Data collection generates a cross-period effect, through which the platform lowers the level of privacy protection, while consumers lose privacy and become gradually worse off.

These studies treat the processed consumer data as an informative signal and provide a microfoundation for the analysis of the value of consumer data. By contrast, we separate data and processing as two strategic inputs to emphasize the different nature of benefits and costs for these inputs. The studies also consider consumer privacy costs, which are treated as a reduced form of utility loss from data provision and are assumed homogeneous across consumers, whereas we consider heterogeneous consumer privacy costs. Finally, these papers study the optimal data collection strategies in the presence of consumer privacy concerns, while our paper directly examines the impact of GDPR on social welfare.

Concluding Remarks

General Data Protection Regulation aims to protect consumer data privacy, however, its adverse effects have been widely documented. We present a new model for the analysis of consumer data acquisition under privacy regulation. We treat both data and analytics as separate strategic variables and consider the heterogeneity of privacy costs across consumers. Using this model to examine the impact of GDPR, we identify a market failure before GDPR and find that GDPR activates a market for data acquisition by imposing consent requirements on data acquisition. We further study the optimal design of the mechanism for consumer data acquisition and deliver important policy implications for implementing the social optimum. We conclude this paper with a number of remarks outlining the model's limitations.

Several papers including Choi et. al (2019) and Acemoglu et. al (2021) focus on information externalities in data collection. We isolate these effects for the simplicity of analysis. This modelling approach is also justified by distinguishing the effects of within-user and cross-user data mining.

A large proportion of big data applications are focused on marketing and product/service recommendations, in which the digital platform uses the predictive data analytics for forecasting and estimating the probability that a consumer will accept a recommended product/service and his/her willingness to pay. Such data analytics rely heavily on a consumer's historical data (behavioural data) and her other available activity profiles (i.e., within-user data mining) and uses behavioural-based machine learning to provide personalized product recommendations and prices. For instance, Fitbit premium service provides users with a personalized service on health, sleep and fitness based on the data collected by their Fitbit device (as well as their Google accounts after it has been acquired by Google). Likewise, Amazon.com recommends a best-matching product for a premium customer based on her historical searching and purchasing data as well as on other available personal data. In other applications, however, data analytics adopts machine learning by leveraging data sets across different users as well. Such applications include Grammarly for spelling, grammar, and tone, Cruise for autopilot, Deep sentinel for home security, etc..[30] This paper focuses on the applications of data analytics that rely mainly on the within-user data mining, which may well suit business models for personalized advertising, product recommendation, pricing, and so on.

We believe that our main insights and results do not change qualitatively when we consider cross-user data mining as well. Suppose the revenue and benefits are also increasing with the aggregate scale of data collected from all consumers, m. It is reasonable to assume that consumer privacy cost is not related to m. Before GDPR, there is a market failure and as a result, digital platforms collect the maximum scale of data. The presence of data externalities does not affect this result. After GDPR, digital platforms will collect more data than without externalities, which further leads to a higher level of data analytics. When digital platforms use the uniform policy, the characterization of equilibrium becomes more complicated since the cut-off threshold τ cannot be expressed explicitly since $m = s\tau$ appears in the benefit function $b(s, m, \phi)$. Since $b(s, m, \phi)$ increases with τ, the presence of data externalities could reduce the digital platform's

[30] See Hagiu and Wright (2021) for further discussion on within-user vs. cross-user learning.

transfer t and increase opt-in population, which is a welfare-enhancing effect. When the digital platform offers a type-dependent policy, the usual mechanism design approach does not apply here since the benefit function $b(s, m, \phi)$ depends on the policy offered to not only an individual consumer but to all consumers. We cannot characterize the optimal mechanism with such an aggregation effect.

Consumer data consists of multiple dimensional attributes and categories. We use a variable of data scale as integration of all attributes, which is another modelling limitation. Choi et. al (2019) emphasize the heterogeneity of consumer privacy sensitivity on different categories of data.[31] Combining their methodology with ours could provide a comprehensive characterization of privacy costs and offer important policy implications. However, handling two dimensional heterogeneity is technically demanding, and which we leave as our next research topic.

[31] Prince and Wallsten (2021) report survey results documenting heterogeneous valuation of online privacy across different data types and different countries.

References

Acemoglu, D., Makhdoumi, A., Malekian, A. and A. Ozdaglar (2021). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, forthcoming.

Acquisti, A., Taylor, C. and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature*, 54(2): 442-492.

Alcobendas, M., S. Kobayashi, and M. Shum (2021). The Impact of Privacy Protection Measures on Online Advertising Markets, *working paper.*

Aridor, G., Che, Y.-K. and T. Salz (2020). The economic consequences of data privacy regulation: Empirical evidence from GDPR. NBER Working Paper 26900.

Becker, G. S. (1980). Privacy and malfeasance: A comment. *Journal of Legal Studies*, 9(4): 823-826.

Bergemann, D., Bonatti, A. and T. Gan (2021). The economics of social data. Working paper, http://www.mit.edu/~bonatti/social.pdf

Choi, J. P., Jeon, D.-S. and B.-C. Kim (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics*, 173: 113-124.

Ebert, N., Ackermann, K. A., and B. Scheppler (2021). Bolder is better: Raising user awareness through salient and concise privacy notices. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, article no: 67, 1-12.

Fainmesser, I. P., Galeotti, A. and R. Momot (2019). Digital privacy. HEC Paris Research Paper No. MOSI-2019-1351.

Hagiu, A. and J. Wright (2021), Data-Enabled Learning, Network Effects and Competitive Advantage", *working paper.*

Ichihashi, S. (2020). Online privacy and information disclosure by consumers. *American Economic Review*, 110(2): 569-595.

Ichihashi, S. (2021a). The Economics of Data Externalities", *Journal of Economic Theory*, 197 (1)

Ichihashi, S. (2021b). Dynamic privacy choices. *American Economic Journal: Microeconomics*, forthcoming.

Jia, J., Jin, G. Z. and L. Wagman (2021). The short-run effects of the General Data Protection Regulation on technology venture investment. *Marketing Science*, 40(4): 661-684.

Johnson, G. A., Shriver, S. K. and S. Du (2020). Consumer privacy choice in online advertising: Who opts out and at what cost to industry? *Marketing Science*, 39(1): 33-51.

Ke, T. T. and K. Sudhir (2020). Privacy rights and data security: GDPR and personal data driven markets. Working paper, https://ssrn.com/abstract=3643979.

Li, H., L. Yu and W. He, (2019). The Impact of GDPR on Global Technology Development. *Journal of Global Information Technology Management*, Vol 22: 1-6.

Libert, T., L. Graves, and R. Nielsen (2018). Changes in Third-Party Content on European News Websites after GDPR. https://ora.ox.ac.uk/objects/uuid:5a5d4eea-6e74-49b4-8c77-71ec6760f127

Lin, T. (2021). Valuing intrinsic and instrumental preferences for privacy. *Marketing Science*, forthcoming.

Prince, J. and S. Wallsten (2021). How much is privacy worth around the world and across platforms? Working paper, https://ssrn.com/abstract=3528386.

Provost, F., and T. Fawcett (2013). *Dat Science for Business.* O'Reilly Media.

Online Appendix
(Not for Publication)

A: Condition for Assumption C

We illustrate the condition for Assumption C. Totally differentiating both sides of $B_s\left(s\left(\theta\right),\phi\right) = \theta$ with respect to θ, we have

$$B_{ss}\left(s\left(\theta\right),\phi\right)s'\left(\theta\right) = 1.$$

Thus, $B_{ss}\left(s\left(\theta\right),\phi\right) < 0$ implies $s'\left(\theta\right) < 0$. That is, the first-best data scale is decreasing in θ. Thus, a sufficient condition for the interior optimum $s^*\left(\theta\right) < \bar{s}$ is $s^*\left(\underline{\theta}\right) < \bar{s}$.

Note that $s^*\left(\underline{\theta}\right)$ is given by

$$B_s\left(s^*\left(\underline{\theta}\right),\phi^*\right) = \underline{\theta}.$$

As $B_s\left(s^*\left(\underline{\theta}\right),\phi^*\right)$ is decreasing in s, $s^*\left(\underline{\theta}\right) < \bar{s}$ is equivalent to

$$B_s\left(\bar{s},\phi^*\right) < B_s\left(s^*\left(\underline{\theta}\right),\phi^*\right) = \underline{\theta},$$

which amounts to

$$\underline{\theta} > B_s\left(\bar{s},\phi^*\right).$$

Assume that $\lim_{s\to\infty} B_s\left(s,\phi\right) = 0$ for any ϕ. The right-hand-side tends to zero when $\bar{s} \to \infty$. When $\bar{s}$ is sufficiently large, $B_s\left(\bar{s},\phi\right)$ becomes arbitrarily small. Thus, the condition $\underline{\theta} > B_s\left(\bar{s},\phi\right)$ is satisfied for any lower bound $\underline{\theta}$ not close to zero.

B: Leading example

Suppose that the per-consumer revenue and benefit functions take the form of Cobb-Douglas with $0 < \rho < 1$:

$$r\left(s,\phi\right) = \alpha s^{\rho}\phi^{1-\rho},\ b\left(s,\phi\right) = \left(1-\alpha\right)s^{\rho}\phi^{1-\rho}.$$

It is straightforward to check that $B\left(s,\phi\right) = s^{\rho}\phi^{1-\rho}$ is concave. Assume further that $I\left(\phi\right) = \phi^2/2$.

Before GDPR

The digital platform collects the maximum amount of consumer data before GDPR: $s = \bar{s}$. The optimal data analytics before GDPR is given by

$$r_{\phi}\left(s,\phi\right) = \alpha\left(1-\rho\right)s^{\rho}\phi^{-\rho} = I'\left(\phi\right) = \phi.$$

This FOC defines the equilibrium ϕ as a function of s:

$$\phi^b\left(s\right) = \left(\alpha\left(1-\rho\right)\right)^{1/(1+\rho)}s^{\rho/(1+\rho)},$$

which is increasing in s.

After GDPR

We can rewrite the FOCs for s and τ as $\rho s^{\rho-1}\phi^{1-\rho}=\tau$ and $s^{\rho}\phi^{1-\rho}=s\left(\tau+h(\tau)\right)$ respectively. Combining these two equations determines the optimal threshold τ^{u} as given by

$$\rho h(\tau^{u})=\left(1-\rho\right)\tau^{u}.$$

Assume further the function $k\left(x\right)\equiv h\left(x\right)/x$ is monotonic, then $\tau^{u}=k^{-1}\left(\frac{1-\rho}{\rho}\right)$.

In addition, the optimal data analytics is given by $F\left(\tau\right)\left(1-\rho\right)s^{\rho}\phi^{-\rho}=\phi$. Combining it with the FOC for s, $\rho s^{\rho-1}\phi^{1-\rho}=\tau$, we have

$$s=\phi^{2}\frac{\rho}{\left(1-\rho\right)F\left(\tau\right)\tau}.$$

Then, solving for s and ϕ leads to

$$\phi^{u}=\left(1-\rho\right)F\left(\tau^{u}\right)\left(\frac{\rho}{\tau^{u}}\right)^{\rho/(1-\rho)},$$

and

$$s^{u}=\left(1-\rho\right)F\left(\tau^{u}\right)\left(\frac{\rho}{\tau^{u}}\right)^{(1+\rho)/(1-\rho)}.$$

First-Best Outcome

The first-best data scale $s\left(\theta\right)$ is determined by

$$B_{s}\left(s\left(\theta\right),\phi\right)=\rho s\left(\theta\right)^{\rho-1}\phi^{1-\rho}=\theta.$$

Solving for $s\left(\theta,\phi\right)$ gives

$$s\left(\theta,\phi\right)=\left(\frac{\rho}{\theta}\right)^{1/(1-\rho)}\phi.$$

Substituting into the FOC for ϕ,

$$\int_{\underline{\theta}}^{\bar{\theta}}\left(1-\rho\right)s\left(\theta\right)^{\rho}\phi^{-\rho}dF\left(\theta\right)=\phi,$$

we obtain

$$\phi^{*}=\left(1-\rho\right)\int_{\underline{\theta}}^{\bar{\theta}}\left(\frac{\rho}{\theta}\right)^{1/(1-\rho)}dF\left(\theta\right),$$

and

$$s^{*}\left(\theta\right)=\left(\frac{\rho}{\theta}\right)^{1/(1-\rho)}\phi^{*}.$$

It follows that $s^{*}\left(\theta\right)$ decreases in θ.

Second-Best Outcome

Similarly, the second-best data policy is given by

$$\begin{aligned}
\phi^{**} &= (1-\rho)\int_{\underline{\theta}}^{\bar{\theta}}\left(\frac{\rho}{\theta+h(\theta)}\right)^{\rho/(1-\rho)} dF(\theta), \\
s^{**}(\theta) &= \left(\frac{\rho}{\theta+h(\theta)}\right)^{1/(1-\rho)}\phi^{**}.
\end{aligned}$$

Compared to the first-best data analytics, $\phi^{**} < \phi^{*}$. Moreover, $s^{**}(\theta) < s^{*}(\theta)$ for any given ϕ. It follows that $s^{**}(\theta) < s^{*}(\theta)$.

C: Equilibrium after GDPR under Uniform Policy

The digital platform's profit function is

$$\Pi^{u} = F(\tau)\left(r(s,\phi)-t\right) - I(\phi) = F(\tau)\left(B(s,\phi)-s\tau\right) - I(\phi),$$

where the second equality comes by substituting $t = s\tau - b(s,\phi)$. Given s and ϕ, choosing the optimal t is equivalent to choosing the optimal threshold τ. The optimization program can be decomposed into two steps. First, given data analytics ϕ, the digital platform chooses s and τ to maximize its revenue $R(s,\tau;\phi) \equiv F(\tau)\left(B(s,\phi)-s\tau\right)$. Second, given the optimal s and τ, the digital platform chooses ϕ to maximize $\Pi^{u} = R(\phi) - I(\phi)$. We now solve for the equilibrium following two steps.

Step 1: Since $B(s,\phi) - s\tau$ is concave in s given ϕ, the maximization of R with respect to s has a unique interior solution, which is determined by the following first-order condition:

$$B_s(s,\phi) = \tau.$$

Meanwhile, it is straightforward to check that $R(s,\tau;\phi)$ is also concave in τ, as well as under the assumption of increasing $h(\tau)$. Thus, optimization with respect to τ has a unique interior solution as given by the FOC:[32]

$$B(s,\phi) = s\left(\tau + h(\tau)\right).$$

Substituting $\tau = B_s(s,\phi)$ into the above equation, we have

$$B(s,\phi) = s\left(B_s(s,\phi) + h(B_s(s,\phi))\right).$$

[32] It is easy to check that the Hessian matrix is negative definite and the optimum s^{a} and τ^{a} are the solutions to the above two FOCs.

Solving for the above equation determines the optimal data scale $s^u(\phi)$, and then the optimal threshold $\tau^u(\phi) = B_s(s^u(\phi), \phi)$. Differentiating both sides of $B_s(s, \phi) = \tau$ with respect to ϕ, it is straightforward to see that $s^u(\phi)$ increases in ϕ: $ds^u(\phi)/d\phi = -B_{s\phi}/B_{ss} > 0$.

Step 2: Substitute $s^u(\phi)$ and $\tau^u(\phi)$ into the $R(s, \tau; \phi)$. Maximizing $\Pi^u = R(\phi) - I(\phi)$ with respect to ϕ has a unique interior solution as $R(\phi)$ is concave and $I(\phi)$ is convex. Differentiating Π^u with respect to ϕ and using the envelope theorem, the optimal data analytics is the solution of the following FOC

$$F(\tau)B_\phi(s, \phi) = I'(\phi).$$

We now compare the equilibrium with the first-best outcome. For opt-in consumers with $\theta < \tau^u$, fixing τ^u, we conduct comparative statics as follows. Recall that the first-best data scale $s^*(\theta, \phi)$, is given by $B_s(s^*(\theta, \phi), \phi) = \theta$, whereas the optimal data scale $s^u(\phi)$ is the solution of $\tau^u = B_s(s^u(\phi), \phi)$. Then, $B_s(s^*(\theta, \phi), \phi) = \theta < \tau^u = B_s(s^u(\phi), \phi)$ implies $s^*(\theta, \phi) > s^u(\phi)$ for any given ϕ, since $B_{ss} < 0$.

Moreover, the optimal data analytics ϕ^u is the solution of $F(\tau^u)B_\phi(s^u(\phi), \phi) = I'(\phi)$, whereas the first-best data analytics ϕ^* is given by

$$\int_{\underline{\theta}}^{\bar{\theta}} B_\phi(s^*(\theta, \phi), \phi)\, dF(\theta) = I'(\phi).$$

Comparing the left-hand sides of the two FOCs and noting that $B_{\phi s} > 0$, we have

$$\int_{\underline{\theta}}^{\bar{\theta}} B_\phi(s^*(\theta, \phi), \phi)\, dF(\theta) > \int_{\underline{\theta}}^{\bar{\theta}} B_\phi(s^u(\phi), \phi)\, dF(\theta) = B_\phi(s^u(\phi), \phi) > F(\tau^u)B_\phi(s^u(\phi), \phi).$$

It follows that $\phi^* > \phi^u$ since $I'(\phi)$ is increasing in ϕ.

Note that both $s^*(\theta, \phi)$ and $s^u(\phi)$ increase in ϕ. Then, $s^*(\theta, \phi) > s^u(\phi)$ for any given ϕ and $\phi^* > \phi^u$ together imply $s^*(\theta) = s^*(\theta, \phi^*) > s^u(\phi^*) > s^u(\phi^u) = s^u$, for any $\theta < \tau^u$.

D: Mechanism Design

We now provide full characterization of the equilibrium for the second best outcomes. Recall that

$$V(\theta) = b(s(\theta), \phi) + t(\theta) - s(\theta)\theta.$$

Note that

$$\frac{\partial V}{\partial s} = b_s(s(\theta), \phi) - \theta, \ \frac{\partial V}{\partial t} = 1,$$

then

$$\frac{\partial}{\partial\theta}\left[\frac{\partial V/\partial s}{\partial V/\partial t}\right]=\frac{\partial}{\partial\theta}\left[b_s\left(s\left(\theta\right),\phi\right)-\theta\right]=-1<0.$$

It follows that the single-crossing condition is satisfied.

We first characterize the Incentive Compatibility (IC) constraint. A consumer with type θ who mis-reports her type as $\tilde{\theta}$ will receive

$$V\left(\tilde{\theta},\theta\right)=t\left(\tilde{\theta}\right)+b\left(s\left(\tilde{\theta}\right),\phi\right)-s\left(\tilde{\theta}\right)\theta.$$

The data policy is incentive compatible if $V\left(\theta\right)\geq V\left(\tilde{\theta},\theta\right)$ for any $\tilde{\theta}\neq\theta$. Differentiating $V\left(\tilde{\theta},\theta\right)$ with respect to $\tilde{\theta}$, we have

$$\frac{\partial V\left(\tilde{\theta},\theta\right)}{\partial\tilde{\theta}}=t'\left(\tilde{\theta}\right)+\left(b_s\left(s\left(\tilde{\theta}\right),\phi\right)-\theta\right)s'\left(\tilde{\theta}\right).$$

The IC constraint requires $V\left(\tilde{\theta},\theta\right)$ be maximized at $\tilde{\theta}=\theta$, which implies

$$t'\left(\theta\right)+\left(b_s\left(s\left(\theta\right),\phi\right)-\theta\right)s'\left(\theta\right)=0.$$

By the envelope theorem, we obtain

$$\frac{dV\left(\theta\right)}{d\theta}=-s\left(\theta\right).$$

At the optimum the participation constraint of the highest type is binding, i.e., $V\left(\bar{\theta}\right)=0$, which implies

$$V\left(\theta\right)=V\left(\bar{\theta}\right)+\int_{\theta}^{\bar{\theta}}s\left(x\right)dx=\int_{\theta}^{\bar{\theta}}s\left(x\right)dx.$$

The digital platform's total profits is given by

$$\Pi=\int_{\underline{\theta}}^{\bar{\theta}}\left(r\left(s\left(\theta\right),\phi\right)-t\left(\theta\right)\right)dF\left(\theta\right)-I\left(\phi\right).$$

Using

$$t\left(\theta\right)=V\left(\theta\right)-b\left(s\left(\theta\right),\phi\right)+s\left(\theta\right)\theta,$$

we can rewrite Π as

$$\begin{aligned}\Pi &= \int_{\underline{\theta}}^{\bar{\theta}}\left(r\left(s\left(\theta\right),\phi\right)-V\left(\theta\right)+b\left(s\left(\theta\right),\phi\right)-s\left(\theta\right)\theta\right)dF\left(\theta\right)-I\left(\phi\right)\\ &= \int_{\underline{\theta}}^{\bar{\theta}}\left[B\left(s\left(\theta\right),\phi\right)-s\left(\theta\right)\theta\right]dF\left(\theta\right)-\int_{\underline{\theta}}^{\bar{\theta}}\left(\int_{\theta}^{\bar{\theta}}s\left(x\right)dx\right)dF\left(\theta\right)-I\left(\phi\right).\end{aligned}$$

Using integration by parts, we obtain

$$\int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\theta}^{\bar{\theta}} s\left(x\right) dx \right) dF\left(\theta\right) = \left[\left(\int_{\theta}^{\bar{\theta}} s\left(x\right) dx \right) F\left(\theta\right) \right]_{\underline{\theta}}^{\bar{\theta}} + \int_{\underline{\theta}}^{\bar{\theta}} F\left(\theta\right) s\left(\theta\right) d\theta = \int_{\underline{\theta}}^{\bar{\theta}} F\left(\theta\right) s\left(\theta\right) d\theta.$$

Substituting this into Π, we have

$$\Pi = \int_{\underline{\theta}}^{\bar{\theta}} \left[B\left(s\left(\theta\right), \phi\right) - s\left(\theta\right)\left(\theta + h\left(\theta\right)\right) \right] dF\left(\theta\right) - I\left(\phi\right) = \int_{\underline{\theta}}^{\bar{\theta}} W\left(\theta, \phi\right) dF\left(\theta\right) - I\left(\phi\right),$$

where

$$W\left(\theta, \phi\right) \equiv B\left(s\left(\theta\right), \phi\right) - s\left(\theta\right)\left(\theta + h\left(\theta\right)\right)$$

is the net social benefit with type θ and the extra negative term $s\left(\theta\right) h\left(\theta\right)$ is the information rent due to the IC constraint.

The maximization of Π with respect to $s\left(\theta\right)$ requires that the term under the integral $W\left(\theta, \phi\right)$ be maximized with respect to $s\left(\theta\right)$ for all θ. That is, the optimal data policy $s\left(\theta\right)$ must maximize the net social benefit for each type of consumers. It is straightforward to check that

$$\frac{\partial^2 W\left(\theta\right)}{\partial s^2\left(\theta\right)} = B_{ss}\left(s\left(\theta\right), \phi\right) < 0.$$

Then, the optimal data policy is the (unique) solution of the following FOC:

$$B_s\left(s\left(\theta\right), \phi\right) = \theta + h\left(\theta\right).$$

The above FOC determines the equilibrium path $s^{**}\left(\theta, \phi\right) = B_s^{-1}\left(\theta + h\left(\theta\right)\right)$. It is easy to check that $s^{**}\left(\theta, \phi\right)$ decreases in θ (since $B_{ss} < 0$) and increases in ϕ (as $B_{s\phi} < 0$).

Under full participation, the optimal data analytics is determined when the marginal cost equal to the expected marginal social benefit. That is,

$$I'\left(\phi\right) = \int_{\underline{\theta}}^{\bar{\theta}} B_{\phi}\left(s\left(\theta\right), \phi\right) dF\left(\theta\right).$$

Substituting $s^{**}\left(\theta, \phi\right)$ into the above equation, the equilibrium data analytics is the solution of

$$\int_{\underline{\theta}}^{\bar{\theta}} B_{\phi}\left(s^{**}\left(\theta, \phi\right), \phi\right) dF\left(\theta\right) = I'\left(\phi\right).$$

Finally, the optimal payment is given by

$$t^{**}\left(\theta\right) = V\left(\theta\right) - b\left(s^{**}\left(\theta\right), \phi^{**}\right) + s^{**}\left(\theta\right)\theta = \int_{\theta}^{\bar{\theta}} s\left(x\right) dx + s^{**}\left(\theta\right)\theta - b\left(s^{**}\left(\theta\right), \phi^{**}\right).$$

E: Proof of Proposition 4

We first characterize the equilibrium with third-party cookies after GDPR and then compare it with the counterfactural benchmark.

Equilibrium with third-party cookies

The digital platform's profit from displaying the third-party's advertisements is

$$\Pi_t^u = F(\tau)\left(\hat{B}(s_t,\phi_t) - (1-\beta)\,\hat{r}(s_t,\phi_t) - \gamma s_t \tau_t\right).$$

The digital platform chooses s_t and τ_t to maximize the above profit. Simultaneously, the third-party chooses ϕ_t to maximize $\pi_t = (1-\beta)\,\hat{r}(s_t,\phi_t) - \hat{I}\,(\phi_t)$.

The digital platform's optimal data scale s_t equates the marginal revenue to its marginal cost, which is given by

$$\hat{B}_s(s_t,\phi_t) - (1-\beta)\,\hat{r}_s(s_t,\phi_t) = \gamma\tau_t. \tag{13}$$

In addition, the digital platform chooses the optimal threshold τ_t such that

$$\frac{\hat{B}(s_t,\phi_t) - (1-\beta)\,\hat{r}(s_t,\phi_t)}{\gamma s_t} = \tau_t + h\,(\tau_t)\,. \tag{14}$$

The third-party's optimal data analytics ϕ_t^u is determined by

$$(1-\beta)\,\hat{r}_\phi(s_t,\phi_t) = \hat{I}'\,(\phi_t)\,. \tag{15}$$

The equilibrium data scale s_t^u, data analytics ϕ_t^u, and threshold τ^u are determined by the FOCs (13), (15), and (14), which can be characterized using the same approach as in the baseline model. The equilibrium profits for the digital platform and the third-party are given respectively by $\Pi_t^u = \gamma s_t^a F(\tau^u) h\,(\tau^u)$ and $\pi_t = (1-\beta)\,\hat{r}(s_t^u,\phi_t^u) - \hat{I}\,(\phi_t^u)$.

Equilibrium without third-party cookies

Consider now the counterfactural scenario in which the digital platform replaces the third-party's advertisements. Consumer data collection through its first-party cookies reduces consumer privacy cost to $s\theta$, but the digital platform needs to incur the cost of data analytics $I\,(\phi)$. Its profit from such substitution is

$$\Pi^u = F(\tau)\left(\hat{B}(s,\phi) - s\tau\right) - I\,(\phi)\,,$$

where $\tau = \left(\hat{b}(s,\phi) + t\right)/s$. The equilibrium data scale s^u, analytics ϕ^u, and threshold τ^u can be derived using exactly the same approach as in the baseline model, by replacing B with $\hat{B}$.

We use a variant of the leading example to compare equilibrium outcomes. Assume $\hat{r}\left(s,\phi\right)=\alpha s_t^{\sigma}\phi_t^{1-\sigma}$, $\hat{b}(s_t,\phi_t)=\left(1-\alpha\right)s_t^{\sigma}\phi_t^{1-\sigma}$, $I\left(\phi\right)=\phi^2/2$, and $\hat{I}\left(\phi_t\right)=j\phi_t^2/2$ with $j<1$.

Equilibrium with third-party cookies

Suppose the digital platform keeps the third-party cookies. The equilibrium threshold τ^u is given by

$$\frac{\tau^u}{\tau^u+h\left(\tau^u\right)}=\varepsilon_{s_t}=\sigma.$$

The optimal third-party data analytics is determined by

$$\alpha\left(1-\beta\right)\left(1-\sigma\right)s_t^{\sigma}\phi_t^{-\sigma}=j\phi_t$$

and the optimal third-party data scale is given by

$$\sigma\left(1-\left(1-\beta\right)\alpha\right)s_t^{\sigma-1}\phi_t^{1-\sigma}=\gamma\tau_t.$$

Combining these two FOCs, we get

$$s_t=\phi_t^2\frac{j}{\gamma\tau_t}\frac{\left(1-\alpha\left(1-\beta\right)\right)\sigma}{\alpha\left(1-\beta\right)\left(1-\sigma\right)}.$$

Substituting it into the FOC for s_t and then solving for ϕ_t and s_t, we obtain

$$\phi_t^u=\frac{\left(\alpha\left(1-\beta\right)\left(1-\sigma\right)\right)}{j}\left(\frac{\left(1-\alpha\left(1-\beta\right)\right)}{\gamma}\right)^{\sigma/(1-\sigma)}\left(\frac{\sigma}{\tau^u}\right)^{\sigma/(1-\sigma)},$$

and

$$s_t^u=\frac{\alpha\left(1-\beta\right)\left(1-\sigma\right)}{j}\left(\frac{\left(1-\alpha\left(1-\beta\right)\right)}{\gamma}\right)^{(1+\sigma)/(1-\sigma)}\left(\frac{\sigma}{\tau^u}\right)^{(1+\sigma)/(1-\sigma)}.$$

The digital platform's equilibrium profit is

$$\begin{aligned}\Pi_t^u &= \gamma s_t^a F(\tau^u)h\left(\tau^u\right)\\ &= \frac{\gamma^{-2\sigma/(1-\sigma)}}{j}\alpha\left(1-\beta\right)\left(1-\sigma\right)^2\left(\left(1-\alpha\left(1-\beta\right)\right)\right)^{(1+\sigma)/(1-\sigma)}\left(\frac{\sigma}{\tau^u}\right)^{2\sigma/(1-\sigma)}F(\tau^u),\end{aligned}$$

where we have used $h(\tau^u)=\left(1-\sigma\right)\tau^u/\sigma$ to derive the second line.

Equilibrium without third-party cookies

Suppose the digital platform replaces the third-party's advertisements with its own business. The equilibrium outcome is exactly the same as in the leading example for the baseline model, by replacing ρ with σ, which are given by

$$\phi^u=\left(1-\sigma\right)F\left(\tau^u\right)\left(\frac{\sigma}{\tau^u}\right)^{\sigma/(1-\sigma)},$$

and

$$s^u=\left(1-\sigma\right)F\left(\tau^u\right)\left(\frac{\sigma}{\tau^u}\right)^{(1+\sigma)/(1-\sigma)},$$

where τ^u is given by

$$\frac{\tau^u}{\tau^u + h(\tau^u)} = \sigma.$$

The digital platform's profit is

$$\begin{aligned}
\Pi^u &= s^u F(\tau^u) h(\tau^u) - I(\phi^u) \\
&= (1-\sigma)\left(\frac{\sigma}{\tau^u}\right)^{(1+\sigma)/(1-\sigma)} F^2(\tau^u) h(\tau^u) - \frac{1}{2}(\phi^u)^2 \\
&= \frac{1}{2}(1-\sigma)^2 \left(\frac{\sigma}{\tau^u}\right)^{2\sigma/(1-\sigma)} F^2(\tau^u).
\end{aligned}$$

Comparing the digital platform's profits in two scenarios and noting that $\tau^u = \tau^u$, we have $\Pi^u > \Pi^u_t$ if and only if

$$\gamma > \bar{\gamma} \equiv \left(\frac{2\alpha(1-\beta)}{jF(\tau^u)}\right)^{(1-\sigma)/(2\sigma)} ((1-\alpha(1-\beta)))^{(1+\sigma)/(2\sigma)}.$$

Recall that the extra surplus for opt-in consumers is $V_t(\theta) = \gamma s^a_t (\tau^u - \theta)$ with three-party cookies and becomes $V(\theta) = s^u(\tau^u - \theta)$ without them. Thus, the replacement benefits consumers if $s^u > \gamma s^a_t$, which amounts to

$$\gamma > \hat{\gamma} \equiv \left(\frac{\alpha(1-\beta)((1-\alpha(1-\beta)))^{(1+\sigma)/(1-\sigma)}}{jF(\tau^u)}\right)^{(1-\sigma)/(2\sigma)}.$$

Since $\hat{\gamma} < \bar{\gamma}$, such replacement improves consumer surplus when it happens $(\gamma > \bar{\gamma})$. Finally, comparing the levels of data analytics, we have $\phi^u > \phi^u_t$ if and only if

$$\gamma > \tilde{\gamma} \equiv \left(\frac{\alpha(1-\beta)}{jF(\tau^u)}\right)^{(1-\sigma)/\sigma} (1-\alpha(1-\beta)).$$

F: Data Acquisition with Personalization

Here, we characterize the digital platform's optimal data policy with personalization. Recall that an opt-in consumer receives an extra benefit $m(s,\phi)x$ from sharing data and the digital platform charges the personalized prices $p(x; s, \phi)$. The analysis before GDPR is given in the main context. We provide the analysis after the GDPR here.

Uniform Policy

Suppose the digital platform is committed to the uniform data policy$\{s, \phi\}$. The digital platform's profit is

$$\Pi^u = \int_0^1 p(x) F(\tau(x))\, dx - I(\phi),$$

where the cut-off threshold is

$$\tau(x) = \frac{m(s,\phi)x - p(x)}{s}.$$

Given $\{s,\phi\}$, we first solve for the digital platform's personalized pricing. Substituting $p(x) = m(s,\phi)x - s\tau(x)$ into the above profit function, we obtain

$$\Pi^u = \int_0^1 (m(s,\phi)x - s\tau(x)) F(\tau(x)) dx - I(\phi).$$

Thus, given $\{s,\phi\}$, choosing $p(x)$ is equivalent to choosing $\tau(x)$ in the maximization. The maximization of Π with respect to $\tau(x)$ requires that the term under the integrand $(m(s,\phi)x - s\tau(x)) F(\tau(x))$ be maximized with respect to $\tau(x)$ for all x. Differentiating the integrand with respect to $\tau(x)$ and solving for the FOC leads to

$$m(s,\phi)x = s(\tau(x) + h(\tau(x))).$$

Using $l(\tau) \equiv \tau + h(\tau)$ ($l(\tau)$ increases in τ), we have

$$\tau(x) = l^{-1}\left(\frac{m(s,\phi)x}{s}\right).$$

We show that $\tau(x)$ increases in x. Differentiating both sides of the FOC with respect to x, we have

$$m(s,\phi) = s\left(1 + h'(\tau)\right)\tau'(x),$$

which implies

$$\tau'(x) = \frac{m(s,\phi)}{s(1 + h'(\tau))} > 0.$$

Next, we solve for the optimal policy $\{s,\phi\}$. Differentiating Π^u with respect to ϕ, we obtain

$$\frac{\partial \Pi^u}{\partial \phi} = \int_0^1 m_\phi(s,\phi) x F(\tau(x)) dx - I'(\phi).$$

The optimal data analytics ϕ^u is given by

$$m_\phi(s,\phi) \int_0^1 x F(\tau(x)) dx = I'(\phi). \tag{16}$$

In addition, differentiating Π^u with respect to s, we have

$$\frac{\partial \Pi^u}{\partial s} = \int_0^1 (m_s(s,\phi)x - \tau(x)) F(\tau(x)) dx.$$

The second-order derivative is negative

$$\frac{\partial^2 \Pi^u}{\partial s^2} = \int_0^1 m_{ss}(s,\phi) x F(\tau(x)) dx < 0.$$

Hence, the optimal s in an interior solution and satisfies

$$m_s(s,\phi)\int_0^1 xF(\tau(x))\,dx=\int_0^1\tau(x)\,F(\tau(x))\,dx. \tag{17}$$

A consumer's net surplus is then given by

$$V(x,\theta)=m(s,\phi)\,x-p(x)-s\theta=s(\tau(x)-\theta)\,.$$

Hence, opt-in consumers are better off after GDPR.

Mechanism Design

Suppose the digital platform offers type-dependent data policy. Without loss of generality, we focus on the direct mechanism in which the digital platform requires a consumer to report her true type θ and then recommends the policy $\{s(\theta)\,,t(\theta,x)\}$ accordingly, where $t(\theta,x)$ is the type-dependent payment to the digital platform. A consumer with type θ and taste x receives a net surplus from opt-in

$$V(\theta;x)=m(s(\theta)\,,\phi)\,x-t(\theta,x)-s(\theta)\,\theta.$$

The digital platform knows the exact value of x but does not observe θ. The policy must be incentive compatible for type θ.

A type θ consumer who mis-reports her type as $\tilde{\theta}$ will receive

$$V\left(\tilde{\theta},\theta;x\right)=m\left(s\left(\tilde{\theta}\right),\phi\right)x-t\left(\tilde{\theta},x\right)-s\left(\tilde{\theta}\right)\theta,$$

and the policy is incentive compatible if $V(\theta;x)\geq V\left(\tilde{\theta},\theta;x\right)$ for any $\tilde{\theta}\neq\theta$. Differentiating $V\left(\tilde{\theta},\theta;x\right)$ with respect to $\tilde{\theta}$, we have

$$\frac{\partial V\left(\tilde{\theta},\theta;x\right)}{\partial\tilde{\theta}}=\left(m_s\left(s\left(\tilde{\theta}\right),\phi\right)x-\theta\right)s'\left(\tilde{\theta}\right)-\frac{\partial t\left(\tilde{\theta},x\right)}{\partial\tilde{\theta}}.$$

The IC constraint requires $V\left(\tilde{\theta},\theta;x\right)$ be maximized at $\tilde{\theta}=\theta$, which implies

$$\left(m_\phi(s(\theta)\,,\phi)\,x-\theta\right)s'(\theta)=\frac{\partial t(\theta,x)}{\partial\theta}.$$

By the envelope theorem, we obtain

$$\frac{dV(\theta;x)}{d\theta}=-s(\theta)\,.$$

At the optimum, the participation constraint of the highest type is binding such that $V\left(\bar{\theta};x\right)=0$. Then

$$V(\theta;x)=V\left(\bar{\theta};x\right)+\int_\theta^{\bar{\theta}}s(y)\,dy=\int_\theta^{\bar{\theta}}s(y)\,dy.$$

Note that $V(\theta; x)$ does not depend on x, since the digital platform knows the exact value of x and can extract full consumer surplus related to taste x.

The digital platform signs up all consumers under the policy $\{s(\theta), t(\theta, x)\}$. Its total profits are given by

$$\Pi = \int_0^1 \int_{\underline{\theta}}^{\bar{\theta}} t(\theta, x)\, dF(\theta)\, dx - I(\phi).$$

Using

$$t(\theta, x) = m(s(\theta), \phi)\, x - s(\theta)\,\theta - V(\theta; x),$$

we can rewrite the digital platform's profits as

$$\begin{aligned} \Pi &= \int_0^1 \int_{\underline{\theta}}^{\bar{\theta}} (m(s(\theta), \phi)\, x - s(\theta)\,\theta - V(\theta; x))\, dF(\theta)\, dx - I(\phi) \\ &= \int_0^1 \left[\int_{\underline{\theta}}^{\bar{\theta}} [m(s(\theta), \phi)\, x - s(\theta)\,\theta]\, dF(\theta) - \int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\theta}^{\bar{\theta}} s(y)\, dy \right) dF(\theta) \right] dx - I(\phi). \end{aligned}$$

Using integration by parts, we have

$$\int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\theta}^{\bar{\theta}} s(y)\, dy \right) dF(\theta) = \left[\left(\int_{\theta}^{\bar{\theta}} s(y)\, dy \right) F(\theta) \right]_{\underline{\theta}}^{\bar{\theta}} + \int_{\underline{\theta}}^{\bar{\theta}} F(\theta)\, s(\theta)\, d\theta = \int_{\underline{\theta}}^{\bar{\theta}} F(\theta)\, s(\theta)\, d\theta.$$

Substituting into Π, we obtain

$$\begin{aligned} \Pi &= \int_0^1 \int_{\underline{\theta}}^{\bar{\theta}} [m(s(\theta), \phi)\, x - s(\theta)(\theta + h(\theta))]\, dF(\theta)\, dx - I(\phi) \\ &= \int_{\underline{\theta}}^{\bar{\theta}} \left[\frac{m(s(\theta), \phi)}{2} - s(\theta)(\theta + h(\theta)) \right] dF(\theta) - I(\phi) \\ &= \int_{\underline{\theta}}^{\bar{\theta}} \tilde{W}(\theta, \phi)\, dF(\theta) - I(\phi), \end{aligned}$$

where

$$\tilde{W}(\theta, \phi) \equiv \frac{m(s(\theta), \phi)}{2} - s(\theta)(\theta + h(\theta)),$$

is the net social benefit with type θ, in which the extra negative term $s(\theta)\, h(\theta)$ is the information rent due to the IC constraint.

The maximization of Π with respect to $s(\theta)$ requires that the term under the integral $\tilde{W}(\theta, \phi)$ be maximized with respect to $s(\theta)$ for all θ. That is, the optimal data policy $s(\theta)$ must maximize the net social benefit for each type of consumers. It is straightforward to check that

$$\frac{\partial^2 W(\theta, \phi)}{\partial s^2(\theta)} = \frac{m_{ss}(s(\theta), \phi)}{2} < 0.$$

Then, the optimal data policy is the (unique) solution of the following FOC:

$$\frac{m_s\left(s\left(\theta\right),\phi\right)}{2}=\theta+h\left(\theta\right).$$

The optimal data policy is determined when the marginal social benefit $\frac{m_s(s(\theta),\phi)}{2}$ is equal to the adjusted marginal cost of privacy $\theta+h\left(\theta\right)$.

Under full participation, the optimal data analytics is determined when the marginal cost equates to the expected marginal social benefit. That is,

$$I'\left(\phi\right)=\int_{\underline{\theta}}^{\bar{\theta}}\frac{m_\phi\left(s\left(\theta\right),\phi\right)}{2}dF\left(\theta\right).$$

Finally, the optimal payment is given by

$$t^{**}\left(\theta,x\right)=m\left(s^{**}\left(\theta\right),\phi^{**}\right)x-s^{**}\left(\theta\right)\theta-V\left(\theta,x\right)=m\left(s^{**}\left(\theta\right),\phi^{**}\right)x-\int_{\theta}^{\bar{\theta}}s\left(y\right)dy-s^{**}\left(\theta\right)\theta.$$

Since the digital platform can use personalized pricing, each consumer's extra surplus from data sharing is independent of x:

$$V\left(\theta;x\right)=V\left(\theta\right)=\int_{\theta}^{\bar{\theta}}s\left(y\right)dy.$$

G: Cookies[33]

Cookies are small text files that are downloaded into a user's device by the web browser when visiting a particular website. Cookies are tiny packets of modulated information transmitted between a server and a browser,[34] and their primary role is to establish a connection and simultaneously retrieve useful information about the user's activity during subsequent page visits. Removing cookie will make it difficult for merchants to keep tracking a user's website browsing record.

A. Purpose of cookies

Over time internet cookies have been repurposed to serve three main objectives:[35]

Session management. This is a process of holding and relaying information of the user across various pages on the merchant's website. For example, a merchant uses a cookie with a unique identifier to map a user with their shopping cart. The user's browser sends a session identifier

[33] I thank my research assistant Ratul Das Chaudhury for completing this supplementary reading material. I made some editorial changes on Ratul's original note.

[34] See Park and Sandhu (2000).

[35] https://developer.mozilla.org/en-US/docs/Web/HTTP/Cookies

to the merchant's server every time the user visits a page or clicks on a link on the merchant's website. Session cookies also reduce the page loading time of the user.

Tracking. Cookies are also used to monitor the browsing preferences of the user. Some cookies purposefully track user activities on social media, news, or shopping websites to generate and analyze user behaviour. For example, web-based advertising platforms like Google AdSense, Amazon Native Shopping Ads, and Adversal use tracking cookies to analyze the browsing behaviour of target users, which can be used for marketing.

Personalization. Some cookies help merchants recall relevant user details and browsing habits. These cookies can benefit the user by enabling certain features on the website to improve the user experience. Personalization enhances the user's web experience by providing tailored product layouts based on previous choices, pre-set preferences, browsing history, and display style.

B. Types of cookies

The General Data Protection Regulation (GDPR) categorizes cookies into three broad groups based on three attributes – duration, provenance, and purpose.[36]

Duration:

- *Session cookies* are temporary and automatically deleted after the termination of the browsing session. For example, if a user logs into their e-commerce account, session cookies help a user stay logged in their account across multiple pages;
- *Persistent cookies* are stored in the user's device and are not deleted once the session terminates. Persistent cookies are utilized for user authentication, tracking, or other related purposes.

Provenance:

- *First-party cookies* are placed directly in the user's device by the website the user is browsing;
- *Third-party cookies* are generated by websites that differ from the site the user is browsing. These cookies are placed by external domains, such as advertisers or web analytics providers. Unlike first-party cookies, third-party cookies can track a user across different websites.

[36] https://gdpr.eu/cookies/

Purpose:

- *Strictly necessary cookies* are crucial for the proper operation of the website, for example, access a secure page or holding items in a shopping cart. Generally, these cookies do not require a user's consent before proceeding;

- *Preference cookies* ensure that the webpage is well managed and suited to a user's choices during the previous page visits, e.g., auto log-in, display preferences, languages, etc. These cookies improve functionality and enhance the user experience;

- *Statistics cookies* help identify the important performance indicators of users and user activity in the domain website. These cookies record clicks on links on a particular page, duration of stay, and other web usage statistics to optimize the user experience;

- *Marketing cookies* are applied to provide users with relevant marketing and advertisement campaigns. These cookies can track users across websites where advertisements are placed and used by marketing agencies to build a user profile and present targeted advertisements.

C. Cookie banners and options

GDPR has prompted websites to display cookie banners and disclaimers. We provide some examples of cookie disclaimers:

- *No banner:* Websites like better.com (a company primarily based in New York) provide users with minimal information about cookie policy on their webpages and only offer a link to their privacy policy.

- *Inform-only:* Some websites such as termsfeed.co and Los Angeles Times (https://www.latimes.com/) only inform users about cookie usage and privacy policies. These websites inform users about using different types of cookies to enhance user experience, improve web functionality, and site analytics. These are examples of implied consent.

- *Confirmation-only:* Several websites, like The Mirror (https://www.mirror.co.uk/) and Essentra (https://www.essentra.com/en), offer users a banner with an 'Accept' / 'Agree' button and make it difficult for users to further access the page without agreeing to the policy. This is classified as forced opt-in consent.

- *Dual option:* Some websites such as GDPR (https://gdpr.eu/cookies/) offer users a binary choice: choose all cookies or only strictly necessary cookies.

- *Two-step granular option*: Some websites, such as Premiere League (https://www.premierleague.com/) and Prolific (https://prolific.co/) also offer a two-step option for their users. They can either accept all cookies or select from a menu of options. The website provides a multitude of granular options (functional, performance, analyticalm and marketing) from which the user can accept cookies of their own choice.

D. GDPR's impact on Cookies

GDPR's rollout has a considerable impact on cookies. Some of the observed changes are as follows:

- Cookie banners: Degeling et. al. (2018) examined 500 popular websites in EU nations post-GDPR. They observed a significant increase of 16% in the display of cookie consent banners, from 46.1% in January 2018 to 62.1% in May 2018.

- Third-party cookies: Libert et. al. (2018) examined the popular news websites in seven EU countries and found that 98% of those sites contain at least one third-party cookie, with an average of 81 cookies per page. The average count of third-party cookies per page went down by 22% in these news websites after GDPR's rollout, with 14% decrease in advertising cookies and 9% reduction in social media cookies.

- Fines for non-compliance: More than 24% of the popular websites in the Baltic states do not display privacy policies after GDPR. The EU's data protection authorities have issued over 800 fines by May 2018 for GDPR non-compliance.[37] Luxembourg National Commission for Data Protection fined Amazon a €746 million for violating data processing procedures, forcing users to comply with cookie policies, and making the opt-out process too challenging. WhatsApp and Google were also found in violation of the GDPR guidelines and fined for the lack of transparency in cookie policies and data processing guidelines.[38]

References

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2018). "We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy". https://link.springer.com/article/10.1007/s00287-019-01201-1

[37] https://www.tessian.com/blog/biggest-gdpr-fines-2020/

[38] https://www.enforcementtracker.com/

Libert, T., L. Graves, and R. Nielsen (2018). Changes in Third-Party Content on European News Websites after GDPR. https://ora.ox.ac.uk/objects/uuid:5a5d4eea-6e74-49b4-8c77-71ec6760f127

Park, J. S., & Sandhu, R. (2000). "Secure cookies on the Web". *IEEE internet computing*, Vol 4(4): 36-44.

H: Data Analytics[39]

Computer scientists define data as bits of information that can be structured, processed, or analyzed to gather insights or make meaningful predictions about relevant topics. However, raw data does not have much value per se; it needs to be processed and analyzed to create value. Data analytics is the process of logically and systematically analyzing raw data to assess, anticipate or forecast various scenarios.[40] The term "data analytics" is an umbrella term, as it incorporates the acts of capturing, processing, storing, analyzing, and using the data.

Data analytics is a complex process of inspecting, cleaning, transforming, and analyzing unstructured data to yield effective business perceptions. Data analytics is slowly evolving to incorporate the processing of big data with analytics.[41] Traditionally, data and analysis were treated as two separate processes that were integrated and analyzed using statistical tools before they would be worthy of offering up valuable insights. The conventional method of collecting data and analyzing was expensive, time-consuming, and computationally challenging. In the past few decades, advancements in storage and sensor technology have contributed to transforming big data into analysis-worthy information[42]. Modern data analytics rely on algorithms and machine learning processes to increase efficiency and optimize decision-making processes. Advancements in computational sciences have contributed significantly to reducing data processing times. Data analytics is integrated in transforming raw and unstructured data into usable knowledge that has the potential to improve competitiveness among businesses, increase productivity, analyze risk, detect fraud to name a few.

Steps of Data Analytics

The primary purpose of any analytics tool is to process and convert unstructured inputs into

[39] I am grateful to my research assistant Ratul Das Chaudhury for completing this supplementary reading material. I made some editorial changes on Ratul's original note.

[40] https://www.investopedia.com/terms/d/data-analytics.asp

[41] See Tsai et. al (2015).

[42] See Bloem et. al (2013) and Kambatla et. al (2014) for detailed discussion.

useful output that can be used to make precise predictions and draw valuable inferences. The steps involved in the development of a data analytics pipeline can be classified as follows:[43]

- *Data collection.* The preliminary step is a proper data collection system and a suitable infrastructure to ingest the data from the source;
- *Data processing.* Providing some structure to a crude dataset is essential in creating a proper data analytics pipeline. Data is usually collected from a multitude of sources. It must undergo extraction, cleaning, and transformation into the requisite layout to make the data worthy of analysis;
- *Data storage.* This steps involves refining and storing the data using a proper data storage service – (a "data warehouse" or "data lake") – to be used for further processing;
- *Data analysis.* The stored data is either achieved or used for analytical purposes to generate useful insights. The type of analyses applied to the database is determined by research objectives and client requirements.
- *Data visualization.* Use the findings from the analysis to create visual or graphical representations and employ the available visualization toolkit to identify the trends and patterns in the data.

Types of Data Analytics

According to Bloem et. al (2013), data Analytics is broadly classified into four subcategories:

- a) *Descriptive analytics* is the study of information that illustrates what has happened over time. This type of analytical method involves analyzing a dataset to gather valuable insights about the past. It is primarily used to investigate whether and when something has been misconstrued and/or misinterpreted in the past. Descriptive analytics as a standalone method is incomplete in that it helps identify a problem without offering any justification;
- b) *Diagnostic analytics* investigates why something has happened. This method involves understanding whether there exists a causal relationship between two events. While descriptive analytics provide insights about specific trends in the past, diagnostic analytics investigates the factors that have contributed to those trends;

[43] See https://www.freecodecamp.org/news/scalable-data-analytics-pipeline/

- c) *Predictive analytics* provides the likelihood of the occurrence of an event. This method utilizes past data to make meaningful future predictions. Predictive analytics is useful to forecast future trends by utilizing descriptive and diagnostic analytics and other available modelling techniques. The precision of the forecasts is largely reliant on the quality of the data and other exogenous factors;

- d) *Prescriptive analytics* is the method of prescribing a plan of action to eliminate a problem or benefit from a trend.

Applications of Data Analytics

Enterprises utilize decision-driven analytics to better understand whether and how analytical thinking benefits the organization[44]. A recent survey highlights that 94 percent of the surveyed business analytics professionals believe that data and analytics are crucial for the growth of their businesses.[45] A significant portion also believe that data analytics increases the productivity of their firm, improves cost efficiency, engenders faster decision-making, and financially benefits the company. The survey also reveals that 65 percent of companies are planning to increase their investment in data analytics in the upcoming years. A McKinsey report highlights that companies that had infused creativity and purpose in conjunction with data analytics had experienced 2.7 times average annual revenue growth in 2020 compared to their peers.[46] Another IBM report indicates that the use of big data analytics gives banking and financial institutions a considerable competitive advantage over their peers.[47] More recent assessments by Allied Market Research reveal that the market for big data and analytics was valued at approximately $198 billion in 2020 and is expected to rise to $684 billion by 2030.[48]

Firms are utilizing consumer data to streamline operations, access newer markets, and to better serve existing customers. Companies such as Google Analytics, Amazon Redshift, Salesforce's Einstein Analytics, and Adobe Analytics provide inexpensive analytics solutions to enterprises. The Economist Intelligence Unit assessment found that a significant portion of the

[44] See Provost and Fawcett (2013) and Brynjolfsson and McElheran (2016).

[45] https://www.microstrategy.com/getmedia/db67a6c7-0bc5-41fa-82a9-bb14ec6868d6/2020-Global-State-of-Enterprise-Analytics.pdf

[46] https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-growth-triple-play-creativity-analytics-and-purpose

[47] See Turner et. al (2013).

[48] https://www.alliedmarketresearch.com/big-data-and-business-analytics-market

surveyed companies underutilize their data.[49] An MIT Solan survey finds that 87 percent of the respondents believe that their organization 'needs to step up' the use of analytics.[50] Cost-effective data analytics solutions provide opportunities to small and medium-sized businesses to harness their sales and consumer data to better understand buying habits, distinguish trends and manage finances. One study finds that 67% of the surveyed small businesses spend more than $10000 a year on data analytics solutions.[51]

Adobe Analytics is an analytics solutions platform that provides real-time analytics, customer behaviour, sales data to their clients. Adobe analytics provide their clients with the ability to collect and analyze large datasets to forecast consumer behaviour, predict sales etc.[52] For example, the packages of Adobe Analytics provides their clients with a multitude of analytics tools like – reporting, dashboarding, data repository services, ad-hoc analysis, tag management, customer analytics, predictive modeling, etc. An HG Insights study reveals that about 47 percent of the 53,532 companies that rely on Adobe Analytics earn annual revenues of around $1 million to $10 million.[53]

Although it is difficult to determine the actual return on investment from data analytics due to the complexity and the indirect effects it generates, both researchers and businesses agree that data analytics fosters innovation, provides efficient problem-solving techniques, better manages risk, and reduces cost.[54] Currently, some enterprises target a return of three and a half times the initial spending on data analytics projects.[55] Current market research has identified the use of analytics has improved efficiency in fraud detection among the top 50 riskiest providers in healthcare.[56] A report of eight selected clients of Adobe analytics reveals that after the adaption of Adobe Analytics there was a 3 percent increase in site traffic, and the average ROI from using Adobe Analytics services is 224 percent.[57]

We provide several business cases of using data analytics:

[49] Unit, E. I. (2011). Big data: Harnessing a game-changing asset. The Economist, 1-32.

[50] See Kiron et. al (2014).

[51] https://smallbiztrends.com/2020/03/data-analytics-trends.html

[52] https://www.adobe.com/content/dam/acom/au/marketing-cloud/playbook/Adobe-Analytics.pdf

[53] https://discovery.hgdata.com/product/adobe-analytics

[54] https://www.pwc.com/us/en/services/consulting/analytics.html

[55] See Shim et. al (2015).

[56] https://www.elderresearch.com/blog/can-data-analytics-really-deliver-1300-roi/

[57] https://business.adobe.com/content/dam/dx/us/en/resources/reports/forrester-tei-adobe/total-economic-impact-analytics-audience-manager.pdf

- *Mango* - a high-end clothing retailer, noticed an influx in web traffic from mobile devices but was struggling to translate the increased traffic to sales. Mango utilized the analytical tools in Google Analytics 360 to analyze consumer behaviour by device type. This led to a 49 percent increase in shoppers adding a product to their cart and a 3.9 percent rise in mobile revenue.[58]

- *Sigma Sport* – a sports gear retailer, while reviewing Google Analytics data noticed a lack of consumer engagement in their website. Combining the results from data analytics with insights from a consumer journey experiment led the company to provide customers with a personalized homepage that resulted in a 28 percent increase in revenue.[59]

- *Netflix* – an online streaming platform, leveraged viewer information and predictive analytics modelling and filtering to improve TV show/film recommendations, customer experience etc. This helped the streaming platform attain a show success rate of 80 percent.[60]

- *McDonald* acquired a data analytics firm for $300 million.[61] The global fast-food chain is investing in data analytics to optimize its food delivery service, reducing operations costs and improving customer experience.[62]

- *Shazam* – a sound recognition application, sought to *Einstein Analytics* for data analytic solutions. Implementing the self-service analytic tools has led to an increase in employee productivity, improved data quality, customer segmentation, etc., which resulted in a 752 percent ROI.[63]

References:

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). "How to improve firm performance using big data analytics capability and business strategy alignment?".

[58] https://services.google.com/fh/files/misc/case-study-mango-dresses-up-its-mobile-results-with-google-optimize-360.pdf

[59] https://services.google.com/fh/files/misc/case-study-sigma-sport-spins-up-28percent-higher-revenue-with-google-optimize-360.pdf

[60] See Srivatsa et. al (2019).

[61] https://www.bringg.com/blog/insights/mcdonalds-300m-acquisition-big-data-retail/

[62] https://digital.hbs.edu/platform-digit/submission/big-mac-to-big-data-why-mcdonalds-is-betting-its-future-on-digital-innovation/

[63] https://www.salesforce.com/news/press-releases/2017/08/09/shazam-increases-sales-productivity-with-salesforce-einstein-analytics-2/

International Journal of Production Economics, Vol 182: 113-131.

Bloem, J., Doorn, M. van, Duivestein, S., Manen T. van, Ommeren, E. van, & Sackdeva, S. (2013). "No more secrets with big data analytics". https://www.bain.com/insights/the-value-of-big-data#

Brynjolfsson, E., & McElheran, K. (2016). "The rapid adoption of data-driven decision-making". *American Economic Review*, Vol 106(5): 133-39.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). "Trends in big data analytics". *Journal of Parallel and Distributed Computing*, Vol 74(7): 2561-2573.

Kiron, D., Prentice, P. K., & Ferguson, R. B. (2014). "Raising the bar with analytics". *MIT Sloan Management Review*, Vol 55(2): 29-33.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc..

Shim, J. P.; French, Aaron M.; Guo, Chengqi; and Jablonski, Joey (2015) "Big Data and Analytics: Issues, Solutions, and ROI," *Communications of the Association for Information Systems*, Vol 37, Article 39.

Srivatsa Maddodi, & Krishna Prasad, K.(2019). "Netflix Bigdata Analytics-The Emergence of Data Driven Recommendation". *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, Vol 3(2):41-51.

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). "Big data analytics: a survey". *Journal of Big Data*, Vol 2(1): 1-32.

Turner, D., Schroeck, and M., & Shockley, R. (2013). "Analytics: The real-world use of big data in financial services." *IBM Institute for Business Value in collaborations with the Saïd Business School*, University of Oxford.

Zikopoulos, P., Eaton, C., Deroos, D., Deutsch, T., & Lapis, G. (2012). *Understanding Big Data: Analytics for Enterprise-Class Hadoop and Streaming Data.* McGraw-Hill. ISBN 978-0-07-179053-6. New York, USA.