Tracing Out International Data Flow: the Value of Data and Privacy

Junjun Quan

September 4, 2023

[Please Click Here for the Latest Version]

Abstract

The paper studies the value of data and privacy by analyzing the impact of the EU's General Data Protection Regulation (GDPR) on US multinational firms and their customers. The GDPR limits firms' access to EU consumers' data, prompting US companies to reallocate their businesses and scale back their EU operations by 10%. Smaller firms have a harder time adapting and experience a more significant and persistent impact. In response to the regulation, large firms hire more AI-related talents. Although GDPR provides better privacy protection for EU consumers, it also leads to a 6% drop in user ratings of digital services, showing the trade-off between privacy protection and data-dependent user experiences. The paper provides a tractable estimation framework and derives moment conditions that can be matched with the empirical findings. This framework combines the value of data and privacy in an equilibrium model and speaks to the welfare impact of a regional privacy regulation like GDPR.

Keywords: Data Economy, Consumer Privacy Protection, Multinational Firms, Value of Data and Privacy

JEL classifications: G30, D12, D22, O34, D62, D18, F20

1 Introduction

Consumer data has become an important form of intangible capital in the digital era. The fast advancement in computing power and artificial intelligence has led to a massive leap in data processing capacity. The development of the data economy[1] is not limited to the information industry, and it is rapidly broadening into all sectors, including the retail and automotive industries.

Assessing the value of intangible capital has always been a challenging task, and it is even more difficult when it comes to valuing data. As companies offer goods and services to their customers, they collect a large amount of information and gain insights into their clients' preferences. The valuable data can be utilized to enhance product quality and deliver customized services that cater to consumer preferences. This feature of data creates a feedback loop and a multiplier effect—improved products and services lead to increased consumption, which in turn results in more data (Farboodi and Veldkamp 2021; Jones and Tonetti 2020). The volume of data that companies can gather is influenced by consumers' privacy preferences and regulatory measures. As such, the multiplier effect in the data feedback loop is closely tied to consumers' privacy preferences. Furthermore, consumers' privacy choices will be affected by the benefits of sharing data. Sharing data with digital service providers (e.g., Netflix and Facebook) can lead to more personalized recommendations, which enhances user experiences at the individual level. If all consumers choose to share more data, firms will have more data to train their algorithms, and this improves the welfare for all users on the platform. However, consumers may not internalize the positive externality of data sharing on others. I visualize my research questions in Figure 1.

[Insert Figure 1 Here.]

This paper proposes a framework where the value of data and privacy are considered jointly. The paper leads with two empirical sections, the demand for data by firms and the demand for privacy by consumers, where I show reduced-form evidence on the impact

[1] As defined in the European Commission's 2017 Communication on Building a European Data Economy, the "data economy" is characterised by an ecosystem of different types of market players — such as manufacturers, researchers and infrastructure providers — collaborating to ensure that data is accessible and usable. This enables the market players to extract value from this data, by creating a variety of applications with a great potential to improve daily life.

of the General Data Protection Regulation (GDPR) on firms and consumers. The paper then builds a two-economy general equilibrium model and provides a tractable estimation framework that combines the value of data and privacy. The paper also speaks to the welfare impact of a regional privacy regulation, i.e., GDPR.

I first focus on the demand for data by firms and study how GDPR, a regional privacy regulation enacted in the European Union (EU), affects US multinational firms. GDPR is a comprehensive privacy protection framework [2] aimed at giving EU residents more control over their personal data. It was approved by the European Parliament in April 2016 and came into effect in May 2018. It superseded its predecessor, the EU Data Protection Directive (DPD), with more specific data protection requirements, a global perspective, tougher enforcement, and high penalties in case of violation.[3] After GDPR's enactment, when firms (data controllers or processors) want to collect and process data from EU residents (data subjects), they will need to ask for explicit consent and inform the consumers how their data will be used. GDPR has a unique global perspective because non-EU companies that collect and process EU consumers' data must also comply. As a result, US multinational firms like Meta face the headwinds from GDPR through their EU business segments.[4]

[Insert Figure 2 Here.]

Since 2012, when the discussion around a new privacy protection framework in the EU started, there has been an increasing number of US public firms disclosing privacy-related risk factors in their 10-K filings[5], as shown by the black line in Figure 2. Since 2016, the disclosure of such risks has become more specific by mentioning privacy regulations like GDPR (passed in 2016 and enacted in 2018), shown by the red line, and CCPA (passed in 2018 and enacted

[2]GDPR also applies to Iceland, Norway, and Liechtenstein, which belong to the European Economic Area (EEA), not EU. As of 2021, the United Kingdom retains the law in identical form despite no longer being an EU member state.

[3]For severe violations, as listed in Art. 83(5) GDPR, a company can be fined up to 20 million euros or 4% of their total global turnover of the preceding fiscal year, whichever is greater. For less severe violations, as defined in Art. 83(4) GDPR, a company will still face fines of up to 10 million euros or 2% of its entire global turnover of the preceding fiscal year, whichever is greater. Please see https://gdpr-info.eu/issues/fines-penalties/ for details.

[4]In Meta's 10-K filing for the fiscal year 2021 released in February of 20202, it says, "we will likely be unable to offer a number of our most significant products and services, including Facebook and Instagram, in Europe," due to GDPR compliance issues.

[5]Starting 2006, US public firms are required by SEC to disclose any risk factors that may materially affect their core business operations in their annual 10-K filings under Item 1A.

in 2020), demonstrated by the blue line. Even though risk factor disclosures are supposed to be forward-looking, firms only started recognizing such risks when the regulations officially came into effect. Many of them only acknowledged the impact one year after the enactment date.

GDPR creates regulatory differences across countries, and it is unique because it concerns the digital "oil", i.e., consumer data. Under this new regulatory framework, EU consumers are better protected than their US counterparts and have more control over the data they share with firms. Furthermore, GDPR imposes high compliance costs on US firms with an EU presence. If EU consumers place a high value on their privacy, they may share less than the optimal amount of data that US firms desire. Consequently, if US firms value the data they collect from consumers, they will instead turn to the US market, a much more fertile ground to reap data. The intuition is illustrated in Figure 1.

The empirical findings confirm this hypothesis. I use the privacy regulation as a supply shock to data and study the demand for data by US multinational firms. I find a compositional shift in the fraction of revenue that US multinational firms derive from each part of the world, with data-intensive[6] firms shifting away from the EU market. Employing a difference-in-differences (DID) design, I observe an 10% drop in the fraction of revenue generated from the European market among US data-intensive firms. This is a rational response of US firms to mitigate the negative effects caused by GDPR. However, their actions may have important implications for the welfare of US consumers. US consumers do not have the proper protection by a comprehensive privacy framework at the federal level. They may be exploited by these "data-hungry" US firms, which want to compensate for their loss of data from the EU market. These observations call for a thorough look into the issues of privacy regulations from a general equilibrium (GE) perspective, where we account for both the efficiency and distributional effects on multinational firms and the impact on consumers' welfare.

Moreover, I show that the business shifting results are not driven by lower profitability in the European market or tech-driven confounding trends. In a further analysis of the market

[6]Data-intensive firms' business models rely heavily on the data they collect from consumers. I introduce a measure of data-intensiveness in Section 2.2.1. This measure captures the variation in the hiring of AI and data management talents and the market and scientific value of computing patents (CPC G06) across firms.

dynamism, I find that the effects are much bigger for small firms than for large firms. The effects for smaller firms also tend to deepen and persist. This resonates with the anecdotal evidence that large firms are better positioned to cope with this regulation shock because they can use their existing legal teams and IT resources. Data-intensive firms with EU exposure mitigate the loss of data by hiring more AI-related talents and developing more data processing technology.

I then use GDPR as a supply shock of privacy to EU consumers and study the demand for privacy by consumers. Privacy protection is never about eliminating data sharing and reaching a state of secrecy. It is about giving consumers the choice to share more or less data as they desire. On the one hand, consumers value privacy and want to limit the amount of data they share with firms; on the other hand, consumers have the incentive to share some data to improve their own user experiences. For example, I want Netflix to know my preferences so that it can recommend TV shows and movies tailored to my tastes. Still, I want to avoid Netflix exploiting my data beyond providing essential services. This is especially relevant when regulations like GDPR let customers decide how much data they are willing to share with firms and how their data can be used. Moreover, a lot of digital apps are provided for "free," and we are essentially bartering our data or attention for access to these services.

By sharing less data with firms, EU consumers face a less satisfying user experience than their US counterparts, implying consumers trade off the benefits of privacy protection and data-dependent user experiences. Firms that engage in targeted advertising put in more advertisements to compensate for the loss of advertising effectiveness. Firms also switch to other sources of revenue, e.g., in-app purchases and subscriptions.

The paper is then followed by a theoretical part. The goal is to build a two-economy general equilibrium model where multinational firms offer goods and services to both domestic and foreign consumers. The model captures both the data feedback loop and the inter-dependency between value of data and privacy. In the model, consumers' consumption behaviors generate valuable data that firms can use to improve their production technology. The extent of data collection is subject to consumers' privacy preferences and regulatory mandates in each regime. In the current version of the paper, I set up a simplified theoreti-

cal framework to help rationalize the empirical findings and provide guidance for calibration. In the final part of the paper, I combine the findings from the two empirical sections, the demand for consumer data and the demand for consumer privacy and calibrate the model and perform preliminary welfare analysis. The moments I target in the model include the share of revenue generated from the European market pre-GDPR, the shifting in revenue from the EU to the US after GDPR, and the decline in service quality for EU users post-GDPR.

Related Literature: Topics on the data economy are gaining traction in recent years. The papers in this literature embody the notion that data is a by-product of economic activities, data can be traded as an asset, and data may enter the production process as an input (Acemoglu et al. 2019; Admati and Pfleiderer 1990; Bergemann et al. 2019; Choi et al. 2019; Cong et al. 2020; Fajgelbaum et al. 2017; Farboodi and Veldkamp 2021; Jones and Tonetti 2020; Ordonez 2013; Veldkamp 2005). In particular, some papers focus on the integration of data technology/AI and human labor and its impact on firms' behaviors (Abis and Veldkamp 2020; Cao et al. 2021, 2020). The literature has been trying to come up with a measure of the value of data. My paper provides a tractable estimation framework that combines the value of data and privacy in an equilibrium model and properly accounts for the feedback loop and the multiplier effect of data.

Past literature has also shed light on the impact of privacy regulation on digital marketing, VC funding, and firm performance (Aridor et al. 2020; Benkler et al. 2018; Bleier et al. 2020; Canayaz et al. 2022; Choi et al. 2019; Evans 2009; Goldfarb and Tucker 2011; Jia et al. 2018, 2020; Johnson et al. 2020; Lenard and Rubin 2013; Martin et al. 2019). Goldfarb and Tucker (2011) study the effects of the European E-Privacy Directive, which limited firms' ability to track users' online behavior and show that online display advertisements in the EU became less effective than other areas after the directive was enacted. Jia et al. (2018) find that following the enactment of GDPR, EU startups experienced adverse effects on financing in terms of overall dollar amount raised, number of deals, and the dollar amount raised per individual deal. Canayaz et al. (2022) study the negative impact of CCPA on the profitability of conversational AI firms. I provide further evidence on the impact of a regional privacy regulation (GDPR) from a global perspective and focus on both firms and

consumers. I show that regional regulation can exert externality on other parts of the world through international businesses.

The literature has also been trying to put a monetary value on consumers' privacy preferences. Tang (2019) runs a lending experiment on a Chinese fintech platform. The paper links loan application completion rate with borrowers' privacy preferences, and measures the value of loans that borrowers are willing to give up in order not to disclose sensitive information (social network ID or employer). Bian et al. (2021) studies how Apples' app privacy disclosures affect app users' willingness to download an app, and its negative impact on revenue. My paper empirically documents that consumers trade off the benefits of privacy protection and data-dependent user experiences, and the value of privacy is estimated from the structural model.

The rest of the paper is structured as follows. In Section 2, I describe the data and the measurement methods used in the paper. In Section 3, I analyze the demand for data by firms. In Section 4, I study the demand for privacy by consumers. In Section 5, I set up a theoretical framework and perform a preliminary model calibration and welfare analysis. Section 6 concludes.

2 Data and Measurement

2.1 Data Sources

2.1.1 US Online Job Posting Data

US Online Job Postings data covers more than 200 million electronic job postings in the US from Jan 1, 2010 to May 31, 2020. Burning Glass web-scraped job posting information from around 40,000 company websites and online job boards, and they apply a de-duplication algorithm to avoid counting the same job posting multiple times. They parse the raw textual data and extract detailed information on the Employer, location, occupation, industry, wages, and skills required. Carnevale et al. (2014) estimate that the job posting data covers around 60% - 70% of all vacancies in the United States. The detailed skill requirements in the job posting data will enable me to measure US firms' demand for different types

of talent. Following Abis and Veldkamp (2020), Acemoglu et al. (2020), and Babina et al. (2020), I classify jobs into AI-related postings and data-management-related postings.[7] Firms' demand for data managers, data scientists, and machine learning engineers can help me measure how a firm's business model depends on consumers' data. I can also study how the workforce composition of US firms changes in response to privacy regulations.

2.1.2 Accounting, Financial, and Geographical Segment Data

I obtain accounting and financial data of US public firms from Compustat North America Fundamentals Quarterly and CRSP, including total assets, total debt, total sales, gross profits, net profits, market capitalization, daily stock prices, etc.

Furthermore, Compustat Geographical Segment data supplements the firm-level accounting data with revenue, costs, investment compositions by geographical regions. FASB[8] 131, effective December 15, 1997, requires public business enterprises to report financial information and descriptive information about their Operating segments.[9] This Statement requires that a public business enterprise report a measure of segment profit or loss, certain specific revenue and expense items, and segment assets. It requires reconciliations of total segment revenues, total segment profit or loss, total segment assets, and other amounts disclosed for segments to corresponding amounts in the enterprise's general-purpose financial statements. It requires that all public business enterprises report information about the revenues derived from the enterprise's products or services (or groups of similar products and services), about the countries in which the enterprise earns revenues and holds assets, and about major customers regardless of whether that information is used in making operating decisions. However, this Statement does not require an enterprise to report information that is not prepared for internal use if reporting it would be impracticable.

The S&P Global Market Intelligence parses the 10-K filing textual data and tabulates the

[7] The keyword list used for classification can be found in Appendix A.2.

[8] Financial Accounting Standards Board.

[9] This Statement supersedes FASB Statement No.14, Financial Reporting for Segments of a Business Enterprise, but retains the requirement to report formation about major customers. It amends FASB Statement No.94, Consolidation of All Majority-Owned Subsidiaries, to remove the special disclosure requirements for previously unconsolidated subsidiaries. This Statement does not apply to nonpublic business enterprises or to not-for-profit organizations. See https://www.fasb.org/page/PageContent?pageId=/reference-library/superseded-standards/summary-of-statement-no-131.html for more details.

segment disclosure in a structured format. The Compustat Business Information files were designed to allow for restated data in conjunction with changes in disclosure requirements. The Segment Item Value File provides the historical data and up to 2 data source years of restated data back to 1998. The number of records for each data year depend on whether the company restates the period with a subsequent source. [10] For each year, I keep the data when it was first reported (historical data). During the sample period 2010-2021, around 72% of US public firms disclose their geographical revenue compositions each year, and 60% of US public firms generate revenue from international sources.

The segment data enables me to measure the fraction of revenue coming from and the strategic importance of each geographical region for US public firms. I am particularly interested in how US multinational firms reallocate their businesses across geographical segments.

2.1.3 Risk Disclosures in Annual 10-K Filing

Under Regulation S-K Item 105, US public firms are required to provide, under the caption "Risk Factors" in their 10-K filings to the SEC, a discussion of the material factors that make an investment in the registrant or offering speculative or risky. They need to concisely explain how each risk affects the registrant or the securities being offered. Campbell et al. (2014) find that managers faithfully disclose the risk they face, and firms facing greater risk disclose more risk factors.

I use textual analysis tools to extract corporate risk disclosures from their annual 10-K filings. First, I obtain the cleaned 10-K filings from Software Repository for Accounting and Finance[11]. Second, I use regular expressions to identify the "Item 1A Risk Factors" section. It is implementable because 10-K filings are structured format-wise. If it exists, "Item 1A" is always followed by "Item 1B" or "Item 2". I can then easily identify the sections of text

[10] For example, if XYZ Corp reported their 1998 business segment data on the 1998 10K, there would be one record for that year. In 1999, XYZ Corp restates their 1998 data with the 1999 10K, there would be one record for 1999 and two records for 1998: one with the Source Year of 1998 and the other with 1999. In 2000, they restate both 1999 and 1998 data. There would be one record for 2000, two records for 1999 (one historical [Source Year = 1999] and one restated [Source Year = 2000]), and three records for 1998 (one historical [Source Year = 1998] and two restated [Source Year = 1999, 2000].

[11] Tim Loughran, Bill McDonald, and their team retrieved the 10-K filings of US public firms from 1993-2021 from the SEC. They parsed the raw filings to easily machine-readable text files. Their parsing procedures are detailed here, https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/10x-stage-one-parsing-documentation/.

that are sandwiched by the tag "Item 1A" and "Item 1B" or "Item 2". Third, since in the previous step, I may also include sections from the index part at the beginning of each 10-K filing, I only keep the longest section among the risk sections identified in the second step. In total, my sample covers 84,369 10-K filings of 18,018 filers with unique Central Index Key (CIK) from 2006 to 2020.

2.1.4 Innovation

Patent data are from the United States Patent and Trademark Office. Kogan et al. (2017) have introduced a new measure of the economic value of patents. They use the stock market response to patent granting to estimate the economic value of patents. They have made the data available online thorough a GitHub repository[12]. They have also matched the patent data to the the CRSP firm/security level identifier.

2.1.5 Google Play Store Data

I collect app review data from Google Play Store to measure user experiences. The review data contains both numerical ratings and textual comments. The numerical rating is on a scale of 1 (low) to 5 (high). In the textual comments, consumers share details about their experiences while using the apps. When we rank the reviews by relevance, the ones at the top are usually very informative about apps' main products or services. By switching the region of the Google Play Store, I collect the data separately for US users and EU users. Since companies often offer different versions of products in different markets, app user experiences can differ across countries. Moreover, the quality of digital services will be affected by the amount of data users share with app developers.

My analysis focuses on 4,883 popular apps on Google Play Store. To compile this list of apps, I start with the 250 most popular apps recommended by Google in each app category, including Art and Design, Auto and Vehicles, Beauty, Books and Reference, Business, Comics, Communication, Dating, Education, Entertainment, Events, Finance, Food and Drink, Health and Fitness, House and Home, Libraries and Demo, Lifestyle, Maps and

[12] https://github.com/KPSS2017/Technological-Innovation-Resource-Allocation-and-Growth-Extended-Data

Navigation, Medical, Music and Audio, News and Magazines, Parenting, Personalization, Photography, Productivity, Shopping, Social, Sports, Tools, Travel and Local, Video Players and Editors, and Weather. Then I extend from this initial list and search for relevant apps associated with each app, and this process brings me to around 20,401 apps.

In 2021, Google announced that all developers on the Google Play platform are required to disclose their apps' privacy and security practices in a Data Safety section of their apps' store listing page. The measure is aimed at helping Google Play users understand how the apps collect and share their data before they download[13]. This information helps users make more informed choices when deciding which apps to install. Figure A1 provides several screenshots from Instagram's Data Safety section on what information it collects from users and for what purposes. By July 20, 2022, all developers must declare how they collect and handle user data for the apps they publish on Google Play and provide details about how they protect this data through security practices like encryption. This includes data collected and handled through any third-party libraries or SDKs used in their apps.

To be included in my sample, an app needs to have a valid Data Safety disclosure and have at least ten reviews before and after GDPR came into effect. These two criteria bring the sample from 20,402 apps to 4,883 in the main analysis.

2.2 Measurement

2.2.1 Data Intensiveness

The data intensiveness measure assesses the degree to which a firm's business operations depend on consumer data collection and the extent to which this data can be used to improve its products, technology, and marketing strategies. Notable examples include information technology firms such as Google, Meta, and Netflix. These companies gather vast amounts of data to refine their algorithms, enhance their products, and function as digital platforms that facilitate advertising campaigns for smaller businesses.

However, the digital economy extends far beyond these well-known tech giants. Rapid advancements in computing power and artificial intelligence have enabled a growing number

[13]Apple App Store also has a similar change in 2021, named privacy nutrition labels. These labels fall into three categories: "Data Used to Track You", "Data Linked to You", and "Data Not Linked to You".

of firms to collect, process, and exploit large volumes of consumer data, sparking a digital transformation across various industries. Retail giants like Walmart and Target, while not traditionally seen as technology firms, have started hiring data scientists and machine learning engineers in response to the increasing need for consumer data analysis. Likewise, the automotive industry is experiencing a digital revolution, with Alphabet's Waymo and GM's Cruise heavily investing in AI talent for their research and development teams working on autonomous vehicles.

It is clear that relying solely on industry classifications is inadequate for understanding the digital economy. Investment in digital assets has shifted from physical infrastructure to talent acquisition in data management and analysis, as well as research and development of data processing technology. I propose a measure of data intensiveness based on the talent employed by firms and the market and scientific value of their data processing technologies. Data processing technology refers to patents with the Cooperative Patent Classification (CPC) code G06, which pertains to computing technology. I assess their scientific value by the number of forward citations these patents receive (adjusted for patent "age") and measure market value using the method proposed by Kogan et al. (2017). In each year-quarter, I compute the scientific value of data processing technology using the following formula:

$$\text{Scientific Value}_{i,t} = \frac{\text{Forward Citations (Newly Granted G06 Patents)}_{i,t}}{\text{Total Forward Citations of All Newly Granted Patents}_{i,t}} \tag{1}$$

and compute the market value of data processing technology as

$$\text{Market Value}_{i,t} = \sum_{p} \frac{\text{Market Value of Patent}_{i,p,t}}{\text{Market Capitalization}_{i,t}} \tag{2}$$

In each year-quarter, the market value of patent p is scaled by the market capitalization of firm i. These two variables capture the first dimension of data intensiveness: data processing technology.

For the second dimension of data-intensiveness, I use the keywords identified by Abis and Veldkamp (2020), Acemoglu et al. (2020), and Babina et al. (2020) and classify jobs that

require AI and data management skills. The list of AI skills includes machine learning, computer vision, deep learning, virtual agents, image recognition, natural language processing, speech recognition, and neural networks, among others. Data management skills encompass Apache Hive, information retrieval, data warehousing, SQL Server, data visualization, database management, data governance, and database administration, among others. The complete list of keywords for AI skills and data management skills can be found in Appendix A.2. For each year-quarter, I compute the percentage of job postings that require AI related skills and data management related skills.

$$\text{AI Talent Demand}_{i,t} = \frac{\text{Job Postings Requiring AI Related Skill}_{i,t}}{\text{Total Job Posting}_{i,t}} \tag{3}$$

$$\text{Data Management Demand}_{i,t} = \frac{\text{Data Management Related Posting}_{i,t}}{\text{Total Job Posting}_{i,t}} \tag{4}$$

I integrate information from these multiple dimensions of data intensiveness, scale them, extract the first principal component from the scaled vectors, and generate a comprehensive measure for data intensiveness.

I compute the pre-2018 average (prior to GDPR implementation) of this data-intensive measure. Firms are ranked based on this comprehensive measure, with the median serving as the cutoff. Firms above the median are classified as data-intensive, while those below the median are categorized as non-data-intensive. Table 1 displays the industry average of this data intensiveness measure, as well as the fraction of firms classified as data-intensive in each industry.

[Insert Table 1 Here.]

3 The Demand for Data by Firms

In Section 2.2.1, we observe that the demand for data scientists and machine learning engineers varies among firms. If data is combined with talent to create knowledge and enhance production technology (Abis and Veldkamp 2020), a negative shock to the data available to

firms is likely to impact their production processes.

This section examines how US multinational corporations react to the General Data Protection Regulation (GDPR), a regional privacy regulation. US multinational firms have access to both EU and non-EU markets. GDPR stands as the most comprehensive and stringent privacy regulation worldwide. In the US, there is no federal-level comprehensive privacy law, aside from industry-specific privacy standards such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA). [14] GDPR grants EU consumers greater control over their data and enhances their role as the supplier of data. Prior research (Aridor et al. 2020; Goldberg et al. 2019) demonstrates that following GDPR's implementation, European households shared less data with firms and made it more difficult for firms to track them online. As a result, this regulation has limited US firms' access to European data. In this sense, GDPR acts as a data supply shock, enabling us to examine the data demand of US multinational firms.[15]

3.1 Cross-Market Business Adjustment

For US multinational firms, the European Union represents a significant foreign market, accounting for a substantial portion of their internationally originated sales. Specifically, when considering US firms with an EU segment, the region contributes to 16% of their total sales. Historically and culturally, consumer preferences in the European market closely align with those in the US domestic market. Consequently, acquiring insights into EU consumers' preferences enables US technology firms to better understand their domestic customers. Thus, the EU market serves as a crucial data source for US firms.

Following the implementation of the GDPR, it is anticipated that US firms' access to EU consumer data will be constrained. In response to this, US multinational firms may strategically shift portions of their businesses away from the European market and towards other regions, particularly the US domestic market, to capitalize on the more lenient regulatory

[14] Several US states have passed state-level privacy laws, including California (effective January 1, 2020), Virginia (effective January 1, 2023), Colorado (effective July 1, 2023), and Utah (effective December 31, 2023).

[15] The regulation was drafted by EU legislators and passed by the European Parliament, making it less likely to be influenced by lobbying efforts from US corporations.

environment. This section tests this hypothesis, examining the potential impact of GDPR on the operations and strategies of US multinational firms in the context of data access and market presence.

[Insert Figure 3 Here.]

Figure 3 shows the fraction of revenue generated from the European market by US firms that had European market operations prior to 2018. We can see a clear decline in the share of EU business for data-intensive firms after 2018, when GDPR came into effect. This decline deepens and persists till 2021. I employ a difference-in-differences identification strategy and formalize the observations in Figure 3 in a regression framework.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_{\text{data}} \cdot \text{GDPR-Effective}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \tag{5}$$

where $Y_{i,t}$ is the fraction of revenue generated from the European market by US firm i in year t, α_t is year fixed-effect, ϕ_i is firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. GDPR_t is a binary variable that equals one if time t is after GDPR's enactment year, 2018. Data-Intensive_i is a binary variable that equals one if firm i is in the data-intensive category.

[Insert Table 2 Here.]

The results are shown in Table 2, where our primary interest lies in the coefficient β_{data} before the interaction term in equation 5. Column (1) reveals that the EU sales percentage (EU sales/total sales) for data-intensive firms experienced a decrease of 1.55 percentage points following GDPR's implementation. Considering the unconditional mean of EU sales percentage before 2018 stands at 16 percentage points, this coefficient corresponds to an 10 (1.55/16) percent decline in EU business. In column (1), I use a binary measure of data-intensiveness. In column (2), I show the results from an alternative specification where I use the original continuous measure of data-intensiveness. The results help us understand the effects at the intensive margin. Column (4) shows no differential change in total sales between data-intensive and non-data-intensive firms. In column (3), it becomes evident that the effects are almost entirely attributable to the reduction in EU sales. In Table A1,

an additional interaction term is introduced, involving the time indicator GDPR-Effective$_t$ and a binary variable Tech$_i$, which equals one when a firm i's North American Industry Classification System Code (NAICS) begins with 51. This inclusion helps alleviate the concern that the findings in Table 2 arise from a common trend within the tech sector, as opposed to firms' reliance on data. The broader issue of data dependence is further discussed in section 2.2.1, which is the focus of this paper.

To further examine the impact of GDPR on EU sales, I extend the regression in equation 5 to a dynamic difference-in-differences framework. This approach allows me to check for pre-trends and investigate when the effect of GDPR begins and how persistent it is. I run the following regression:

$$Y_{i,t} = \alpha_t + \phi_i + \sum_{\tau \neq 2018} \beta_{\text{data},\tau} \cdot \boldsymbol{I}(t = \tau) \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \tag{6}$$

The notations in the above equation are similar to those in equation 5, with the exception that we now include by-period interaction terms and analyze the coefficients $\beta_{\text{data},\tau}$. Figure 4 plots the coefficients from the regression in equation 6, along with a 95 percent confidence band. The figure clearly shows no pre-trend, and the negative impact of GDPR only emerges after 2018, gradually deepening over time.

[Insert Figure 4 Here.]

The drop in EU sales may be attributed to either a reduction in business size in real terms or a decline in the profitability of the EU segment. To investigate whether the effect is primarily driven by US firms actively reallocating their businesses across geographical segments rather than decreasing profitability in the EU market, I examine the profitability of the EU segment and the firm for both data-intensive and non-data-intensive firms. The results are displayed in Table A2, where I consider two measures of profitability: gross profit margin (GPM) and operating profit margin (OPM). As evident from the table, the coefficient preceding the interaction term is nearly zero and lacks statistical significance. Consequently, no discernible change in profitability exists between data-intensive firms and non-data-intensive firms, either for the EU segment or at the firm level.

3.2 The Decreasing Return to Data

In the previous section, we established that data-intensive firms experienced a decrease in the size of their EU business following the implementation of GDPR. However, the cause of this decrease could be either supply-driven or demand-driven. On the supply side, firms may actively reallocate internal resources due to the limited access to EU consumers' data, which affects the value of the EU market segment. For example, firms might allocate fewer resources to improving product or service quality in the EU. On the demand side, EU consumers might opt for fewer products or services from the EU due to a decline in their quality. As discussed in the introduction, the quality of digital services can be influenced by both platform-wide service quality, which depends on the aggregate level of data sharing, and the individual level of data sharing. When consumers have the option to decide how much data to share with firms, they will balance the benefits of privacy protection against data-dependent user experiences. By sharing less data with firms, consumers may encounter a diminished user experience, prompting them to substitute away from digital consumption.

[Insert Table 3 Here.]

To verify that the observed effects are predominantly supply-driven, we will consider existing research that theoretically demonstrates diminishing returns to data (Farboodi and Veldkamp 2021). It is implied that larger firms, possessing a more substantial stock of data, might be less affected by privacy regulations that limit their access to new data. In this subsection, we test this hypothesis by running the same regression as in equation 5, but we separate the analysis for large firms (above median market capitalization) and small firms (below median market capitalization). The results are shown in Table 3. We can see that small firms face a much bigger impact than large firms. The coefficient is around twice as big as the coefficient with the full sample. In Figure 5, we plot the dynamic effects for small firms and see that the effects on small firms gradually deepen and persist till 2021. In Figure 6, we can see that the effect on large firms is not statistically significant from 0. This confirms our hypothesis that the drop in EU sales are largely supply-driven. As the EU segment delivers less value in terms of the consumer data, US multinational firms

strategically reallocate their resources across geographical segments and switch to markets with more lenient privacy regulations.

[Insert Figure 5 and Figure 6 Here.]

3.3 Effects on Talent Hiring

If data is combined with talent to create knowledge and improve production technology (Abis and Veldkamp 2020), a negative shock to the amount of data available to firms will likely change their production process. In this section, I look into the complementarity and substitutability between data and talents. I examine how the demand for AI and data management talents changes for data-intensive versus non-data intensive firms after GDPR. Since it usually takes 1-2 years for job posting to be reflected in actual hiring, I use the passage time of GDPR (April 2016) as the time cutoff in this analysis. We adopt a difference-in-differences identification strategy and run the following regression.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR-Pass}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \tag{7}$$

where $Y_{i,t}$ is the fraction of job postings that require AI or data management related skills, α_t is year fixed-effect, ϕ_i is firm fixed-effect, and $\boldsymbol{X}_{i,t}$ captures time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. GDPR-Pass_t equals one if time t is after GDPR's passage year, 2016. Data-Intensive_i is a binary variable that equals one if firm i is classified into the data-intensive category.

[Insert Table 4 Here.]

As shown in Table 4, data-intensive firms hire less data-management related talents after GDPR was passed in 2016. The effects are bigger for small firms (below median market capitalization) than large firms (above medial market capitalization). There are multiple ways to interpret the results. One interpretation is that when firms have less access to data in the European market, their demand for data management related talents will also be less.

[Insert Table 5 Here.]

In Table 5, we can see that data-intensive firms hire more AI-related talents after GDPR was passed in 2016. The effects are muted for small firms but both statistically and economically significant for large firms. The unconditional average of the percent of job postings that require AI-related skills is 1.2 percent. Given that AI hiring started to pick up steam after 2016, the results in column (1) is not very surprising. The more interesting part is the differential impact on large firms versus small firms. Large firms aggressively ramped up their AI talent hiring after GDPR is expeceted to limit their access to consumer data from the European market while small firms lag behind.

4 The Demand for Privacy by Consumers

Privacy protection is not solely about limiting data sharing, but about granting consumers the autonomy to decide the extent of data sharing. Indeed, sharing data often comes with rewards, either pecuniary or otherwise. When it comes to monetary rewards, many are happy to provide phone numbers and email addresses in exchange for discounts. For instance, we might share our contact information to get 10 percent off on an online shopping site. On the non-monetary side, we permit platforms like social media and streaming services to access our personal data and online behavior. This, in turn, allows them to refine and personalize our user experiences. The allure is clear: imagine a TikTok stream impeccably tailored to one's taste or a Netflix dashboard highlighting favorite shows. Yet, the balance between data sharing and privacy is delicate. When companies push boundaries or misuse personal data, consumer welfare might be impaired.

In this section, I explore how consumers weigh the benefits of data sharing against the need for privacy protection. As with section 3, the introduction of GDPR acts as a natural experiment, altering the "supply of privacy." This regulatory change empowers EU consumers with greater data autonomy, letting them decide how much data they share with companies. By analyzing review data from the Google Play Store, I aim to understand how the user experiences of EU and non-EU consumers change post-GDPR.

4.1 Google Play Store Review Data

Consumers evaluate Apps along three primary dimensions. Firstly, they look at an app's functionality, placing emphasis on how well it performs its intended tasks and the intuitiveness of its interface. Secondly, they consider the advertisements present within the app, with a keen eye on their relevance and intrusiveness to the user experience. Lastly, any additional offerings such as in-app purchases or subscription options are also taken into account.

In Google Play Store, app users can leave both numerical ratings (on a scale of 1-5) and textual comments. People comment on all three aspects of user experiences as mentioned above. We can visit different versions of the Google Play Store by changing the country and language options. This provides us with a way to differentiate between the comments left by EU and non-EU users. For example, when you use the url, "https://play.google.com/store/apps/details?id=com.instagram.android&hl=en_US&gl=US," you can visit the US version of the Instagram page. The Ratings and Reviews section will show the reviews left by US users. When you change the language and country option, from "&hl=en_US&gl=US" to "&hl=fr&gl=FR", you can visit the French version of the Instagram page. The comments section will only show the reviews from French users. Since GDPR applies to all EEA countries, I gather reviews from the EEA countries in one subsample, while US reviews are compiled separately.

Apps vary in their reliance on consumer data. Drawing from the data safety disclosures discussed in Section 2.1.5 from the Google Play Store, I have classified apps into two categories: those that are heavily data-driven for personalization and those that operate with minimal user information. Users interacting with the former are likely to notice a significant change in their experience if they opt to share less data, while the impact is much more limited for users of the latter group.

Before delving into an in-depth analysis of this review data, I will first present some summary statistics to set the context. To be included in my sample, an app needs to have a valid Data Safety disclosure and have at least ten reviews before and after GDPR came into effect. There are 4,883 apps in the main analysis.

[Insert Table 6 Here.]

4.2 App Ratings

EU and US users operate under distinct privacy regulatory frameworks. For EU users, choosing to share less data with mobile app providers can have implications on their user experience. This is especially the case for apps that rely heavily on data for personalization. Such apps often seek a diverse range of information to tailor user experiences. This can encompass basic details like names and email addresses, but may extend to more sensitive data such as political or religious beliefs, sexual orientations, and health metrics. Additional data, like browsing histories and in-app activities, also contribute to this personalization process.

To test for this hypothesis, I employ a difference-in-differences identification strategy and study how limited access to data in the European market affects the quality of service provided by mobile apps, measured by the daily average user numeric ratings. I run the following regression.

$$Y_{i,m,t} = \alpha_m + \phi_i + \beta_{\text{service}} \cdot \text{GDPR}_m \times \text{Personalization Collected}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,m,t} + \varepsilon_{i,m,t} \quad (8)$$

where $Y_{i,m,t}$ is the daily average rating for app i on day t. α_m is the year-month fixed-effect, ϕ_i is the app fixed-effect. GDPR_m is a binary variable that equals one if time t is after GDPR's enactment month, May 2018. $\boldsymbol{X}_{i,m,t}$ is a vector of time-varying app characteristics, including the total number of daily review. $\text{Personalization Collected}_i$ is a binary variable that equals one if app i collects user data for personalization purposes. I analyze the reviews by the EU and US users separately.

[Insert Table 7 Here.]

The results are shown in Table 7. Columns (1) and (2) show that, for apps that collect user information for personalization purposes, the user rating of EU users declined while the user rating of US users did not change after GDPR came into effect. User numeric ratings are concentrated around 4.0, and the inter-quartile range for EU users is 3.67-5.0. Therefore, a 0.08 decline in user rating translates into a 6 percent decrease of inter-quartile range.

In column (3), I run a triple difference regression.

$$\begin{aligned} Y_{i,m,k,t} =& \alpha_m + \phi_i + \psi_k + \beta_1^* \cdot \text{GDPR}_m \times \text{Personalization Collected}_i \times \text{EU}_k \\ &+ \beta_2 \cdot \text{GDPR}_m \times \text{Personalization Collected}_i + \beta_3 \cdot \text{GDPR}_m \times \text{EU}_k \\ &+ \beta_4 \cdot \text{Personalization Collected}_i \times \text{EU}_k + \boldsymbol{\gamma} \boldsymbol{X}_{i,m,k,t} + \varepsilon_{i,m,k,t} \end{aligned}$$

where $Y_{i,m,k,t}$ is the average daily rating by users from region k for app i on day t. ψ_k is the region (US or EU) fixed effect. EU_k is an indicator variable that equals one if the reviews come from the EU users. $\boldsymbol{X}_{i,m,k,t}$ is a vector of time-varying app characteristics, including the total number of daily review. The coefficient β_1^* before the triple interaction term captures the differential change in user quality between the EU and US users after GDPR for apps collecting personalization information.

4.3 Advertisement Complaints

As of March 2023, 97% of the apps on Google Play Store, and 94.5% of the apps on Apple App Store are free to download.[16] Moreover, for most of the apps, we can enjoy basic functions without paying a penny. Then how do app developers make money?

Of course, app owners are not running charities. There are multiple ways for them to monetize their users, including advertisements and in-app purchases and subscriptions. When we use an app, we devote our attention to the content displayed in the user interface. Like the television industry, user or viewer attention can be exploited for advertising. App owners can incorporate and auction off advertisement slots in their apps. Common types of mobile advertisements include banners, pop-up windows, native ads, and rewarded videos. As advertisement publishers, app owners work with advertisement networks and delegate the advertisement auctions to them.

However, most of people do not like advertisements and find them extremely annoying. The average numeric rating is 3.0 when mobile app users mention advertisement related keywords in their reviews, compared to 4.0 for all types of reviews.

In-app advertisement is one important income source for most free apps. Oftentimes, we

[16] https://www.statista.com/statistics/263797/number-of-applications-for-mobile-phones/

are also asked to register for an account with our email addresses or phone numbers. And for a lot of social media apps, we willingly provide personal information like names, genders, birthdays, home addresses, places of births, etc. The list goes on, and sometimes we would be surprised at how much we have shared with the internet. When we use these apps, we also reveal our own preferences through app activities. These are all valuable data that can be used by these apps to build a digital profile of us. Moreover, the data we shared with different apps can also be linked together using either our device unique identifiers or other individual identifiers like email addresses or phone numbers. All these valuable information can further be used for targeted advertising, which shows different advertisements to people with different preferences.

When users choose to share less information with apps and they choose to block third-party advertisement tracking, we might see a decrease in advertisement effectiveness. I define apps that engage in target advertising as the ones that collect personal information for advertising and marketing purposes and apps that collect device specific IDs for cross-app tracking. I do not have micro-level data on the number of advertisements each app puts into their user interface, but I do observe the comments that are related to complaints about advertisements.

[Insert Table 8 Here.]

If app owners try to make up for the loss in advertisement effectiveness, they might put in more advertisements. If the number of user complaints about advertisements are proportional to the number of ads being put into these apps, we will very likely see an increase in the advertisement related complaints. Table 8 confirms this hypothesis, we see a significant increase in advertising related complaints for both EU and US consumers. However, the increase is much larger for EU users. The results translate into a 10 percent increase in advertisement intensity for EU mobile apps after GDPR came into effect.

[Insert Table 9 Here.]

Another way apps can generate revenue is through in-app purchases and subscriptions. I do not have data on in-app purchases and subscriptions, but users often write reviews about

these type of expenses. I identify these type of comments through textual review data. The analysis results are shown in Table 9. We can see that there is a much larger increase in purchase related comments among EU users than their US counterparts. This implies that mobile apps are switching to other sources of revenue after data privacy regulations render advertisement less effective.

[Insert Table 10 Here.]

In Table 10, we can see that the results in Table 8 and Table 9 are not driven by larger increase in active user base in the European market. The growth in total number of reviews are very similar across the two markets.

5 Model

This study endeavors to bridge the gap between the empirical evidence and theoretical work on data and privacy. I provide a framework to quantitatively measure the value of data, the importance of privacy, and the welfare implications of privacy regulations. To accomplish this, I develop a two-economy general equilibrium model to better understand the strategic choices made by US multinational firms when facing regional privacy regulations like the GDPR in the European Economic Area.

The model's structure is illustrated in Figure 1. US firms provide goods and services to both European customers and US customers (or, more accurately, customers from the rest of the world). Data is a byproduct of economic activities, and firms collect and analyze consumers' data to learn about their preferences, inspire new concepts or products, and boost productivity. Notably, Figure 1 underscores the data feedback loop within a two-economy framework.

While consumers enjoy the advantages of personalized recommendations and enhanced service quality, they have concerns about sharing personal data with firms. Privacy concerns may stem from the psychological costs or social stigma of disclosing excessive personal information, as well as from predatory advertising or pricing tactics employed by firms.

5.1 Households

At time t, a continuum of households, denoted by $i \in [0,1]$, exists within each economy k, where $k \in \{\text{US}, \text{EU}\}$. Household i chooses between two consumption types: digital goods c_{ijk} and non-digital goods c'_{ik}. Examples of digital goods include social media platforms, streaming services, online shopping, and any other digital services that will potentially document your digital footprints. In contrast, non-digital goods represent other outside options, such as purchasing groceries at a local market, attending concerts, or engaging in offline entertainment services. Households select from a wide array of digital products, with $j \in [0,1]$. The utility of household i is determined by the following equation, which consists of three components: digital consumption, non-digital consumption, and privacy concerns.

$$u_{it,k} = \int_0^1 K \left(\underbrace{(\xi \bar{x}_{jt} + x_{ijt,k}) \ln c_{ijt,k}}_{\text{digital consumption}} - \underbrace{\delta_i x_{ijt,k}^2 \ln c_{ijt,k}}_{\text{privacy concerns}} \right) dj + \underbrace{\ln c'_{it,k}}_{\text{other consumption}} \tag{9}$$

Data and Service Quality As households engage in digital consumption activities, they generate data ($d_{ijk} = c_{ijk}$). Firms providing these services collect a portion (x_{ijk}) of the generated data, using it to refine marketing strategies, product offerings, and technologies. $\bar{x}_{jt}$ is the average level of data sharing among all customers of firm j.

$$\bar{x}_{jt} = \frac{\int_i x_{ijt,us} c_{ijt,us} di + \int_{i'} x_{i'jt,eu} c_{i'jt,eu} di'}{\int_i c_{ijt,us} di + \int_{i'} c_{i'jt,eu} di'} \tag{10}$$

Digital consumption benefits households when they share a reasonable amount of data with service providers. Personalized recommendations from platforms like Netflix and Amazon are highly sought after, as they elevate user experiences. Since the product market is subject to search frictions, allowing firms to learn about our preferences can help reduce search costs. While this paper does not delve into the micro-foundations of search friction in the product market, the concept is captured using a simple multiplicative form $(\xi \bar{x}_{jt} + x_{ijkt}) \ln c_{ijk}$. We can view it as the data-augmented user experiences. $x_{ijkt} \ln c_{ijk}$ captures the private benefit of data sharing. Individual consumption experience is affected by individual data sharing. The quality of digital services increases in proportion to the amount of data shared with

firms. $\xi \bar{x}_{jt} \ln c_{ijk}$ represents the social benefit of data sharing. ξ is a parameter that governs the social value of data. More data sharing by all consumers of firm j improves the product quality and recommendation algorithm at the firm level. K serves as a scale factor, determining the relative importance of digital consumption to non-digital consumption.

Privacy Preferences The second term in equation 9 captures the disutility from sharing data with firms.[17] Within the context of this paper, I abstain from distinguishing among the various mechanisms that underlie consumers' privacy preferences. In my model, households place a premium on safeguarding their privacy, incurring a disutility denoted by the expression $\delta_i x_{ijk}^2 \ln c_{ijk}$ when they choose to share a fraction x_{ijk} of their data with firms. It is important to note that households exhibit variations in their privacy preferences, with certain individuals displaying a heightened awareness of privacy concerns, while others exhibit a greater willingness to share their data. I assume that $\delta_i \in \{0, \delta\}$. There is a fraction α of people in the population that are privacy-sensitive for whom $\delta_i = \delta > 0$. Conversely, the remaining fraction (1-α) represents individuals who exhibit complete indifference towards data sharing, thereby $\delta_i = 0$. By employing a quadratic functional form with respect to x_{ijk}, I aim to capture the intuitive notion that when consumers share a reasonable amount of data with firms, it can lead to utility improvements. However, excessive data collection and violations of consumers' privacy by firms can tip the scales, making increased data sharing detrimental to consumer welfare.

5.2 Digital Firms

Multinational Digital Firms Multinational digital firms strategically allocate their resources across two geographical segments, the US and EU markets. Firm j's total production at time t is Y_{jt}, and it sells $Y_{jt,us}$ of them to the US market and $Y_{jt,eu}$ to the EU market. Firms combine technology A_{jt} and labor $L_{jt,us}$, $L_{jt,eu}$ to produce final products. The profits

[17] Firms can potentially track consumers across platforms and learn about every aspect of their preferences beyond the reasonable use of data. Excessive data collection by firms can lead to predatory advertising and pricing practices. There is also a social cost associated with the revelation of sensitive personal information, e.g. medical records, marital status, sexual orientation, religious beliefs, and immigration status, especially for people from disadvantaged socioeconomic backgrounds. Sometimes, firms do not even need to directly obtain such information because machine learning algorithms can make inferences from other observable personal traits, including but not limited to searching, browsing, and shopping histories.

at time t is given by

$$\Pi_{jt} = p_{jt,us}Y_{jt,us} + p_{jt,eu}Y_{jt,eu} - w_t\left(L_{jt,us} + L_{jt,eu}\right) \tag{11}$$

subject to

$$\begin{aligned} A_{jt} =& (D_{j,t-1})^{\eta} \\ Y_{jt,us} =& A_{jt}(L_{jt,us})^{1-\eta} \\ Y_{jt,eu} =& A_{jt}(L_{jt,eu})^{1-\eta} \\ D_{jt} =& (1-\lambda)D_{j,t-1} + x_{jt,us}Y_{jt,us} + x_{jt,eu}Y_{jt,eu} \end{aligned}$$

where

$$x_{jt,us} = \frac{\int_i x_{ijt,us}c_{ijt,us}di}{\int_i c_{ijt,us}di}, \quad x_{jt,eu} = \frac{\int_i x_{ijt,eu}c_{ijt,eu}di}{\int_i c_{ijt,eu}di} \tag{12}$$

$x_{jt,us}$ is the average level of data sharing in the US market, and $x_{jt,eu}$ is the average level of data sharing in the EU market.[18] US multinational firms can price discriminate across two markets but not at the individual level. They charge the US and EU markets different effective prices, $p_{jt,us}$ and $p_{jt,eu}$, adjusting for the value of data they collect from each market.[19]

The production in the two regional markets, $Y_{jt,us}$ and $Y_{jt,eu}$, share the same technology, A_{jt}. Firm j uses accumulated data $D_{j,t-1}$ from time $t-1$ to create technology, where the exponent η captures the output elasticity of data. Data is generated as a byproduct of economic activities, and one unit of consumption generates one unit of data. Therefore, $Y_{jt,us}$ units of consumption from the US markets generate $Y_{jt,us}$ units of data and $x_{jt,us}$ fraction of these data are collected by firm j. In each period, $x_{jt,us}Y_{jt,us} + x_{jt,eu}Y_{jt,eu}$ of new

[18] Notice here I have already used the market clearing condition for the goods market. $Y_{jt,us} = \frac{\int_i x_{ijt,us}c_{ijt,us}di}{\int_i c_{ijt,us}di}$ and $Y_{jt,eu} = \frac{\int_i x_{ijt,eu}c_{ijt,eu}di}{\int_i c_{ijt,eu}di}$.

[19] A salient feature of the data economy is that a lot of digital services are provided for free, which is essentially data barter. Firm j sells products to different markets at different prices, $p_{jt,us}$ and $p_{jt,eu}$. We can think of $p_{jt,us}$ and $p_{jt,eu}$ as the quality adjusted product price. We pay for free apps like Facebook, Instagram, and Tiktok with our attention and participation. We contribute valuable user data to these digital platforms. Suppose quality adjusted price $p_{jt,k} = P_{jt,k}/Q_{jt,k}$. If firm j wants to attract more customers from country k, it can do so by either lowering its listing price $P_{jt,k}$ or improving the quality of the platform $Q_{jt,k}$, customer support or infrastructure.

data are collected, and data depreciates at rate λ.

Non-Digital Firms Non-digital goods serve as the numeraire and outside options for consumers. For simplicity, I abstract away from the production of non-digital products and assume they are supplied elastically at fixed price 1. Prices of digital products are expressed relative to the non-digital product.

5.3 Equilibrium Definition

An equilibrium consists of quantities $\{c_{ijt,us}, c_{ijt,eu}, x_{ijt,us}, x_{ijt,eu}, c'_{it,us}, c'_{it,eu}, Y_{jt}, Y_{jt,us}, Y_{jt,eu}, D_{jt}, L_{jt,us}, L_{jt,eu}\}$ and prices $\{p_{jt,us}, p_{jt,eu}\}$ such that

1. A digital firm chooses a sequence of production decisions $\{L_{jt,us}, L_{jt,eu}\}$ to maximize the discounted value of all future profits.

$$\sum_{t=1}^{\infty} p_{jt,us} Y_{jt,us} + p_{jt,eu} Y_{jt,eu} - w_t \left(L_{jt,us} + L_{jt,eu}\right) \tag{13}$$

   $\{L_{jt,us}, L_{jt,eu}, Y_{jt,us}, Y_{jt,eu}, D_{jt}\}$ solve the firm problem.

2. US households choose a sequence of consumption decisions $\{c_{ijt,us}, c'_{it,us}\}$ to maximize her utility each period.

3. EU households choose a sequence of consumption decisions $\{c_{ijt,eu}, c'_{it,eu}\}$ to maximize her utility each period.

4. Conditional on who owns the data, data sharing choices $\{x_{ijt,us}, x_{ijt,eu}\}$ are incentive-compatible.

5. $\{p_{jt,us}, p_{jt,eu}\}$ clear the goods market.

$$\int_i c_{ijt,us} = Y_{it,us}, \quad \int_i c_{ijt,eu} = Y_{it,eu}, Y_{it,us} + Y_{it,eu} = Y_{it} \tag{14}$$

6. Data depreciates and accumulates from period to period.

$$D_{jt} = (1-\lambda) D_{j,t-1} + x_{jt,us} Y_{jt,us} + x_{jt,eu} Y_{jt,eu} \tag{15}$$

5.4 Two Data Sharing Regimes

5.4.1 Pre-GDPR

First, I consider a baseline scenario where there are not significant regulatory differences across the two markets regarding data privacy. I call this the pre-GDPR regime. In this regime, firms control how much data they want to collect from consumers, and households only choose their consumption bundles. $x_{jt,us}$ and $x_{jt,eu}$ are firms' choice variables. If there are no costs associated with collecting or storing data, firms will collect all the generated data. As a results, firms set $x_{jt,us} = 1$ and $x_{jt,eu} = 1$. For simplicity, I assume a symmetric setting in the baseline model, where all households make the same consumption choices, and all firms make the same production choices. I also assume that all households are privacy-sensitive, $\alpha = 1$.

Households For households in region k $\in \{us, eu\}$, their optimization problem is given by

$$\max_{\{c_{ijt,k}\},c'_{it,k}} u_{it,k} = K \int_0^1 \left(\xi \bar{x}_{jt} + x_{ijt,k} - \delta x_{ijt,k}^2\right) \ln c_{ijt,k} dj + \ln c'_{it,k} \tag{16}$$

subject to

$$W_{it,k} \geq \int_0^1 p_{jt,k} c_{ijt,k} dj + c'_{it,k}$$

where $W_{it,k}$ is the endowment of households in country k.[20] $p_{jt,k}$ is the price of the digital good i in country k. Households choose how many digital goods $c_{ijt,k}$ or non-digital goods $c'_{it,k}$ to consume. Households are hand-to-mouth, and they neither save nor make inter-temporal consumption decisions. This is a reasonable assumption given my main focus is on consumers' digital consumption and data sharing choices.

We can set up the Lagrangian of the households' optimization problem.

$$\begin{aligned}\mathcal{L}_{it,k} =& K \int_0^1 \left(\xi \bar{x}_{jt} + x_{ijt,k} - \delta x_{ijt,k}^2\right) \ln c_{ijt,k} dj + \ln c'_{it,k} \\ &+ \mu_{it,k} \left(W_{it,k} - \int_0^1 p_{jt,k} c_{ijt,k} dj - c'_{it,k}\right)\end{aligned} \tag{17}$$

[20] In an extension, I will also consider the case where firms redistribute all the profits back to households

We can derive the first order conditions

$$\text{US digital goods: } \frac{\partial \mathcal{L}_{it,k}}{\partial c_{ijt,k}} = K\left(\xi \bar{x}_{jt} + x_{ijt,k} - \delta x_{ijt,k}^2\right) c_{ijt,k}^{-1} - \mu_{it,k} p_{jt,k} = 0 \tag{18}$$

$$\text{non-digital goods: } \frac{\partial \mathcal{L}_{it,k}}{\partial c'_{it,k}} = c'^{-1}_{it,k} - \mu_{it,k} = 0 \tag{19}$$

$$\text{budget constraint: } \frac{\partial \mathcal{L}_{it,k}}{\partial \mu_{it,k}} = W_{it,k} - \int_0^1 p_{jt,k} c_{ijt,k} dj - c'_{it,k} \tag{20}$$

From equation 18 and 19, the optimal digital consumption for households from country k follows

$$c_{ijt,k} = \frac{K\left(\xi \bar{x}_{jt} + x_{ijt,k} - \delta x_{ijt,k}^2\right) c'_{it,k}}{p_{jt,k}} \tag{21}$$

Along with the budget constraint, we can get

$$\begin{aligned} c'_{it,k} =& \frac{W_{it,k}}{K \int_0^1 \left(\xi \bar{x}_{jt} + x_{ijt,eu} - \delta x_{ijt,eu}^2\right) dj + 1} \\ =& \frac{W_{it,k}}{K X_{it,k} + 1} \end{aligned} \tag{22}$$

where

$$X_{it,k} = \int_0^1 \left(\xi \bar{x}_{jt} + x_{ijt,k} - \delta x_{ijt,k}^2\right) dj \tag{23}$$

We can solve for the optimal digital consumption

$$c_{ijt,eu}^{us} = \frac{K\left(\xi \bar{x}_{jt} + x_{ijt,k} - \delta x_{ijt,k}^2\right) W_{it,k}}{p_{jt,k}\left(K X_{it,k} + 1\right)} \tag{24}$$

Digital Firms For US multinational digital firms, they decide how much to produce and what fraction of products sell to which markets. Their optimization problems are given by

$$\max_{\{\{L_{jt,k}\},\{x_{jt,k}\}\}} V(D_{j0}) = \sum_{t=1}^{\infty} \left(p_{jt,us} Y_{jt,us} + p_{jt,eu} Y_{jt,eu} - w_t (L_{jt,us} + L_{jt,eu})\right) \tag{25}$$

subject to

$$\begin{aligned} Y_{jt,us} =& (D_{j,t-1})^{\eta} (L_{jt,us})^{1-\eta} \\ Y_{jt,eu} =& (D_{j,t-1})^{\eta} (L_{jt,eu})^{1-\eta} \\ D_{jt} =& (1-\lambda) D_{j,t-1} + x_{jt,us} Y_{jt,us} + x_{jt,eu} Y_{jt,eu} \\ x_{jt,k} \in& [0,1] \end{aligned}$$

The HJB equation can be written as

$$V(D_{j,t-1}) = \max_{\{\{L_{jt,k}\},\{x_{jt,k}\}\}} (p_{jt,us} Y_{jt,us} + p_{jt,eu} Y_{jt,eu} - w_t (L_{jt,us} + L_{jt,eu})) + \frac{1}{1+r} V(D_{jt}) \quad (26)$$

The first order conditional w.r.t. $L_{jt,us}$

$$\underbrace{p_{jt,us}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,us})^{-\eta}}_{\text{current marginal product of capital}} + \underbrace{\frac{1}{1+r} V'(D_{jt}) \frac{\partial D_{jt}}{\partial L_{jt,us}}}_{\text{future value of data}} = w_t \quad (27)$$

where

$$\frac{\partial D_{jt}}{\partial L_{jt,us}} = x_{jt,us}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,us})^{-\eta} \quad (28)$$

Substitute the above expression into equation 27 and we can get

$$p_{jt,us}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,us})^{-\eta} + \frac{1}{1+r} V'(D_{jt}) x_{jt,us}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,us})^{-\eta} = w_t \quad (29)$$

Similarly, we can get the first order conditional w.r.t $L_{jt,eu}$

$$\underbrace{p_{jt,eu}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,eu})^{-\eta}}_{\text{current marginal product of capital}} + \underbrace{\frac{1}{1+r} V'(D_{jt}) \frac{\partial D_{jt}}{\partial L_{jt,eu}}}_{\text{future value of data}} = w_t \quad (30)$$

where

$$\frac{\partial D_{jt}}{\partial L_{jt,eu}} = x_{jt,eu}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,eu})^{-\eta} \quad (31)$$

Substitute the above expression into equation 30 and we can get

$$p_{jt,eu}(D_{j,t-1})^{\eta}(1-\eta)(L_{jt,eu})^{-\eta}+\frac{1}{1+r}V'(D_{jt})x_{jt,eu}(D_{j,t-1})^{\eta}(L_{jt,eu})^{-\eta}=w_t \tag{32}$$

On the balanced growth path, suppose the stock of data grows at the constant rate b_{digital}.

$$D_{jt}=(1+b_{\text{digital}})D_{j,t-1} \tag{33}$$

We guess and verify the value function on the balanced growth path. Suppose

$$V(D_{j,t-1})=B_{\text{digital}}\cdot(D_{j,t-1})^{\eta} \tag{34}$$

Then we can solve for $L_{jt,us}$ and $L_{jt,eu}$

$$(L_{jt,us})^{\eta}=\frac{(D_{j,t-1})^{\eta}\left(B_{\text{digital}}(1+b_{\text{digital}})^{\eta}x_{jt,us}+(1-\eta)(1+r)p_{jt,us}\right)}{(1+r)w_t} \tag{35}$$

and

$$(L_{jt,eu})^{\eta}=\frac{(D_{j,t-1})^{\eta}\left(B_{\text{digital}}(1+b_{\text{digital}})^{\eta}x_{jt,eu}+(1-\eta)(1+r)p_{jt,eu}\right)}{(1+r)w_t} \tag{36}$$

Take the first order derivative of the value function w.r.t. $D_{j,t-1}$. By the Envelope Theorem

$$V'(D_{j,t-1})=p_{jt,us}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,us}\right)^{1-\eta}+p_{jt,eu}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,eu}\right)^{1-\eta}+\frac{1}{1+r}V'(D_{j,t})\frac{\partial D_{j,t}}{\partial D_{j,t-1}} \tag{37}$$

where

$$\frac{\partial D_{j,t}}{\partial D_{j,t-1}}=1-\kappa+x_{jt,us}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,us}\right)^{1-\eta}+x_{jt,eu}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,eu}\right)^{1-\eta} \tag{38}$$

Therefore,

$$\frac{1}{1+r}V'(D_{jt})=\frac{V'(D_{j,t-1})-p_{jt,us}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,us}\right)^{1-\eta}-p_{jt,eu}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,eu}\right)^{1-\eta}}{1-\kappa+x_{jt,us}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,us}\right)^{1-\eta}+x_{jt,eu}\eta\left(D_{j,t-1}\right)^{\eta-1}\left(L_{jt,eu}\right)^{1-\eta}} \tag{39}$$

We also have

$$V'(D_{j,t-1}) = \eta B_{\text{digital}}(D_{j,t-1})^{\eta-1} \tag{40}$$

and

$$V'(D_{j,t}) = \eta B_{\text{digital}}(1+b_{\text{digital}})^{\eta-1}(D_{j,t-1})^{\eta-1} \tag{41}$$

I will solve for b_{digital} on the balanced growth path numerically in the calibration section.

Market Clearing The market clearing condition for digital goods in each region k is

$$\begin{aligned}(D_{j,t-1})^{\eta}\left(L_{jt,k}\right)^{1-\eta} &= \int_0^1 c_{ijt,k}di \\ &= \int_0^1 \frac{K\left(\xi\bar{x}_{jt}+x_{ijt,k}-\delta x_{ijt,k}^2\right)W_{it,k}}{p_{jt,k}\left(KX_{it,k}+1\right)}di \\ &= \frac{C_{jt,k}}{p_{jt,k}}\end{aligned} \tag{42}$$

where

$$C_{jt,k} = \int_0^1 \frac{K\left(\xi\bar{x}_{jt}+x_{ijt,k}-\delta x_{ijt,k}^2\right)W_{it,k}}{\left(KX_{it,k}+1\right)}di$$

$C_{jt,k}$ is the total expenditure on digital product j in region k. Since

$$(D_{j,t-1})^{\eta}\left(L_{jt,k}\right)^{1-\eta} = D_{j,t-1}\left(\frac{B_{\text{digital}}(1+b_{\text{digital}})^{\eta}x_{jt,us}+(1-\eta)(1+r)p_{jt,us}}{(1+r)w_t}\right)^{\frac{1-\eta}{\eta}} \tag{43}$$

we have

$$\frac{C_{jt,k}}{p_{jt,k}} = D_{j,t-1}\left(\frac{B_{\text{digital}}(1+b_{\text{digital}})^{\eta}x_{jt,k}+(1-\eta)(1+r)p_{jt,k}}{(1+r)w_t}\right)^{\frac{1-\eta}{\eta}} \tag{44}$$

To see how firms adjust their pricing based on the value of data they collect from consumers, we consider a hypothetical scenario where consumers can share data and personalize their user experience in the current period but firms cannot use the data for future production.

That is, $x_{ijt,k} > 0$ and $x_{jt,k} = 0$. The pricing when firms are barred from using consumers' data for production is

$$p^*_{jt,k} = \frac{(w_t)^{1-\eta} C^{\eta}_{jt,k}}{(1-\eta)^{1-\eta} \left(D^{us}_{j,t-1}\right)^{\eta}} \tag{45}$$

When data are allowed for future production, the pricing becomes

$$\begin{aligned} p_{jt,k} =& \frac{((1+r)w_t)^{1-\eta} C^{\eta}_{jt,k}}{(D_{j,t-1})^{\eta} \left(B_{\text{digital}}(1+b_{\text{digital}})^{\eta} x_{jt,k} + (1-\eta)(1+r)p_{jt,k}\right)^{1-\eta}} \\ =& \underbrace{\frac{1}{\left(B_{\text{digital}}(1+b_{\text{digital}})^{\eta} x_{jt,k} / \left((1-\eta)(1+r)p_{jt,k}\right) + 1\right)^{1-\eta}}}_{\text{price discount, paying for data}} \cdot p^*_{jt,k} \end{aligned} \tag{46}$$

Equation 46 shows us that firms "pay" for the data they collect from consumers. When firms' access to data in one region is limited, they will raise the price in that market to compensate for the loss in value that can be derived from data. Clarifying this point helps us understand firms' actions under the alternative regime, post-GDPR. One thing to notice is that we assume that firms cannot price discriminate at the individual level.

5.4.2 Post-GDPR

In the second scenario, GDPR is effective in the European Union. The major difference from the earlier regime is that EU households regain control of their own data and choose $x_{ijt,eu}$. As a result, digital firms only set the data sharing decisions for households from the US (the rest of the world), $x_{ijt,us} = 1$. While choosing their desired level of data sharing, EU households do not incorporate the positive externality they have on other households. Since each individual is atomistic, $\frac{\partial \bar{x}_{jt}}{\partial x_{ijt,k}} = 0$. Given other people's data sharing choice, EU households' optimal level of data sharing is

$$x^*_{ijt,eu} = \frac{1}{2\delta} \tag{47}$$

When $\delta > \frac{1}{2}$, $x^*_{ijt,eu} < 1$. This is reflected in one extra set of control variable $\{x_{ijt,eu}\}$ for EU households.

$$\max_{\{c_{ijt,eu}\},\{x_{ijt,eu}\},c'_{it,eu}} u_{it,eu} = K \int_0^1 \left(\xi \bar{x}_{jt} + x_{ijt,eu} - \delta x^2_{ijt,eu}\right) \ln c_{ijt,eu} dj + \ln c'_{it,eu} \tag{48}$$

subject to

$$W_{it,eu} \geq \int_0^1 p_{jt,eu} c_{ijt,eu} dj + c'_{it,eu}$$

For US households, their optimization problems remain the same.

$$\max_{\{c_{ijt,us}\},c'_{it,us}} u_{it,us} = K \int_0^1 \left(\xi \bar{x}_{jt} + x_{ijt,us} - \delta x^2_{ijt,us}\right) \ln c_{ijt,us} dj + \ln c'_{it,us} \tag{49}$$

subject to

$$W_{it,us} \geq \int_0^1 p_{jt,us} c_{ijt,us} dj + c'_{it,us}$$

For US multinational digital firms, they lose one control variable, $x_{jt,eu}$. Alternatively, we can view that EU consumers' decisions put an upper bound on the amount of data US firms can collect. Their optimization problem becomes

$$\max_{\{\{L_{jt,k}\},\{x_{jt,k}\}\}} V(D_{j0}) = \sum_{t=1}^{\infty} \left(p_{jt,us} Y_{jt,us} + p_{jt,eu} Y_{jt,eu} - w_t (L_{jt,us} + L_{jt,eu})\right) \tag{50}$$

subject to

$$\begin{aligned}
Y_{jt,us} &= (D_{j,t-1})^{\eta} (L_{jt,us})^{1-\eta} \\
Y_{jt,eu} &= (D_{j,t-1})^{\eta} (L_{jt,eu})^{1-\eta} \\
D_{jt} &= (1-\lambda) D_{j,t-1} + x_{jt,us} Y_{jt,us} + x_{jt,eu} Y_{jt,eu} \\
x_{jt,us} &\in [0,1], \quad x_{jt,eu} \in [0, 1/(2\delta)]
\end{aligned}$$

Following the same procedure as in the previous section, we can solve for equilibrium outcomes under the alternative post-GDPR regime. The setup of the model is in part inspired by the empirical section. There are two key findings in the empirical section. First, after GDPR came into effect, US multinational firms in the data-intensive category reduced their exposure to the European market. They experienced a 10% reduction in the revenue coming

from the European Union, but their total revenue did not change. Second, EU consumers saw a decline in their user experience on digital platforms. These empirical findings match the predictions of the model. We can use the quantitative findings to estimate the two key parameters in the model, the output elasticity of data η and privacy preferences of consumers δ. I will explain in the following section how I plan to estimate these two parameters.

5.5 Calibration and Welfare Analysis

In this subsection, I will outline the theoretical and empirical moments that can be used to estimate η and δ. In Section 3, I find that US data-intensive firms strategically reallocate their businesses across geographical segments, scaling back their EU operations by 10%. In Section 4, I show that, EU consumers face a less satisfactory consumer experience than their US counterparts after sharing less data with firms. This implies that consumers trade off the benefits of privacy protection and data-dependent user experiences. In this section, I perform a numerical analysis and calibrate the baseline model.

The targeted moments correspond to the main empirical findings from Section 3 and Section 4. These include the shifting in revenue from the EU to US after GDPR, and the decline in service quality for EU users after GDPR. There are two parameters from the model that need to be internally calibrated, including the output elasticity of data η, and the privacy preference δ. I match the two moment conditions mentioned above and jointly pin down these two parameters.

First, we define:

$$\psi_{jt} = \frac{Y_{jt,eu}}{Y_{jt,us} + Y_{jt,eu}} \tag{51}$$

Then the business reallocation post-GDPR is:

$$\frac{\psi_{j,\text{post}}}{\psi_{j,\text{pre}}} - 1 = -0.10 \tag{52}$$

The empirical moment on the right-hand side is the percent change in EU market share after

GDPR, as identified from Section 3. Second, we define service quality as:

$$s_{ijt,k} = \xi \bar{x}_{jt} + x_{ijt,k} \tag{53}$$

Here, I assume there is equal contribution from platform algorithm accuracy and personalization with individual data sharing. Then the change in service quality post-GDPR for EU consumers is:

$$\frac{s_{ij,eu,post}}{s_{ij,eu,pre}} - 1 = -0.06 \tag{54}$$

The empirical moment on the right-hand side is the percent change in user satisfaction, as identified from Section 4. Below are internally calibrated and externally chosen parameters for this numerical analysis.

[Insert Table 11 Here.]

In the baseline model, I assume that households are endowed with exogenous income, and the labor supply is perfectly elastic. I calibrate the model under the two regimes as specified in Section 5.4.1 and Section 5.4.2. I compute the changes in product prices in each region, the change in data sharing among EU users, and the change in data growth rate for digital firms.

[Insert Table 12 Here.]

where Δp_{eu} is the percent change in product price in the EU, and Δp_{us} is the percent change in product price in the US, and Δx_{eu} is the percent change in the fraction of date shared by European users, and $\Delta b_{\text{digital}}$ is the change in data growth rate for digital platforms. Below are the results from preliminary welfare analysis. Welfare is measure as the consumers' total utility from consumption in each geographical region. That is, I focus on consumer surplus.

[Insert Table 13 Here.]

We can see that GDPR leads to welfare loss for both US and EU consumers. The welfare loss for EU consumers comes from both price adjustment of the digital platforms and the loss from the social value of data. The welfare loss for US consumers mainly comes from the loss in social value of data even though there is a gain from the price decrease in the US market. The above table shows us the contribution of price impact on welfare by shutting down price impact in column (1). Column (2) shows us the price impact, and column (3) shows us the net change in consumer surplus. For example, -1.9% - 7.6% = - 9.5%.

[Insert Table 14 Here.]

The second table shows us the contribution of the externality of data sharing. Since individual consumer fail to internalize the positive externality they have on other consumers. This friction leads to suboptimal outcomes when EU households are given the choice to determine how much data to share with firms.

6 Conclusion

Because of the data feedback loop, the paper proposes that we should jointly evaluate the value of data for firms and the value of privacy for consumers. The paper first provides a detailed analysis of the impact of GDPR on US data-intensive firms and their customers. I find that US multinational firms acknowledge the negative impact of GDPR and strategically reallocate their businesses across geographical segments, scaling back their EU operations by 10% while keeping the total business size unchanged. Smaller firms struggle to respond and experience a bigger and more persistent effect. Firms also respond by hiring more AI-related talents. Firms do not price in the reduced access to data by increasing product markups but by providing lower-quality service and monetizing consumers by putting in more advertisements. By sharing less data with firms, EU consumers face a less satisfactory consumer experience than their US counterparts, implying consumers trade off the benefits of privacy protection and data-dependent user experiences.

The paper then provides a tractable estimation framework that combines the value of data and privacy in an equilibrium model and speaks to the welfare impact of a regional

privacy regulation like GDPR. Even though GDPR only applies to EU residents, all firms with business operations in the EU must comply. Through the business operations and adjustments of US multinational firms, GDPR will also have complicated implications for the welfare of US consumers. The project contributes to the broader discussion on the data economy and privacy regulations from an international perspective. Several US states follow the EU's footsteps and have passed similar privacy regulation frameworks, including California, Colorado, Connecticut, Utah, and Virginia. More US states are considering such regulations, and a new iteration of federal-level privacy regulation, the American Data Privacy and Protection Act, is on the agenda. Understanding the potential impact of data privacy regulations on the delicate dynamics between firms and consumers is essential.

References

Abis, S. and Veldkamp, L. (2020). The changing economics of knowledge production. *Available at SSRN 3570130*.

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2020). Ai and jobs: Evidence from online vacancies. Technical report, National Bureau of Economic Research.

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2019). Too much data: Prices and inefficiencies in data markets.

Admati, A. R. and Pfleiderer, P. (1990). Direct and indirect sale of information. *Econometrica: Journal of the Econometric Society*, pages 901–928.

Aridor, G., Che, Y.-K., and Salz, T. (2020). The economic consequences of data privacy regulation: Empirical evidence from gdpr. Technical report, National Bureau of Economic Research.

Babina, T., Fedyk, A., He, A. X., and Hodson, J. (2020). Artificial intelligence, firm growth, and industry concentration. *Firm Growth, and Industry Concentration (November*, 22:2020.

Benkler, Y., Faris, R., and Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bergemann, D., Bonatti, A., and Gan, T. (2019). The economics of social data.

Bian, B., Ma, X., and Tang, H. (2021). The supply and demand for data privacy: Evidence from mobile apps. *Available at SSRN 3987541*.

Bleier, A., Goldfarb, A., and Tuckerc, C. (2020). Consumer privacy and the future of data-based innovation and marketing. *International Journal of Research in Marketing*.

Campbell, J. L., Chen, H., Dhaliwal, D. S., Lu, H.-m., and Steele, L. B. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1):396–455.

Canayaz, M., Kantorovitch, I., and Mihet, R. (2022). Consumer privacy and value of consumer data. *Swiss Finance Institute Research Paper*, (22-68).

Cao, S., Jiang, W., Wang, J. L., and Yang, B. (2021). From man vs. machine to man+ machine: The art and ai of stock analyses. Technical report, National Bureau of Economic Research.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2020). How to talk when a machine is listening: Corporate disclosure in the age of ai. Technical report, National Bureau of Economic Research.

Carnevale, A. P., Jayasundera, T., and Repnikov, D. (2014). Understanding online job ads data. *Georgetown University, Center on Education and the Workforce, Technical Report (April)*.

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

Cong, L. W., Xie, D., and Zhang, L. (2020). Knowledge accumulation, privacy, and growth in a data economy. *Privacy, and Growth in a Data Economy (October 8, 2020)*.

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of economic perspectives*, 23(3):37–60.

Fajgelbaum, P. D., Schaal, E., and Taschereau-Dumouchel, M. (2017). Uncertainty traps. *The Quarterly Journal of Economics*, 132(4):1641–1692.

Farboodi, M. and Veldkamp, L. (2021). A model of the data economy. Technical report, National Bureau of Economic Research.

Goldberg, S., Johnson, G., and Shriver, S. (2019). Regulating privacy online: An economic evaluation of the gdpr. *Available at SSRN 3421731*.

Goldfarb, A. and Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404.

Jia, J., Jin, G. Z., and Wagman, L. (2018). The short-run effects of gdpr on technology venture investment. Technical report, National Bureau of Economic Research.

Jia, J., Jin, G. Z., and Wagman, L. (2020). Gdpr and the localness of venture investment. *Available at SSRN 3436535*.

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

Jones, C. I. and Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–58.

Kogan, L., Papanikolaou, D., Seru, A., and Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, 132(2):665–712.

Lenard, T. M. and Rubin, P. H. (2013). The big data revolution: Privacy considerations. *Technology Policy Institute*, pages 1–2.

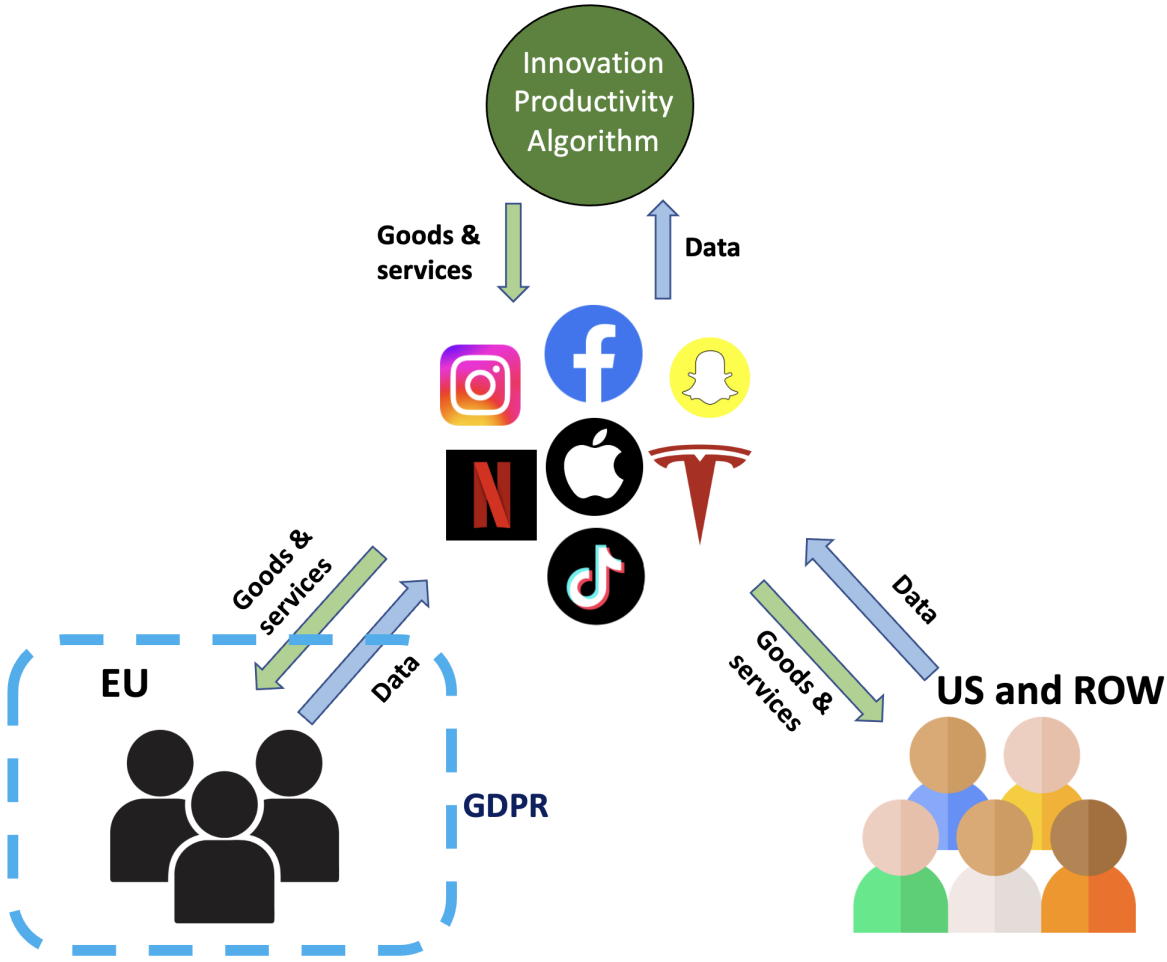
Martin, N., Matt, C., Niebel, C., and Blind, K. (2019). How data protection regulation affects startup innovation. *Information systems frontiers*, 21(6):1307–1324.

Ordonez, G. (2013). The asymmetric effects of financial frictions. *Journal of Political Economy*, 121(5):844–895.

Tang, H. (2019). The value of privacy: Evidence from online borrowers. *Available at SSRN*.

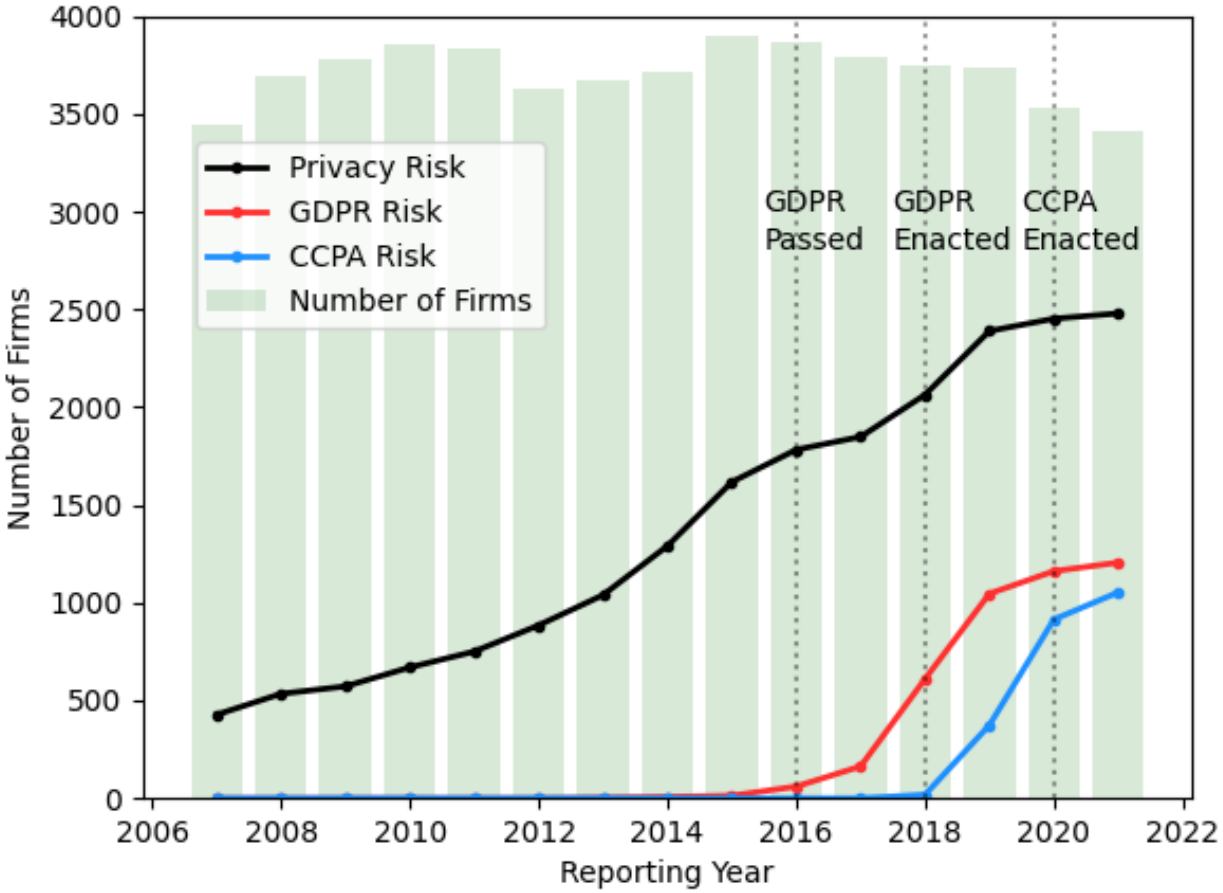
Veldkamp, L. L. (2005). Slow boom, sudden crash. *Journal of Economic theory*, 124(2):230–257.

Figure 1: Data Feedback Loop in a Two-Economy Setting



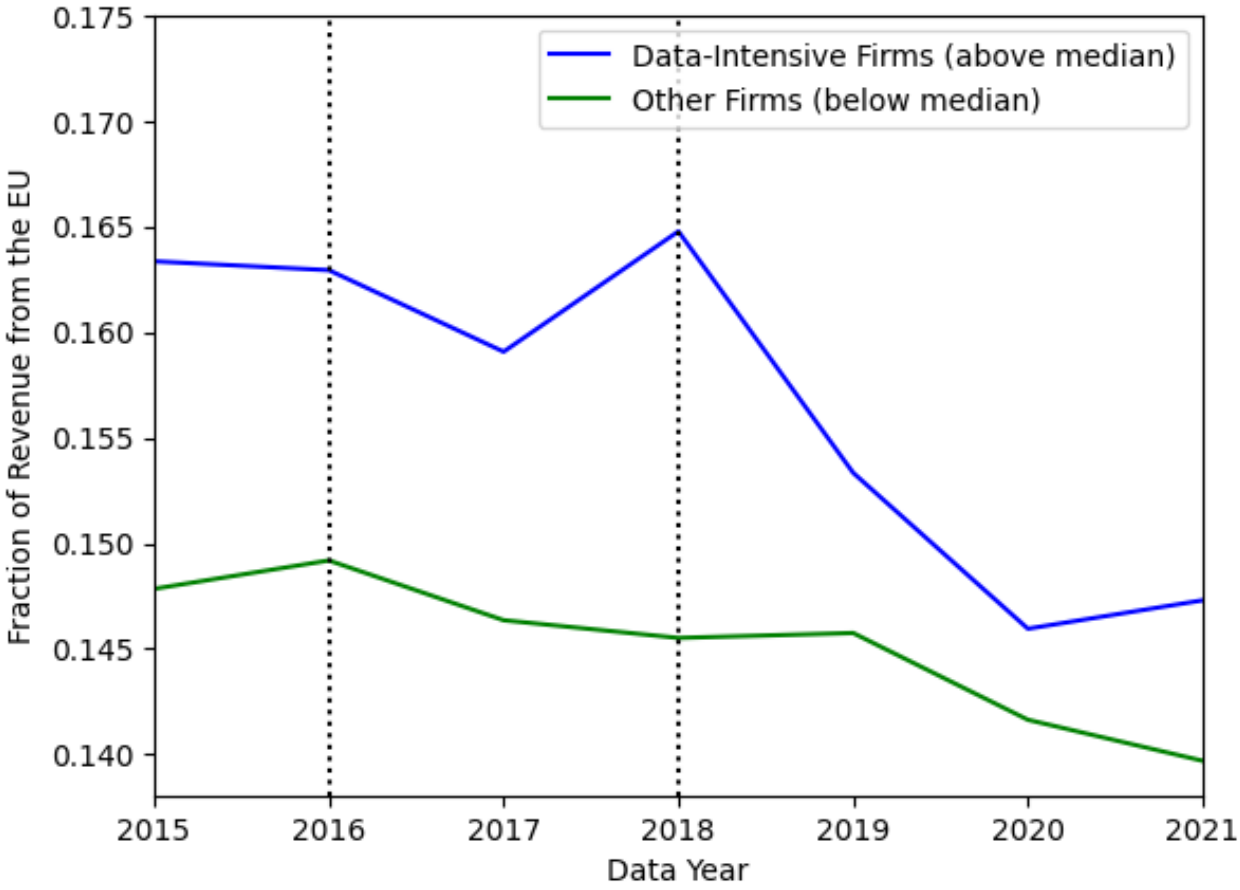
Notes: US firms provide goods and services to both European customers and US customers (or, more accurately, customers from the rest of the world). Data is a byproduct of economic activities, and firms collect and analyze consumers' data to learn about their preferences, inspire new concepts or products, and boost productivity. While consumers enjoy the advantages of personalized recommendations and enhanced service quality, they have concerns about sharing personal data with firms. Privacy concerns may stem from the psychological costs or social stigma of disclosing excessive personal information, as well as from predatory advertising or pricing tactics employed by firms.

Figure 2: The Number of US Public Firms Disclosing Privacy-Related Risk Factors in 10-K Filings



Notes: The light green bar shows the number of US public firms with valid risk factor disclosures (Item 1A) in their annual 10-K filings. The number of US public firms is around 3500-4000 from 2006 to 2021, so the sample covers most of the US public firms. The black line shows the number of US public firms that disclose any privacy related risk; the red line shows the number of US public firms that disclose GDPR related risk; and the blue line shows the number of US public firms that disclose CCPA related risk.

Figure 3: Fraction of Revenue from the EU



Notes: The figure illustrates the proportion of revenue generated by US firms from the European market, with particular emphasis on firms that had substantial European market operations prior to 2018. The sample is divided into two distinct groups based on their data intensiveness: the more data-intensive group (above median) and the less data-intensive group (below median). The measure of data intensiveness, as detailed in Section 2.2.1, serves as a crucial factor in assessing the potential impact of regulatory changes on firm behavior.

Figure 4: The Dynamics of Cross-Market Adjustment

Notes: I extend the regression in equation 5 to a dynamic difference-in-difference setting so that I can check for the pre-trend and examine when the effect of GDPR kicks in. I run the following regression.

$$Y_{i,t} = \alpha_t + \phi_i + \sum_{\tau \neq 2018} \beta_\tau \cdot \boldsymbol{I}(t = \tau) \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t}$$

α_t is the year fixed-effect, ϕ_i is the firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. $\boldsymbol{I}(t = \tau)$ is a binary variable that equals one if year $t = \tau$. Data-Intensive$_i$ is a binary variable that equals one if firm i is in the data-intensive category. I focus on the firms with significant European market operations before 2018 and divide the sample into the more data-intensive group (above median) and the less data-intensive group (below median). I define data intensiveness in section 2.2.1. The standard errors are clustered at the industry level.

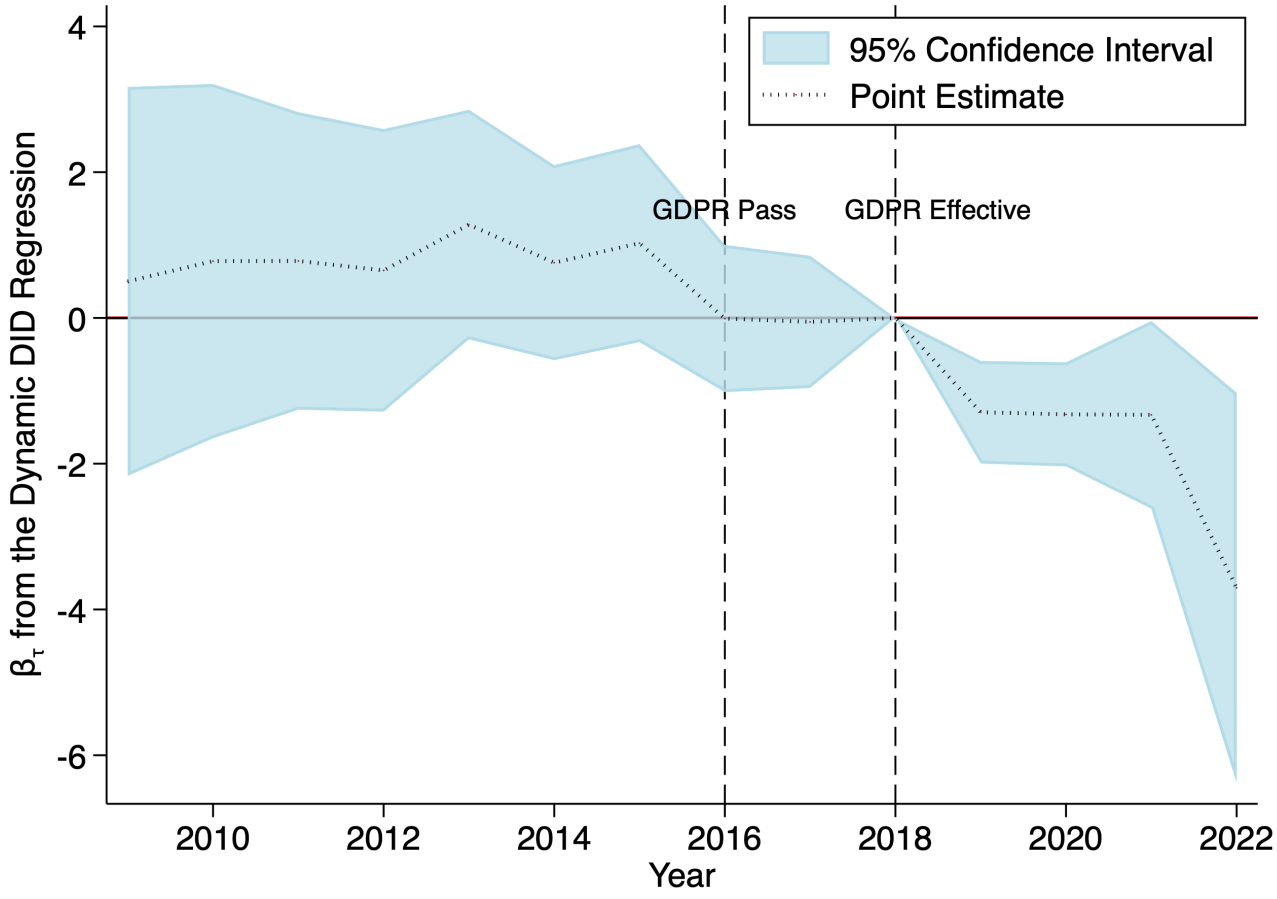


Figure 5: Dynamic Effects for Small Firms

Notes: Similar to what we did section 3.1, I extend the regression in equation 5 to a dynamic setting and study the dynamics of small firms. Small firms are defined as the ones with market capitalization below sample median.

$$Y_{i,t} = \alpha_t + \phi_i + \sum_{\tau \neq 2018} \beta_\tau \cdot \boldsymbol{I}(t = \tau) \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t}$$

α_t is the year fixed-effect, ϕ_i is the firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. $\boldsymbol{I}(t = \tau)$ is a binary variable that equals one if year $t = \tau$. Data-Intensive$_i$ is a binary variable that equals one if firm i is in the data-intensive category. I focus on the firms with significant European market operations before 2018 and divide the sample into the more data-intensive group (above median) and the less data-intensive group (below median). I define data intensiveness in section 2.2.1.

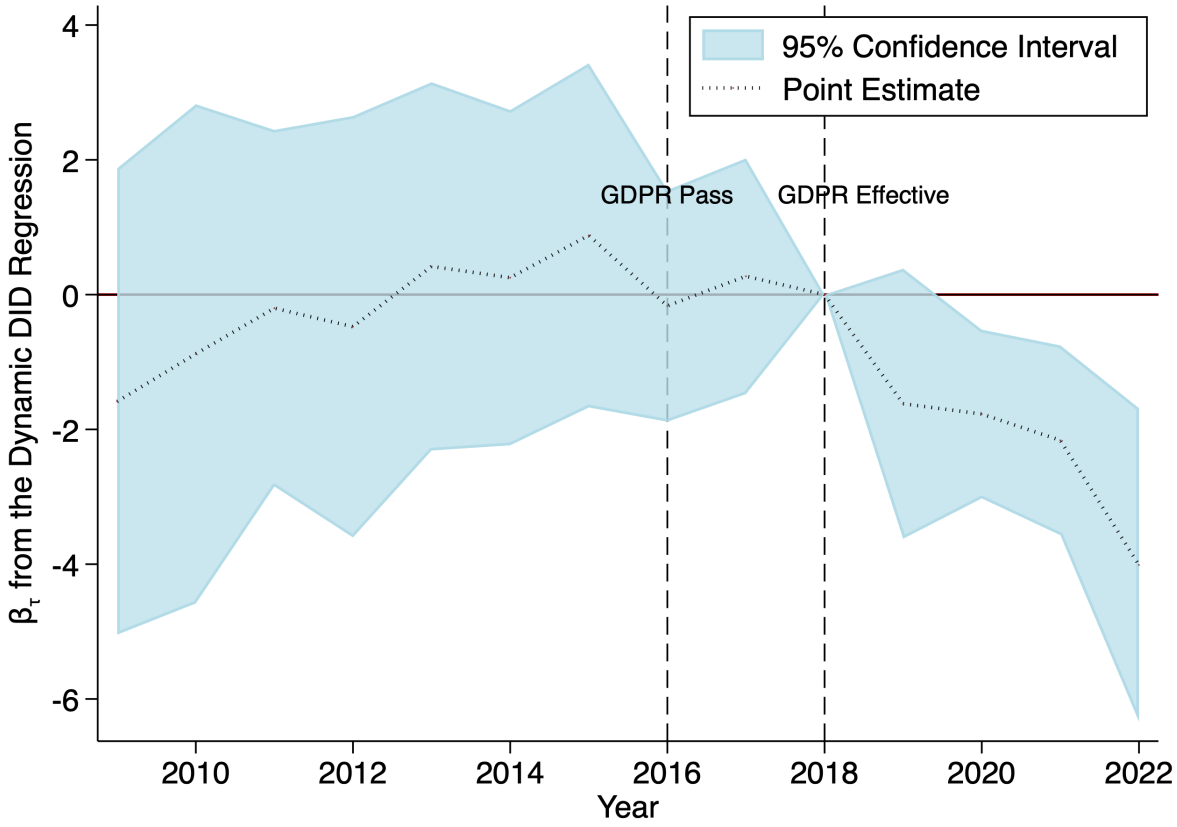


Figure 6: Dynamic Effects for Large Firms

Notes: Similar to what we did section 3.1, I extend the regression in equation 5 to a dynamic setting and study the dynamics of large firms. Large firms are defined as the ones with market capitalization above sample median.

$$Y_{i,t} = \alpha_t + \phi_i + \sum_{\tau \neq 2018} \beta_\tau \cdot \boldsymbol{I}(t = \tau) \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t}$$

α_t is the year fixed-effect, ϕ_i is the firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. $\boldsymbol{I}(t = \tau)$ is a binary variable that equals one if year $t = \tau$. Data-Intensive$_i$ is a binary variable that equals one if firm i is in the data-intensive category. I focus on the firms with significant European market operations before 2018 and divide the sample into the more data-intensive group (above median) and the less data-intensive group (below median). I define data intensiveness in section 2.2.1.

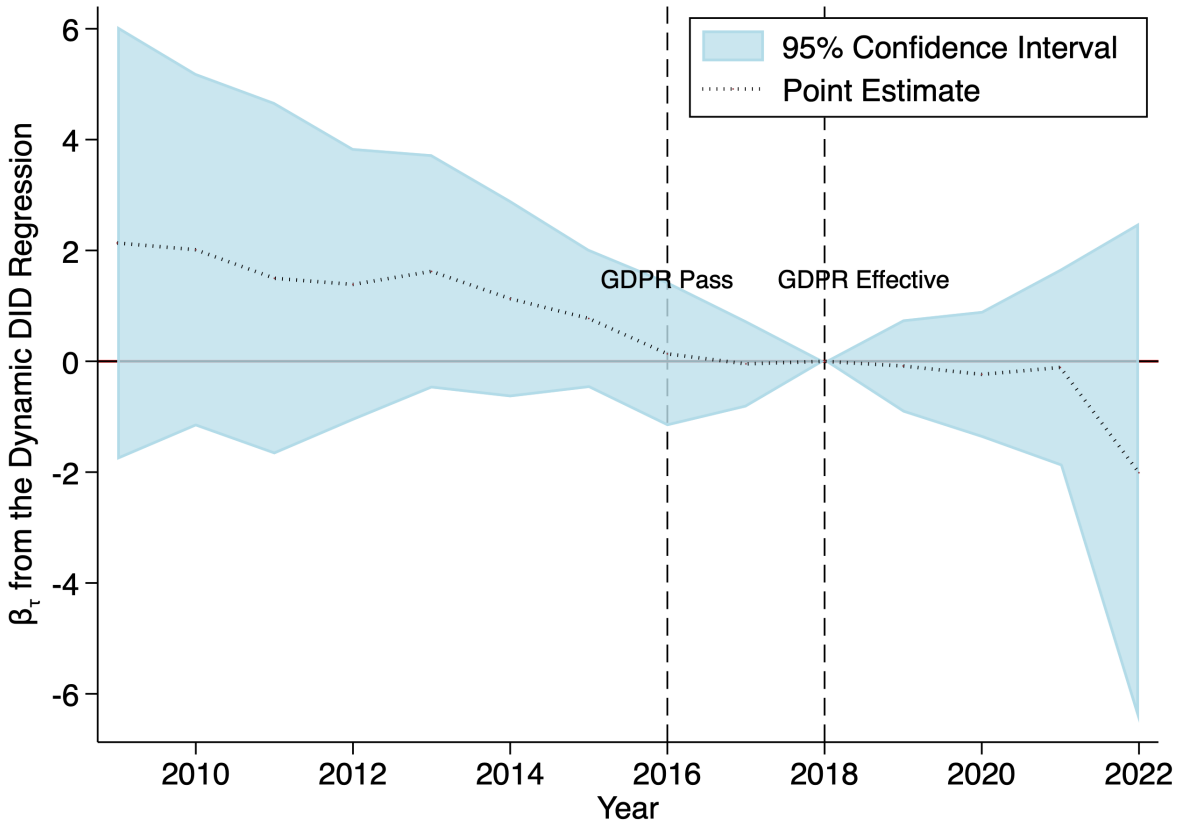


Table 1: Data Intensiveness by Industries

NAICS	Name	# Firms	Data-Intensiveness	% Data-Intensive
51	Information	521	1.21	0.84
54	Professional, Scientific, and Technical Services	180	0.78	0.75
33	Motor, Semiconductor, and Equipment Manufacturing	956	0.69	0.60
45	General Merchandise, Personal Care, and Clothin...	72	0.45	0.43
56	Administrative and Support and Waste Management...	82	0.38	0.55
52	Finance and Insurance	681	0.36	0.42
62	Health Care and Social Assistance	98	0.28	0.49
42	Wholesale Trade	116	0.25	0.38
32	Wood, Chemical, Materials Manufacturing	693	0.23	0.47
61	Educational Services	21	0.22	0.62
53	Real Estate and Rental and Leasing	233	0.22	0.25
48	Transportation and Warehousing	89	0.21	0.38
71	Arts, Entertainment, and Recreation	27	0.18	0.26
21	Mining, Quarrying, and Oil and Gas Extraction	185	0.15	0.34
44	Equipment and Grocery Retail Trade	126	0.15	0.16
31	Food, Tobacco, and Textile Manufacturing	136	0.13	0.21
23	Construction	59	0.10	0.15
72	Accommodation and Food Services	90	0.05	0.08

Table 2: Cross-Market Business Adjustment

Notes: I employ a difference-in-differences identification strategy to examine how US multinational firms respond to restricted access to EU consumers' data due to the enactment of GDPR in May 2018. Since most US firms report their geographical revenue compositions at an annual frequency, the observations of the sample used in this table are at the firm-year level. I run the following regression.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_{\text{data}} \cdot \text{GDPR-Effective}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t}$$

α_t is the year fixed-effect, ϕ_i is the firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. GDPR-Effective$_t$ is a binary variable that equals one if time t is after GDPR's enactment date, May 2018. Data-Intensive$_i$ is a binary variable that equals one if firm i is in the data-intensive category. I also examine one specification where I include the continuous measure of data-intensiveness. I define data intensiveness in section 2.2.1. For the dependent variable, $Y_{i,t}$, I first look at the fraction of revenue generated from the European market by US firms in column (1) and (2). In columns (3) and (4), I look at total sales and EU sales scaled by total assets. The standard errors are clustered at the industry level, and t-stats are reported in parentheses.

Dependent Variable:	EU Sales Percentage		EU Sales/Assets	Sales/Assets
	(1)	(2)	(3)	(4)
GDPR Effective × Data-Intensive (dummy)	-1.548**		-2.453**	-1.220
	(-2.370)		(-2.378)	(-0.417)
GDPR Pass × Data-Intensive (dummy)	-0.713			
	(-1.208)			
GDPR Effective × Data-Intensive (value)		-1.086**		
		(-2.138)		
Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R^2	0.765	0.765	0.793	0.846
obs	8,648	8,648	8,655	9,122

Table 3: Cross-Market Business Adjustment (Large versus Small Firms)

Notes: I employ a difference-in-differences identification strategy and study how US multinational firms respond when their access to EU consumers' data is restricted. I group firms into large firms (above median market capitalization) and small firms (below median market capitalization) and look at their effects separately. Since most US firms report their geographical revenue compositions at an annual frequency, the observations of the sample used in this table are at the firm-year level. I run the following regression.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR-Effective}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t}$$

α_t is the year fixed-effect, ϕ_i is the firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. GDPR-Effective$_t$ is a binary variable that equals one if time t is after GDPR's enactment date, May 2018. Data-Intensive$_i$ is a binary variable that equals one if firm i is in the data-intensive category. I focus on the firms with significant European market operations before 2018 and divide the sample into the more data-intensive group (above median) and the less data-intensive group (below median). I define data intensiveness in section 2.2.1. In all specifications, I control for firm-level time-varying characteristics, including book-to-market ratio, leverage, and size. The standard errors are clustered at the industry level, and t-stats are reported in parentheses.

	Small Firms		Large Firms	
	EU Sale Percentage	Sales/Asset	EU Sale Percentage	Sales/Asset
	(1)	(2)	(3)	(4)
GDPR Effective × Data-Intensive	-1.891**	-0.667	-1.041	-1.128
	(-2.086)	(-0.170)	(-1.190)	(-0.360)
Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R^2	0.754	0.767	0.775	0.913
obs	3,822	4,121	4,826.000	5,001

Table 4: Hiring Demand Change in Data Management Skills

Notes: I employ a difference-in-differences identification strategy and study how US multinational firms change their demand for data-management skills when their access to EU consumers' data is expected to be restricted. The observations are at the firm-year level. I have defined data intensiveness in section 2.2.1.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR-Pass}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \tag{55}$$

where $Y_{i,t}$ is the fraction of job postings that require data management related skills, α_t is year fixed-effect, ϕ_i is firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and size. GDPR-Pass$_t$ is a binary variable that equals one if time t is after GDPR's passage year, 2016. In all specifications, I control for firm-level time-varying characteristics, including book-to-market ratio, leverage, and firm size. The standard errors are clustered at the industry level, and t-stats are reported in parentheses.

Dependent Variable:	Full	Small	Large
Data Management Skill Percentage	(1)	(2)	(3)
GDPR Pass × Data-Intensive	-4.926***	-7.750***	-3.082***
	(-7.274)	(-4.802)	(-3.587)
GDPR Effective × Data-Intensive	-4.569***	-6.953***	-3.056***
	(-6.214)	(-4.792)	(-3.043)
Book to Market	-3.335**	-5.470***	1.414
	(-2.509)	(-2.718)	(0.932)
Leverage	4.695**	6.723***	-0.821
	(2.646)	(2.849)	(-0.382)
Market Valuation ($ bn)	-0.001	-0.058	-0.002
	(-0.210)	(-0.086)	(-0.307)
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.505	0.444	0.603
obs	7,726	3,262	4,464

Table 5: Hiring Demand Change in AI Skills

Notes: I employ a difference-in-differences identification strategy and study how US multinational firms change their demand for AI-related skills when their access to EU consumers' data is expected to be restricted. The observations are at the firm-year level. I have defined data-intensiveness in section 2.2.1.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR-Pass}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \tag{56}$$

where $Y_{i,t}$ is the fraction of job postings that require AI-related skills, α_t is year fixed-effect, ϕ_i is firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and size. GDPR-Pass$_t$ is a binary variable that equals one if time t is after GDPR's passage year, 2016. In all specifications, I control for firm-level time-varying characteristics, including book-to-market ratio, leverage, and firm size. The standard errors are clustered at the industry level, and t-stats are reported in parentheses.

Dependent Variable:	Full	Small	Large
AI Skill Percentage	(1)	(2)	(3)
GDPR Pass × Data-Intensive	0.942***	0.378	1.282***
	(2.867)	(0.520)	(4.129)
GDPR Effective × Data-Intensive	1.056***	0.581	1.312***
	(2.991)	(0.755)	(4.001)
Book to Market	-0.744***	-1.099***	-0.444
	(-3.909)	(-3.392)	(-1.131)
Leverage	0.086	-0.113	0.633
	(0.289)	(-0.208)	(0.750)
Market Valuation ($ bn)	0.008	0.126	0.007
	(1.470)	(0.342)	(1.411)
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.568	0.524	0.642
obs	7,726	3,262	4,464

Table 6: Reviews Summary Statistics

Notes: I show the summary statistics of the review data below, including average daily rating, annual total reviews with advertisement complaints, annual totals reviews that mention in-app purchases or subscriptions.

	Daily Average Score		Total Annual Ads Complaints		Total Annual Purchase Comments	
	EU	US	EU	US	EU	US
count	6,457,975	7,373,199	36,635	41,081	36,635	41,081
mean	4.04	3.91	39.05	20.34	46.20	34.74
std	1.08	1.19	389.93	181.72	260.74	177.32
min	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.67	3.33	0.00	0.00	0.00	0.00
50%	4.40	4.30	1.00	1.00	1.00	2.00
75%	5.00	5.00	7.00	5.00	10.00	14.00
max	5.00	5.00	34,723	25,527	11,774	16,571

Table 7: Daily Average Rating

Notes: I employ a difference-in-differences identification strategy and study how limited access to data in the European market affects the quality of service provided by mobile apps, measured by the daily average user numeric ratings. The observations of the sample used in this table are at the app-day level. In columns (1) and (2), I run the following regression.

$$Y_{i,m,t} = \alpha_m + \phi_i + \beta_{\text{service}} \cdot \text{GDPR}_m \times \text{Personalization Collected}_i + \boldsymbol{\gamma X}_{i,m,t} + \varepsilon_{i,m,t}$$

where $Y_{i,m,t}$ is the daily average rating for app i on day t. α_m is the year-month fixed-effect, ϕ_i is the app fixed-effect. GDPR_m is a binary variable that equals one if time t is after GDPR's enactment month, May 2018. $\text{Personalization Collected}_i$ is a binary variable that equals one if app i collects user data for personalization purposes. In column (3), I run a triple difference regression.

$$\begin{aligned} Y_{i,m,k,t} =& \alpha_m + \phi_i + \psi_k + \beta_1^* \cdot \text{GDPR}_m \times \text{Personalization Collected}_i \times \text{EU}_k \\ &+ \beta_2 \cdot \text{GDPR}_m \times \text{Personalization Collected}_i + \beta_3 \cdot \text{GDPR}_m \times \text{EU}_k \\ &+ \beta_4 \cdot \text{Personalization Collected}_i \times \text{EU}_k + \boldsymbol{\gamma X}_{i,m,k,t} + \varepsilon_{i,m,k,t} \end{aligned}$$

where $Y_{i,m,k,t}$ is the average daily rating by users from region k for app i on day t. ψ_k is the region (US or EU) fixed-effect. EU_k is an indicator variable that equals one if the reviews come from the EU users. The coefficient β_1^* before the triple interaction term captures differential change in user quality between the EU and US users after GDPR for apps collecting personalization information.

Dependent Variable:	EU Users	US Users	All
Daily Average Rating	(1)	(2)	(3)
GDPR Effective × Personalization Collected	-0.081***	-0.016	-0.005
	(-4.201)	(-0.816)	(-0.274)
GDPR Effective × Personalization Collected × EU			-0.072***
			(-4.325)
Total Daily Review #	0.002***	0.004***	0.002***
	(13.762)	(18.955)	(17.193)
GDPR Effective × EU			0.045***
			(4.273)
Personalization Collected × EU			0.072***
			(4.838)
Year Month FE	Yes	Yes	Yes
App FE	Yes	Yes	Yes
Region FE	No	No	Yes
R^2	0.243	0.239	0.233
obs	5,757,960	6,548,836	12,306,796

Table 8: Annual Advertisement Complaints

Notes: I employ a difference-in-differences identification strategy and study how limited access to data in the European market affects the number of advertisements complaints by mobile app users. The observations in this analysis are at the app-year level. In columns (1) and (2), I run the following regression:

$$\ln Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR}_t \times \text{Target Advertising}_i + \varepsilon_{i,t}$$

where $Y_{i,t}$ is the total number of advertisement related complaints for app i in year t. I take the logarithm of the annual number of complaints to the natural base. α_t is the year fixed effect. ϕ_i is the app fixed effect. GDPR_t is a binary variable that equals one if time t is after GDPR's enactment year, 2018. $\text{Target Advertising}_i$ is a binary variable that equals one if app i collects user data for targeted advertising purposes. I analyze reviews left by the EU and US users separately. In column (3), I run a triple difference regression:

$$\begin{aligned}\ln Y_{i,k,t} =& \alpha_t + \phi_i + \psi_k + \beta_1^* \cdot \text{GDPR}_t \times \text{Target Advertising}_i \times \text{EU}_k \\ &+ \beta_2 \cdot \text{GDPR}_t \times \text{Target Advertising}_i + \beta_3 \cdot \text{GDPR}_t \times \text{EU}_k \\ &+ \beta_4 \cdot \text{Target Advertising}_i \times \text{EU}_k + \varepsilon_{i,k,t}\end{aligned}$$

where $Y_{i,k,t}$ is the total number of advertisement related complaints for app i in year t. ψ_k is the region (US or EU) fixed effect. EU_k is an indicator variable that equals one if the reviews come from the EU users. The coefficient β_1^* before the triple interaction term captures the differential change in advertisement intensity between the EU and US mobile app markets.

Dependent Variable:	EU Users	US Users	All
ln(Annual # of Advertising Complaints)	(1)	(2)	(3)
GDPR Effective × Target Advertising × EU			0.098***
			(3.308)
GDPR Effective × Target Advertising	0.259***	0.189***	0.179***
	(7.942)	(7.302)	(6.719)
GDPR Effective × EU			-0.016
			(-0.923)
Target Advertising × EU			-0.206***
			(-4.950)
Year FE	Yes	Yes	Yes
App FE	Yes	Yes	Yes
Region FE	No	No	Yes
R^2	0.804	0.800	0.703
obs	33,328	37,247	70,575

Table 9: Annual Purchase and Subscription Related Comments

Notes: I employ a difference-in-differences identification strategy and study how limited access to data in the European market affects the number of comments related to purchase and subscriptions. The observations in this analysis are at the app-year level. In columns (1) and (2), I run the following regression:

$$\ln Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR}_t \times \text{Target Advertising}_i + \varepsilon_{i,t}$$

where $Y_{i,t}$ is the total number of purchase related comments for app i in year t. α_t is the year fixed effect, ϕ_i is the app fixed effect. GDPR_t is a binary variable that equals one if time t is after GDPR's enactment year, 2018. $\text{Target Advertising}_i$ is a binary variable that equals one if app i collects user data for targeted advertising purposes. I analyze the reviews left by the EU and US users separately. In column (3), I run a triple difference regression:

$$\begin{aligned}\ln Y_{i,k,t} =& \alpha_t + \phi_i + \psi_k + \beta_1^* \cdot \text{GDPR}_t \times \text{Target Advertising}_i \times \text{EU}_k \\ &+ \beta_2 \cdot \text{GDPR}_t \times \text{Target Advertising}_i + \beta_3 \cdot \text{GDPR}_t \times \text{EU}_k \\ &+ \beta_4 \cdot \text{Target Advertising}_i \times \text{EU}_k + \varepsilon_{i,k,t}\end{aligned}$$

where $Y_{i,k,t}$ is the total number of purchase related comments by users in region k for app i in year t. ψ_k is the region (US or EU) fixed effect. EU_k is an indicator variable that equals one if the reviews come from the EU users. The coefficient β_1^* before the triple interaction term captures the differential change in the prevalence of paid services and subscriptions between the EU and US mobile app markets.

Dependent Variable:	EU Users	US Users	All
ln(Annual # of Purchase Comments)	(1)	(2)	(3)
GDPR Effective × Target Advertising × EU			0.090**
			(2.383)
GDPR Effective × Target Advertising	0.205***	0.117***	0.115***
	(5.318)	(3.834)	(3.645)
GDPR Effective × EU			0.030
			(1.221)
Target Advertising × EU			-0.200***
			(-3.667)
Year FE	Yes	Yes	Yes
App FE	Yes	Yes	Yes
Region FE	No	No	Yes
R^2	0.791	0.813	0.652
obs	33,328	37,247	70,575

Table 10: Total Annual Reviews

Notes: I employ a difference-in-differences identification strategy and study how limited access to data in the European market affects the total number of mobile app reviews. The observations in this analysis are at the app-year level. In columns (1) and (2), I run the following regression:

$$\ln Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR}_t \times \text{Target Advertising}_i + \varepsilon_{i,t}$$

$Y_{i,t}$ is the total number of reviews for app i in year t. α_t is the year fixed effect, ϕ_i is the app fixed effect. GDPR_t is a binary variable that equals one if time t is after GDPR's enactment year, 2018. $\text{Target Advertising}_i$ is a binary variable that equals one if app i collects user data for targeted advertising purposes. I analyze the reviews left by the EU and US users separately. In column (3), I run a triple difference regression:

$$\begin{aligned}\ln Y_{i,k,t} =& \alpha_t + \phi_i + \psi_k + \beta_1^* \cdot \text{GDPR}_t \times \text{Target Advertising}_i \times \text{EU}_k \\ &+ \beta_2 \cdot \text{GDPR}_t \times \text{Target Advertising}_i + \beta_3 \cdot \text{GDPR}_t \times \text{EU}_k \\ &+ \beta_4 \cdot \text{Target Advertising}_i \times \text{EU}_k + \varepsilon_{i,k,t}\end{aligned}$$

where $Y_{i,k,t}$ is the total number of reviews by users in region k for app i in year t. EU_i is an indicator variable that equals one if the reviews come from the EU users. The coefficient β_1^* before the triple interaction term captures the differential change in total reviews between the EU and US mobile app markets.

Dependent Variable:	EU Users	US Users	All
ln(Annual # of Reviews)	(1)	(2)	(3)
GDPR Effective × Target Advertising × EU			0.067
			(1.563)
GDPR Effective × Target Advertising	0.033	-0.014	-0.025
	(0.643)	(-0.327)	(-0.569)
GDPR Effective × EU			0.209***
			(6.942)
Target Advertising × EU			-0.047
			(-0.632)
Year FE	Yes	Yes	Yes
App FE	Yes	Yes	Yes
Region FE	No	No	Yes
R^2	0.827	0.808	0.694
obs	33,328	37,247	70,575

Table 11: Calibration Parameters

Parameter	Value	Notes
η	0.137	Output Elasticity of Data
δ	0.547	Privacy Preferences
Π_{us}	7	US Household Income Level
Π_{eu}	5	EU Household Income Level
K	2	Relative Importance of Digital Consumption
w_{us}	0.4	Labor Cost
r	0.05	Discount Rate
κ	0.2	Data Depreciation Rate
D_0	0.6	Initial Data stock

Table 12: Equilibrium Outcome Change

Δp_{eu}	Δp_{us}	Δx_{eu}	$\Delta b_{\text{digital}}$
+ 17.2%	-2.0%	-8.6%	-9.0%

Table 13: Consumer Surplus - Price Impact

	w/o price impact	w. price impact	net change
	(1)	(2)	(3)
EU consumers	-1.9%	-7.6%	-9.5%
US consumers	-2.1%	+0.8%	-1.3%

Table 14: Consumer Surplus - Data Sharing Externality

	w/o externality	w. externality	net change
	(1)	(2)	(3)
EU consumers	-5.7%	-3.8%	-9.5%
US consumers	0.6%	-1.9%	-1.3%

A Appendix

A.1 Additional Tables and Figures

Table A1: Cross-Market Business Adjustment with Tech Controls

I employ a difference-in-differences identification strategy and study how US multinational firms respond when their access to EU consumers' data is restricted. Since most US firms report their geographical revenue compositions at an annual frequency, the observations of the sample used in this table are at the firm-year level. I run the following regression.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t} \tag{57}$$

α_t is the year fixed effect, ϕ_i is the firm fixed effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. GDPR_t is a binary variable that equals one if time t is after GDPR's enactment date, May 2018. Data-Intensive_i is a binary variable that equals one if firm i is in the data-intensive category. I focus on the firms with significant European market operations before 2018 and divide the sample into the more data-intensive group (above median) and the less data-intensive group (below median). I define data intensiveness in section 2.2.1. For the dependent variable, $Y_{i,t}$, I first look at the fraction of revenue generated from the European market by US firms in column (1). In column (2), I include one extra interaction term, GDPR Pass × Data-Intensive, which captures the time period between GDPR's passage and enactment. In columns (3) and (4), I look at total sales and sales scaled by total assets.

Dependent Variable:	EU Sale Percentage	Sales/Assets	EU Sales/Assets
	(1)	(2)	(3)
GDPR Effective × Data-Intensive	-1.020***	0.019	-1.717***
	(-2.606)	(0.014)	(-3.415)
GDPR Effective × Tech Industry	-1.088*	1.951	-0.695
	(-1.816)	(1.076)	(-1.177)
Controls	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
R^2	0.767	0.850	0.733
obs	10,352	10,950	10,362

Table A2: EU Segment and Firm-Level Profitability

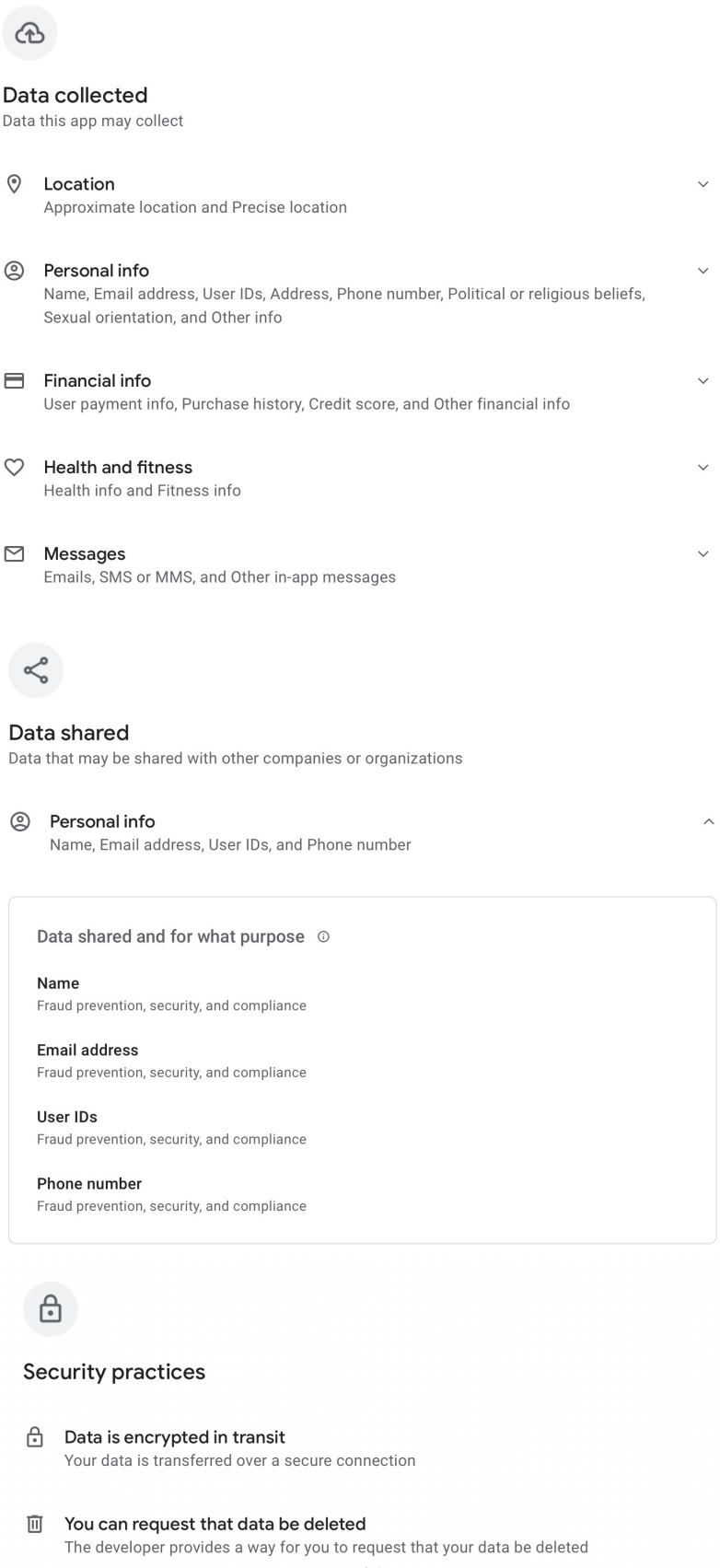
Notes: In this table, I look at how the profitability of US multinational firms changes after GDPR' enactment. I run the following regression.

$$Y_{i,t} = \alpha_t + \phi_i + \beta_1 \cdot \text{GDPR}_t \times \text{Data-Intensive}_i + \boldsymbol{\gamma} \boldsymbol{X}_{i,t} + \varepsilon_{i,t}$$

α_t is the year fixed-effect, ϕ_i is the firm fixed-effect, and $\boldsymbol{X}_{i,t}$ is a vector of time-varying firm-level characteristics, including book to market ratio, leverage, and firm size. GDPR_t is a binary variable that equals one if time t is after GDPR's enactment date, May 2018. Data-Intensive_i is a binary variable that equals one if firm i is in the data-intensive category. I focus on the firms with significant European market operations before 2018 and divide the sample into the more data-intensive group (above median) and the less data-intensive group (below median). I define data intensiveness in section 2.2.1. For the dependent variable, $Y_{i,t}$, I look into two measures of profitability, gross profit margin (GPM) and operating profit margin (OPM) in percentage points. The profitability measures are winterized at the 0.5% level on both ends. In all specifications, I control for firm-level time-varying characteristics, including book-to-market ratio, leverage, and size. The standard errors are clustered at the industry level.

	EU GPM	EU OPM	Firm GPM	Firm OPM
	(1)	(2)	(3)	(4)
GDPR Effective × Data-Intensive	-0.224	0.018	0.062	0.021
	(-1.384)	(0.886)	(0.507)	(0.251)
Controls	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes
R^2	0.919	0.907	0.685	0.696
obs	139	604	9,085	9,085

Figure A1: Data Safety Section on Google Play



A.2 Skill Keywords

AI Skills: Sentiment Analysis, Random Forests, Maximum Entropy Classifier, LDA, TensorFlow, Deep Learning, Classification Algorithms, Machine Learning, Libsvm, Latent Semantic Analysis, Backpropagation, Text Mining, Convolutional Neural Network, Geospatial Intelligence, Xgboost, Torch, NLP, Speech Recognition, Gradient Boosting, Neural Network, Long Short-Term Memory, Platfora, Latent Dirichlet Allocation, Nearest Neighbor, Reinforcement Learning, Neuroscience, Neural Nets, Recurrent Neural Network, Lasso, Pattern Recognition, Semi-Supervised Learning, Conditional Random Field, Natural Language Processing, Computer Vision, Artificial Intelligence, ND4J, Kernel Methods, Instance-Based Learning, Microsoft Cognitive Toolkit, Xgboost, Sentiment Classification, Long Short-Term Memory, LSTM, Libsvm, RNN, Word2Vec, MXNet, Caffe Deep Learning Framework, Autoencoders, MLPACK, Keras, Theano, Torch, Wabbit, Boosting, TensorFlow, Vowpal, Convolutional Neural Network, CNN, JUNG framework, OpenNLP, Natural Language Toolkit, NLTK, Unsupervised Learning, Dlib, Scikit-learn, Latent Semantic Analysis, Latent Dirichlet Allocation, Stochastic Gradient Descent, SGD, Dimensionality Reduction, Deep Learning, DBSCAN, Density-Based Spatial Clustering of Applications with Noise, AI ChatBot, Recommender Systems, Random Forests, Deeplearning4j, AdaBoost Algorithm, Support Vector Machines, SVM, Unstructured Information Management Architecture, Apache UIMA, Maximum Entropy Classifier, Pybrain, Computational Linguistics, Naive Bayes, H2O (software), WEKA, Clustering Algorithms, Matrix Factorization, Object Recognition, Classification Algorithms, Information Extraction, Image Recognition, Bayesian Networks, Supervised Learning, OpenCV, K-Means, Opinion Mining, Neural Networks, Support Vector Machine, Computer Vision, DBSCAN, Image Recognition, Mahout, Computational Linguistics, Object Recognition, Opinion Mining, Caffe Deep Learning Framework, Automatic Speech Recognition, Artificial Intelligence, Evolutionary Algorithm, Virtual Agents, Decision Trees, Predictive Models, Genetic Algorithm, Chatbot, OpenCV, Random Forest, Scikit-learn, Machine Translation, Elastic-Net, Keras, Ridge Regression, Image Processing, Big Data Analytics.

Data Management Skills: Apache Hive, Information Retrieval, Data Management Platform, DMP, Data Collection, Data Warehousing, SQL Server, Data Visualization, Database

Management, Data Governance, Data Transformation, Extensible Markup Language, XML, Data Validation, Data Architecture, Data Mapping, Oracle PL, SQL, Database Design, Data Integration, Teradata, Database Administration, BigTable, Data Security, Database Software, Data Integrity, File Management, Splunk, Relational DataBase Management System, Teradata DBA, Data Migration, Information Assurance, Enterprise Data Management, SSIS, Sybase, jQuery, Data Conversion, Data Acquisition, Master Data Management, Data Capture, Data Verification, MongoDB, Data Warehouse Processing, SAP HANA, Data Loss Prevention, Data Engineering, Database Schemas, Database Architecture, Data Documentation, Data Operations, Oracle Big Data, Domo, Data Manipulation, Data Management Platform, DMP, HyperText Markup Language, Data Access Object, DAO, Structured Query Reporter, SQR, Data Dictionary System, Data Entry, Data Quality, Data Collection, Information Systems, Information Security, Change data capture, Data Management, Data Governance, Data Encryption, Data Cleaning, Semi-Structured Data, Data Evaluation, Data Privacy, Dimensional and Relational Modeling, Data Loss Prevention, Data Operations, Relational Database Design, Database Programming, Information Systems Management, Database Tuning, Object Relational Mapping, Columnar Databases, Datastage, Data Taxonomy, Informatica Data Quality, Data Munging, Data Archiving, Warehouse Operations, Solaris, Data Modeling, Data Feed management, Data discovery, Exporting Large Datasets, Exporting Datasets, Database Performance, Designing Relational databases, Implementing Relational Databases, Designing and Implementing Relational Databases, Database Development, Data Production Process, Normalize Large Datasets, Normalize Datasets, Create Database, Develop Database, Data Onboarding, Data Sourcing, Data Purchase, Data Inventory, Cloud Security, Negotiating Data, Data Attorney, Data and Technology Attorney, Reliability Engineering, Reliability Engineer, Data Specialist, Enable Vast Data Analysis, Enable Data Analysis, Data Team, Capturing Data, Processing Data, Supporting Data, Error Free Data Sets, Error Free Datasets, Live Streams of Data, Data Accumulation, Kernel Level Development, Large Scale Systems, Hadoop, Distributed Computing, Multi Database Web Applications, Connect Software Packages to Internal and External Data, Explore Data Possibilities, Architect Complex Systems, Build Scalable Infrastructure for Data Analysis, Build Infrastructure for Data Analysis, Solutions for at Scale Data Exploration, Solutions for

Data Exploration, Information Technology Security, Security Engineer, Security Architect.