The Effects of Privacy Regulation on the Supply of Stolen Data

Anderson Frailey*

April 30, 2025

Latest version available here

Abstract

Individuals are constantly generating streams of data collected by businesses, educational institutions, data brokers, and many other organizations. These organizations are regularly targeted by cyber criminals attempting to steal that data in order to exploit or sell it in online markets. In this paper I propose a model of the stolen data economy to show how privacy regulations may affect the market. I then introduce a novel dataset of data breaches to study the effects of the European Union's General Data Protection Regulation (GDPR), a policy governing the collection and storage of user data, on the quantity of data available in the illicit market. Using a difference-in-differences design, I find that the GDPR caused a 60 percent reduction in the number of data breaches traded, but no reduction in the aggregate amount of data available. Analyzing the contents of the individual breaches, I find a nearly 70 percent increase in the amount of data they contain. These results are consistent with the model's prediction that low-value hacking targets becoming disproportionally less valuable after the GDPR, which in turn causes higher-value targets to make up a larger portion of post-GDPR data breaches.

*Email: af3wt@virginia.edu. I would like to thank my advisors: Amalia Miller, Lee Lockwood, and Denis Nekipelov for their guidance and support on this paper. Conversations with Sarah Turner, Daniel Kwiatkowski, Danielle Citron and Diego J. Jiménez Hernández were incredibly helpful in this paper's development. I received valuable comments and feedback from participants in the Public and Labor Student Workshop and Economics Research Colloquium at the University of Virginia, as well as the 2024 AEA Summer Mentoring Pipeline Conference. Finally, this paper would not have been possible without the data provided by SpyCloud. For that I thank Pablo Maceda, Ronak Patel, Wallis Romzek, and Trevor Hilligoss. The conclusions and views expressed in this paper do not reflect those of SpyCloud or any persons affiliated with SpyCloud. All remaining mistakes are my own.

1 Introduction

When individuals interact with businesses, schools, and almost any other modern organization, they generate streams of data containing their names, financial information, address, religious and political views, and more. While producing all of this information has the presumed benefit of allowing the organizations collecting it to provide better services or more relevant advertising, it has also subjected those whose data are collected to the risk of that data being improperly accessed and misused. One study found that the average digital identity appeared in nine separate data breaches and over one billion emails and passwords could be found online in 2023 alone (SpyCloud, 2024).

Exposed data is a valuable commodity for cyber criminals. It can be used to commit identify theft, fraud, and as the starting ground for future data breaches. Online markets for the trade of stolen data have developed where bundles of data are swapped for money, reputation, and bragging rights. Trades are conducted in Telegram channels, on the dark web, and niche forums on the clear web, making it possible for even those who lack technical skills to gain access to stolen data.[1]

In this paper, I propose a model of the stolen data economy to show how data privacy regulations may affect the market. I then estimate how the European Union's General Data Protection Regulation (GDPR)—one of the most comprehensive data privacy laws in the world—changed aggregate outcomes in the market and the size and contents of the data packages sold.

The GDPR is a broad reaching regulation that governs the collection and processing of personal data by covered organizations. It explicitly states when data collection is considered lawful, and prohibits the processing of sensitive data with few exceptions. Additionally, the GDPR gives individuals the right to have their data deleted, transferred, or rectified; requires detailed record keeping on data collection, impact assessments prior to data processing, and the designation of a Data Protection Officer; and increases cybersecurity investment requirements. Data breach notification requirements and large fines also significantly increase the cost of suffering a data breach. Previous research estimates the GDPR increased the cost of data storage by 20 percent, resulting in a 26 percent decrease in data storage among firms in the EU relative to comparable American organizations by (Demirer et al., 2024). In the context of the stolen data market, the GDPR is a negative supply shock. By reducing the amount of data collected and requiring increased cybersecurity, it reduces the availability of

[1]The dark web is the portion of the web that is intentionally obfuscated and only accessible through specialized internet browsers. The clear web consists of websites that can be reached by anyone and will be indexed by search engines. Clear web forums that facilitate the trade of stolen data typically require a user creating an account to view and participate in the market, technically making them part of the deep web.

the market's primary input good: data.

At a high level, the stolen data supply chain can be broken down into two components: legal data collection and data theft. The organizations we interact with regularly collect data on their customers, employees, and users for marketing, internal efficiency, and general day-to-day operations. By reducing the amount of data that is collected, the GDPR also reduces the amount of data that can be stolen.

Data is stolen by cyber criminals through a variety of means. Phishing attacks attempt to trick members of targeted organizations into revealing login information. Ransomware attacks have shifted to threatening victims with data exposure, in addition to the encryption of their data, if they do not pay the ransom (Cong et al., 2023). Software vulnerabilities or improperly configured databases may unknowingly expose databases to the outside world, making it possible for those outside the organization to access data on customers and employees. Privacy regulations impact this section of the supply chain through minimum security requirements, which, if binding, decrease the probability of successfully breaching a compliant organization.

For hackers, each potential target has an expected value and cost of hacking. Assuming they are profit maximizers, hackers will only try to hack those with a positive net value of hacking. This creates a set of profitable targets that is a subset of all potential targets. By reducing the amount of data collected and requiring organizations to invest in security, privacy regulations should decrease the value and increase the cost of breaching regulated organizations. This will shrink the profitable target set, and change the expected value of breaches that still occur. Depending on the relative changes in value and cost, relatively low-valued targets may be disproportionally removed from the profitable target set. As a result, the expected value of the targets that remain could increase.

Actions taken by the agents throughout the supply chain manifest themselves in the stolen data market. In this market, sellers are at least semi-anonymous and there is some degree of opaqueness regarding product quality, creating significant risk of adverse selection in the market.[2] I model this market following Akerlof (1970) and show that, under the right conditions, the GDPR may actually alleviate the adverse selection problem by causing higher quality products to be sold in the market.

Empirically, I employ a unique dataset of stolen data packages traded in the market. Each observation is of an individual data package and contains information on the organization the data originated from, as well as the amount and types of data included. It is important to note that these data only cover what is available online, not necessarily everything that

[2]For a discussion on how online illicit markets attempt to solve this issue with contracts and reputation building tools, see Vu et al. (2020).

was stolen in a given data breach. The two may differ if a hacker decides to keep some data for themselves or that some of the data is not worth selling. Each package is labeled as being available before or after the GDPR, and whether the data it contains should have been protected by the GDPR. To the best of my knowledge, this is the first use of such data in the economics literature.

To determine aggregate effects, I combine the individual data packages to create a country-quarter level panel spanning from January 2017 to November 2023 that tracks the number of data breaches and records available that originate in a given country. I use this panel to estimate a difference-in-differences model measuring the effect of the GDPR on those two outcomes. I also break the post-GDPR period into short-run and long-run periods to measure if and how the effects changed overtime. Short-run is defined as one year after the regulation went into effect and long-run is anytime after that.

I find that the GDPR caused the number of data breaches originating from regulated countries to decrease by approximately 60 percent overall, with the long-run decrease being slightly larger than the short-run (61 percent versus 54 percent). Despite this, I find no statistically significant change in the number of records available. The granularity of my data allow me to estimate how the composition of the individual data packages changes to explain the lack of change in number of records.

At the individual data package level, I estimate how the contents of the data packages—the number of records, amount of personally identifiable information (PII), and number of unique types of data—changes after the GDPR. As with the aggregate effects, I estimate both an overall change and separate short-and long-run changes. I find that the size of data packages, in terms of number of records, originating in regulated countries increases nearly 70 percent in the long-run, while there is no statistically significant short-run change. The fraction of those records that are considered PII and number of unique data types in these packages do not change in any measured time period.

The increase in size of the data packages explains how the number of data packages could fall without an accompanying decrease in the number of records. The theoretical model I present suggests this is due to a shift towards more data rich targets after the GDPR changed the viable target set. Additional empirical evidence of a shift towards larger targets is in the UK cyber security breach survey, which shows that small organizations (those with fewer than 50 employees) make up 95 percent of reported breaches in the 2017 survey, but only 48 percent in the 2022 survey. Large organizations (those with 250 or more employees) increase their share from less than one percent to approximately 24 percent (Department for Digital, Culture, Media and Sport, 2022).

The effects of data privacy and security legislation have been studied in a number of

contexts including healthcare (Miller and Tucker, 2009, 2011, 2018) and online advertising (Goldfarb and Tucker, 2011). The GDPR specific literature covers its effects on firm performance (Koski and Valmari, 2020; Chen et al., 2022; Goldberg et al., 2024), competition (Johnson et al., 2023), investment (Jia et al., 2021; Kircher and Foerderer, 2021; Janßen et al., 2022), and data collection (Aridor et al., 2021; Lukic et al., 2023; Demirer et al., 2024). These papers typically find negative effects of the regulation: decreases in competition, investment, and firm performance. Or changes whose net welfare effects are more ambiguous, such as decreased data storage. While I do not attempt to calculate the overall welfare effects of the policy, this paper is the first to show a seemingly unambiguous benefit of the GDPR: the reduction in the number of data packages online. But even with this reduction, the extent to which individuals benefit is unclear given that there was not an accompanying reduction in the number of records available. It is possible that, while there are fewer breaches, those that remain contain enough information to leave the affected individuals no better off than before.

Significant work studying stolen data markets has been conducted by criminologists, who have derived some estimates of their sizes and products offered (Franklin et al., 2007; Holt and Lampke, 2010; Holt et al., 2016). These papers conduct in depth, descriptive studies of a handful of individual forums where data is sold. They do not study how public policy and new technologies can have trickle-down effects on these markets. My unique dataset allows me to fill that gap in this are of the criminology literature, and extend the contribution into the economics of crime.

The model I present conceptually aligns with Becker (1968). The decisions of the hackers to attempt a data breach is based on the perceived costs and benefits of doing so. When the costs increase and the benefits decrease, there are fewer breaches. The costs and benefits are not perfectly observable, requiring hackers to base their actions on their beliefs of data collection practices and how well potential targets have protected themselves. This is similar to the mechanisms in Ayres and Levitt (1998) and Braakmann et al. (2024). In Ayres and Levitt, car thieves could not observe which vehicles had tracking devices installed, but were aware of which areas had higher installation rates. The higher likelihood of stealing a car that could be tracked caused them to steal fewer cars in those areas. In Braakmann et al., the price of gold increasing motivated burglars to target homes in areas where homeowners were expected to store more gold. The GDPR has the opposite effect of Braakmann et al., but a similar effect to Athey et al.. By causing a reduction in the expected value and increase in the expected cost of breaching European organizations, the regulation incentivizes hackers to change who they target.

The remainder of this paper is structured as follows. Section 2 formally presents the

model of the stolen data market discussed earlier. In Section 3 I describe the data used in this study. My empirical strategy is defined in Section 4 and the results are presented in Section 5. I provided concluding remarks and paths for future research in Section 6.

2 A Model of Stolen Data Production

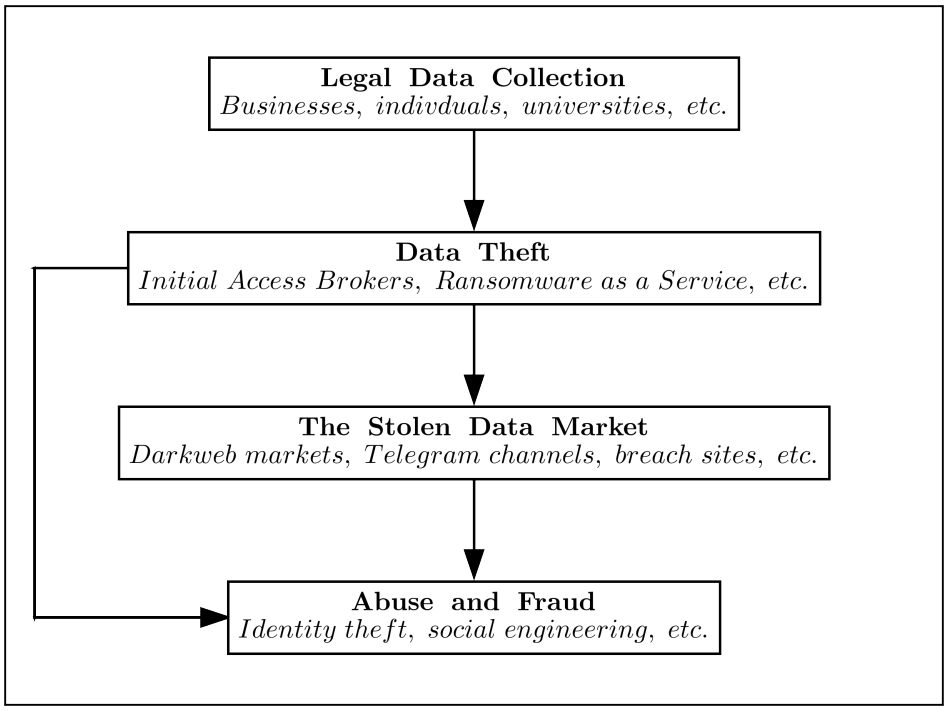


Figure 1: The Stolen Data Supply Chain

The production of stolen data can be described by a two-part "supply chain", depicted in figure 1. It begins with data collectors deciding what information to collect. Data collectors are companies, schools, governments, and any other entity that holds customer, user, and employee data. Collecting data comes at a cost. They must pay to gather it, keep it stored, and respond to user requests regarding their data. There is also the persistent risk that they suffer a data breach and incur additional costs as a result. These include sending notifications to those affected, offering credit monitoring, performing security audits, legal costs, and fines imposed by governments. To mitigate this risk, organizations can invest in security measures. Some are technical, such as consistently patching software vulnerabilities and encrypting data. Others are non-technical, such as teaching employees to detect phishing emails or improve their password management. For both types, the goal is to make it more

difficult for data to fall into the wrong hands.[3]

In the second stage, data theft, hackers target a subset of data collecting organizations based on the expected cost and benefit of doing so. Assuming that they are profit maximizing agents, they will only want to hack an organization if the expected profit from doing so is positive. Once they have the data, they can either keep it for themselves or sell it in the market.

Stolen data is traded in Telegram channels and other online black markets. Suppliers may advertise their products by describing what is in the data package and where it originated from (Holt and Lampke, 2010). Because they are anonymous, online, and illegal, these markets are vulnerable to adverse selection problems.[4] I model this part of the market following Akerlof (1970) and describe the conditions necessary for the market to exist.

Privacy regulations are a negative supply shock along two dimensions. First, they reduce the amount of data stored by organizations, as discussed in Demirer et al. (2024). Second, they typically require increased investment in cybersecurity, making it more difficult to breach a regulated organization. Data from the United Kingdom Cyber Security Breaches Survey shows that nearly two thirds of respondents made operational changes in response to the GDPR. Among those that made changes, 100 percent reported making changes related to the cybersecurity policies and practices (table 1). Both effects increase the cost of acquiring the key input to the market: the data itself.

This changes the incentives of the attackers. The marginal value of the data that can be extracted from a regulated organization decreases, while the marginal cost of breaching one has increases, encouraging changes in the optimal effort allocation. In equilibrium, this may increase or decrease the expected value of the data packages still sold, which will influence demand for the goods. In the remainder of this section, I present a model that describes the behavior of both agents, and the effects of privacy regulation on their choices and the final market equilibrium.

2.1 Legal Data Collection—Organizational Behavior

Organizations in this framework choose what types and how much data to collect. With J total types of data available, each individual type of data, j, is used to generate information. Denoting the total amount of each type of data collected as d_j, the function $I(d_1, \ldots, d_J)$

[3]This goal is not always achieved. Miller and Tucker (2011) find that use of encryption technology is actually associated with an increase in reports of data loss.

[4]Users and platforms now rely heavily on reputation to facilitate trade. Some platforms have created contract systems that set expectations for the parties involved in a transaction and help build supplier's reputation (Vu et al., 2020). Often, suppliers will give away their stolen data rather than sell it to help build their reputation.

Table 1: Percentage of Organizations Reporting Operational Changes in Response to the GDPR

	Any Change		Change in Cybersecurity	
Survey Year	2018	2019	2018	2019
Overall	12.75%	63.71%	100.00%	100.00%
Small	9.91%	61.56%	100.00%	100.00%
Medium	28.71%	90.66%	100.00%	100.00%
Large	52.91%	95.81%	100.00%	100.00%

Source: Department for Digital, Culture, Media and Sport (2022), author's calculations. Respondents were asked *"Has your organisation made any changes or not to the way you operate in response to GDPR?"* and *"Have any of these changes been related to your cyber security policies or processes, or not?"* The fraction of respondents answering yes to the first question is in the first two columns. The fraction answering yes to the second question, among those answering yes to the first, is in the last two columns.

determines the total information generated. The total cost of collecting these data is given by the function $C(d_1, \ldots, d_J)$. I assume that the information function takes the form:

$$I(d_1, \ldots, d_J) = A\left(\alpha_1 d_1^{\rho} + \ldots + \alpha_J d_J^{\rho}\right)^{\frac{\nu}{\rho}}$$

where ν determines returns to scale and ρ the level of substitutability between data types. A is an organization specific productivity term.[5]

For simplicity, I assume linear data collection cost: $C(d_1, \ldots, d_J) = \sum_{j=1}^{J} \omega_j d_j$, where ω_j is the cost of collecting a unit of type j data. Cost of collection can vary between data types due to laws governing how certain types of data are stored. Examples include additional encryption or security requirements for data that are particularly sensitive such as health and financial information. Additionally, some privacy regulations give individuals the right to have their data corrected for mistakes or deleted upon request. The frequency with which those requests are made may vary by data type. For example, a customer of a credit rating agency is more likely to notice and request correction of an error that greatly affects their credit score than they are a smaller error, such as an in incorrect address.

Each organization also invests some amount in security, S, to prevent data breaches. A unit of security costs ω_S to purchase and directly reduces the probability of suffering a breach. Regardless of the size of the investment, breach probability never reaches zero because, no

[5] Demirer et al. (2024) use a similar information function. Rather than include a term for each type of data, they use a singular term for the total amount of data stored and add the amount of computation used a choice variable.

matter how much security an organization has, there is always the possibility for human error or a previously unknown software vulnerability that could expose their data. I adopt the breach probability function introduced in Gordon and Loeb (2002). Given an intrinsic level of risk r, the probability of a breach after accounting for security investment is:

$$\mathbb{P}(S) = \frac{r}{S+1}, \quad r \in [0,1].$$

Security investment decreases the probability of a breach, but at a decreasing rate.[6]

If they suffer a data breach, the organization will incur losses $L(d_1, \ldots, d_J)$. These damages include lost sales, restoring their computer systems, lawsuits, and fines. Again for simplicity I assume that total losses are linear in data collection and include a fixed loss ℓ: $L(d_1, \ldots, d_J, \ell) = \ell + \sum_{j=1}^{J} \gamma_j d_j$. Like the ω terms, the γ terms vary by data type because some data will result in bigger losses than others if stolen.

The organization faces the optimization problem:

$$\max_{d_1,\cdots,d_J,S} \quad A\left(\alpha_1 d_1^{\rho} + \ldots + \alpha_J d_J^{\rho}\right)^{\frac{\nu}{\rho}} - \sum_{j=1}^{J}(\omega_j d_j) - \omega_s S - \frac{r}{S+1}\left(\ell + \sum_{j=1}^{J}\gamma_j d_j\right).$$

As an example, assume there are just two data types, making the problem:

$$\max_{d_1,d_2,S} \quad A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu}{\rho}} - \omega_1 d_1 - \omega_2 d_2 - \omega_S S - \frac{r}{S+1}\left(\ell + \gamma_1 d_1 + \gamma_2 d_2\right). \tag{1}$$

Taking the first order conditions with respect to S, d_1, and d_2 yields:

$$\frac{r}{(S+1)^2}\left(\ell + \gamma_1 d_1 + \gamma_2 d_2\right) = \omega_S \tag{2}$$

$$\alpha_1 d_1^{\rho-1} \nu A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu-\rho}{\rho}} = \omega_1 + \frac{r}{S+1}\gamma_1 \tag{3}$$

$$\alpha_2 d_2^{\rho-1} \nu A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu-\rho}{\rho}} = \omega_2 + \frac{r}{S+1}\gamma_2 \tag{4}$$

Simply put, they will invest in security until the marginal benefit, the reduction in expected losses due to a data breach, equals the cost of an additional unit of security (equation 2). Similarly, they will collect data until the marginal benefit—the additional information generated—equals the marginal cost—the cost of collecting and the increased cost of a breach (equations 3 and 4).

[6]The more general form in Gordon and Loeb includes measures for security productivity, making the function $\frac{r}{(\varsigma S+1)^{\beta}}$. I have assumed that $\varsigma = \beta = 1$. This does not meaningfully change the interpretation of my results.

Rearranging equation 2 reveals that the optimal S is:

$$S^* = \sqrt{\frac{r\left(\ell + \gamma_1 d_1^* + \gamma_2 d_2^*\right)}{\omega_S}} \tag{5}$$

Intuitively, optimal security investment will be increasing in fundamental risk and the various costs associated with a breach.

Using equations 3 and 4, the optimal levels of data collection are described by the equations:

$$d_1^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2}\right)^{\frac{\rho}{1-\rho}}\right]^{\frac{\nu-\rho}{\rho(1-\nu)}} \tag{6}$$

and:

$$d_2^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2}\right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2}\right)^{\frac{\rho}{1-\rho}}\right]^{\frac{\nu-\rho}{\rho(1-\nu)}}. \tag{7}$$

Full derivations are in section A.1 of the appendix. The primary takeaway from the above equations is that data collection decreases as the cost of collection increases.

Privacy regulations increase the cost of collecting data in numerous ways. In the case of the GDPR, criteria that must be met for any data collection to be legal are defined in Article 6, and Article 9 prohibits the collection of particularly sensitive data. Also on the cost of collection side, the GDPR gives individuals the right to have their data deleted, transferred, or rectified (Articles 12-13); requires record keeping of data processing (Article 30), conducting impact assessments prior to processing data (Article 35), and the designation of a Data Protection Officer (Article 37). Each of these provisions increases the costs of collecting data, ω_j, for each type of data and the size of that increase may vary by type. Finally, the cost of being breached increases because of notification requirements (Article 33) and the potential for fines after the breach (Article 83). This increases both the fixed

costs of a breach ℓ, and the costs associated with each type of data stolen, γ_j.

In addition to governing when data collection is legal, the GDPR requires implementing a minimum level of cybersecurity appropriate for the organization's risk level (Article 32), effectively setting a lower bound, $\underline{S}$, on security investment. If $S^* < \underline{S}$, organizations will need to increase their spending on security beyond their unregulated choice. Together, the organizational decisions derived in this section will determine their value as targets in the next section.

2.2 Data Theft

Once data has been collected, organizations become potential targets for breaches. Each target i has an expected value of the data that can be stolen from them and cost of hacking denoted V_i and C_i, respectively. Quality is based on the amount and type of data they collect, and cost is a function of their security investment. The expected profit of hacking target i is

$$\pi_i = V_i - C_i.$$

A profit maximizing hacker will only target a given organization if $\pi_i \geq 0$, or $V_i \geq C_i$. This creates a threshold that splits targets into those that get hacked and those that do not, shown by the 45 degree zero-profit line in figure 2. Targets that fall above the line, the profitable set, will be hacked, those below will not. With this delineation, the expected value of a hacked target is

$$\mathbb{E}\left[V_i | V_i \geq C_i\right] = \int_{\underline{C}}^{\bar{C}} \int_{C_i}^{\bar{V}} V_i dF(V_i, C_i)$$

where $\underline{C}, \bar{C}$, and $\bar{V}$ are the lower and upper bounds for C_i and V_i.

Privacy regulations will both decrease the value and increase the cost of hacking regulated entities. For target i, the new value and cost are

$$\begin{aligned} V_i^{Post} &= (1-\phi)V_i \quad 0 < \phi < 1 \\ C_i^{Post} &= \xi C_i \quad \xi \geq 1 \end{aligned}$$

creating a new zero-profit condition: $(1-\phi)V_i = \xi C_i$ that must be satisfied for the target to be hacked. The new expected value of breaches is then:

$$\mathbb{E}\left[V_i \middle| \frac{\xi}{1-\phi} C \leq V_i\right] = \int_{\underline{C}}^{\bar{C}} \int_{\frac{\xi}{(1-\phi)}C_i}^{\bar{V}} V_i dF(V_i, C_i).$$

The total number of breaches decreases for all valid values of ξ and ϕ, but whether the post-GDPR expected value is higher or lower than pre-GDPR expected value depends on the correlation of ϕ and ξ with V and C, and the joint distribution of V and C.

Suppose that $(V, C) \sim Uniform[0, 1]^2$. If ϕ and ξ are constants, then each potentially targeted organization experiences the same proportional decrease in value and increase in cost. They will fall out of the profitable target set proportionately, and the expected breach value is unchanged.

If instead ϕ or ξ are correlated with V or C, the slope of the zero profit line will no longer be constant and either high or low value targets will be disproportionally removed from the profitable target set.

In the case where ξ is positively correlated with V, the marginal return to security investment will be higher for high-value targets than low. This will cause the zero profit line to become steeper at high values of V, disproportionally removing high-value targets from the subset of targets that are worth hacking. The same is true if ϕ were to be positively correlated with V. A positive correlation between V and ϕ would mean that the decrease in value caused by the GDPR would be larger for high-value targets than low. In either case, high-value targets are disproportionally removed from the profitable target set and the expectation of V falls.

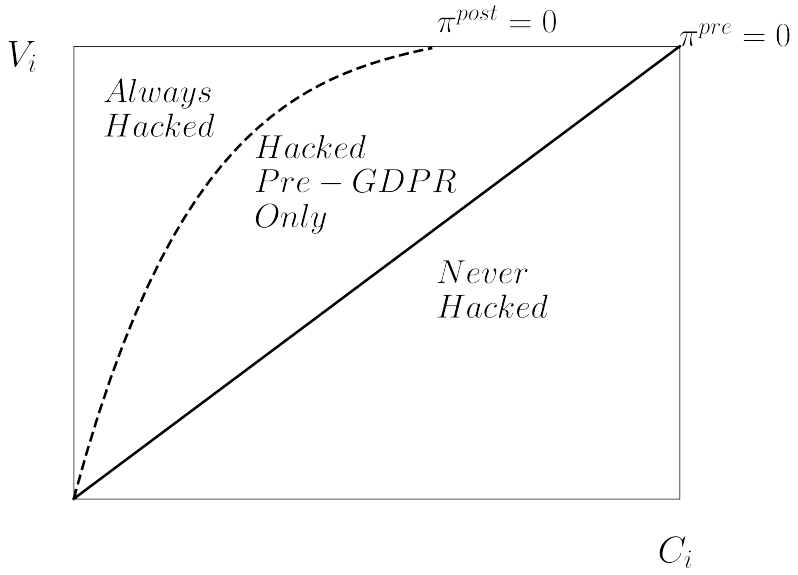
If ξ is negatively correlated with V, the marginal return to security investment is lower for high-value targets than low. Similarly, if ϕ is negatively correlated with V, the GDPR reduced value less for high-value targets than low. In either case, the zero profit line flattens out at higher realizations of V and low-value targets are disproportionally removed from the set of hacked targets. The expectation of V will be higher post-GDPR than prior to the regulation because fewer low-value organizations are in the profitable target set. This is shown by the curved line in figure 2.

Equation 5 in the previous section shows that a data collector's optimal security investment is increasing in the amount and value of data they collect. Since the value of a breach is an increasing function of the amount and value of data that are collected, high-value targets will also have more and better security than low-value targets pre-GDPR under this model. Assuming the marginal return to security is decreasing, the increase in hacking cost caused by the GDPR's security requirements will be relatively smaller for high-value targets than low, meaning ξ and V are negatively correlated.

If there is a correlation between ϕ and V, it is likely to also be negative. Demirer et al. (2024) find that IT-intensive industries have a smaller response—in terms of reducing data collection—to the GDPR that less IT-intensive industries. They also find the increase in data collection costs the GDPR caused was smaller for larger organizations. Assuming

that IT-intensive and large organizations make for high-value targets, ϕ and V will also be negatively correlated. This causes the slope of the zero-profit line to flatten more as V increases, resulting in an even more disproportionate removal of low-quality targets.

Figure 2: Conceptual Model



Notes: Both pre-and post-GDPR the zero-profit lines split the potential target set into groups that are and are not hacked. Those to the right of the line would be unprofitable due to high costs and low qualities, while those to the right are worth breaching. After the GDPR, low-quality targets get disproportionally excluded from the target set, increasing the expected quality of those still breached.

2.3 The Stolen Data Market

After stealing data from the original data collectors, hackers have the option of keeping or selling it in the market. Participants in this market are at least semi-anonymous and only the sellers know the true quality of the data they hold until it is sold, making it ripe for issues of adverse selection. I use the lemons model from Akerlof (1970) as the foundation of this section of the model.

Suppose hacker utility is given by

$$U^H = M + \sum_{i=1}^{\mathcal{B}^H} V_i$$

where M is non-data consumption, whose price is normalized to one, and $\mathcal{B}^H$ is the set of data packages, which come from the individual breaches, they hold. This is not the entire

set of potential targets, only those that are breached. V_i is the value of the data from breach i, as described in the previous section.

Hackers will only sell the data they have stolen if the price they get is higher than the utility they gain from holding it: $V_i \leq p$. Market supply is then:

$$S(p) = \mathcal{B}\mathbb{P}\left(V_i \leq p \middle| C \leq V\right) \tag{8}$$

where $\mathcal{B}$ is the set of all hacked targets.

Buyers have a similar utility function:

$$U^B = M + \sum_{i=1}^{\mathcal{B}^B} \kappa V_i.$$

The parameter κ allows for buyers and sellers to have different values of the same bundle of data. This can occur if the skill sets needed to steal the data and profit from it are different, meaning there are comparative advantages between buyers and sellers. If $\kappa > 1$, the buyers of stolen data are more productive in their use of stolen data than those who steal it. The larger κ, the larger that gap in ability. $\mathcal{B}^B$ is the set of data packages held by the buyer, and all other parameters in the buyer's utility function are the same as in the hacker's

Buyers cannot observe the true quality of the data packages sold and thus make their purchase decisions based on the expected value: $\mu \equiv \mathbb{E}\left[V | C \leq V \leq p\right]$. They will only purchase data packages if $\kappa\mu \geq p$. With an income of Y, total demand for stolen data is

$$D(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa\mu \geq p \\ 0 & \text{Otherwise} \end{cases}$$

The expected value of the data provided at a given price is mechanically less than the price, meaning a market will only exist if κ is sufficiently large. The difference in ability to obtain and exploit stolen data leads to labor specialization in the market. Those who are most adept at stealing data sell at least a portion of their data to those who are better at exploiting the information in it. With a sufficiently large κ, there will be an equilibrium price p^* that clears the market.

After the GDPR, hacker utility becomes

$$U^{H,Post} = M + \sum_{i=1}^{\mathcal{B}^{H,Post}} (1-\phi)V_i$$

where $\mathcal{B}^{H,Post} \leq \mathcal{B}^{H}$ is the number of breaches the hold post-GDPR. They will now sell if $(1-\phi)V_i \leq p$. Which creates the new supply curve:

$$S^{Post}(p) = \mathcal{B}^{Post}\mathbb{P}\left((1-\phi)V_i < p\right)$$

where $\mathcal{B}^{Post}$ is the set of targets hacked post-GDPR.

On the buyer side, their new expected value of the packages sold is

$$\mu^{Post} = \mathbb{E}\left[V \middle| \frac{\xi}{1-\phi}C \leq V \leq \frac{p}{1-\phi}\right].$$

Where it exists, demand remains unchanged, but the minimum κ needed for it to exist changes to satisfy $\kappa(1-\phi)\mu^{Post} \geq p$.

If lower-value targets disproportionally fall out of the target set, μ^{Post} may be higher than μ, depending on the exact value of ϕ. This will lower the minimum κ needed for demand to exist. The decrease in supply will also increase the price, making hackers more willing to sell their higher-value breaches. As a result, even though there are fewer breaches, the value of what is traded may increase. Given that the amount of data is one aspect of value, it is theoretically possible that the GDPR actually increases the amount of data traded online.

2.4 Stylized Example

To demonstrate how expected value and the size of the market change in response to privacy regulations, suppose again that $(V, C) \sim Uniform[0,1]^2$. Prior to the GDPR,

$$\mathbb{E}\left[V \middle| C \leq V\right] = \frac{2}{3}.$$

Hackers will only sell their data if the price they get is higher than their utility gain should they keep it, making supply:

$$\begin{aligned} S(p) &= \mathcal{B}\mathbb{P}(V \leq p) \\ &= \mathcal{B}p^2 \end{aligned} \tag{9}$$

The expected quality of a breach given that it is being sold, μ, is

$$\mathbb{E}\left[V | C \leq V \leq p\right] = \frac{2}{3}p.$$

Demand only exists in this market if $\kappa\mu \geq p$, so the minimum κ required is $\kappa = 3/2$ and the demand curve is:

$$D(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3}{2} \\ 0 & \text{Otherwise} \end{cases} \tag{10}$$

Equations 9 and 10 yield the pre-GDPR equilibrium:

$$\begin{aligned} p^* &= \left(\frac{Y}{\mathcal{B}}\right)^{\frac{1}{3}} \\ Q^* &= Y^{\frac{2}{3}}\mathcal{B}^{\frac{1}{3}} \end{aligned} \tag{11}$$

Full derivations can be found in section A.2 of the appendix.

Post-GDPR, let $\xi_i = \theta V_i^{\sigma}$ and for simplicity assume that ϕ is constant. The zero profit line is now

$$V = \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}}$$

And the expectation of V in this range is

$$\mathbb{E}\left[V \middle| \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}} \leq V\right] = \frac{2-\sigma}{3-\sigma}$$

As can be seen, the change in expected quality depends entirely on σ. If $\sigma = 0$, then $\xi = \theta$ and is constant across all values of V. While hackers will be worse off than before because their utility from each hack is $(1-\phi)V$, $\mathbb{E}[V]$ will be unchanged. In other words, the composition of the remaining breaches, in terms of the distribution of value, will remain the same. There will just be fewer of them. If σ is positive, ξ grows with V and the expected value of breaches will fall. Finally, if σ is negative, ξ is smaller for high levels of V, and the expectation of V will be higher than pre-GDPR levels.

Given that the utility they attain from holding onto any given data package has fallen, hackers will be more willing to sell what they steal. Specifically, they will now sell if $(1-\phi)V \leq p$. The expected value of goods sold in the market at any given price is now

$$\mathbb{E}\left[V \middle| \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}} \leq V \leq \frac{p}{1-\phi}\right] = \frac{2-\sigma}{3-\sigma}\frac{p}{1-\phi}. \tag{12}$$

The supply of data packages on the market also changes:

$$\begin{aligned} S^{Post}(p) &= \mathcal{B}^{Post}\mathbb{P}\left(V \leq \frac{p}{1-\phi}\middle|\left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}} \leq V\right) \\ &= \mathcal{B}^{Post}\left(\frac{p}{1-\phi}\right)^{2-\sigma} \end{aligned} \tag{13}$$

Although the total number of packages sold will fall because fewer organizations are hacked, the portion of hacks being sold at a given price will increase.

While hackers are more willing to sell their goods, for buyers κ must now be large enough for $\kappa(1-\phi)\mu^{Post} \geq p$ to hold true. Given the expectation of V in equation 12, the new minimum κ required for the market to exist is

$$\kappa \geq \frac{3-\sigma}{2-\sigma}.$$

Demand is now

$$D^{Post}(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3-\sigma}{2-\sigma} \\ 0 & \text{Otherwise} \end{cases} \tag{14}$$

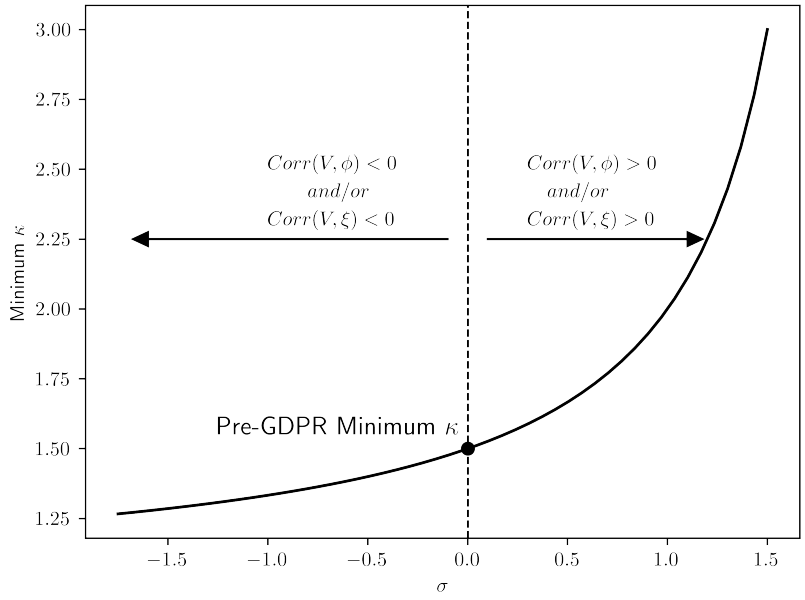
Figure 3 shows how the minimum κ needed for a market to exist changes with σ. When σ is negative, low-value targets are disproportionally removed from the profitable target set. This increases the expected quality of the remaining targets in the set, which also increases buyer's quality expectations, μ^{post}. As a result, the market can be supported with a smaller κ. The opposite is true when σ is positive. In this case, high-value targets are disproportionally removed from the profitable target set, reducing μ. For a market to exist, κ must be large enough to counteract this change.

If κ is sufficiently large, the new post-GDPR equilibrium price and quantity are

$$\begin{aligned} p^*_{Post} &= \left(\frac{Y}{\mathcal{B}^{Post}}\right)^{\frac{1}{3-\sigma}}(1-\phi)^{\frac{2-\sigma}{3-\sigma}} \\ Q^*_{Post} &= Y^{\frac{2-\sigma}{3-\sigma}}\left(\frac{\mathcal{B}^{Post}}{(1-\phi)^{2-\sigma}}\right)^{\frac{1}{3-\sigma}}. \end{aligned} \tag{15}$$

How the post-GDPR equilibrium compares to the pre-GDPR equilibrium will depend on the values of ϕ and σ. To demonstrate, I simulate the model under pre-GDPR conditions and two potential post-GDPR states of the world. In the first, $Corr(\xi, V) < 0$, i.e., there are diminishing returns to security investment. In the second, $Corr(\xi, V) > 0$, i.e., there are

Figure 3: Minimum κ



Notes: The above figure shows the minimum κ needed for a market to exist given σ.

increasing returns to security investment. For simplicity, I make ϕ a constant equal to 0.26.[7] Table 2 lists the full set of parameters in the simulation. The pre-GDPR parameters are set to create the original, linear, zero-profit line, while both sets of post-GDPR parameters create non-linear zero-profit lines. In all cases, I assume κ is at least 1.5 since that is the smallest value possible for the market to have existed prior to the GDPR. If κ must be larger than 1.5 for the market to exist, I set it equal to $(3-\sigma)/(2-\sigma)$.

Table 2: Simulation Parameters

	Pre-GDPR Baseline	Post-GDPR	
Parameter		$Corr(\xi, V) < 0$	$Corr(\xi, V) > 0$
Y	55,000	55,000	55,000
N	1,000,000	1,000,000	1,000,000
ϕ	0	0.26	0.260
θ	1	1	$1 + (\frac{1}{V^{\sigma}})$
σ	0	-3.0	0.200

With $(V, C) \sim Uniform[0, 1]^2$, half of all the potential targets are breached pre-GDPR, and the expected quality of those breaches is 2/3. In this market, the price equals $\kappa\mu$ as

[7] I chose $\phi = 0.26$ because Demirer et al. (2024) find the GDPR reduced data storage by 26 percent in the long-run. This number could be changed and the general findings of the model would remain the same.

buyers will pay up to their expected utility gain (table 3, column one).

Table 3: Simulation Outcomes

	Pre-GDPR	Post-GDPR	
		$Corr(\xi, V) < 0$	$Corr(\xi, V) > 0$
% Targets Hacked	0.501	0.148	0.194
$\mathbb{E}[V\|Hacked]$	0.666	0.833	0.655
Minimum κ	1.500	1.250	1.667
% of Hacked Data Packages Sold	0.230	0.564	0.522
Equilibrium Price	0.479	0.660	0.525
Equilibrium Quantity	115,353	83,446	101,363
$\mathbb{E}[V\|Sold]$	0.320	0.742	0.465
$\mathbb{E}[(1-\phi)V\|Sold]$	0.320	0.549	0.344
U^B	55,056	68,693	59,960
U^H	351,797	100,336	112,434

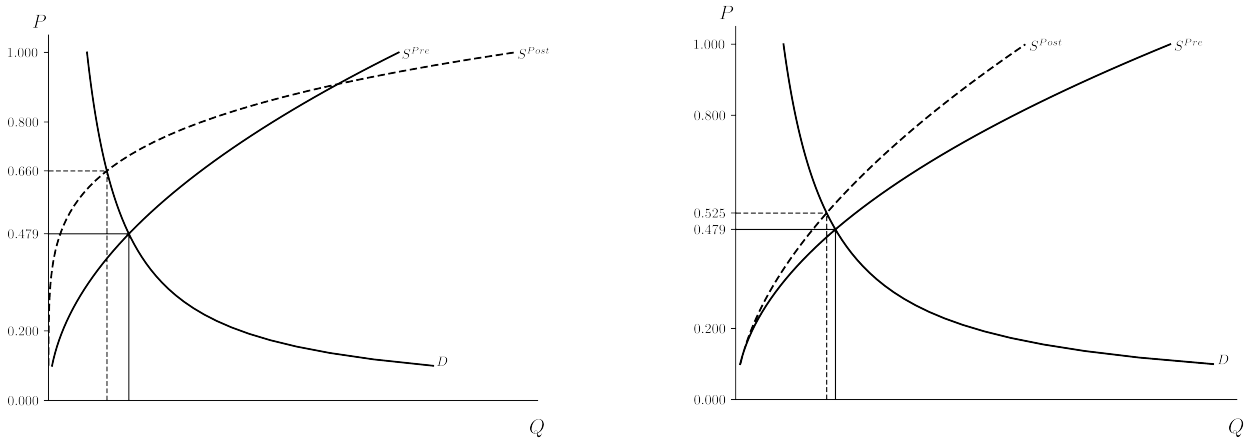
Notes: This table presents the results of the main simulation exercise in section 2.4. The simulations in columns one and two use $\kappa = 1.5$ while in column three κ is increased to 1.667 in order for demand to exist.

In the first post-GDPR simulation, where $Corr(\xi, V) < 0$, the expected profitability of hacking falls for all value levels, resulting in only 15 percent of all targets being hacked. But because of the diminishing returns to security investment, the increase in hacking cost is smaller for high-value targets than low. As a result, a higher portion low-value targets fall out of the profitable target set than high-value. This raises the value buyers expect to receive, which lowers the minimum κ needed for the market to exist to 1.25. As is expected with a decrease in supply, equilibrium price rises while equilibrium quantity falls. The increase in price incentivizes hackers to sell higher quality data packages, as shown in figure 5, further increasing $\mathbb{E}\left[V \middle| Sold\right]$. The results from this simulation are in the second column of table 3. Figure 4a plots this market equilibrium relative to the pre-GDPR period.

The second post-GDPR simulation sets θ and σ to make ξ increase with V. The results of this simulation are in column three of table 3 and plotted in figure 4b. As before, there is a decrease in supply with a higher equilibrium price and quantity. The expected value of the targets that are still hacked with their breaches being sold is lower than that in column two, requiring a higher κ for the market to exist. To run the model, it is necessarily to raise κ to 1.667 to satisfy this condition. In table 4 I instead leave κ equal to 1.5 for all simulations. While that is sufficient for a pre-GDPR and the first post-GDPR market to exist, demand will be zero in the second post-GDPR condition.

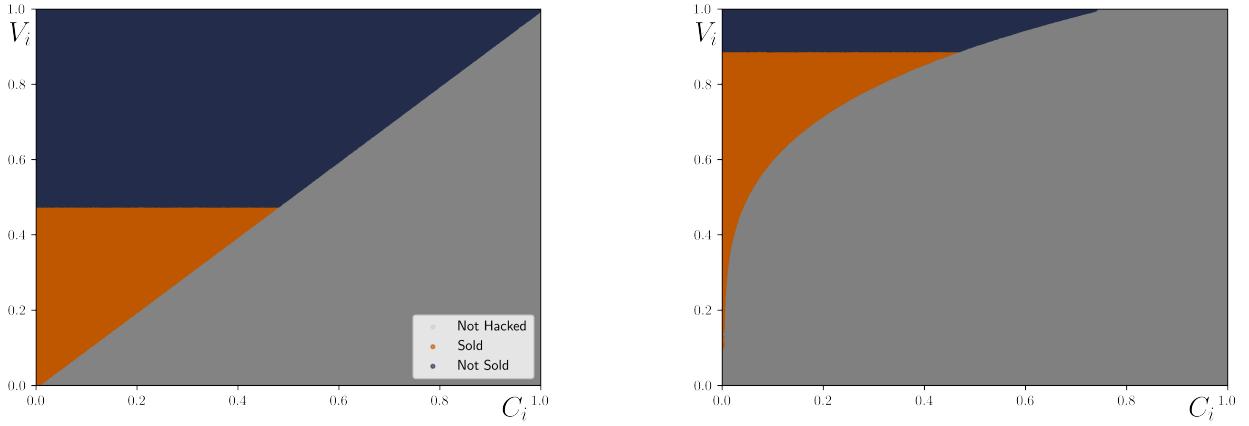
These simulations show that under the right conditions privacy regulations may actually

Figure 4: Simulated Equilibrium



(a) $Corr(\xi, V) < 0$ *(b)* $Corr(\xi, V) > 0$

Figure 5: Data Packages Sold and Not Sold



(a) Pre-GDPR *(b)* Post-GDPR

increase the expected value of data packages stolen and traded. This reduces the adverse selection problem in the market and increases buyer utility.

3 Stolen Data Market Observations

Data for this study come primarily from SpyCloud, a private cybersecurity company specializing in identity threat protection. They have constructed a catalog of data breaches gathered from a number of online stolen data marketplaces. Each observation is of a data package traded in the market, containing information on which organizations the data were taken from, what types of data were stolen, and the total number of records included. The

Table 4: Simulation Outcomes: Fixed κ

	Pre-GDPR	Post-GDPR	
		$Corr(\xi, V) < 0$	$Corr(\xi, V) > 0$
% Targets Hacked	0.501	0.148	0.194
$\mathbb{E}[V\|Hacked]$	0.666	0.833	0.655
κ	1.500	1.500	1.500
% of Hacked Data Packages Sold	0.230	0.564	0
Equilibrium Price	0.479	0.660	-
Equilibrium Quantity	115,353	83,446	0
$\mathbb{E}[V\|Sold]$	0.320	0.742	-
$\mathbb{E}[(1-\phi)V\|Sold]$	0.320	0.549	-
U^B	55,056	68,693	55,000
U^H	351,797	100,336	94,118

Notes: This table presents the results of the second simulation exercise in section 2.4. For each simulation $\kappa = 1.5$, which results in there being no demand in the model in column three.

data packages were available online between 2015 and 2023, though the breaches they originate from may have occurred as early as 2002. To the best of my knowledge, this is the first time such a dataset has been used to quantitatively study the effect of any policy change on the stolen data market. The details of the data allow me to go deeper than the aggregate and summary statistics previous research has depended on to see what is actually being traded. Unfortunately, I do not observe prices for all but a handful of data packages, restricting this paper to measuring just quantity effects.

Table 5: Data Package Summary Statistics

	Number of Records	PII Fraction	# of Data Types
Observations	4,394	4,394	4,394
Mean	3,544,186	0.690	6.220
Std. Dev.	28,996,342	0.191	5.208
Min.	1	0	1
25%	5,164	0.500	2
50%	46,748	0.667	4
75%	288,555	0.855	9
Max.	716,409,393	1	55

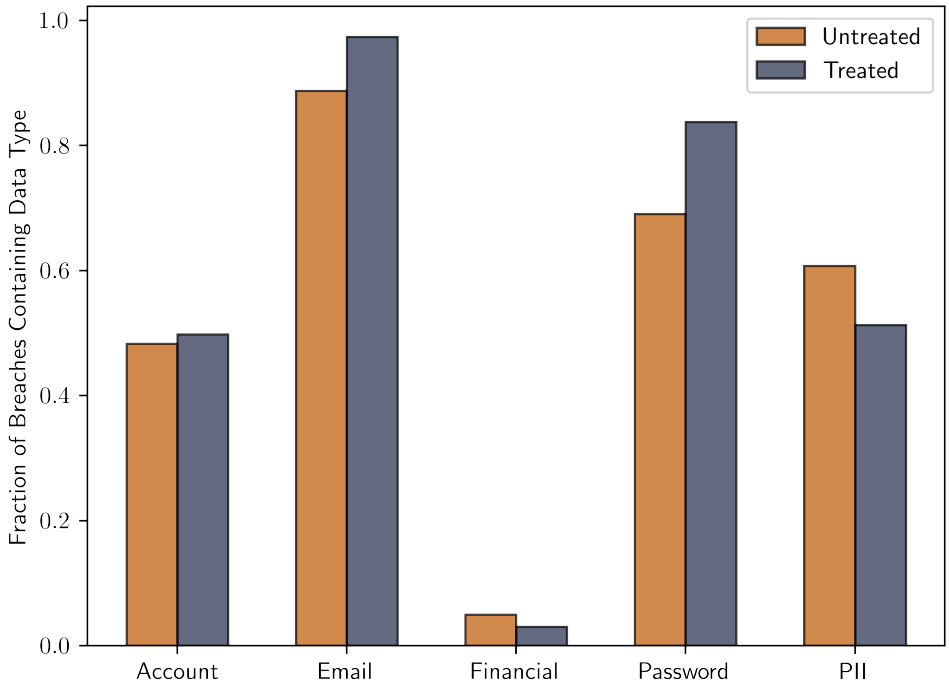
Notes: PII fraction is the fraction of records in a data package that are considered PII. A discussion of what constitutes PII is in the appendix.

Table 5 displays summary statistics for the three outcomes of interest in the study at

the data package level: the total number of records in a data package; the fraction of the data in a data package that is personally identifiable information (PII); and the number of data types in each package. A data type is, tautologically, a type of data. Examples are email addresses, credit card information, or whether the identified person owns a cat. PII has multiple legal definitions, but can be thought of as information that can identify and individual and may not be publicly known. A more in depth discussion of the definition of PII is in section B of the appendix.

Data packages vary greatly in terms of size, measured by the number of records. The largest contains over 700 million data points, while the smallest only one. Similarly, they range from having only one type of data to 55. Where they are more alike is in the fraction of records in the breach that are personally identifiable information. The 25th percentile breach is 50 percent PII, and 75th percentile breach has 85 percent PII. Emails, considered PII under the GDPR, are the most common type of data in these breaches, closely followed by passwords (figure 6).

Figure 6: Fraction of Data Packages Containing Each Data Type



Notes: The password category includes both individual passwords themselves, and information related to passwords, such as the salt used to help obscure them. Financial information includes bank, credit card, and loan data, Treated refers to all data packages originating in the EU, untreated to those originating outside the EU.

Because the GDPR applies to any entity collecting data on EU residents, not just those

in the EU, identifying treatment and control groups is difficult. For each data package, I observe either the country from which the data originate or the name of organization that was breached, and both for a subset of the observations. When I observe only the originating country, I assign the breach to that country. This makes treatment categorization simple: if the data originates in the EU, the package is treated. In the cases where I only observe the organization from which the data were stolen, I use one of two processes. First, I determine where the organization is headquartered. If they are an EU-based organization, the package is treated. If they are not, I search their privacy policy (where available) to see whether it has a section on European privacy laws. Those that do are categorized as treated. In cases where the organization is based outside the EU and lacks any indicators that they conduct business in the EU, I use a method similar to Demirer et al. (2024), who use firm's data server locations to categorize them into treatment and control units. I cannot observe data center locations, so instead I use the server locations for where they host their websites, as that location is endogenous to the location of an organization's users and customers.

Although it has largely been abstracted away, the internet is fundamentally a physical network. Data flows through fiber optic cables that span across oceans and continents to deliver content to users. This means the further a user is physically located from the server hosting the content, the longer it takes content to be delivered. The difference in time may only be fractions of a second, but that can still have a noticeable effect on outcomes organizations care about. Previous research has found that a 0.1 second improvement in website load time can increase spending on retails sites by almost 10% (Deloitte, 2020). For streaming and gaming sites, decreasing lag time improves user experience and can be used as a differentiating factor. Together, this creates an incentive for organizations to host their website on servers that are physically near their users to minimize load time.

There are two pieces of the internet's architecture key to connecting users with websites that allow me to observe where sites are physically located: DNS and GeoDNS. A Domain Name System (DNS) is essentially a phonebook for the internet. When a user types a domain name (e.g., www.fangraphs.com) into their web browser, it sends a query to the DNS, which then finds the IP address of the server hosting that website and connects it to the user. A GeoDNS does the same while taking into account the location of the user sending the initial query. For websites hosted in multiple locations, it will respond with the IP address of the server hosting the requested website that is closest to the user. As an example, suppose a website is hosted on one server in San Francisco, California and another Berlin, Germany. A user in Los Angeles will be connected with the San Francisco server, and a user in Frankfurt will be connected to the Berlin server.

To find where an organization hosts their website, I use the GeoNet API tool from Shodan,

an internet devices research company.[8] The GeoNet API allows me to send GeoDNS queries from six locations around the world to any website and record the IP addresses that respond to each request.[9] I conduct these queries for the website of each organization with a data package in my sample. After collecting the IP addresses of the responding servers, I use Shodan's IP address lookup tool to find the physical location of each one. Under this method, I categorize a breach as having come from a regulated entity if the organization hosts their website on at least one server in the EU. An organization that hosts their websites both in and outside the EU will also be considered regulated. For those packages that have not been manually assigned to a specific country, they are assigned to the country in which their originating organization hosts a majority of their servers. Table 6 breaks down the number of data packages that fall into each category before and after the GDPR.

Table 6: Group Counts

	Pre-GDPR N=1,621		Post-GDPR N=2,773	
	Non-EU	EU	Non-EU	EU
N	1,175	446	2,293	480

Data packages are only included in the final dataset if I can determine whether the originating organization is subject to the GDPR (or was in the cases where the organization is no longer active), and when the data are available online (pre- or post-GDPR). This sample represents only data that are posted online, not necessarily all data that is collected or stolen. It is possible that some stolen data packages are not traded, in which case I cannot observe them.[10] There are many reasons why a package may be unobservable. The hacker may decide they can profit more from using the data themselves than from publishing it. Or the hacker may have full access to the organization's data, but decide only a subset is worth taking and selling. In the case of ransomware, the victim organization may decide to pay the ransom to prevent their data from being leaked.

Using the country assigned to each package and the date it was available online, I aggregate the individual packages into a country-quarter level panel. The panel spans January 2017 to November 2023. Each country can be thought of as a market in the theoretical model in Section 2. The value and cost of hacking organizations in regulated countries will be affected by the GDPR, and remain unchanged in unregulated countries. Choosing Jan-

[8] https://geonet.shodan.io

[9] Requests are sent from servers in the United States, England, the Netherlands, Germany, India, and Singapore

[10] If only part of a data package is traded, I only observe what is traded, not everything that was stolen.

uary 2017 to November 2023 as the study's time frame means that data packages available online prior to 2017 are not included in the panel, even though I can observe full information on them. I make the choice to exclude pre-2017 periods because SpyCloud was started in 2016. This padding removes any bias that may occur if the packages collected early in their operation are fundamentally different from the ones discovered later. For consistency in the sample I also exclude these observations from the primary data package level analysis. To control for population in the analysis, I add annual population data from the World Bank to the panel.

For robustness checks, I construct additional panels excluding any period after March 2020—to remove any bias introduced by the COVID-19 pandemic, and any data package originating from a multinational organization. The latter removes any bias that may arise due to partially treated organizations.

Not every country experiences a breach in every period. For those observations, I assign a value of zero to the two outcome variables: number of data breaches and number of records. As shown in figures 7 and 8, this creates mass points at zero for both variables. I discuss the implications of this for my estimation strategy in section 4.

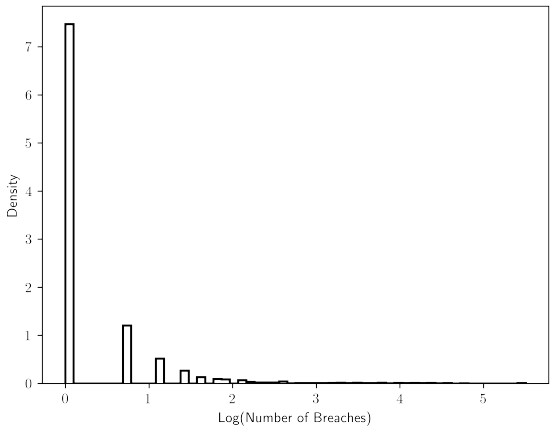
Roughly 27 percent of all country-quarters have a positive number of breaches (table 7), but among the positive observations there are an average of six breaches and 21.6 million records stolen (table 8). There is a large variation in both outcomes with as many as 245 breaches occurring and over one billion records being available in a quarter. Figures 9a and 9b show how the number of data breaches and number of records trended over time. There is a clear decline in the number of data breaches immediately after the GDPR went into effect, but no obvious and persistent change in the number of records becoming available in each quarter.

Table 7: Panel Summary Statistics

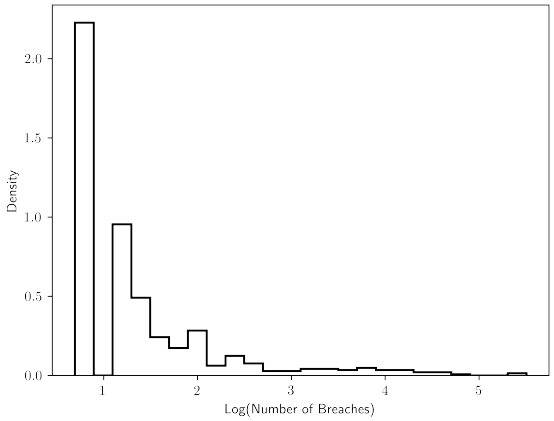
	Number of Breaches	Number of Records (M)	> 0 Breaches
Observations	2,716	2,716	2,716
Mean	1.618	5.73	0.265
Std. Dev.	9.522	49.73	0.442
Min.	0	0.00	0
25%	0	0.00	0
50%	0	0.00	0
75%	1	0.00	1
Max.	245	1,009.74	1

The observations dropped from the final datasets because either their country of origin

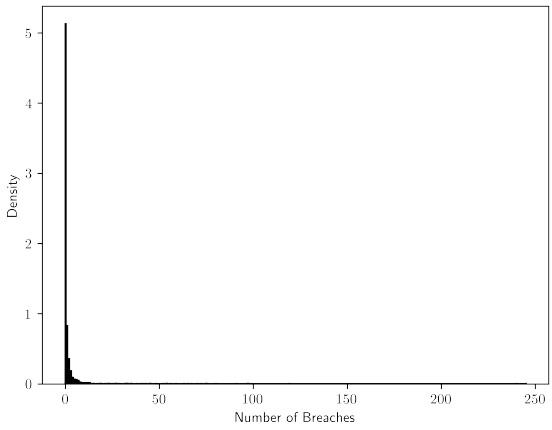
Figure 7: Distribution of the Aggregate Number of Data Breaches Per Country and Period



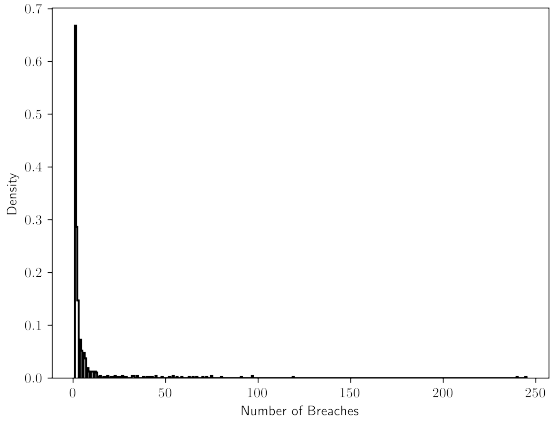
(a) Logged Using All Observations



(b) Logged Using Only Positive Observations



(c) In Levels Using All Observations

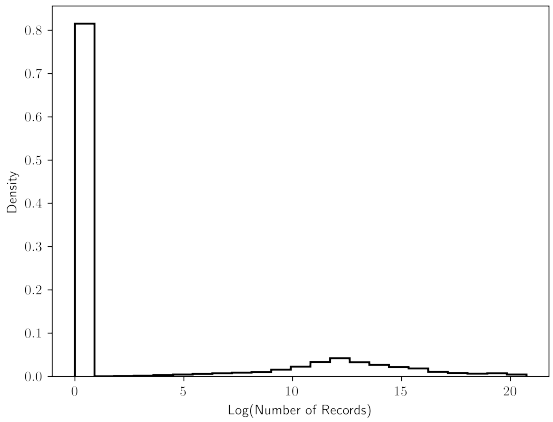


(d) In Levels Using Only Positive Observations

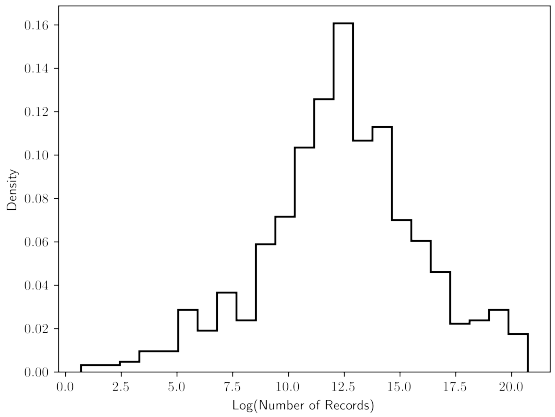
Table 8: Panel Summary Statistics - Non-Zero Periods Only

	Number of Breaches	Number of Records (M)
Observations	721	721
Mean	6.094	21.60
Std. Dev.	17.735	94.78
Min.	1	0.00
25%	1	0.04
50%	2	0.26
75%	4	1.90
Max.	245	1,009.74

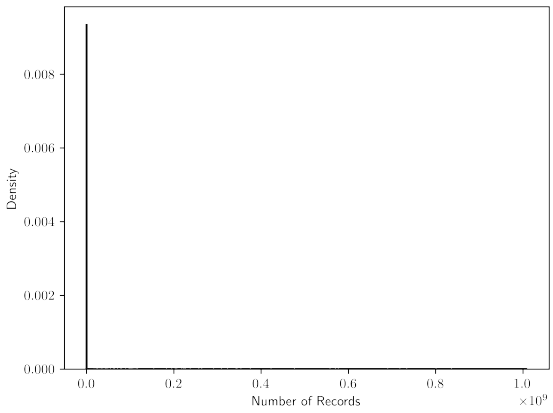
Figure 8: Distribution of the Aggregate Number of Records by Country



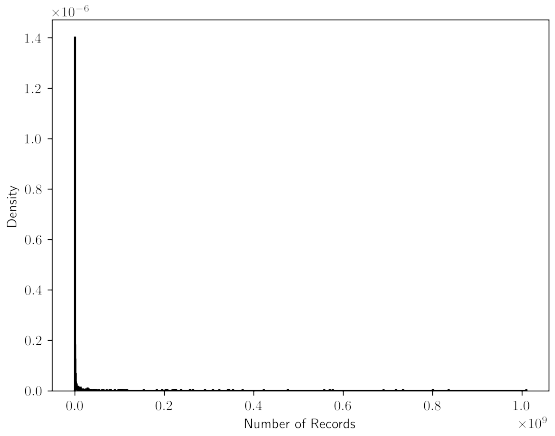
(a) Distribution of Log(Number of Records) Using All Observations



(b) Distribution of Log(Number of Records) Using Only Positive Observations



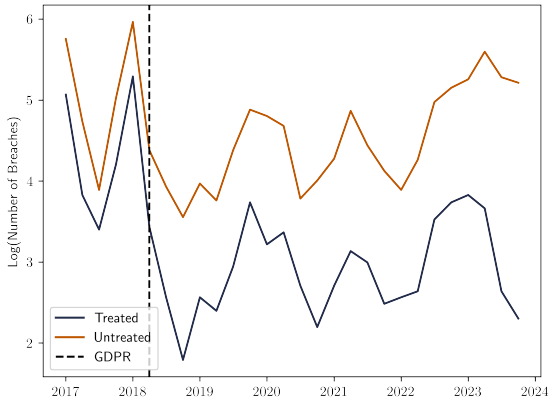
(c) Distribution of Number of Records Using All Observations



(d) Distribution of Number of Records Using Only Positive Observations

or breach date could not be determined tend to contain fewer records than those included in the study (figure 10). There are three periods where a large number of breaches were dropped: The first and second quarters of 2018, and the fourth quarter of 2020. In each of these periods there was a data breach whose contents were an amalgamation of data from many other smaller breaches. The 2020 breach specifically, known as the Cit0Day breach, was a collection of over 23,000 breaches websites bundled together. The Cit0Day website collected each of those smaller breaches and offered access to the information they contained for a fee. These observations are dropped because it is not possible to identify when these smaller breaches occurred. It is possible they were breaches that occurred years prior to the larger breach, or right before. Figure 11 plots the number of breaches and records included and excluded from the final sample over time.

(a) Number of Data Breaches



(b) Number of Records Online

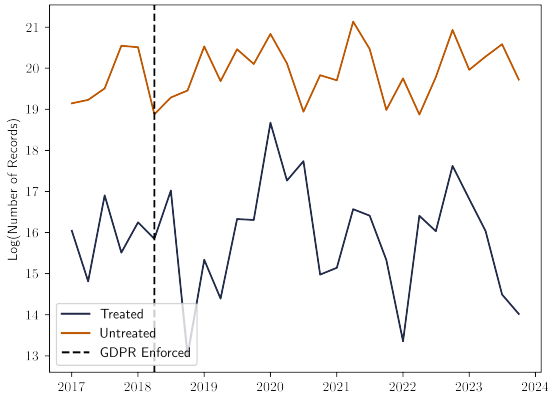
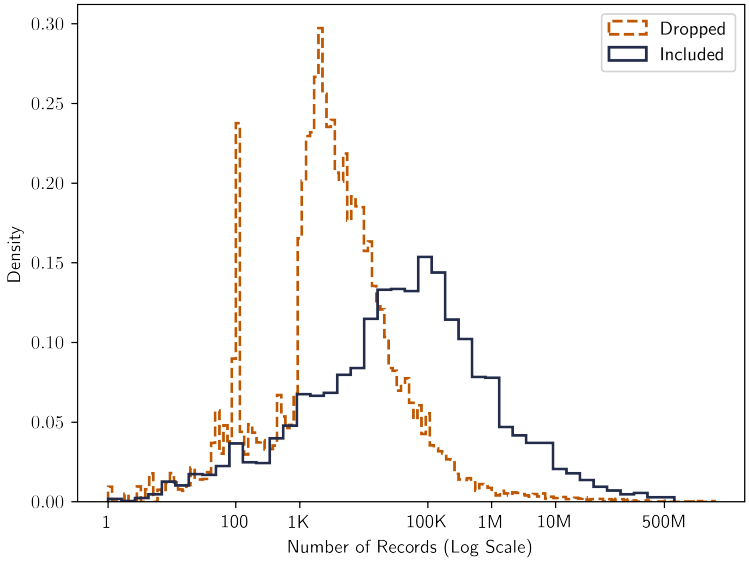


Figure 9: Number of Breaches and Records Time Series

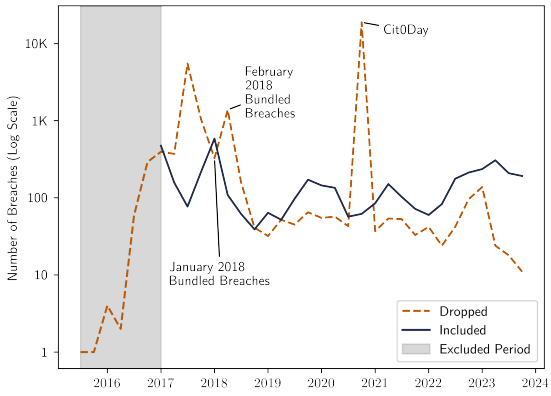
Figure 10: Distribution of Records Per Breach: Dropped vs. Included



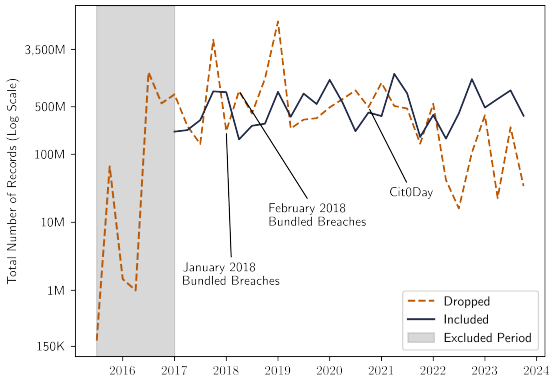
4 Empirical Strategy

I estimate the effects of the GDPR on aggregate quantities in the stolen data market, and the contents of the individual data packages traded. This allows me to test both predictions of my model. The model predicts there will be an unambiguous decrease in the number of data breaches after the GDPR—which will be tested by the aggregate analysis—and that any observed changes in the expected value of a breach will depend on whether the GDPR had a larger effect on high or low-valued targets. If the GDPR changed the costs and benefits of hacking low-valued targets more than high-valued, the expected value of a breach will increase. If high-valued targets are more affected, the expected value of a breach will fall.

Figure 11: Comparison of Dropped and Included Breaches Over Time



(a) Number of Breaches



(b) Number of Records

The individual data package analysis will test this by examining the effect of the GDPR on the amount and types of data included in the packages.

Treatment status in all cases is determined by where the data was originally collected, as discussed in Section 3, and the date the data package was available online. A data package is in the treatment group if it originated in the EU or was stolen from an organization subject to the GDPR and became available in June 2018 or later. This definition includes multinational organizations, such as large social media organizations, as treated if they have any users in the EU. As discussed in Demirer et al. (2024), this may complicate identification because these organizations may respond differently to the GDPR. The data they hold is partially treated since they likely hold information on individuals inside and outside the EU.[11]

4.1 Aggregate Effects

Aggregate effects are estimated using the country-quarter panel described in section 3. Each observation of country i is the aggregate of the individual data packages originating from that country in time period t. Most countries do not have a positive number of breaches in each period, creating a mass point at zero (figures 7 and 8). The model I present in Section 2 suggests that privacy regulations could affect the extensive margin because they change the relative value of breaching organizations in regulated countries, making them less likely to have a positive number of breaches in a given period. To measure the extensive margin effect, I estimate the linear probability model:

[11] Robustness checks excluding data packages from multinationals are in section C.2 of the appendix, and their findings are discussed in sections 5.1 and 5.2.

$$Positive_{it} = \gamma_i + \tau_t + \delta D_i \times Post\text{-}GDPR_t + \varepsilon_{it}.$$

where γ_i and τ_t are country and quarter fixed effects. D_i equals one if the country is in the EU, and $Post\text{-}GDPR_t$ equals one if the period is after the second quarter of 2018. The dependent variable, $Positive_{it}$, is an indicator for whether country i has at least one breach in period t.

To measure the impact of the GDPR on the number of breaches and total number of records available, I estimate the average treatment effect in levels as a percentage of the baseline mean:

$$\delta^{Agg\%} = \frac{E[Y(1) - Y(0)|D]}{E[Y(0)|D]}$$

where $Y(1)$ and $Y(0)$ are the outcomes with and without treatment, respectively. This is interpreted as the percentage change in the average outcome between regulated and unregulated countries.

The parameter δ^{Agg} is found using a Poisson model:

$$Y_{it} = \exp\left(\gamma_i + \tau_t + \delta^{Agg} D_i \times Post\text{-}GDPR_t - log(population_{it})\right) \varepsilon_{it} \tag{16}$$

where γ_i, τ_t, D_i, and $Post\text{-}GDPR_t$ are all defined as in the extensive model. To explicitly obtain the percentage change in the outcome, δ^{Agg} must be transformed: $\delta^{Agg\%} = \exp(\delta^{Agg}) - 1$. Standard errors are clustered at the country level. The offset, $log(population)$ is used to account for difference in sizes among the countries.

To test whether the effect changes over time, I break the $Post\text{-}GDPR_t$ term into short- and long-run effects, estimating:

$$Y_{it} = \exp\left(\gamma_i + \tau_t + \delta_{SR}^{Agg} Short\text{-}Run_t \times D_i + \delta_{LR}^{Agg} Long\text{-}Run_t \times D_i - log(population_{it})\right) \varepsilon_{it} \tag{17}$$

where $Short\text{-}Run_t$ equals one when $t \in \{\text{June 2018 – May 2019}\}$ and $Long\text{-}Run_t$ equals one for all periods after May 2019.[12]

The identifying assumptions underlying these models are conditional no anticipation, and that the growth rate between periods the treated group would have realized in the absence of treatment is the same as that experienced by the control group, i.e., there are parallel trends in the ratio of outcomes between periods (Wooldridge, 2023):

[12]The short and long-run definitions follow Demirer et al. (2024)

$$\frac{E[Y^{Post}(0)|D=1]}{E[Y^{Pre}(0)|D=1]} = \frac{E[Y^{Post}(0)|D=0]}{E[Y^{Pre}(0)|D=0]}.$$

To test this assumption, I estimate an event study model:

$$Y_{it} = \exp\left(\gamma_i + \tau_t + \sum_{t \neq -1} \delta_{it}^{Agg} D_i \times \textit{Post-GDPR}_t - log(population)\right)\varepsilon_{it} \tag{18}$$

where all notation is defined as before and standard errors are once again clustered at the country level.

Under the model in section 2, the increase in cost and decrease in value of breaching regulated organizations caused by the GDPR should cause the number of data breaches originating in regulated countries to decrease. All else equal, the number of records should decrease as well, but changes in which targets are hacked and which data packages are subsequently sold may blunt this effect. If high-value targets are less affected by the GDPR than low-value, the expected value of the remaining breaches increases, which could result in more data being available despite the decrease in the number of breaches.

I use a Poisson model rather than a log-like transformation because of the mass points at zero. In order to use log-like transformations on the outcomes, it would be necessary to either add a constant to each observation or use a transformation that is defined at zero, such as the inverse hyperbolic sine, to include the full sample in the estimation.

Mullahy and Norton (2024) show that log-like transformations significantly change the estimated marginal effects when zero mass points are present. Further, Chen and Roth (2023) find that, in the presence of zero mass points, if the treatment has extensive margin effects, the estimated average treatment effect is sensitive to the units of the outcome variable, making the interpretation of the estimates difficult. The framework I present in section 2 makes clear that privacy regulations should affect the extensive margin as it changes the relative value of breaching organizations in regulated countries, making them more or less likely to experience a positive number of breaches.

4.2 Data Package Effects

Effects on individual data packages are estimated using the linear model:

$$y_i = \gamma_i + \tau_t + \delta^{DP} D_i \times \textit{Post-GDPR}_t + \epsilon_{it} \tag{19}$$

where D_i equals one if the package originated from a regulated organization, and $\textit{Post-GDPR}_t$ indicates whether the data package was available June 1, 2018 or later. I use June 1, 2018

as the beginning treatment date, rather than the day the GDPR was enforced, to allow for a lag between when data became available online and when it was stolen.

The three outcomes of interest are the log of the total number of records in the package, amount of personally identifiable information (PII), and the number of unique types of data in the package. The parameter of interest is δ^{DP}.

I once again break the *Post-GDPR* term into short- and long-run effects and estimate:

$$y_i = \gamma_i + \tau_t + \delta^{DP}_{SR} D_i \times \textit{Short-Run}_t + \delta^{DP}_{LR} D_i \times \textit{Long-Run}_t + \varepsilon_i \tag{20}$$

where $\textit{Short-Run}_t$ and $\textit{Long-Run}_t$ are defined as they were in the aggregate effects section. This allows for changes in the behavior of both those collecting and stealing data. The former may increase their compliance with the regulation. The latter may change who they decide to target in response to changes in data collection and security practices.

The expected effects on individual data packages are ambiguous under the model in section 2. All breaches are expected to be less valuable after the GDPR, This would imply they contain fewer records, PII, and data types. However, if the GDPR disproportionally drives low-value targets out of the profitable target set, then the expected value of a breach may increase, even if the total number of data breaches falls. Given the restrictions on collecting PII, the fraction of all records that are PII might be expected to decrease, but that will also depend on the effects of the regulation on non-PII data collection as well.

5 Results

The main results are presented in sections 5.1 and 5.2. I discuss the results in the context of the model along with the limitations of this paper in section 5.3. Robustness checks and alternative model specifications are discussed in section 5.4.

5.1 Aggregate Effects

On the extensive margin, I find the GDPR is associated with a roughly 21 percent decline in the probability of finding a data breach that originates from a regulated country online (table 9). This effect is larger in the long-run than short-run (-22 percent versus -17 percent, respectively).

The aggregate treatment effects on the number of breaches and total amount of data being taken from a country are presented in table 10. I find that the number of data breaches fell approximately 54 percent and 61 percent in the short and long run, respectively. This result is consistent with the predicted effects of both a decrease in the amount of data collected

Table 9: Extensive Margin Effects

	Dependent Variable: Positive Number of Breaches	
	(1)	(2)
Post x Treatment	-0.209***	
	(0.040)	
SR x Treatment		-0.171***
		(0.051)
LR x Treatment		-0.218***
		(0.040)
Observations	2,716	2,716
R^2	0.469	0.469
Period Fixed Effects	Y	Y
Country Fixed Effects	Y	Y

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level.

and an increase in security investment by regulated organizations on the market. Fewer organizations are worth hacking, so there is a decrease in the number of data breaches. The same logic applies to my extensive margin findings.

Despite the large decrease in the number of breaches, I find no statistically significant change in the number of records in the market. Mechanically this only possible if the remaining breaches contain significantly more data, which my data package-level analysis finds. This could be caused by higher-value targets with more data becoming a larger share of the breaches traded in the market.

Event study plots to provide evidence that the parallel trends assumption holds are in figure 12. For number of data breaches, the coefficient estimates for each period prior to the GDPR have zero in the 95 percent confidence interval, while post GDPR there is a clear decrease in the number of data breaches (figure 12a). Each period of the event study shows no statistically significant effect on the number of records traded (12b).

5.2 Individual Data Package Content Effects

At the individual breach level, I find that data packages originating in regulated organizations increased in size nearly 70 percent, as measured by number of records they contain (column four of table 11). This effect is driven by long run changes, with there being a positive but statistically insignificant change in the number of records in the short run. An increase in the size of the data packages is counterintuitive on its face. If data privacy legislation

Table 10: Aggregate Effects

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-0.921***		0.345	
	(0.265)		(0.430)	
SR x Treatment		-0.782***		-0.217
		(0.299)		(0.590)
LR x Treatment		-0.934***		0.410
		(0.283)		(0.430)
$\hat{\delta}$	-0.602		0.412	
	(0.105)		(0.606)	
$\hat{\delta}^{SR}$		-0.543		-0.195
		(0.137)		(0.475)
$\hat{\delta}^{LR}$		-0.607		0.507
		(0.111)		(0.647)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716
Pseudo R^2	0.792	0.792	0.847	0.847

**p<0.1, **p<0.05, ***p<0.01*

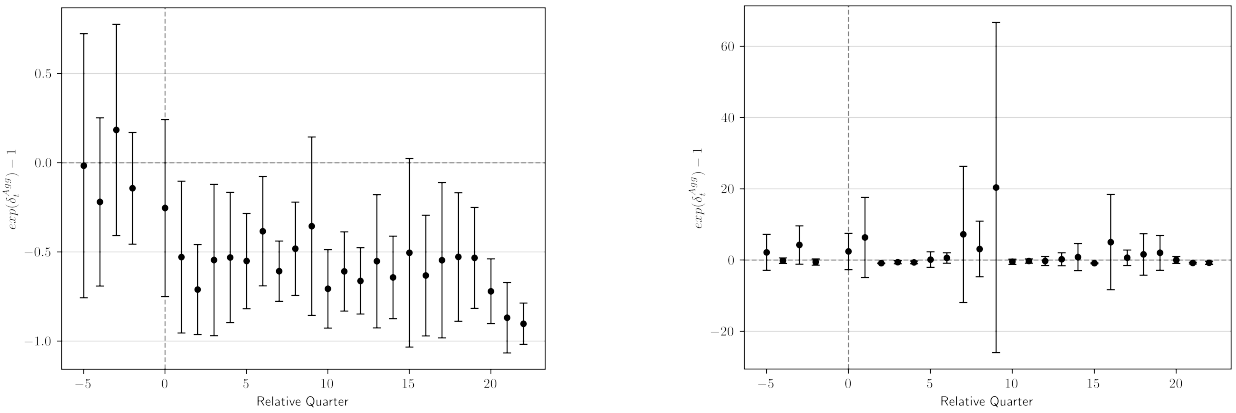
Notes: Standard Errors are clustered at the country level.

successfully reduces data collection, which it appears to do, then it seems natural that there would be a corresponding reduction in the number of records included in the packages. Less data collected means there is less data to steal. But if, as discussed in section 2.4, the GDPR drives low-value breaches out of the market and brings more high-value breaches into the market, then the expected value of the remaining breaches increases even after accounting for the change in value caused by the GDPR. These breaches would contain larger amounts of data, increasing the expected number of records in any given breach. Figure 13 shows that the distribution of the number of records in a breach shifted right after the GDPR.

Looking specifically at the amount of PII in a breach, I find that the number of records that constitute PII increased by 63 percent in the long-run (table 12). Given that most of the data in the packages qualifies as PII (table 5), this is expected with the increase in the overall number of records per package.

These are the only statistically significant change at the data package level. I find no change in the fraction of records that are PII (table 13) or number of unique types of data in the packages (table 14). One potential explanation for this is that only certain types

Figure 12: Aggregate Effect Event Studies



(a) Number of Data Breaches

(b) Number of Records Available

Notes: The figures present estimates of the δ_{it}^{Agg} coefficients in equation 18 converted to percentage changes using $exp(\delta_{it}^{Agg}) - 1$. The bars are the 95 percent confidence intervals with standard errors clustered at the country level. Period $t = -1$, the first quarter of 2018, is normalized to be zero.

of data have value in the stolen data market. If the data no longer collected by regulated organizations is not considered valuable in this other market, it is unlikely that there would be an effect on data package contents beyond their size. Higher-value targets becoming a larger share of the market also explain these findings.

Figure 13: Number of Records Density

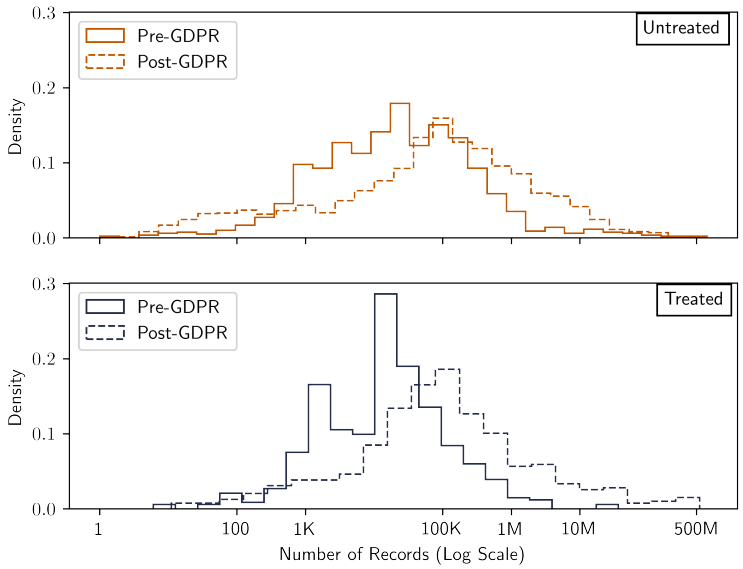


Table 11: Data Package Effects: Number of Records

	Dependent Variable: Log(Number of Records)			
	(1)	(2)	(3)	(4)
Post x Treatment	0.959**		0.513*	
	(0.427)		(0.260)	
SR x Treatment		0.470		0.090
		(0.379)		(0.266)
LR x Treatment		0.931**		0.508**
		(0.398)		(0.249)
Multinational			1.380***	1.398***
			(0.248)	(0.253)
$\hat{\delta}$	1.610		0.670	
	(1.114)		(0.435)	
$\hat{\delta}^{SR}$		0.600		0.095
		(0.606)		(0.291)
$\hat{\delta}^{LR}$		1.538		0.662
		(1.011)		(0.413)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	4,394	4,394	4,394	4,394
R^2	0.268	0.268	0.276	0.276

*p<0.1, **p<0.05, ***p<0.01

Notes: Standard errors are clustered at the country level.

Table 12: Data Package Effects: Number of PII Records

	Dependent Variable: Number of PII Records			
	(1)	(2)	(3)	(4)
Post x Treatment	0.937**		0.486*	
	(0.424)		(0.266)	
SR x Treatment		0.571		0.190
		(0.409)		(0.282)
LR x Treatment		0.914**		0.490*
		(0.392)		(0.255)
Multinational			1.394***	1.402***
			(0.268)	(0.270)
$\hat{\delta}$	1.552		0.626	
	(1.082)		(0.432)	
$\hat{\delta}^{SR}$		0.770		0.210
		(0.723)		(0.341)
$\hat{\delta}^{LR}$		1.495		0.632
		(0.978)		(0.417)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	4,394	4,394	4,394	4,394
R^2	0.270	0.270	0.277	0.277

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level.

Table 13: Data Package Effects: PII Fraction

	Dependent Variable: PII Fraction			
	(1)	(2)	(3)	(4)
Post x Treatment	-0.010		-0.013	
	(0.010)		(0.016)	
SR x Treatment		-0.006		-0.008
		(0.023)		(0.021)
LR x Treatment		-0.008		-0.010
		(0.013)		(0.018)
Multinational			0.009	0.008
			(0.020)	(0.020)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	4,394	4,394	4,394	4,394
R^2	0.415	0.415	0.415	0.415

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. PII fraction is the number of records in a data packages considered PII divided by the total number of records in the data package.

Table 14: Data Package Effects: Number of Data Types

	Dependent Variable: Number of Unique Data Types			
	(1)	(2)	(3)	(4)
Post x Treatment	0.383		0.350	
	(0.238)		(0.264)	
SR x Treatment		0.397		0.376
		(0.512)		(0.506)
LR x Treatment		0.426		0.403
		(0.270)		(0.285)
Multinational			0.102	0.077
			(0.234)	(0.228)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	4,394	4,394	4,394	4,394
R^2	0.280	0.280	0.280	0.280

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level.

5.3 Discussion and Limitations

The above results are consistent with what the model in section 2 predicts should happen after a privacy policy goes into effect. On the aggregate side, the GDPR reduces the value and increases the cost of hacks, causing there to be fewer breaches. At 60 percent, the reduction I find is large, suggesting the combined value and cost effects are substantial. The model predicts that, if the change in value and cost of hacking disproportionally affects low-value targets, high-value targets will make up a larger share of post-GDPR breaches, resulting in an increase in the expected value of the breaches that remain. My data package-level findings support this. The value of a breach is a function of both the types of data and size of the breach. Given that I find no change in the fraction of records that are PII or number of unique data types in these breaches, and a large increase in the number of records they include, the results suggest that value increased on average. While I cannot directly estimate the parameters of the model, this would imply that the change in breach cost, ξ, is smaller among high-value targets than low. If the change in value ϕ also varies with V, the two are likely to be negatively correlated as well.

Implicit in my model is the assumption that hacker skill remains constant. If hackers were to become more productive, the cost of hacking would decrease, resulting in more breaches, but the expected value may decrease as relatively low-value targets become viable marks now that they are cheaper to hack. That I find a decrease in the number of and increase in the quality of breaches suggests this is not a concern. However, I do not observe any direct measure of hacker ability and therefore cannot fully rule out the possibility that it has changed.

Finally, estimating the overall welfare impact of the GDPR with regard to its effects on cyber crime is beyond the scope of this paper. That said, reducing the number of data breaches is likely beneficial to those not looking to buy or sell them. The extent of that benefit may be limited given that the number of records did not change. With the same amount of data available, individuals may be no better off than they were before. To test this, one would need to calculate how the probability of a person's data being online has changed, or at least count the number of unique individuals with data in each breach, which I am unable to do with my data.

Another factor that will determine individual welfare effects is by whom their stolen data are used. Returning to the model, this market only exists if buyers have a sufficiently high comparative advantage over sellers in exploiting data. Reducing the number of traded breaches may therefore also reduce data access for those who are particularly adept at data exploiting it. If each person only appears once in each data package, then even as the data packages grow larger and include more people, each individual is made better off because of this.

On the cyber criminal side of the problem, the GDPR may have made them better off in some cases. As shown in my simulated experiment, if the GDPR alleviates part of the adverse selection problem in the market, buyers of stolen data are actually better off after the policy. Hackers are universally worse off after the GDPR, though they do receive a higher price for what they sell. If more detailed price data become available, future research could attempt to assess whether reality matches the simulation.

5.4 Robustness

To check the robustness of my results, I re-estimated the aggregate effects using a number of alternative samples and model specifications.

On the extensive margin, to test whether the extensive margin findings are driven by small countries with few breaches, I split the analysis into two groups: countries with populations below the median in 2018 and countries with populations above the median. I find that the

extensive margin effect is slightly larger in the small country panel than the large country panel. The former experiences a 22 percentage point decrease in the probability of having a positive number of breaches while the latter has a 17 percent decrease (column one of appendix tables A6 and A7). The short-run extensive margin effects for large countries are also statistically insignificant while there was a 23 percent decrease among small countries in this period (column two of appendix tables A6 and A7). These results suggest that some of the extensive margin effects are driven by smaller countries.

In my main specification, I use the log of the country's population as an offset in the Poisson model to account for differences in population size. Appendix table A17 shows that removing the offset has no effect on the estimation. Using population to weight the model in lieu of the offset increases the estimated decrease in the number of breaches to 67 percent, still within the standard error of the main results, while there is still no statistically significant change in the number of records.

Converting the two outcome variables to be in per capita terms (breaches per capita and number of records per capita) increases the estimated decrease in the number of breaches to 76 percent overall and 77 percent in the long-run. However, converting the outcomes to per capita terms changes them from discrete to continuous variables, making a Poisson model inappropriate to use.

Using the same panel, I compare the Poisson difference-in-differences results to those of linear models with log-like transformations of the outcomes of interest the outcomes in levels in appendix tables A11-A16. The two log-like transformations used are $Log(Y+1)$ and the inverse hyperbolic sine. When the outcome is in levels, I use number of breaches per million and number of records per thousand to make the coefficients more interpretable.

Across all models and outcome specifications, there is a negative and significant effect on the number of breaches. The effect falls from a 61 percent decrease to as low as a 10 percent decrease in the number of breaches when using the $Log(Y+1)$ transformation and breaches per capita as the outcome. Chen and Roth (2023) show that when log-like transformations are used on data with a mass point at zero, the coefficient estimates will be arbitrarily sensitive to the units of the outcome variable, explaining this discrepancy. For all other models where the outcome is not in per capita terms, the estimated treatment effects fall between my estimated extensive margin effects and the treatment effect estimated with the Poisson model. Mullahy and Norton (2024) show that log-like transformations with mass points at zero will be a weighted average of the extensive and intensive margins effects, which likely explains these differences. Finally, the levels outcomes are not directly comparable to the Poisson since they are not percentage changes, but they are negative and significant. The Poisson estimates are the levels change as a percentage of the control mean, so this

result simply confirms that the Poisson effect is valid.

Where model selection matters is in estimating the treatment effect on number of records. The Poisson and levels models show no statistically significant change in the number of records across all specifications. When a log-like transformation is applied to the outcome variable I consistently find large and significant decreases in the number of records. However, as previously discussed, log-like transformations are unreliable when the outcome has a mass point at zero. Additionally, given that there is no effect in levels (table A12) and there is no obvious change in the number of records available overtime (figure 9b), it is unlikely that the results with log-like transformations are dependable.

In the remaining robustness analysis, I change how to panel is constructed. First, I remove all observations from Brazil and China from the panel. Brazil and China implemented data privacy regulations of their own near the end of the study. Removing these observations slightly lowers the estimated treatment effect on the number of breaches to a 56 percent decrease, though this still falls within the standard error of the original estimates. There is still no significant effect on the number of records (table A8).

Next, I excluded all periods after the first quarter of 2020 to remove any noise brought on by the COVID-19 pandemic. During the pandemic, organizations may have been more vulnerable to cybersecurity incidents if they did not have the proper infrastructure in place to safely operate remotely. For example, they may have lowered some of the barriers needed to access company databases in order for their employees to work from home, making it easier for those databases to be improperly accessed. While the pandemic was a global shock, differences in lock down dates and enforcement may have caused some country-level variation that would not be accounted for by the time or country fixed effects. Without the COVID era observations, I find a 48 percent decrease in the number of data breaches. This is still large and statistically significant, but smaller than the result in my main specification. As in the main results, I still find that there is no statistically significant change in the number of records available. The parameters estimated are presented in appendix table A9.

Multinational organizations introduce a challenge to this study because it is not immediately obvious which country to assign their breaches and data. Because the GDPR extends beyond EU borders and applies to all organizations that collect data on EU residents, those who collect data on individuals both in and outside the EU are effectively partially treated. To the best of my knowledge, there is no definitive research on whether these organizations treat all of their data equally, giving the same protections the GDPR provides to EU residents to their non-EU users, or whether they have distinct processes for handling EU data.[13]

[13]In the course of writing this paper I have read the privacy policies of many multinational organizations. Some have a single privacy that applies to all users. These typically include a section specifically for EU

To test whether these organizations are significantly influencing the aggregate outcomes, I remove all breaches of multinational organizations from the data prior to aggregating the individual breaches into the panel. I find a 60 percent decrease in the number of data breaches, roughly the same as my main specification. For the number of records, the total and short-run effects are once again statistically insignificant, but in the long-run I find a 124 percent increase in the number of records, significant at the ten percent level. The full results are in appendix table A10.

Finally, to check whether the results are driven by any one country in the EU, I repeatedly re-estimate the model removing one EU country at a time. As shown in appendix figures A3 and A4, the treatment effect estimates are well within the 95 percent confidence interval of the main model estimates each time.

At the data package level, I removed emails from the definition of PII to see if there was a change in the amount of non-email data as a portion of all the records in a package. I still find no change. Next, I re-estimated the model for each outcome variable using the full sample of breaches, rather than just breaches from 2017 and beyond. These early period breaches were dropped from the main analysis because they happened either before SpyCloud's founding or early in their lifetime, and may be different from the breaches collected after SpyCloud's monitoring infrastructure was well established. I find once again the number of records in a package increases in the long-run. The point estimate shows an 80 percent increase versus 67 percent in the main model, though is still within the standard error (table A27). There is once again no effect on the fraction of records in a breach that are PII, but now the number of data types increases by 0.56 post-GDPR (tables A28 and A29). Though statistically significant, a half of a data type increase holds little economic value.

6 Conclusion

As organizations continue to collect large amounts of data, the risk of that data being stolen and sold with be ever-present. In this paper I have shown that data protection regulations can have a significant effect on the illicit market for data.

I estimate that the GDPR is associated with a 60 percent reduction in the number of data breaches originating in EU countries available in stolen data markets. There is however no accompanying reduction in the number of individual records in these markets, as the size of data packages increased nearly 70 percent as well. I find no other changes in the contents of the data packages. The model of stolen data production I propose shows that one potential

residents. Others have different privacy policies for every country they operate in. The European policies detail the rights those users have over their data, the non-European ones do not.

explanation of these effects is low-value targets disproportionally falling out of the profitable target set, increasing the expected value of all remaining breaches.

This paper partially fills the gap in the literature on privacy regulation, and the GDPR in particular, regarding potential benefits of these regulations. It is the first to study the effects of privacy regulation on the stolen data market and show a causal impact.

There are many paths forward for future research on this topic. My model can be generalized and solved with alternative distributions of target value and hacking cost, or assumptions about how privacy regulations affect both. Additionally, my model suggests only one of many possible explanations for my empirical findings. Qualitative and quantitative work on the abilities and behaviors of hackers could provide insights into whether the effects I observe empirically are driven strictly by the changes in hacker incentives and buyer expectations I propose, or if there are other factors, such as changes in hacker still, at play.

This paper is missing a key component of the market: prices. Although there are many hurdles to collecting quality price data in these markets, doing so would open the door to a more complete analysis of their inner workings and the value hackers place on certain types of information.

Finally, while there have been a number of studies on the effects of the GDPR on specific firm outcomes, changes in data collection, and now cyber crime, there is still no overarching study of its overall welfare effects or how individual components of the policy influence outcomes of interest. With more countries considering and adopting data privacy regulations, research on this subject would have high returns in the debate over how to design future policy.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Aridor, G., Che, Y.-K., and Salz, T. (2021). The effect of privacy regulation on the data industry: Empirical evidence from gdpr. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 93–94, New York, NY, USA. Association for Computing Machinery.

Athey, S., Catalini, C., and Tucker, C. (2017). The digital privacy paradox: Small money, small costs, small talk. (23488). DOI: 10.3386/w23488.

Ayres, I. and Levitt, S. D. (1998). Measuring positive externalities from unobservable vic-

tim precaution: An empirical analysis of lojack. *The Quarterly Journal of Economics*, 113(1):43–77.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(2):169–217.

Braakmann, N., Chevalier, A., and Wilson, T. (2024). Expected Returns to Crime and Crime Location. *American Economic Journal: Applied Economics*, 16(4):144–160.

Chen, C., Frey, C., and Presidente, G. (2022). Privacy regulation and firm performance: Estimating the gdpr effect globally*.

Chen, J. and Roth, J. (2023). Logs with zeros? some problems and solutions*. *The Quarterly Journal of Economics*, 139(2):891–936.

Cong, L. W., Harvey, C. R., Rabetti, D., and Wu, Z.-Y. (2023). An anatomy of crypto-enabled cybercrimes. (30834). DOI: 10.3386/w30834.

Deloitte (2020). Milliseconds make millions. Technical report, Deloitte.

Demirer, M., Jiménez Hernández, D. J., Li, D., and Peng, S. (2024). Data, privacy laws and firm production: Evidence from the gdpr. (32146). DOI: 10.3386/w32146.

Department for Digital, Culture, Media and Sport (2022). Cyber Security Breaches Survey: Combined Dataset, 2016-2022. data collection. SN: 8971, DOI: http://doi.org/10.5255/UKDA-SN-8971-1.

Franklin, J., Paxson, V., Perring, A., and Savage, S. (2007). An inquiry into the nature and causes of the wealth of internet miscreants. In *Proceedings of the 14th ACM conference on Computer and communications security*, CCS '07, page 375–388, New York, NY, USA. Association for Computing Machinery.

Goldberg, S. G., Johnson, G. A., and Shriver, S. K. (2024). Regulating privacy online: An economic evaluation of the gdpr. *American Economic Journal: Economic Policy*, 16(1):325–358.

Goldfarb, A. and Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, 57(1):57–71.

Gordon, L. A. and Loeb, M. P. (2002). The economics of information security investment. *ACM Transactions on Information and System Security*, 5(4):438–457.

Holt, T. J. and Lampke, E. (2010). Exploring stolen data markets online: products and market forces. *Criminal Justice Studies*, 23(1):33–50.

Holt, T. J., Smirnova, O., and Chua, Y. T. (2016). Exploring and estimating the revenues and profits of participants in stolen data markets. *Deviant Behavior*, 37(4):353–367.

Janßen, R., Kesler, R., Kummer, M. E., and Waldfogel, J. (2022). Gdpr and the lost generation of innovative apps. (30028). DOI: 10.3386/w30028.

Jia, J., Jin, G. Z., and Wagman, L. (2021). The short-run effects of the general data protection regulation on technology venture investment. *Marketing Science*, 40(4):661–684.

Johnson, G. A., Shriver, S. K., and Goldberg, S. G. (2023). Privacy and market concentration: Intended and unintended consequences of the gdpr. *Management Science*.

Kircher, T. and Foerderer, J. (2021). Does eu-consumer privacy harm financing of us-app-startups? within-us evidence of cross-eu-effects. (4058437).

Koski, H. and Valmari, N. (2020). *Short-term Impacts of the GDPR on Firm Performance*. Number 77.

Lukic, K., Miller, K. M., and Skiera, B. (2023). The impact of the general data protection regulation (gdpr) on online tracking. (4399388).

Miller, A. R. and Tucker, C. (2009). Privacy protection and technology diffusion: The case of electronic medical records. *Management Science*, 55(7):1077–1093.

Miller, A. R. and Tucker, C. (2018). Privacy protection, personalized medicine, and genetic testing. *Management Science*, 64(10):4648–4668.

Miller, A. R. and Tucker, C. E. (2011). Encryption and the loss of patient data. *Journal of Policy Analysis and Management*, 30(3):534–556.

Mullahy, J. and Norton, E. C. (2024). Why transform y? the pitfalls of transformed regressions with a mass at zero. *Oxford Bulletin of Economics and Statistics*, 86(2):417–447.

Schwartz, P. M. and Solove, D. J. (2011). The pii problem: Privacy and a new concept of personally identifiable information. *New York University Law Review*, 86(6):1814–1894.

SpyCloud (2024). SpyCloud Annual Identity Exposure Report 2024. Technical report, SpyCloud.

Vu, A. V., Hughes, J., Pete, I., Collier, B., Chua, Y. T., Shumailov, I., and Hutchings, A. (2020). Turning up the dial: the evolution of a cybercrime market through set-up, stable, and covid-19 eras. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, page 551–566, New York, NY, USA. Association for Computing Machinery.

Wooldridge, J. M. (2023). Simple approaches to nonlinear difference-in-differences with panel data. *The Econometrics Journal*, 26(3):C31–C66.

Appendices

A Model Derivations

A.1 Legal Data Collection

The objective of organizations is to generate information at the lowest cost. Information is generated by collecting data, which has a cost in itself, and also carries the risk of being stolen. If data is stolen, organizations will face additional costs. These costs are related to sending out breach notifications, conducting post-incident audits, fines imposed by the government, and legal fees.

Each organization faces the optimization problem:

$$\max_{d_1,\cdots,d_J,S} \quad A\left(\alpha_1 d_1^{\rho} + \ldots + \alpha_J^{\rho} d_J\right)^{\frac{\nu}{\rho}} - \sum_{j=1}^{J}(\omega_j d_j) - \omega_s S - \frac{r}{S+1}\left(\ell + \sum_{j=1}^{J}\gamma_j d_j\right).$$

The two data type case presented in the main body is:

$$\max_{d_1,d_2,S} \quad A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu}{\rho}} - \omega_1 d_1 - \omega_2 d_2 - \omega_S S - \frac{r}{S+1}\left(\ell + \gamma_1 d_1 + \gamma_2 d_2\right). \tag{21}$$

The first order conditions with respect to S, d_1, and d_2 are:

$$\omega_S = \frac{r}{(S+1)^2}\left(\ell + \gamma_1 d_1 + \gamma_2 d_2\right) \tag{22}$$

$$\frac{\omega_1 + \frac{r}{S+1}\gamma_1}{\alpha_1} d_1^{1-\rho} = \nu A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu-\rho}{\rho}} \tag{23}$$

$$\frac{\omega_2 + \frac{r}{S+1}\gamma_2}{\alpha_2} d_2^{1-\rho} = \nu A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu-\rho}{\rho}} \tag{24}$$

Equation 22 can be rearranged to obtain the optimal S:

$$S^* = \sqrt{\frac{r\left(\ell + \gamma_1 d_1^* + \gamma_2 d_2^*\right)}{\omega_S}} \tag{25}$$

Setting the left-hand sides of equations 23 and 24 equal and solving for d_2 in terms of d_1 yields:

$$d_2 = \left[\frac{\alpha_2}{\omega_2 + \frac{r}{S+1}\gamma_2} \frac{\omega_1 + \frac{r}{S+1}\gamma_1}{\alpha_1} \right]^{\frac{1}{1-\rho}} d_1. \tag{26}$$

Which can be substituted into equation 23:

$$\frac{\omega_1 + \frac{r}{S+1}\gamma_1}{\alpha_1} d_1^{1-\rho} = \nu A \left(\alpha_1 d_1^{\rho} + \alpha_2 \left[\frac{\alpha_2}{\omega_2 + \frac{r}{S+1}\gamma_2} \frac{\omega_1 + \frac{r}{S+1}\gamma_1}{\alpha_1} \right]^{\frac{\rho}{1-\rho}} d_1^{\rho} \right)^{\frac{\nu-\rho}{\rho}}.$$

Factoring out d_1^{ρ} and $\left[\frac{\omega_1 + \frac{r}{S+1}\gamma_1}{\alpha_1}\right]^{\frac{\rho}{1-\rho}}$ then simplifying the resulting equation gives the optimal selection of d_1:

$$d_1^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1} \right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1} \right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{\rho}{1-\rho}} \right]^{\frac{\nu-\rho}{\rho(1-\nu)}}. \tag{27}$$

which gives the optimal d_2 when inserted into 26:

$$d_2^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1} \right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{\rho}{1-\rho}} \right]^{\frac{\nu-\rho}{\rho(1-\nu)}}. \tag{28}$$

While not a closed form solution, equations 25, 27, and 28 do show that optimal data collection is decreasing in both costs (ω_i and γ_i) and risk (r). The optimal level of security investment is increasing in both fundamental risk and costs associated with a breach.

A.2 Stylized Example

Assuming that $(V, C) \sim Uniform[0,1]^2$, the expected quality of V given $V > C$ is

$$\begin{aligned}\mathbb{E}\left[V\middle|V \geq C\right] &= \int_0^1 2V^2 dV \\ &= \frac{2}{3}\end{aligned}$$

Hackers will only sell the data they steal if the price they receive is higher than the utility they gain from holding the data. With hacker utility given by

$$U^H = M + \sum_{i=1}^{\mathcal{B}^H} V_i,$$

they will only sell data package i if $p \geq V_i$. The expected quality of the breaches they sell is then

$$\begin{aligned}\mathbb{E}\left[V\middle|C \leq V \leq p\right] &= \int_0^p V^2 \frac{2}{p^2} dV \\ &= \frac{2}{3}p \\ &= \mu\end{aligned}$$

where μ is buyer's expectation of quality given that the data are being sold.

Buyer utility is given by

$$U^B = M + \sum_{i=1}^{\mathcal{B}^B} \kappa V_i.$$

They will only buy data if $\kappa\mu \geq p$. In this example, κ must be at least $3/2$ for the market to exist. With a total income of Y, buyer's demand for data is:

$$D(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3}{2} \\ 0 & \text{Otherwise} \end{cases} \tag{29}$$

And supply is

$$\begin{aligned}S(p) &= \mathcal{B}\mathbb{P}\left(V \leq p\right) \\ &= \mathcal{B}p^2.\end{aligned} \tag{30}$$

Setting equations 29 and 30 equal and solving for p gives the equilibrium price:

$$p^* = \left(\frac{Y}{\mathcal{B}}\right)^{\frac{1}{3}}.$$

And equilibrium quantity:

$$Q^* = Y^{2/3}\mathcal{B}^{1/3}.$$

After the GDPR, quality for all targets falls and the cost of hacking increases to

$$\begin{aligned} V_i^{Post} &= (1-\phi)V_i \quad 0 < \phi < 1 \\ C_i^{Post} &= \xi C_i \quad \xi \geq 1 \end{aligned}$$

Assuming $\xi_i = \theta V_i^{\sigma}$ and ϕ is constant, the zero profit line is now given by

$$V^{1-\sigma} = \frac{\theta}{1-\phi}C$$

Integrating the above along the Y-axis shows that the joint probability distribution of V and C is

$$f_{VC}(V,C) = \begin{cases} \frac{\theta(2-\sigma)}{1-\phi} & \text{if } 0 \leq V \leq 1 \text{ and } 0 \leq C \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

And the marginal distribution of V is

$$f_V = (2-\sigma)V^{1-\sigma}$$

The expectation of V among the hacked is now

$$\begin{aligned} \mathbb{E}\left[V \middle| V \geq \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}}\right] &= \int_0^1 (2-\sigma)V^{2-\sigma}dV \\ &= \frac{2-\sigma}{3-\sigma}. \end{aligned}$$

Hackers utility after accounting for the overall decrease in value is

$$U^{H,Post} = M + \sum_{i=1}^{\mathcal{B}^{H,Post}} (1-\phi)V_i.$$

They will only sell what they steal if $(1-\phi)V_i \leq p$. The joint probability distribution over

this area of the curve is

$$f_{VC}(V,C) = \begin{cases} \frac{\theta(2-\sigma)}{1-\phi}\left(\frac{1-\phi}{p}\right)^{2-\sigma} & \text{if } 0 \leq V \leq 1 \text{ and } 0 \leq C \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

Post-GDPR supply is therefore

$$\begin{aligned} S^{Post}(p) &= \mathcal{B}^{Post}\mathbb{P}\left((1-\phi)V_i < p\right) \\ &= \mathcal{B}\left(\frac{p}{1-\phi}\right)^{2-\sigma} \end{aligned} \tag{31}$$

For a given price p, the expected quality of the data packages sold is now

$$\begin{aligned} \mathbb{E}\left[V \middle| \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}} \leq V \leq \frac{p}{1-\phi}\right] &= \int_0^{\frac{p}{1-\phi}} (2-\sigma)\left(\frac{1-\phi}{p}\right)^{2-\sigma} V^{1-\sigma} dV \\ &= \frac{2-\sigma}{3-\sigma}\frac{p}{1-\phi}. \end{aligned}$$

Buyers will only buy if $\kappa\mu^{Post} \geq p$ where μ^{Post} is the above expectation of quality. This changes the minimum κ needed for the market to exist to $\frac{3-\sigma}{2-\sigma}$. The demand curve is now

$$D^{Post}(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3-\sigma}{2-\sigma} \\ 0 & \text{Otherwise} \end{cases} \tag{32}$$

Setting equations 31 and 32 equal yields the post-GDPR equilibrium:

$$\begin{aligned} p^*_{Post} &= \left(\frac{Y}{\mathcal{B}^{Post}}\right)^{\frac{1}{3-\sigma}} (1-\phi)^{\frac{2-\sigma}{3-\sigma}} \\ Q^*_{Post} &= Y^{\frac{2-\sigma}{3-\sigma}}\left(\frac{\mathcal{B}^{Post}}{(1-\phi)^{2-\sigma}}\right)^{\frac{1}{3-\sigma}}. \end{aligned} \tag{33}$$

B Data

B.1 UK Survey Data

The UK survey data referenced throughout the paper are from the United Kingdom Cyber Security Breach Survey: Combined Dataset, 2016-2022 (Department for Digital, Culture, Media and Sport, 2022). I accessed the data through the UK Data Services online portal on March 20, 2023.

Only the 2018 and 2019 survey asked respondents whether they made any changes in response to the GDPR. The survey asked about the types of changes made as well, which I have combined into five groups: human changes (e.g., staff training and hiring), technical changes (e.g., updated system configurations and increased spending on security), policy changes (e.g., conducting more audits and changing who has admin rights), third-party changes (e.g., changing IT service providers), and other changes (e.g., changing the nature of the business).

Table A1: UK Cyber Security Breach Survey Dates and Sample

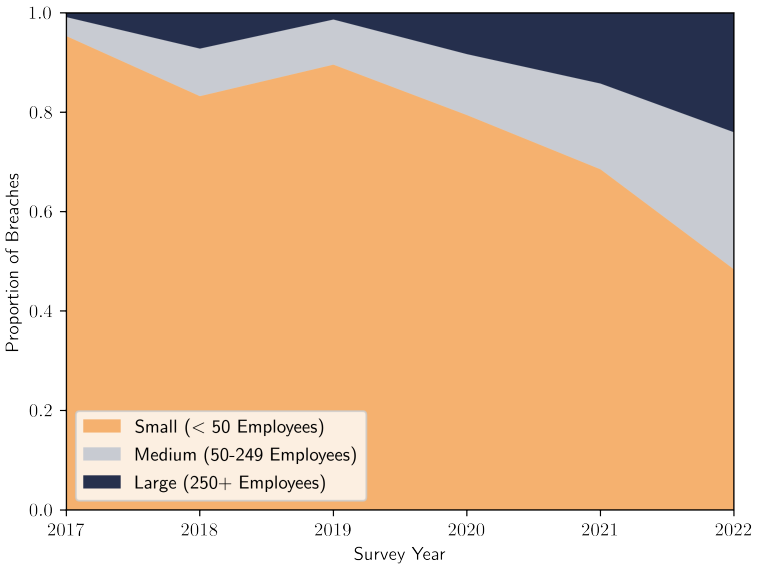
Survey Year	Sample Size	Survey Period
2016	1,008 businesses	November 30, 2015 – February 5, 2016
2017	1,523 businesses	October 24, 2016 – January 11, 2017
2018	1,519 businesses, 569 charities	October 9, 2017 – December 14, 2017
2019	1,566 businesses, 514 charities	October 10, 2018 – December 23, 2019
2020	1,348 businesses, 337 charities	October 19, 2019 – December 23, 2019
2021	1,419 businesses, 487 charities, 378 educational institutions	October 12, 2020 – January 21, 2021
2022	1,243 businesses, 424 charities, 490 educational institutions	September 20, 2021 – January 21, 2022

For figure A1, an organization was considered breached if they reported a ransomware or other malware infection; hacking of bank accounts; phishing attacks; unauthorized file access; or any other breach or attack.

B.2 Breach Data

The individual breach data obtained for this study contains many more data package observations than are included in the final paper. Observations were dropped for one of three reasons. First, any breaches that could not be attributed to an organization or country were removed. Second, any data package that was discovered during a breach of a breach indexing website, or similar "breach of breaches" was dropped. These breaches are of websites and other platforms that bundle access to credentials leaked in other breaches to their users. Essentially, the data being leaked in those breaches had itself been stolen from its original owner. What makes these observations unusable is the lack of a clear date when the data

Figure A1: UK Data Breaches



Source: Department for Digital, Culture, Media and Sport (2022), author's calculations.

were originally stolen. The observed date is of the larger breach, but it is unknown when the smaller breaches that comprise the breach occurred. Finally, data packages that appeared online prior to 2017 were removed. As briefly discussed when the panel data was described, the organization collecting these data was founded in 2016. Dropping these early breaches allows for the possibility that the breaches collected prior to that founding were meaningfully different from those that were collected later.

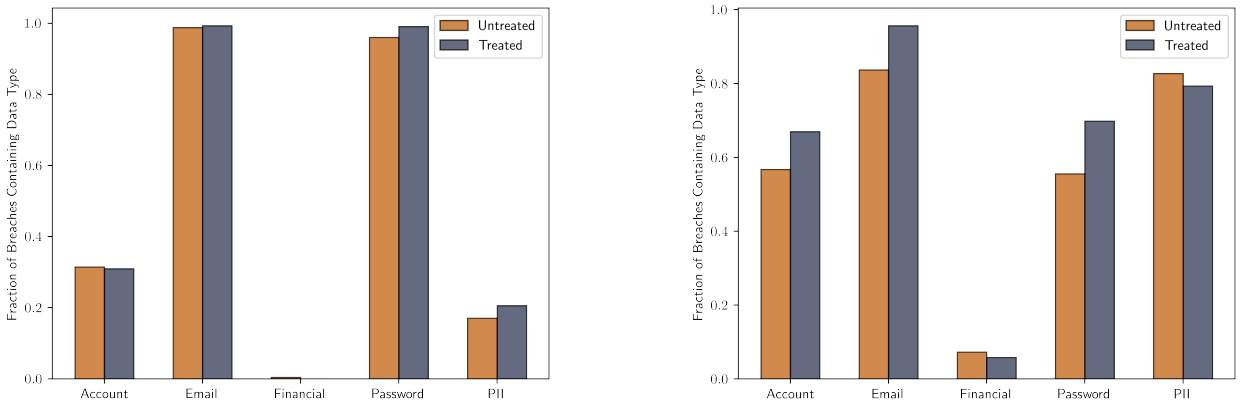
B.3 Defining Personally Identifiable Information

From a legal standpoint, there are three commonly used definitions of "personally identifiable information" (Schwartz and Solove, 2011). The tautological definition used in the Video Privacy Protection Act says that PII is information which identifies a person. The non-public information approach used in the Gramm-Leach-Bliley Act defines PII as non-public personal information. Finally, the specific-types approach explicitly lists the types of data that are considered PII. I borrow from all three approaches.

In the data I am able to observe the specific types of records in a data packages. I classify data as PII if reveals location, financial, contact, user account, or personal information. Account information covers emails, usernames, and passwords. Personal information includes as political and religious views, sexual orientation, and aspects of a person's home life such as if they have children or pets. As most of the data packages included emails and passwords

(figures 6 and A2), this makes the fraction of records in a data package that are PII fairly close to one. As part of my robustness checks, I repeat the data package analysis of the effect of the GDPR on the fraction of records in a data package that are PII using an alternative definition that removes emails and passwords. I find that this did not change the main result that the GDPR had no effect on the portion of records in a breach that are PII (table A25).

Figure A2: Fraction of Data Packages Containing Each Data Type, Pre-and Post-GDPR



(a) Pre-GDPR

(b) Post-GDPR

B.4 Descriptive Information

Tables A2-A5 report unconditional differences in means between various data package groups.

Table A2 compares treated and untreated data packages across the full sample. There are statistically significant differences between the two in the fraction of records that are PII, and the number of unique data types. Although they are statistically significant, they are not particularly meaningful. Given that both types have close to 70 percent PII, a 4 percentage point difference is not particularly large. And the difference in number of unique data types is less than one, making them effectively the same from an interpretation perspective.

Data packages that became available before and after the GDPR are then compared in table A3. The data packages get significantly larger after the GDPR in terms of both the number of records and the number of unique data types. As shown in tables A4 and A5, which compare the packages pre-and post-GDPR for the control and treated groups, respectively, this affect is seen in both, though it is much larger in the treated group. This is consistent with the findings that expected data package size significantly increased after the GDPR, and the theory that attackers may have shifted their efforts towards larger targets.

Table A2: Data Package Means: Treated vs. Untreated

	Means			Differences
	0 N=3,468	1 N=926	Overall Mean N=4,394	Treated - Untreated
Number of Records	3,308,275	4,427,708	3,544,186	1,119,433
	(464,961.454)	(1,129,779.958)	(437,434.656)	(1,221,716.787)
PII Fraction	0.698	0.660	0.690	-0.038***
	(0.003)	(0.006)	(0.003)	(0.007)
# of Data Types	6.365	5.678	6.220	-0.687***
	(0.089)	(0.166)	(0.079)	(0.188)

** p< 0.1, ** p< 0.05, *** p< 0.01*

Table A3: Data Package Means: Pre-GDPR vs. Post-GDPR

	Means			Differences
	0 N=1,621	1 N=2,773	Overall Mean N=4,394	Post - Pre
Number of Records	1,598,365	4,681,646	3,544,186	3,083,281***
	(465,366.681)	(636,601.781)	(437,434.656)	(788,560.699)
PII Fraction	0.550	0.771	0.690	0.221***
	(0.003)	(0.003)	(0.003)	(0.004)
# of Data Types	3.163	8.007	6.220	4.844***
	(0.069)	(0.104)	(0.079)	(0.125)

** p< 0.1, ** p< 0.05, *** p< 0.01*

Table A4: Data Package Means: Pre- vs. Post-GDPR, Untreated

	Means			Differences
	0 N=1,175	1 N=2,293	Overall Mean N=3,468	Post - Pre
Number of Records	2,123,526	3,915,375	3,308,275	1,791,849**
	(640,456.742)	(621,655.452)	(464,961.454)	(892,547.107)
PII Fraction	0.552	0.773	0.698	0.221***
	(0.003)	(0.004)	(0.003)	(0.005)
# of Data Types	3.279	7.946	6.365	4.667***
	(0.090)	(0.113)	(0.089)	(0.145)

** p< 0.1, ** p< 0.05, *** p< 0.01*

Table A5: Data Package Means: Pre- vs. Post-GDPR, Treated

	Means			Differences
	0 N=446	1 N=480	Overall Mean N=926	Post - Pre
Number of Records	214,813	8,342,189	4,427,708	8,127,375***
	(92,543.690)	(2,163,638.826)	(1,129,779.958)	(2,165,617.072)
PII Fraction	0.546	0.765	0.660	0.219***
	(0.004)	(0.009)	(0.006)	(0.010)
# of Data Types	2.859	8.298	5.678	5.439***
	(0.080)	(0.259)	(0.166)	(0.271)

** p< 0.1, ** p< 0.05, *** p< 0.01*

C Results

Additional results from alternative specifications of the estimated models are presented here.

C.1 Extensive Margin Effects

On the extensive margin, I separately estimate equation 4.1 for small and large countries. The former are countries with above median population in 2018, the latter countries with below median population in 2018. Results are in tables A6 and A7.

Table A6: Extensive Margin Effects: Small Countries

	Dependent Variable: Positive Number of Breaches	
	(1)	(2)
Post x Treatment	-0.224***	
	(0.051)	
SR x Treatment		-0.230***
		(0.066)
LR x Treatment		-0.222***
		(0.051)
Observations	1,344	1,344
R^2	0.326	0.326
Period Fixed Effects	Y	Y
Country Fixed Effects	Y	Y

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. Small countries are defined as those with a population below the median in 2018.

Table A7: Extensive Margin Effects: Large Countries

	Dependent Variable: Positive Number of Breaches	
	(1)	(2)
Post x Treatment	-0.168***	
	(0.060)	
SR x Treatment		-0.105
		(0.088)
LR x Treatment		-0.182***
		(0.058)
Observations	1,372	1,372
R^2	0.510	0.511
Period Fixed Effects	Y	Y
Country Fixed Effects	Y	Y

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. Large countries are defined as those with a population above the median in 2018.

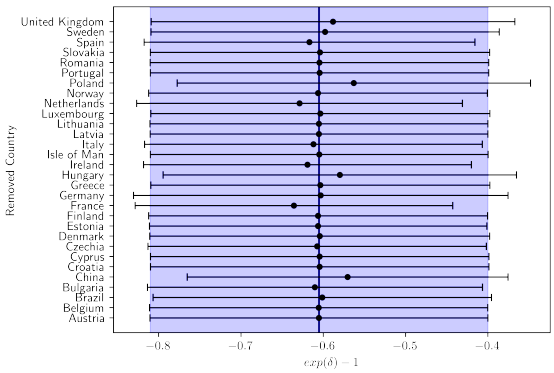
C.2 Aggregate Effects

To test the robustness of my aggregate effect estimates, I first re-estimate each aggregate effect after removing a treated country from the data. For each removed country, the estimate stays well within the 95 percent confidence interval of the estimate with the full sample (figures A3 and A4).

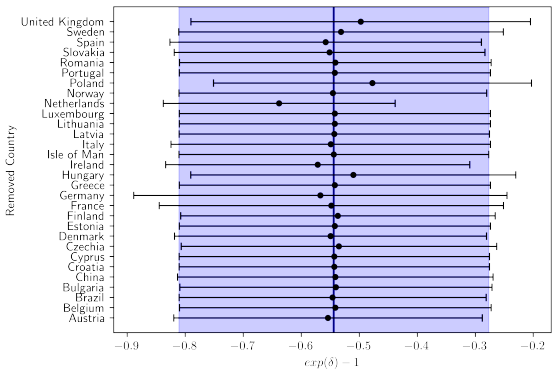
Next, I use different methods to construct the panel. Brazil and China each adopted their own data privacy laws near the end of the study period. Removing them from the sample slightly reduces the estimated effect on the number of breaches, though it is still significant. As before, there are no statistically significant effects on the number of records available (table A8).

Table A9 shows the aggregate results when I exclude data from after the first quarter of 2020 to avoid any pandemic effects. This significantly reduces the number of post-treatment observations. The change in the number of records remains insignificant and in number of data breaches significant, but the long-run effect in the latter case does change. The short-run effects on both outcomes are identical to using the full panel, which is unsurprising since observations in the pre-treatment period and short-run all remain in this new panel. The only change is in the long-run estimates, where the reduction in number of data breaches shrank, though is still statistically significant. In this shorter panel there are only three

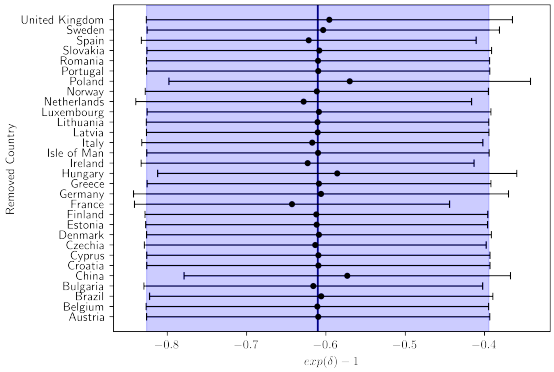
Figure A3: Number of Data Breaches Effects Removing Countries



(a) Overall Effect



(b) Short Run Effect



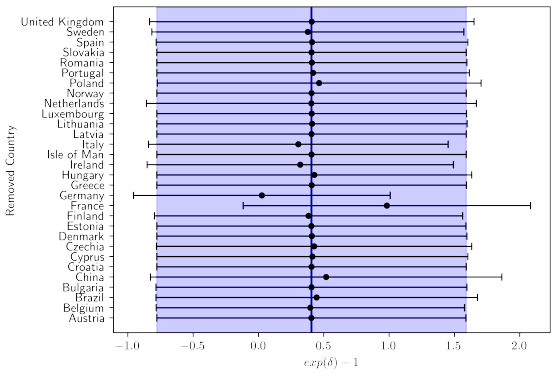
(c) Long Run Effect

Notes: Each point represents the estimated effect on number of data breaches after removing observations from the specified country. The whiskers are the 95 percent confidence interval. The solid line is the point estimate including all countries, and the shaded area is the 95 percent confidence interval around that point.

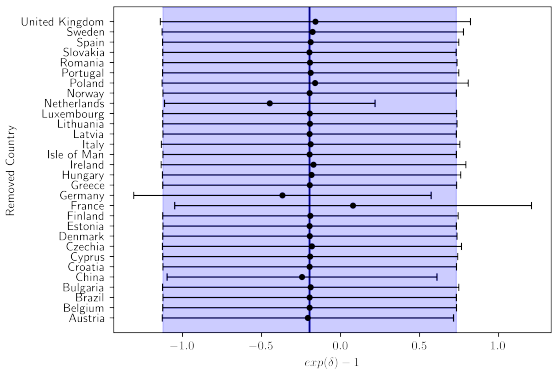
long-run periods: the third and fourth quarters of 2019 and the first quarter of 2020. These results suggest that the long run effect grows as time goes on.

Table A10 shows the quantity results when I exclude data packages originating in multinational organizations from the panel. Whether an organization is a multinational is determined in one of two ways. First, if their website is hosted in more than one country, they are considered multinational. Second, if their website and organizational information, such as privacy policies, discuss having customers or users in more than one country. The argument for excluding these effects is that multinational organizations may be partially treated. The GDPR applies to data specifically from EU residents. A multinational organization would therefore have to comply if they have any users in the EU, but it is not clear whether they

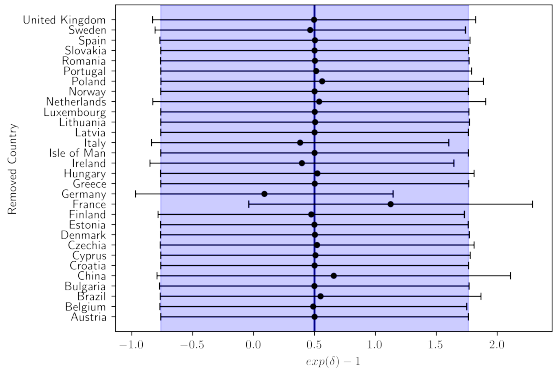
Figure A4: Number of Records Effects Removing Countries



(a) Overall Effect



(b) Short Run Effect



(c) Long Run Effect

Notes: Each point represents the estimated effect on number of records after removing observations from the specified country. The whiskers are the 95 percent confidence interval. The solid line is the point estimate including all countries, and the shaded area is the 95 percent confidence interval around that point.

would change their data collection and protection practices for all their users, or just those in the EU.

When multinational breaches are excluded, there is actually a long-run increase in the number of records available after the GDPR. The effect on the number of data breaches is roughly equivalent to the one found in the main specification.

Each of the previous tests left the definition of the outcome variables unchanged and were estimated with same Poisson regression as in the main paper. Tables A11-A16 test changes in the outcome variable definition, the effect of adding covariates to the equation, and using three other models to derive estimates.

First, I estimate the effect using a linear model, rather than a Poisson model:

Table A8: Aggregate Effects: Dropping Brazil and China

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-0.825***		0.454	
	(0.228)		(0.461)	
SR x Treatment		-0.781***		-0.279
		(0.301)		(0.577)
LR x Treatment		-0.829***		0.549
		(0.242)		(0.455)
$\hat{\delta}$	-0.562		0.575	
	(0.100)		(0.726)	
$\hat{\delta}^{SR}$		-0.542		-0.243
		(0.138)		(0.436)
$\hat{\delta}^{LR}$		-0.564		0.732
		(0.106)		(0.788)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,660	2,660	2,660	2,660
Pseudo R^2	0.811	0.811	0.885	0.885

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. Brazil and China have been removed from the panel

$$Y_{it} = \gamma_s + \tau_t + \delta D_i \times Post\text{-}GDPR_t + \epsilon_{it} \tag{34}$$

where each term is defined as before. In addition to using a linear model, I use two log-like transformations of the outcome variable, $log(Y_{it}+1)$ and the inverse hyperbolic sine (IHS) function $ln(Y_{it} + \sqrt{Y_{it}^2+1})$. These transformations are necessary, rather than just using $log(Y_{it})$ because there are a number of periods in which countries have no breaches. Using these transformations significantly changes the results from the Poisson model. For the number of records, both the $log(Y_{it}+1)$ and IHS transformation give large and statistically significant negative estimates of the treatment effect, unlike the Poisson which showed no change. I believe this is due to a significant extensive margin effect. Chen and Roth (2023) and Mullahy and Norton (2024) both discuss how, when there are mass points at zeros, log-like transformations may greatly influence the estimated coefficients.

As discussed in the main body of the paper, there are significant and negative extensive margin effects (table 9). This is likely the source of the discrepancies in effect sizes between

Table A9: Aggregate Effects: Excluding COVID Years

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-0.639***		-0.214	
	(0.182)		(0.572)	
SR x Treatment		-0.782***		-0.218
		(0.300)		(0.592)
LR x Treatment		-0.533***		-0.210
		(0.152)		(0.577)
$\hat{\delta}$	-0.472		-0.193	
	(0.096)		(0.462)	
$\hat{\delta}^{SR}$		-0.543		-0.196
		(0.137)		(0.476)
$\hat{\delta}^{LR}$		-0.413		-0.189
		(0.089)		(0.468)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	852	852	852	852
Pseudo R^2	0.849	0.849	0.885	0.885

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. All periods after the first quarter or 2020 are excluded from the panel.

the models and the primary reason for using the Poisson model over the linear models with a log-like transformation.

Without the log-like transformation, when it is estimated in levels, the linear model produces results that are in line with, though interpreted differently than, the Poisson model. Specifically, I still find no effect on the number of records and a significant negative effect on the number of data breaches (columns 7 and 8 of each table).

Next, I estimate the models using various measures to account for population size. In tables A13 and A14, I add population in millions as a covariate. It is not included in the levels models because the outcomes are already scaled to be records/data packages per million. In all cases there is no significant change in the estimates and the population coefficient is insignificant.

In tables A15 and A16, I change the outcome for the log-like transformation to also be number of records/data packages per million, and add a population offset to the Poisson model. This noticeably changes the magnitude of both log-like transformations in each

Table A10: Aggregate Effects: Excluding Multinational Organizations

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-0.909***		0.655	
	(0.279)		(0.435)	
SR x Treatment		-0.805***		-0.718
		(0.289)		(0.704)
LR x Treatment		-0.918***		0.806*
		(0.299)		(0.444)
$\hat{\delta}$	-0.597		0.925	
	(0.112)		(0.837)	
$\hat{\delta}^{SR}$		-0.553		-0.512
		(0.129)		(0.343)
$\hat{\delta}^{LR}$		-0.601		1.239
		(0.119)		(0.993)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,632	2,632	2,632	2,632
Pseudo R^2	0.784	0.784	0.824	0.825

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. Data packages originating from multinational organizations are excluded from the panel construction.

outcome. As Chen and Roth (2023) discuss, this is a reflection of the sensitivity of log-like transformations to the scale of the outcome variable when extensive margin effects are present. The offset in the Poisson model effectively changes the outcome to a rate, as in breaches per million. The estimates however are roughly the same as those in the model without the offset.

Table A11: Alternative Models: Number of Data Breaches

	Poisson		Log(Y + 1)		IHS		Levels	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post x Treatment	-0.930***		-0.416***		-0.518***		-0.146***	
	(0.265)		(0.085)		(0.102)		(0.032)	
SR x Treatment		-0.785***		-0.387***		-0.484***		-0.135***
		(0.299)		(0.111)		(0.136)		(0.033)
LR x Treatment		-0.942***		-0.423***		-0.525***		-0.149***
		(0.283)		(0.081)		(0.097)		(0.033)
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716	2,716	2,716	2,716	2,716
R^2			0.692	0.692	0.683	0.683	0.105	0.105
Pseudo R^2	0.793	0.793						

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. IHS: inverse hyperbolic sine transformation of the dependent variable. In the levels regression, the dependent variable is number of data breaches per million. Unlike the main specification, the Poisson model does not include a population offset.

Table A12: Alternative Models: Number of Records

	Poisson		Log(Y + 1)		IHS		Levels	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post x Treatment	0.341		-2.295***		-2.437***		-3.626	
	(0.430)		(0.394)		(0.418)		(20.950)	
SR x Treatment		-0.219		-1.805***		-1.920***		-1.971
		(0.590)		(0.577)		(0.609)		(18.370)
LR x Treatment		0.406		-2.404***		-2.552***		-3.993
		(0.430)		(0.396)		(0.420)		(22.980)
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716	2,716	2,716	2,716	2,716
R^2			0.545	0.545	0.542	0.543	0.182	0.182
Pseudo R^2	0.847	0.847						

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. IHS: inverse hyperbolic sine transformation of the dependent variable. In the levels regression, the dependent variable is number of records per thousand. Unlike the main specification, the Poisson model does not include a population offset.

Table A13: Alternative Models with Covariates: Number of Data Breaches

	Poisson		Log(Y + 1)		IHS		Levels	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post x Treatment	-1.080***		-0.414***		-0.515***		-0.146***	
	(0.308)		(0.086)		(0.103)		(0.032)	
SR x Treatment		-0.835***		-0.390***		-0.488***		-0.135***
		(0.323)		(0.111)		(0.136)		(0.033)
LR x Treatment		-1.107***		-0.421***		-0.522***		-0.149***
		(0.323)		(0.082)		(0.098)		(0.033)
GDP Per Capita	-0.000	-0.000	0.000	0.000	0.000	0.000		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		
Population	-0.004	-0.004	0.004	0.004	0.006	0.006		
	(0.006)	(0.006)	(0.003)	(0.003)	(0.004)	(0.004)		
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2,648	2,648	2,648	2,648	2,648	2,648	2,716	2,716
R^2			0.697	0.697	0.687	0.687	0.105	0.105
Pseudo R^2	0.797	0.797						

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. IHS: inverse hyperbolic sine transformation of the dependent variable. In the levels regression, the dependent variable is number of data breaches per million. Annual population data is provided by the World Bank. Unlike the main specification, the Poisson model does not include a population offset.

Table A14: Alternative Models with Covariates: Number of Records

	Poisson		Log(Y + 1)		IHS		Levels	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post x Treatment	0.482		-2.261***		-2.403***		-3.626	
	(0.509)		(0.393)		(0.416)		(20.950)	
SR x Treatment		-0.147		-1.849***		-1.968***		-1.971
		(0.618)		(0.572)		(0.604)		(18.370)
LR x Treatment		0.568		-2.367***		-2.515***		-3.993
		(0.520)		(0.397)		(0.420)		(22.980)
GDP Per Capita	0.000	0.000	0.000	0.000	0.000	0.000		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		
Population	0.015	0.016	0.021	0.019	0.022	0.019		
	(0.015)	(0.015)	(0.022)	(0.022)	(0.023)	(0.023)		
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2,648	2,648	2,648	2,648	2,648	2,648	2,716	2,716
R^2			0.551	0.551	0.548	0.549	0.182	0.182
Pseudo R^2	0.847	0.848						

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. IHS: inverse hyperbolic sine transformation of the dependent variable. In the levels regression, the dependent variable is number of records per thousand. Annual population data is provided by the World Bank. Unlike the main specification, the Poisson model does not include a population offset.

Table A15: Alternative Models: Number of Data Breaches Scaled by Population

	Poisson		Log(Y + 1)		IHS		Levels	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post x Treatment	-0.921***		-0.101***		-0.124***		-0.146***	
	(0.265)		(0.019)		(0.025)		(0.032)	
SR x Treatment		-0.782***		-0.100***		-0.121***		-0.135***
		(0.299)		(0.021)		(0.026)		(0.033)
LR x Treatment		-0.934***		-0.102***		-0.124***		-0.149***
		(0.283)		(0.019)		(0.025)		(0.033)
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716	2,716	2,716	2,716	2,716
R^2			0.254	0.254	0.234	0.234	0.105	0.105
Pseudo R^2	0.792	0.792						

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. IHS: inverse hyperbolic sine transformation of the dependent variable. The dependent variable is number of breaches per million, except in the Poisson model, where a log(population) offset is used instead. Annual population data is provided by the World Bank.

Table A16: Alternative Models: Number of Records Scaled by Population

	Poisson		Log(Y + 1)		IHS		Levels	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Post x Treatment	0.345		-0.510***		-0.619***		-3.626	
	(0.430)		(0.122)		(0.140)		(20.950)	
SR x Treatment		-0.217		-0.440***		-0.536***		-1.971
		(0.590)		(0.158)		(0.185)		(18.370)
LR x Treatment		0.410		-0.526***		-0.638***		-3.993
		(0.430)		(0.126)		(0.144)		(22.980)
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716	2,716	2,716	2,716	2,716
R^2			0.447	0.447	0.450	0.450	0.182	0.182
Pseudo R^2	0.847	0.847						

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. IHS: inverse hyperbolic sine transformation of the dependent variable. The dependent variable is number of records per thousand, except in the Poisson model, where a log(population) offset is used instead. Annual population data is provided by the World Bank.

For my final model specification tests, I used alternative ways of controlling for population and added covariates. My results were largely unchanged in each case. In table A17, I remove the population offset. Table A18 weights the estimates using population rather than including an offset. In table A19, I again remove the population offset and opt instead for using per capita outcomes variables. Finally, I split the sample into small and large countries in tables A20 and A21, and include indicators for whether the observation is a small or large country in table A22.[14]

A shortcoming of my data is that many variables that would be reasonable to include as covariates, such as the fraction of people with internet access, are not consistently observed for every country. Rather than drop observations and unbalance the panel to account for this, the only covariate I add to the model is GDP per capita. As shown in table A23, this does not have a significant effect on my effect estimates.

Table A17: Aggregate Effects: No Offset

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-0.930***		0.341	
	(0.265)		(0.430)	
SR x Treatment		-0.785***		-0.219
		(0.299)		(0.590)
LR x Treatment		-0.942***		0.406
		(0.283)		(0.430)
$\hat{\delta}$	-0.605		0.406	
	(0.105)		(0.604)	
$\hat{\delta}^{SR}$		-0.544		-0.197
		(0.136)		(0.474)
$\hat{\delta}^{LR}$		-0.610		0.500
		(0.110)		(0.645)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716
Pseudo R^2	0.793	0.793	0.847	0.847

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. No population offset is used

The final aggregate effects test I conduct estimates the effect on the number of small and

[14]Small or large in this context means above or below the median population in 2018.

Table A18: Aggregate Effects: Weighted Estimation

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-1.113***		-0.007	
	(0.384)		(0.473)	
SR x Treatment		-0.932***		-0.476
		(0.212)		(0.558)
LR x Treatment		-1.127***		0.040
		(0.411)		(0.497)
$\hat{\delta}$	-0.671		-0.007	
	(0.126)		(0.469)	
$\hat{\delta}^{SR}$		-0.606		-0.379
		(0.084)		(0.347)
$\hat{\delta}^{LR}$		-0.676		0.041
		(0.133)		(0.518)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716
Pseudo R^2	1.426	1.426	1.031	1.031

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. The population offset is removed and observations are instead weighted by population.

large breaches. Small breaches are those with more than the median number of records, large breaches are those with more than the median number of records. As shown in table A24, the decline in breaches is concentrated entirely among small breaches. This is consistent with the model's prediction that there will be a shift to more data rich targets, and my empirical finding that breach sizes increased after the GDPR.

C.3 Data Package Effects

Estimates of the change in PII fraction using a slightly different definition of PII are in table A25. Under this definition, I remove emails and passwords from PII. I find no significant change, as is the case using the original definition.

As previously discussed, data packages from periods prior to January 2017 were excluded from the main dataset. Tables A27-A29 report the results using the full sample, including those early breaches. In all cases, the signs of the estimated coefficients remain the same. The magnitude of the increase in number of records is larger (comparing table A27 to table

Table A19: Aggregate Effects Per Capita Outcomes

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-1.406***		0.339	
	(0.369)		(0.408)	
SR x Treatment		-0.109		-0.326
		(0.505)		(0.723)
LR x Treatment		-1.474***		0.397
		(0.370)		(0.418)
$\hat{\delta}$	-0.755		0.403	
	(0.090)		(0.573)	
$\hat{\delta}^{SR}$		-0.103		-0.278
		(0.453)		(0.522)
$\hat{\delta}^{LR}$		-0.771		0.487
		(0.085)		(0.621)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716
Pseudo R^2	0.073	0.073	0.473	0.474

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level.

11), but the estimates are still within each other's standard errors. For the number of unique data types, using the full sample does result in a statistically significant increase, unlike the smaller sample. But the increase is still less than a single data type and therefore not economically meaningful.

Finally, I estimated extensive margin effects for each of the data types using the linear probability model

$$Positive_i = \gamma_i + \tau_t + \delta D_{it} + \varepsilon_{it}$$

where $Positive_i$ is one if the data package contains a positive amount of that data; γ_i and τ_t are country and quarter fixed effects, respectively; and D_{it} is an indicator for whether the data package is treated.

There is a short-run increase in the probability of a data package containing email addresses and password information, but neither is maintained into the long-run. Long term, the only data type showing a significant change is account information, which saw an eight percent increase in the likelihood that it is in a data package (table A30).

Table A20: Aggregate Effects: Small Countries

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-1.328***		0.145	
	(0.429)		(0.437)	
SR x Treatment		-0.210		-2.229***
		(0.695)		(0.630)
LR x Treatment		-1.383***		0.286
		(0.427)		(0.431)
$\hat{\delta}$	-0.735		0.156	
	(0.114)		(0.505)	
$\hat{\delta}^{SR}$		-0.189		-0.892
		(0.564)		(0.068)
$\hat{\delta}^{LR}$		-0.749		0.331
		(0.107)		(0.573)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	1,344	1,344	1,344	1,344
Pseudo R^2	0.471	0.472	0.745	0.750

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. Observations are limited to countries with below median populations in 2018.

Table A21: Aggregate Effects: Large Countries

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-0.863***		0.290	
	(0.283)		(0.497)	
SR x Treatment		-0.742**		-0.089
		(0.328)		(0.617)
LR x Treatment		-0.874***		0.338
		(0.300)		(0.504)
$\hat{\delta}$	-0.578		0.336	
	(0.120)		(0.664)	
$\hat{\delta}^{SR}$		-0.524		-0.085
		(0.156)		(0.565)
$\hat{\delta}^{LR}$		-0.583		0.402
		(0.125)		(0.706)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	1,372	1,372	1,372	1,372
Pseudo R^2	0.804	0.804	0.828	0.828

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. Observations are limited to countries with above median populations in 2018.

Table A22: Aggregate Effects: Size Indicators

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Above Median Pop. x Post	-0.276		-0.520	
	(0.357)		(0.411)	
Above Median Pop. x SR		0.925*		-0.598
		(0.486)		(0.585)
Above Median Pop. x LR		-0.262		-0.452
		(0.388)		(0.359)
Post x Treatment	-1.330***		0.141	
	(0.426)		(0.433)	
SR x Treatment		-0.153		-2.279***
		(0.715)		(0.605)
LR x Treatment		-1.243**		0.001
		(0.546)		(0.456)
Above Median Pop. x Post x Treatment	0.467		0.149	
	(0.511)		(0.656)	
Above Median Pop. x SR x Treatment		-0.647		2.231***
		(0.785)		(0.857)
Above Median Pop. x LR x Treatment		0.178		0.471
		(0.566)		(0.837)
$\hat{\delta}$	0.596		0.160	
	(0.815)		(0.761)	
$\hat{\delta}^{SR}$		-0.477		8.311
		(0.411)		(7.981)
$\hat{\delta}^{LR}$		0.195		0.601
		(0.676)		(1.340)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,716	2,648	2,716	2,648
Pseudo R^2	0.793	0.797	0.847	0.847

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. Observations are unweighted.

Table A23: Aggregate Effects: With Covariates

	Number of Breaches		Number of Records	
	(1)	(2)	(3)	(4)
Post x Treatment	-1.042***		0.396	
	(0.285)		(0.471)	
SR x Treatment		-0.822**		-0.186
		(0.321)		(0.596)
LR x Treatment		-1.064***		0.468
		(0.298)		(0.483)
GDP Per Capita	-0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
$\hat{\delta}$	-0.647		0.485	
	(0.101)		(0.700)	
$\hat{\delta}^{SR}$		-0.560		-0.169
		(0.141)		(0.495)
$\hat{\delta}^{LR}$		-0.655		0.597
		(0.103)		(0.771)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	2,648	2,648	2,648	2,648
Pseudo R^2	0.796	0.796	0.847	0.847

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level. Observations are unweighted.

Table A24: Aggregate Effects by Breach Size

	Below Median		Above Median		Total	
	(1)	(2)	(3)	(4)	(5)	(6)
Post x Treatment	-0.864***		0.009		-0.921***	
	(0.284)		(0.378)		(0.265)	
SR x Treatment		-0.608**		0.024		-0.782***
		(0.294)		(0.400)		(0.299)
LR x Treatment		-0.885***		0.007		-0.934***
		(0.299)		(0.422)		(0.283)
$\hat{\delta}$	-0.579		0.009		-0.602	
	(0.119)		(0.381)		(0.105)	
$\hat{\delta}^{SR}$		-0.456		0.024		-0.543
		(0.160)		(0.410)		(0.137)
$\hat{\delta}^{LR}$		-0.587		0.007		-0.607
		(0.123)		(0.425)		(0.111)
Period Fixed Effects	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y
Observations	2,716	2,716	2,716	2,716	2,716	2,716
Pseudo R^2	0.784	0.784	0.700	0.700	0.792	0.792

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard Errors are clustered at the country level.

Table A25: Data Package Effects: PII Fraction - Excluding Emails and Passwords

	Dependent Variable: PII Fraction			
	(1)	(2)	(3)	(4)
Post x Treatment	0.002		-0.015	
	(0.019)		(0.013)	
SR x Treatment		0.047		0.034
		(0.038)		(0.037)
LR x Treatment		0.002		-0.012
		(0.022)		(0.017)
Multinational			0.052***	0.048***
			(0.019)	(0.018)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	4,394	4,394	4,394	4,394
R^2	0.422	0.422	0.423	0.423

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered by country. PII definition excludes emails and passwords.

Table A26: Data Package Effects: Number of PII Records - Excluding Emails and Passwords

	Dependent Variable: Log(Number of PII Records)			
	(1)	(2)	(3)	(4)
Post x Treatment	0.948		0.362	
	(0.642)		(0.392)	
SR x Treatment		1.857**		1.385*
		(0.853)		(0.698)
LR x Treatment		0.897		0.370
		(0.591)		(0.403)
Multinational			1.810***	1.741***
			(0.346)	(0.336)
$\hat{\delta}$	1.579		0.436	
	(1.657)		(0.563)	
$\hat{\delta}^{SR}$		5.405		2.993
		(5.463)		(2.788)
$\hat{\delta}^{LR}$		1.453		0.448
		(1.449)		(0.583)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	4,394	4,394	4,394	4,394
R^2	0.382	0.383	0.386	0.386

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered by country. PII definition excludes emails and passwords.

Table A27: Data Package Effects: Number of Records

	Dependent Variable: Log(Number of Records)			
	(1)	(2)	(3)	(4)
Post x Treatment	0.977**		0.584**	
	(0.413)		(0.259)	
SR x Treatment		0.424		0.089
		(0.359)		(0.251)
LR x Treatment		0.962**		0.589**
		(0.392)		(0.259)
Multinational			1.418***	1.431***
			(0.305)	(0.310)
$\hat{\delta}$	1.657		0.793	
	(1.096)		(0.464)	
$\hat{\delta}^{SR}$		0.529		0.093
		(0.549)		(0.274)
$\hat{\delta}^{LR}$		1.617		0.803
		(1.025)		(0.467)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	5,669	5,669	5,669	5,669
R^2	0.280	0.280	0.289	0.289

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. Estimates use the full sample and do not drop early period data packages.

Table A28: Data Package Effects: PII Fraction

	Dependent Variable: PII Fraction			
	(1)	(2)	(3)	(4)
Post x Treatment	-0.014		-0.016	
	(0.009)		(0.013)	
SR x Treatment		-0.010		-0.012
		(0.024)		(0.022)
LR x Treatment		-0.011		-0.013
		(0.012)		(0.016)
Multinational			0.010	0.009
			(0.015)	(0.015)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	5,669	5,669	5,669	5,669
R^2	0.396	0.395	0.396	0.396

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. Estimates use the full sample and do not drop early period data packages.

Table A29: Data Package Effects: Number of Data Types

	Dependent Variable: Number of Unique Data Types			
	(1)	(2)	(3)	(4)
Post x Treatment	0.490*		0.491*	
	(0.264)		(0.258)	
SR x Treatment		0.377		0.381
		(0.492)		(0.521)
LR x Treatment		0.539*		0.543*
		(0.302)		(0.289)
Multinational			-0.001	-0.015
			(0.202)	(0.210)
Period Fixed Effects	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y
Observations	5,669	5,669	5,669	5,669
R^2	0.277	0.277	0.277	0.277

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. Estimates use the full sample and do not drop early period data packages.

Table A30: Data Types Extensive Margin Effects

	Account		Email		Financial		Passwords		PII	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Post x Treatment	0.083**		0.033		-0.011		0.033		0.012	
	(0.037)		(0.034)		(0.012)		(0.062)		(0.026)	
SR x Treatment		0.147***		0.035*		0.037		0.070*		0.005
		(0.054)		(0.020)		(0.031)		(0.036)		(0.062)
LR x Treatment		0.081*		0.028		-0.014		0.016		0.020
		(0.047)		(0.040)		(0.013)		(0.068)		(0.032)
Observations	4,394	4,394	4,394	4,394	4,394	4,394	4,394	4,394	4,394	4,394
R^2	0.275	0.275	0.378	0.378	0.067	0.067	0.358	0.358	0.477	0.477
Period Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Country Fixed Effects	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

**p<0.1, **p<0.05, ***p<0.01*

Notes: Standard errors are clustered at the country level. The extensive margin is estimated using a linear probability model. The dependent variable is an indicator for whether the data package contains data of each type. PII in columns 9 and 10 does not include emails or passwords.