Essays on Cybercrime, Privacy, and Information

Anderson Joseph Frailey
Katy, Texas

Master of Arts, Economics, University of Virginia, 2020
Bachelor of Arts, Economics, The University of Texas at Austin, 2016

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Economics

University of Virginia
May, 2025

Committee Members:
Amalia Miller
Lee Lockwood
Denis Nekipelov
Bo Sun

Contents

Acknowledgements

I would obviously be lying if I said every moment of the last six years was enjoyable. This experience was an exercise in perseverance and there were many days when I did not think I would make it to the end. I'm grateful to so many people for helping me make it through mostly intact.

It's only natural to first thank my committee: Amalia Miller, Lee Lockwood, Denis Nekipelov, and Bo Sun. Thank you for jumping on board when I decided to write an entirely new dissertation proposal weeks before it was due and sticking with me as I figured this whole thing out. Special thanks to Amalia for being on board with this adventure in privacy and teaching a class with me on it. And thank you to Lee and Denis for always making me write down a model, even when I didn't want to (especially when I didn't want to). I lost some sleep working on them, but this dissertation is better for it.

Thank you to Sarah Turner for believing I had potential as a researcher before I did, for helping fund my final year, and being very patient while I dragged out our work on chapter three.

Many other scholars were generous with their time and knowledge throughout my time in grad school including, but not limited to, Eric Young, Leora Friedberg, Kerem Coşar, Danielle Citron, and Diego J. Jiménez Hernández.

UVA was the last PhD program I applied to, and I only applied because of Ken Elzinga. His chapter on the Spirit and Northwestern Airlines merger in *The Antitrust Revolution* made me think I would enjoy being an antitrust economist. While I did not become an antitrust economist, serving as Head TA under Mr. Elzinga and Lee Coppock was the most rewarding experience of my time at UVA. Thank you both for showing me how to be a proper Professor. Thank you to Bella Hicks for keeping the train running during those two years. Thank you to the hundreds of students in our classes who helped make me a better economist and teacher.

I would not have been able to survive the past six years without the support of my family and I cannot thank them enough. Mom and Dad for supporting me endlessly, always being there help me decompress, and generally being the best parents one could ask for. Austin and Alana for being great pandemic housemates and a relief to be around whenever I came home. And Errington for being the fun new addition to the family.

My friends, old and new, were with me through the highest of highs and the lowest of lows. Cody, Andrew, and Pablo have been there for me for what now seems like my entire life. Thankfully we don't seem to be sick of each other yet. Aspen, thanks for always being down for a hike, terrible movie, ice cream, or vent session. Every summer I would count

down the days to the next body-of-water trip with the Coastal Grandmas: Erin, Max and Val, Alex, and David. I'd be remiss to not thank Savannah, Elizabeth, and Hadley for their friendship. I am grateful to so many of my fellow students who helped me learn, listened to my half-baked ideas, and were generally lovely companions. In no particular order, Max, Camille, Joe, Diego, Sasha, Avantika, Pallavi, Daniel, Anirban, DeShawn, and everyone in the Public and Labor Student Group were all helpful in someway shape or form. You were a lovely group to trauma bond with. There are many, many others I could include in my list of people to thank, but that would be another essay in and of itself.

The American Economic Association Mentoring Program (AEAMP) has been an incredible resource. I thank them for supporting my trips to conferences and introducing me to a group of incredibly talented economists. I received invaluable feedback during each of the summer mentoring meetings I attended as well. Thank you in particular to my mentor in the program, Colin Cannonier. The work AEAMP has done to help underrepresented minorities thrive the economics profession is so valuable, as is the community of diverse economists they have developed.[1]

There are a number of other entities and such that feel worthy of acknowledgement. I lived my life one 48-page Field Notes memo book at a time. Many of the words here were conceived on Amtrak trains. The Northeast Regional and Crescent 20 lines will always hold a special place in my heart. The inevitable delays just gave me more time to work. Finally, the artists behind the Bridgerton soundtrack were also the artists behind many late night writing sessions. I thank them for their talents.

I benefited financially from the Dean's Doctoral Fellowship and Marshall Jevons Fund. Chapter 1 was made possible by SpyCloud's generous provision of data. For that I thank Pablo Maceda, Ronak Patel, Wallis Romzek, and Trevor Hilligoss. The conclusions and views expressed herein do not reflect those of SpyCloud or any persons affiliated with SpyCloud. Syrell Greer and Jalen Mui provided excellent research assistance on chapter 2.

All remaining mistakes are my own.

Hook 'Em 'Hoos
Anderson J. Frailey
Spring 2025
Charlottesville, VA

[1] See Antman et al. (2025) for evidence of the program's impact.

Abstract

It next to impossible to participate in the modern economy without involving data. Whether a person is online shopping or applying for a loan, data is being collected and used to make inferences. This shift to a data rich world has had a number of benefits, but it has also raised many questions about privacy, the use of information, and data protection. The amount and quality of available data directly impacts markets, while the organizations holding valuable data are constantly at risk of suffering a cyberattack. This dissertation examines three topics at the intersection of cybercrime, privacy, and information.

In chapter 1, I study how data privacy regulations affect the market for stolen data. 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.

Chapter 2 continues with the study of cybercrime but with a focus on how it affects targeted firms, and how those effects depend on whether the firm discussed their risk prior to the event. I create a framework for understanding the effects of risk disclosure both at the initial stages and when the risky event actually occurs. Using a collection of cybersecurity incidents affecting publicly traded firms, I test the predictions of the framework by analyzing the market's reaction to the incidents and how it varies by disclosure status. I find that prior risk disclosure is not in and of itself predictive of how the market will respond to an event, but how the market reacts to the initial risk disclosure is. This chapter focuses on the role of information in markets, which transitions into my final chapter

In chapter 3, I shift my focus to the role of information in consumer credit markets, specifically studying how signal noise affects the credit constraints of borrowers and the risks faced by lenders. In consumer credit markets, credit scores are used as a signal of the creditworthiness of a borrower. Higher credit scores send a positive signal, increasing access to credit. As part of the federal response to the COVID-19 pandemic, the Coronavirus Aid, Relief, and Economic Security Act (CARES Act) paused collection on student loans and suspended collections on delinquent loans. The latter provision of the legislation lead

to beneficiaries of the policy seeing large increases in their credit score despite not taking personal action to rectify their delinquencies. This added noise to the signal sent by the credit scores. This paper studies the impact of that noise on consumer credit markets. I show that beneficiaries of the policy were more likely to open auto and credit card loans, and more likely to go delinquent on auto loans relative to two distinct control groups.

List of Tables

List of Figures

Chapter 1.

The Effects of Privacy Regulation on the Market for Stolen Data

1.1 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 chapter, 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

[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.

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 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 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

[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).

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 chapter 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 chapter is structured as follows. Section 1.2 formally presents the model of the stolen data market discussed earlier. In Section 1.3 I describe the data used in this study. My empirical strategy is defined in Section 1.4 and the results are presented in Section 1.5. I provided concluding remarks and paths for future research in Section 1.6.

1.2 A Model of Stolen Data Production

The production of stolen data can be described by a two-part "supply chain", depicted in figure 1.1. 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

[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.

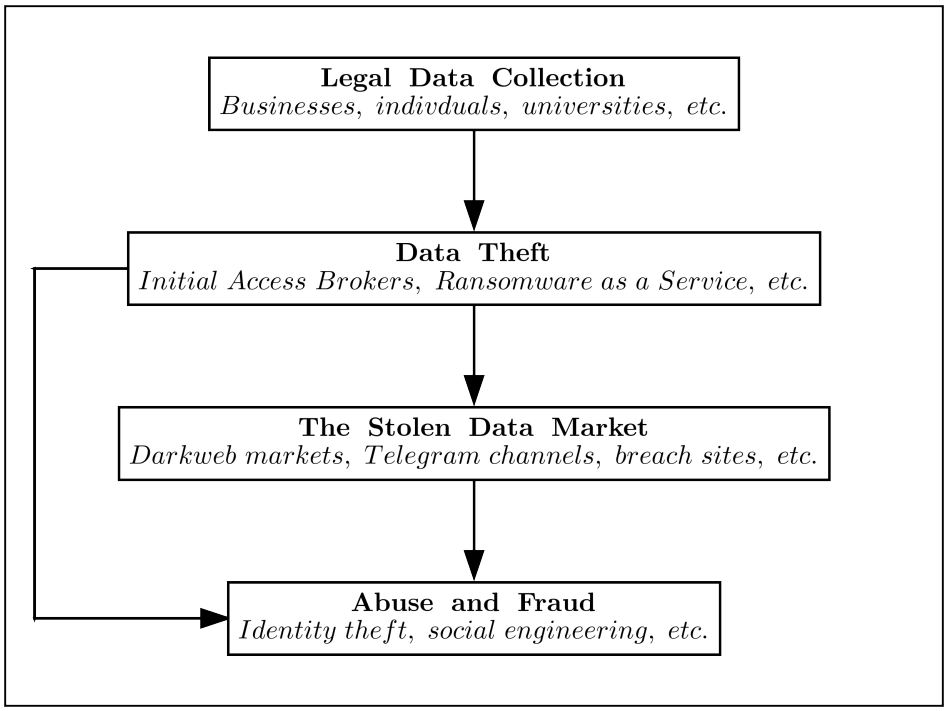


Figure 1.1: The Stolen Data Supply Chain

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 1.1). Both effects increase the cost of acquiring the key input to the market: the data itself.

[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 1.1: 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.

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.

1.2.1 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 $I(d_1, \ldots, d_J)$ 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]

[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

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 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:

a choice variable.

[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.

$$\max_{d_1,d_2,S} \quad A\left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu}{\rho}} - \omega_1 d_1 - \omega_2 d_2 - \omega_S S - \frac{r}{S+1}\left(\ell + \gamma_1 d_1 + \gamma_2 d_2\right). \tag{1.1}$$

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

$$\frac{r}{(S+1)^2}\left(\ell + \gamma_1 d_1 + \gamma_2 d_2\right) = \omega_S \tag{1.2}$$

$$\alpha_1 d_1^{\rho-1} \nu A \left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu-\rho}{\rho}} = \omega_1 + \frac{r}{S+1}\gamma_1 \tag{1.3}$$

$$\alpha_2 d_2^{\rho-1} \nu A \left(\alpha_1 d_1^{\rho} + \alpha_2 d_2^{\rho}\right)^{\frac{\nu-\rho}{\rho}} = \omega_2 + \frac{r}{S+1}\gamma_2 \tag{1.4}$$

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 1.2). 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 1.3 and 1.4).

Rearranging equation 1.2 reveals that the optimal S is:

$$S^* = \sqrt{\frac{r\left(\ell + \gamma_1 d_1^* + \gamma_2 d_2^*\right)}{\omega_S}} \tag{1.5}$$

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

Using equations 1.3 and 1.4, the optimal levels of data collection are described by the equations:

$$d_1^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2}\right)^{\frac{\rho}{1-\rho}}\right]^{\frac{\nu-\rho}{\rho(1-\nu)}} \tag{1.6}$$

and:

$$d_2^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1} \right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{\rho}{1-\rho}} \right]^{\frac{\nu-\rho}{\rho(1-\nu)}}. \tag{1.7}$$

Full derivations are in section A.1.1 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.

1.2.2 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 1.2. 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 1.2.

Equation 1.5 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.

1.2.3 The Stolen Data Market

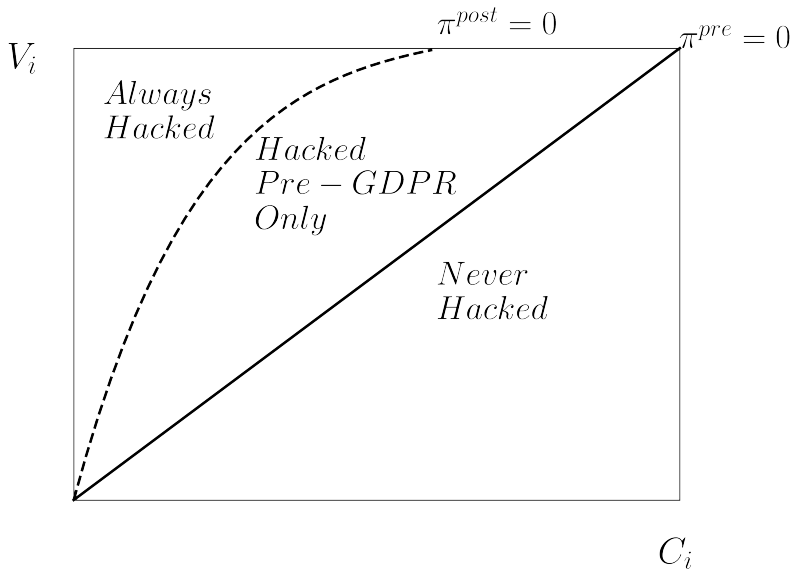
After stealing data from the original data collectors, hackers have the option of keeping or selling it in the market. Participants in this market are at least semi-anonymous and only the sellers know the true quality of the data they hold until it is sold, making it ripe for issues of adverse selection.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

Figure 1.2: 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.

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:

$$S(p) = \mathcal{B}\mathbb{P}\left(V_i \leq p \middle| C \leq V\right) \tag{1.8}$$

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: $\mu \equiv \mathbb{E}[V|C \leq V \leq p]$. 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.

1.2.4 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:

$$\begin{aligned} S(p) &= \mathcal{B}\mathbb{P}(V \leq p) \\ &= \mathcal{B}p^2 \end{aligned} \tag{1.9}$$

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:

$$D(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3}{2} \\ 0 & \text{Otherwise} \end{cases} \tag{1.10}$$

Equations 1.9 and 1.10 yield the pre-GDPR equilibrium:

$$\begin{aligned} p^* &= \left(\frac{Y}{\mathcal{B}}\right)^{\frac{1}{3}} \\ Q^* &= Y^{\frac{2}{3}}\mathcal{B}^{\frac{1}{3}} \end{aligned} \tag{1.11}$$

Full derivations can be found in section A.1.2 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

$$\mathbb{E}\left[V \middle| \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}} \leq V \leq \frac{p}{1-\phi}\right] = \frac{2-\sigma}{3-\sigma}\frac{p}{1-\phi}. \tag{1.12}$$

The supply of data packages on the market also changes:

$$\begin{aligned} S^{Post}(p) &= \mathcal{B}^{Post}\mathbb{P}\left(V \leq \frac{p}{1-\phi} \middle| \left(\frac{\theta}{1-\phi}C\right)^{\frac{1}{1-\sigma}} \leq V\right) \\ &= \mathcal{B}^{Post}\left(\frac{p}{1-\phi}\right)^{2-\sigma} \end{aligned} \tag{1.13}$$

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 1.12, the new minimum κ required for the market to exist is

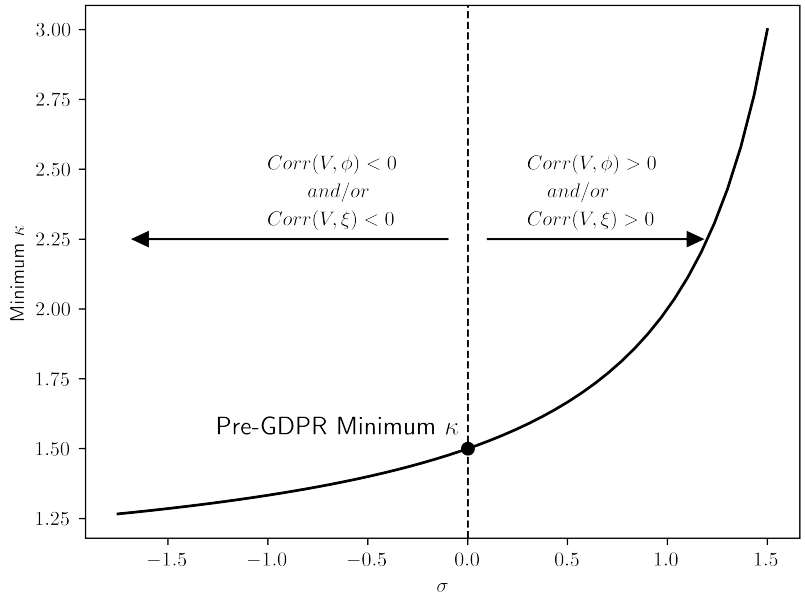
$$\kappa \geq \frac{3-\sigma}{2-\sigma}.$$

Demand is now

$$D^{Post}(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3-\sigma}{2-\sigma} \\ 0 & \text{Otherwise} \end{cases} \tag{1.14}$$

Figure 1.3 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

Figure 1.3: Minimum κ



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

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

$$\begin{aligned} p^*_{Post} &= \left(\frac{Y}{\mathcal{B}^{Post}}\right)^{\frac{1}{3-\sigma}} (1-\phi)^{\frac{2-\sigma}{3-\sigma}} \\ Q^*_{Post} &= Y^{\frac{2-\sigma}{3-\sigma}} \left(\frac{\mathcal{B}^{Post}}{(1-\phi)^{2-\sigma}}\right)^{\frac{1}{3-\sigma}}. \end{aligned} \tag{1.15}$$

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 increasing returns to security investment. For simplicity, I make ϕ a constant equal to 0.26.[7] Table 1.2 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

[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.

than 1.5 for the market to exist, I set it equal to $(3-\sigma)/(2-\sigma)$.

Table 1.2: 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 buyers will pay up to their expected utility gain (table 1.3, column one).

Table 1.3: 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

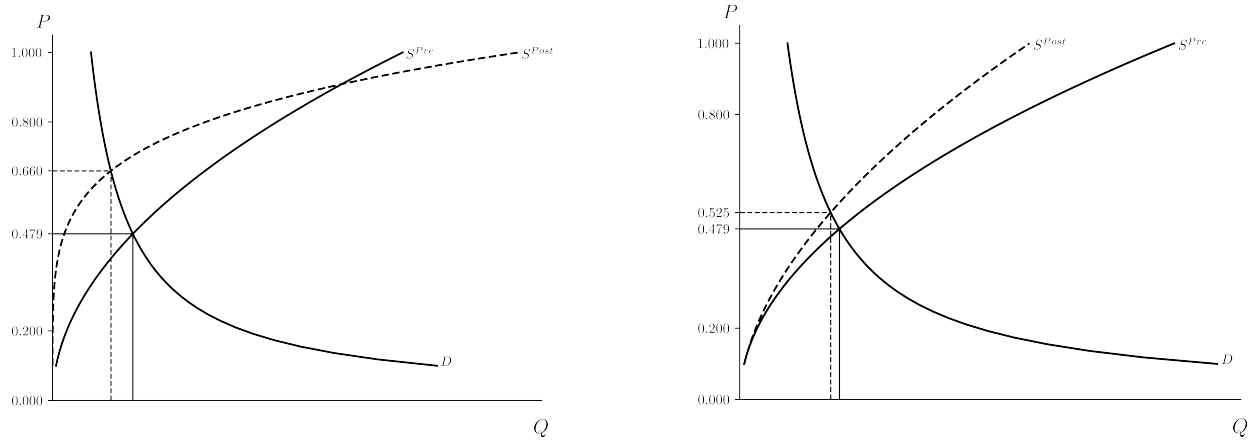
Notes: This table presents the results of the main simulation exercise in section 1.2.4. The simulations in columns one and two use $\kappa = 1.5$ while in column three κ is increased to 1.667 in order for demand to exist.

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 1.5, further increasing $\mathbb{E}\left[V\middle|Sold\right]$. The results from this simulation are in the second column of table 1.3. Figure 1.4a 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 1.3 and plotted in figure 1.4b. 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 1.4 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.

Figure 1.4: Simulated Equilibrium



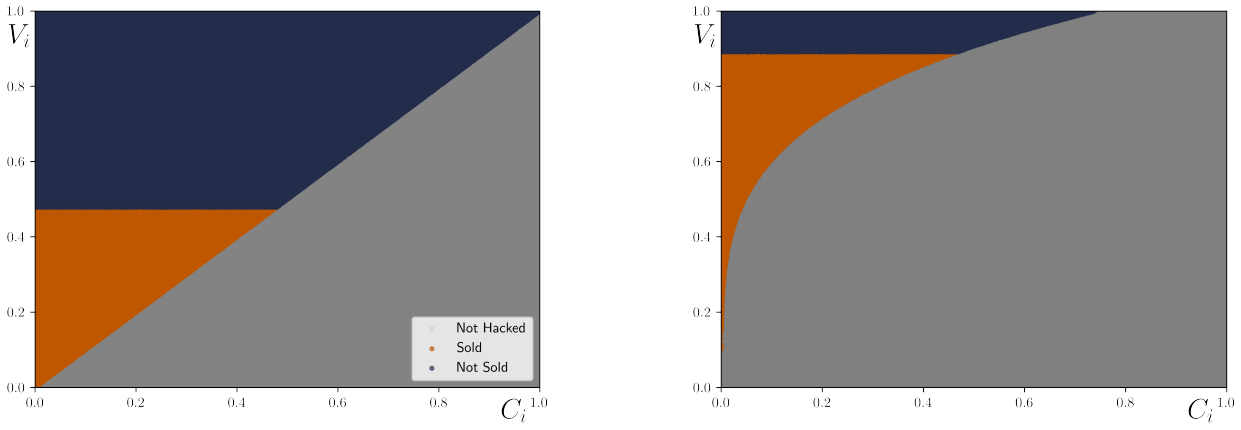
(a) $Corr(\xi, V) < 0$ *(b)* $Corr(\xi, V) > 0$

These simulations show that under the right conditions privacy regulations may actually increase the expected value of data packages stolen and traded. This reduces the adverse selection problem in the market and increases buyer utility.

1.3 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

Figure 1.5: Data Packages Sold and Not Sold



(a) Pre-GDPR

(b) Post-GDPR

Table 1.4: 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 1.2.4. 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 chapter to measuring just quantity effects.

Table 1.5 displays summary statistics for the three outcomes of interest in the study at

Table 1.5: 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

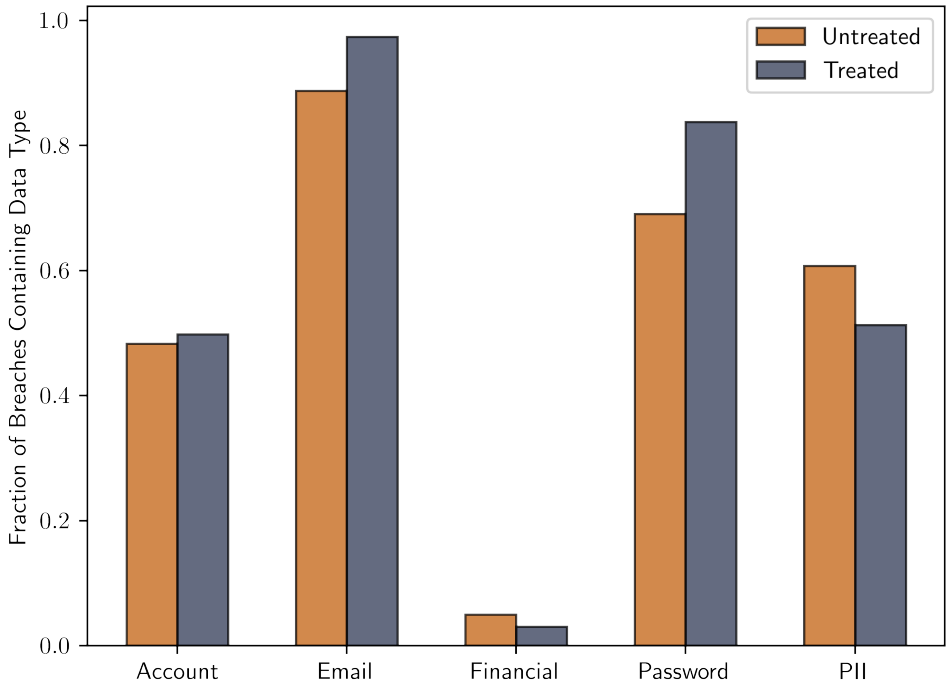
Notes: PII fraction is the fraction of records in a data package that are considered PII. A discussion of what constitutes PII is in the appendix.

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 A.2 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 1.6).

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

Figure 1.6: 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.

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 1.6 breaks down the number of data packages that fall into each category before and after the GDPR.

Table 1.6: 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

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

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

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 1.2. The value and cost of hacking organizations in regulated countries will be affected by the GDPR, and remain unchanged in unregulated countries. Choosing January 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.

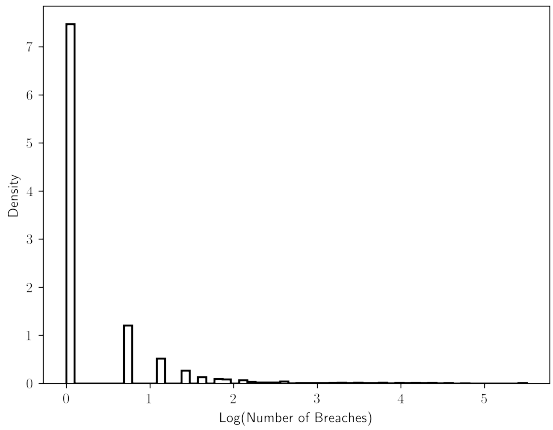
For robustness checks, I construct additional panels excluding any period after March 2020—to remove any bias introduced by the COVID-19 pandemic, and any data package originating from a multinational organization. The latter removes any bias that may arise due to partially treated organizations.

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 1.7 and 1.8, this creates mass points at zero for both variables. I discuss the implications of this for my estimation strategy in section 1.4.

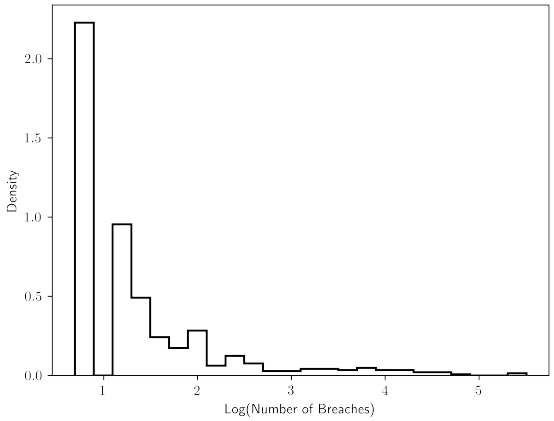
Roughly 27 percent of all country-quarters have a positive number of breaches (table 1.7), but among the positive observations there are an average of six breaches and 21.6 million records stolen (table 1.8). There is a large variation in both outcomes with as many as 245 breaches occurring and over one billion records being available in a quarter. Figures 1.9a and 1.9b 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

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

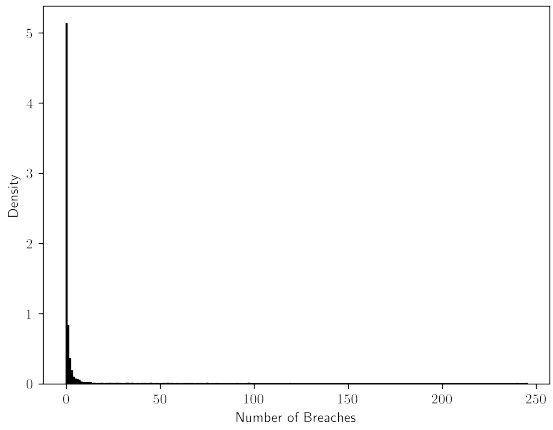
Figure 1.7: Distribution of the Aggregate Number of Data Breaches Per Country and Period



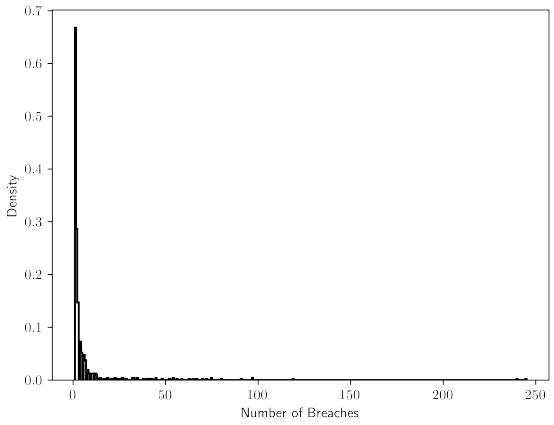
(a) Logged Using All Observations



(b) Logged Using Only Positive Observations



(c) In Levels Using All Observations



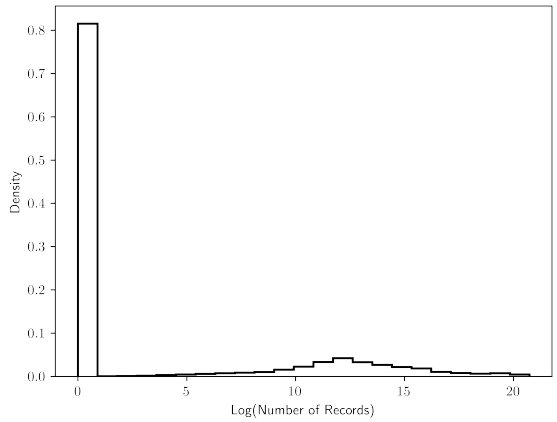
(d) In Levels Using Only Positive Observations

in each quarter.

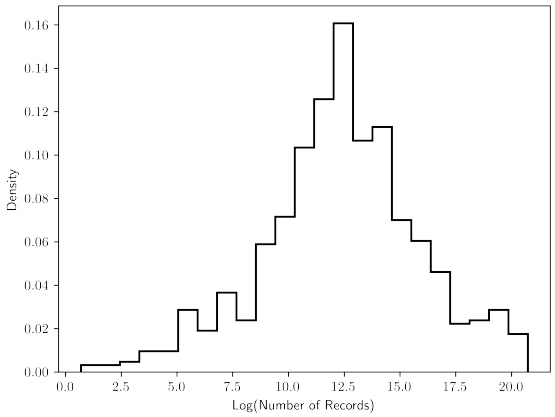
Table 1.7: Panel Summary Statistics

	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

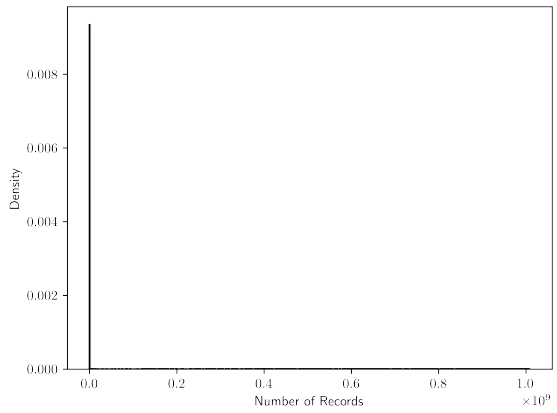
Figure 1.8: Distribution of the Aggregate Number of Records by Country



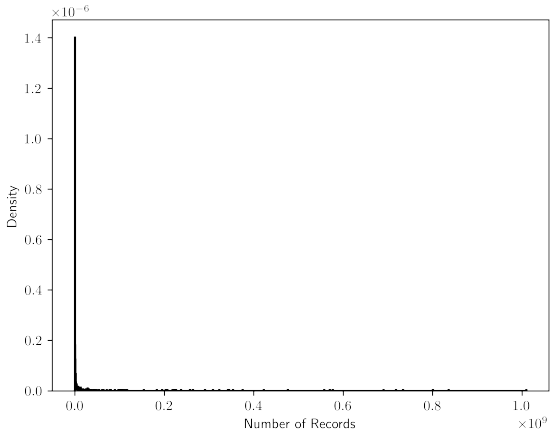
(a) Distribution of Log(Number of Records) Using All Observations



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



(c) Distribution of Number of Records Using All Observations

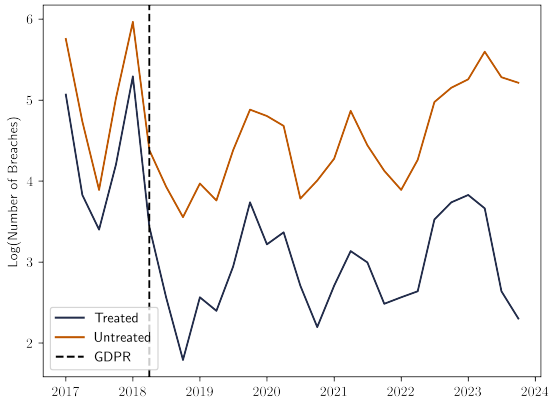


(d) Distribution of Number of Records Using Only Positive Observations

Table 1.8: 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

(a) Number of Data Breaches



(b) Number of Records Online

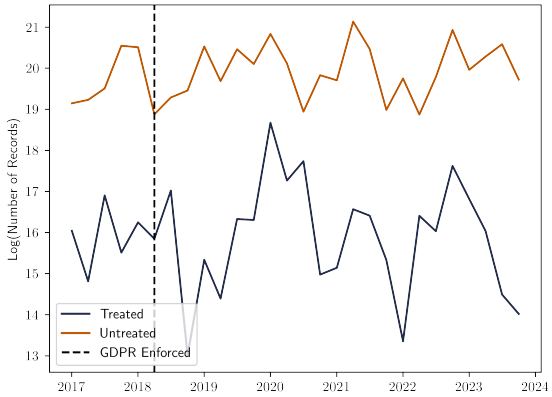


Figure 1.9: Number of Breaches and Records Time Series

The observations dropped from the final datasets because either their country of origin or breach date could not be determined tend to contain fewer records than those included in the study (figure 1.10). 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 1.11 plots the number of breaches and records included and excluded from the final sample over time.

1.4 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. The individual data package analysis will test this by examining the effect of the GDPR on

Figure 1.10: Distribution of Records Per Breach: Dropped vs. Included

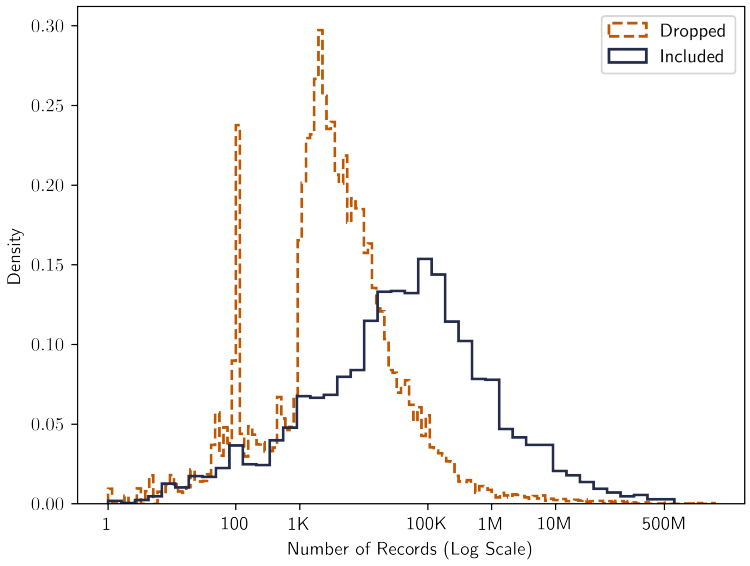
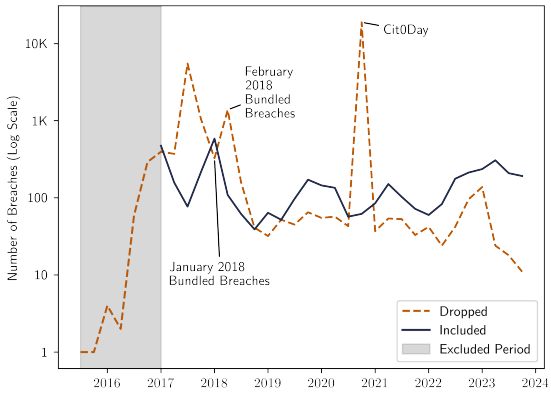
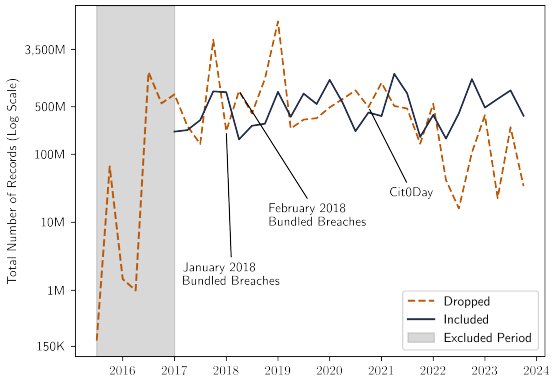


Figure 1.11: Comparison of Dropped and Included Breaches Over Time



(a) Number of Breaches



(b) Number of Records

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 1.3, 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]

1.4.1 Aggregate Effects

Aggregate effects are estimated using the country-quarter panel described in section 1.3. 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 1.7 and 1.8). The model I present in Section 1.2 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:

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

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:

$$Y_{it} = \exp\left(\gamma_i + \tau_t + \delta^{Agg} D_i \times Post\text{-}GDPR_t - log(population_{it})\right) \varepsilon_{it} \tag{1.16}$$

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.

[11] Robustness checks excluding data packages from multinationals are in section A.3.2 of the appendix, and their findings are discussed in sections 1.5.1 and 1.5.2.

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

$$Y_{it} = \exp\left(\gamma_i + \tau_t + \delta_{SR}^{Agg} Short\text{-}Run_t \times D_i + \delta_{LR}^{Agg} Long\text{-}Run_t \times D_i - log(population_{it})\right)\varepsilon_{it} \tag{1.17}$$

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):

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

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

$$Y_{it} = \exp\left(\gamma_i + \tau_t + \sum_{t \neq -1} \delta_{it}^{Agg} D_i \times Post\text{-}GDPR_t - log(population)\right)\varepsilon_{it} \tag{1.18}$$

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

Under the model in section 1.2, 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)

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

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 1.2 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.

1.4.2 Data Package Effects

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

$$y_i = \gamma_i + \tau_t + \delta^{DP} D_i \times \textit{Post-GDPR}_t + \epsilon_{it} \tag{1.19}$$

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 $\textit{Post-GDPR}$ term into short- and long-run effects and estimate:

$$y_i = \gamma_i + \tau_t + \delta^{DP}_{SR} D_i \times \textit{Short-Run}_t + \delta^{DP}_{LR} D_i \times \textit{Long-Run}_t + \varepsilon_i \tag{1.20}$$

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 1.2. 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.

1.5 Results

The main results are presented in sections 1.5.1 and 1.5.2. I discuss the results in the context of the model along with the limitations of this chapter in section 1.5.3. Robustness checks and alternative model specifications are discussed in section 1.5.4.

1.5.1 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 1.9). This effect is larger in the long-run than short-run (-22 percent versus -17 percent, respectively).

Table 1.9: 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.

The aggregate treatment effects on the number of breaches and total amount of data being taken from a country are presented in table 1.10. 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 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

Table 1.10: 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.

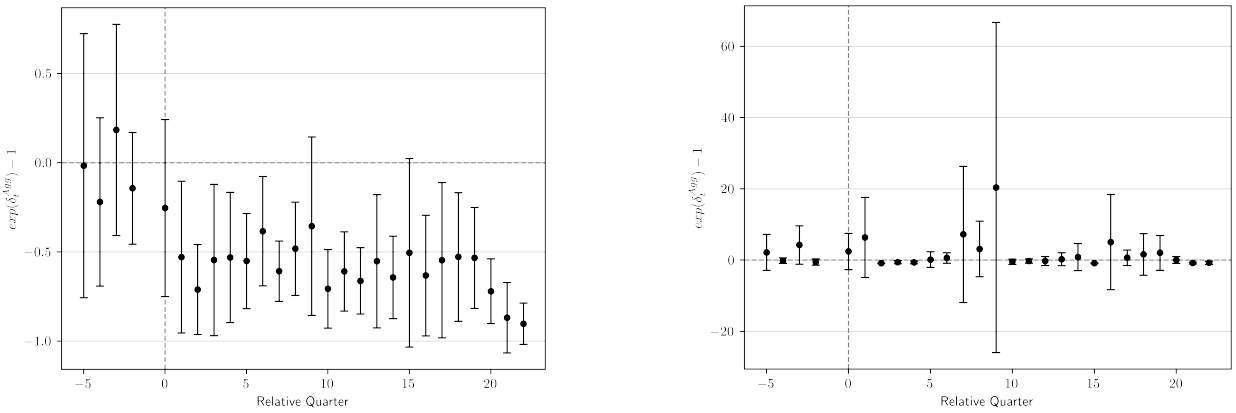
remaining breaches contain significantly more data, which my data package-level analysis finds. This could be caused by higher-value targets with more data becoming a larger share of the breaches traded in the market.

Event study plots to provide evidence that the parallel trends assumption holds are in figure 1.12. 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 1.12a). Each period of the event study shows no statistically significant effect on the number of records traded (1.12b).

1.5.2 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 1.11). 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

Figure 1.12: 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 1.18 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.

size of the data packages is counterintuitive on its face. If data privacy legislation 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 1.2.4, 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 1.13 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 1.12). Given that most of the data in the packages qualifies as PII (table 1.5), 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 1.13) or number of unique types of data in the packages (table 1.14). One potential explanation for this is that only certain types 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 1.13: Number of Records Density

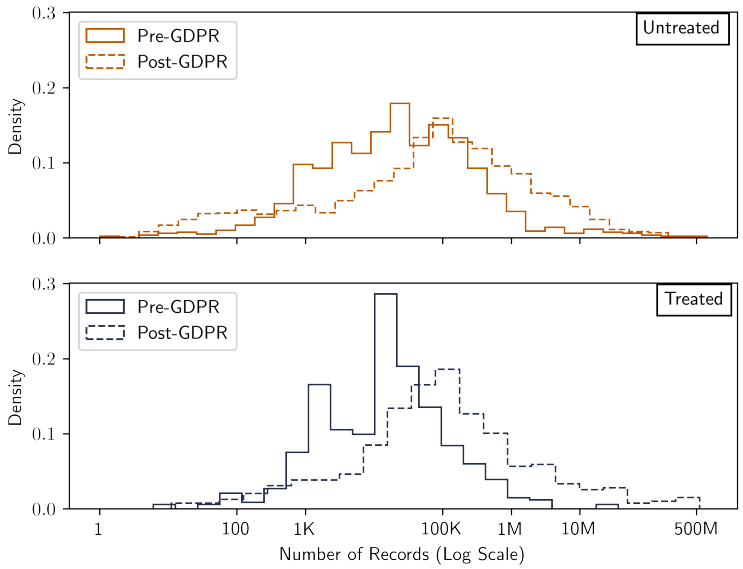


Table 1.11: Data Package Effects: Number of Records

	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 1.12: 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 1.13: 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 1.14: 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.

1.5.3 Discussion and Limitations

The above results are consistent with what the model in section 1.2 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 chapter. 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.

1.5.4 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 A.6 and A.7). 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 A.6 and A.7). 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 A.17 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 A.11-A.16. 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 A.12) and there is no obvious change in the number of records available overtime (figure 1.9b), 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 A.8).

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 A.9.

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] 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 A.10.

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 A.3 and A.4, 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

[13] In the course of writing this chapter 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 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.

percent in the main model, though is still within the standard error (table A.27). 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 A.28 and A.29). Though statistically significant, a half of a data type increase holds little economic value.

1.6 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 chapter 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 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 chapter 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 chapter 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.

Chapter 2.

Information and the Market Reaction to Cybersecurity Incident Disclosures

2.1 Introduction

When markets are efficient, the risk of an adverse event will be priced into firm market values when it becomes public knowledge. Upon realization of the event, firm share prices should change to reflect both the fully realized loss and any new information gained from the event. The total size of this change will therefore depend in part upon how much of the risk was priced in, and the accuracy of the prior loss expectations. Both of these factors can be influenced by the amount and type of information previously known about the firm's risk profile. One avenue companies will use to communicate these risks is their annual and quarterly filings to the Securities and Exchange Commission (SEC) where there are sections specifically dedicated to discussing risks to the firm.

Cybersecurity risks—ransomware attacks, data breaches, and other computer crimes—have become a particularly persistent threat to businesses big and small as the world has shifted to increasingly relying on digital infrastructure to operate. The cost of these incidents include repairing computer systems, lost revenue from forced downtime, and lost reputation. Cybersecurity risk has become so pervasive that in 2023 the SEC adopted a rule requiring firms dedicate a new section of their annual 10-K filings to discussing it.[1] Even before this new regulation, it was becoming increasingly common for firms to discuss cybersecurity risks in the risk factors section of their annual 10-K and quarterly 10-Q financial filings.

This chapter examines the role of those prior information disclosures in determining how the market responds to news of a cybersecurity incident. I first present a framework describing the effects of risk disclosure on firm stock price and the market's response to

[1] United States Code of Regulations Title 17 Chapter II Part 229 Subpart 229.100 §229.106.

the realization of an adverse event. Using a set of publicly traded firms who disclosed cyberattacks, I then test the predictions of the framework. I begin by analyzing each firm's 10-K and 10-Q filings to determine whether they disclosed cybersecurity risks, then estimate abnormal returns around each filing and the cybersecurity incidents themselves. I find that whether a firm disclosed risk prior to suffering a cyberattack is not by itself predictive of how the market will respond to the attack. However, the details of the market's response to the initial risk disclosure is.

The framework herein begins with the simple assumption that publicly known risks faced by firms will be priced in to their market value. When the risk is revealed, the firm's market value falls by the expected loss. The framework centers around this risk being determined primarily through firm disclosures in SEC filings, but its principle implications will apply regardless of whether the information is provided directly by the company or gleaned from outside sources such as investor research or the experiences of comparable firms.

When the event occurs, firm market value again falls. This fall has two components: the unrealized expected losses and a learning effect (Kamiya et al., 2021). Unrealized losses are the remainder of the actual loss less the expected loss already priced in, The learning effect is a result of investors adjusting their beliefs over the loss distribution faced by the firm from that type of risk. If the loss distribution worsens, the learning effect is negative, causing a larger loss of market value.

I show that whether risk disclosure raises or lowers firm market value depends on whether the firm is low or high-risk, relative to prior expectations. Upon realization of the event, the market's response will be more negative for high-risk firms than low-risk firms if the information revealed in the disclosure is about the size of the losses that would be incurred, rather than the probability of the event. In both cases, the size of the response to the event relative to had they not disclosed will differ for high and low-risk firms. With high-risk firms, the disclosure gave the market an opportunity to price in more of the risk prior to the event than if they had not disclosed. Signaling low-risk in the disclosures, however, makes it more surprising when the event occurs, causing a more negative response.

The framework has three implications that can be tested empirically. First, the response to disclosure should be more negative for high-risk (relative to prior expectations) firms than low-risk. In fact, low risk firms should actually see an increase in value because they revealed their risk is lower than what investors expected prior to disclosure. Second, whether the firm discloses risk is predictive of the size of the response to the event itself, but that effect works in opposite directions for low and high-risk firms. Third, the response to disclosure reveals whether investors view the firm as high- or low-risk relative to non-disclosers, which in turn reveals information on the relative size of the response to the event. The details of this

last point depend on whether the firm is disclosing potential losses or the probability of the event.

To test these implications, I conduct event studies around firm 10-K and 10-Q filings and cybersecurity incidents. Filing data are accessed via the SEC EDGAR portal and incidents are both manually collected from news reports and drawn from the Privacy Rights Clearinghouse Data Breach Chronology. For each filing I observe whether the firm discussed cybersecurity risk in the filing's risk-factors section by conducting a keyword search for a number of cybersecurity-related words and phrases. For each incident, I observe the targeted firm and the first date it became public knowledge.

I calculate cumulative abnormal returns around each filing and regress this result on indicator variables for whether the firm discussed cybersecurity risk in the filing. I find that on average there are no abnormal returns around each filing, and that the initial mention of cybersecurity risk is not predictive of the market's response to the filing. Subsequent mentions of risk, however, increase abnormal returns by just under one percent. In the context of the framework, this may imply that the effects of high and low-risk disclosure are simply offsetting, rather than being evidence that there is no effect at all.

Around each incident, I find negative and significant abnormal returns of slightly less than one percent. This is almost entirely driven by the market's response the day the event becomes public knowledge. Regressing these returns on a set of covariates that include an indicator for whether the firm had previously disclosed cybersecurity risk in a filing, I find that whether a firm disclosed their risk prior to the attack seems to have no influence on the market's response. However, when I remove the disclosure indicator from the regression and instead including the actual abnormal returns around the first filing to discuss the firm's risks, I find that the response to the initial filing is highly predictive of the response to the event itself. That the indicator alone is not predictive of the event response but the response to the initial filing is suggests that there may be offsetting effects between firms. As I will show in my framework, the market's response after a high-risk firm suffers an incident will be smaller in magnitude if they had previously disclosed their high-risk status than it would be if they had not. The opposite is true for low-risk firms. The market's response to the event will be larger in magnitude if they disclosed their risk level than if they had not. In expectation, the size of the market's response for disclosing firms relative to if they had not may appear to be zero, depending on the relative portion of each type of firm in the sample. So while the firm-specific filing response will be indicative of the eventual response to the event itself, simply stating whether the firm did or did not disclose their risk may not be.

This chapter joins a series of prior literature on how the stock market responds to announcements of cybersecurity incidents (Acquisti et al., 2006; Spanos and Angelis, 2016;

Smith et al., 2018; Makridis and Dean, 2018; Makridis, 2021; Tosun, 2021; Kamiya et al., 2021; Akyildirim et al., 2024). As with these previous works, I find that the market does respond negatively to news of an incident. While the negative response is both intuitive and well established, there is still much to be learned about what factors determine the size of the response. This is the first paper to extend the analysis to the role prior risk disclosure plays in determining the response magnitude. More broadly, this work contributes to the policy debate over what constitutes sufficient risk disclosure on the part of firms. My finding that the market's response to the initial risk disclosure is positively correlated with its response to the event itself suggests that the information firms disclose even prior to the SEC's new disclosure rules provided valuable information to investors.

Previous research has also studied how the market consumes information. Feroz et al. (1991) for example find that news of SEC enforcement action against a firm for reporting errors cause -13 percent abnormal returns, but even if the firm had previously announced their errors there are still abnormal returns of -6 percent. These results suggest that the full cost of a risk action, even when realization is likely, is not fully incorporated into firm stock prices until it actually happens.

Li and Ramesh (2009) focus specifically on information learned in quarterly reports, finding that there are only abnormal returns if those reports coincide with, or are themselves, the first time earnings information is revealed to the public. Stice (1991) also suggests that investors may not rely on the 10-K and 10-Q filings themselves to learn about companies. Stice finds evidence to suggest that the information contained in these reports was not fully priced in until it was also reported in the media. This result may have been a product of its time. Asthana and Balsam (2001) show that the SEC's adoption of the EDGAR reporting system sped up the diffusion of information after a 10-K filing. That said, an extensive review of the literature by Blankespoor et al. (2020) finds that learning from these disclosures is an "active economic choice", rather than a costless process.

This chapter adds to this section of the literature by showing that the market does seem to be partaking in that active choice to learn about cyber risk from formal filings. Additionally, my finding that the market's response to these incidents is abnormal returns of -0.88 percent is similar in size to previous research opens the door to future work on how exactly the market is learning about cybersecurity risk. As Kamiya et al. (2021) propose, much of the negative response in the market to cybersecurity events can be attributed to investors updating their beliefs about the loss distribution firms face. Overtime, it would not be unreasonable to expect that investors eventually converge on the true loss distribution, reducing the size of the market's response to these events. That I find a similar market response as Kamiya et al. despite using a sample of events that occurred well after the last

of their events calls into question whether that convergence is happening. Future research can explore whether that expectation is simply wrong, if, as in Feroz et al. (1991), they wait until the event actually happens to react, or if the loss distribution has continued to evolve overtime.

The remainder of the paper is structured as follows. My conceptual framework of the influence of risk disclosure in the market reaction to adverse events is presented in section 2.2. Data used for the study are described in section 2.3. My empirical strategy is detailed in section 2.4, and the results presented in section 2.5. I conclude with a discussion of the results, their implications, and future research in section 2.6.

2.2 Risk Disclosure Framework

This framework modifies the model in Kamiya et al. (2021) to show how the market's response to an adverse event still vary based on information the firm previously disclosed. Events occur with probability $p \in (0, 1)$ and cause firms to incur a loss of L. When firms disclose their risk, they are specifically disclosing information on p and/or L. I discuss the implications of each type of disclosure below.

2.2.1 Loss Disclosure

Under this disclosure regime, firm-specific losses from the event, L_i, are announced. For simplicity, I assume that the probability of the event, p, is a constant.

The presence of risk reduces the firm's stock price, P, to

$$P = V_i - p\tilde{L}$$

where V_i is their risk-free value. For disclosing firms, $\tilde{L} = L_i$. For non-disclosing firms, $\tilde{L} = \bar{L}$, the expected losses for non-disclosing firms.

There are two types of firms in the model, high-risk firms where $L_i = L_h$, and low-risk firms where $L_i = L_\ell$. The designation between high and low-risk is relative to prior expectations, rather than a specific threshold value. The loss values are set such that $L_\ell < \bar{L} < L_h$. When a firm switches from not disclosing to disclosing, their market value will change to reflect the change in risk:

$$\Delta P_d = -p\left(L_i - \bar{L}\right). \tag{2.1}$$

For low-risk firms, this disclosure will cause investors to view them as less risky than before,

increasing their share price. The opposite is true for high risk firms. Denoting the share of disclosing firms that are low-risk as π, the expected market response will be dependent on the weighted average of the response to each firm type's disclosures:

$$\begin{aligned}\mathbb{E}\left[\Delta P_d\right] &= \pi\left[-p\left(L_\ell - \bar{L}\right)\right] + (1-\pi)\left[-p\left(L_h - \bar{L}\right)\right] \\ &= p\bar{L} - \left[\pi p L_\ell + (1-\pi)\, p L_h\right].\end{aligned} \tag{2.2}$$

In words, if the disclosing group has a similar risk composition as the non-disclosing group so that the weighted average response is the same as the non-disclosed expected loss, the expected response to disclosure will be zero. Even though both types of firm have a non-zero response to disclosure, positive in the case of low-risk firms and negative in the case of high-risk firms, the two may cancel each other out in expectation unless one type of firm is more likely to disclose than the other.

When the event occurs, it will trigger a second price change to account for the unrealized losses:

$$\Delta P_e = -\left(1-p\right)\tilde{L}. \tag{2.3}$$

Given that $L_\ell < L_h$, high-risk firms will once again see a larger loss than low-type firms. The expectation, however, will once again depend on the composition of the disclosing firms. This value will be:

$$\begin{aligned}\mathbb{E}\left[\Delta P_e \middle| \textit{discloser}\right] &= \pi\left[-\left(1-p\right)L_\ell\right] + (1-\pi)\left[-\left(1-p\right)L_h\right] \\ &= -\left(1-p\right)\left[\pi L_\ell + (1-\pi)\, L_h\right].\end{aligned} \tag{2.4}$$

For firms that did not disclose, the expected response to the event is simply

$$\mathbb{E}\left[\Delta P_e \middle| \textit{non-discloser}\right] = -\left(1-p\right)\bar{L}.$$

As with the expected response to disclosure, the expected response to the event given the firm did disclose will be close to that of the expected response of non-disclosers if $\pi L_\ell + (1-\pi)\, L_h = \bar{L}$.

The above assumes that investors do not update their loss expectations after an event. However, Kamiya et al. (2021) show that, in the case of cybersecurity incidents, there is evidence that much of the market's reaction to the even is a result of learning more about the losses firms face. Learning will affect the previously disclosing firms and non-disclosing firms differently.

Accounting for the possibility of learning requires a slight change in notation. Event

probability before and after the initial occurrence is $p_t, t \in \{0,1\}$. The distinction between p_0 and p_1 allows for the possibility that investors learn the event probability is higher or lower than previously believed. For simplicity, I continue to assume that, even if it is updated, the event probability is the same for both types of firms. The losses for each type of firm are $L_{i,t},\ i \in \{\ell, h\},\ t \in \{0,1\}$.

The change in price after the event with learning is

$$\Delta P_e^l = -\left(1-p_0\right) L_{i,0} + \underbrace{\left(p_0 L_{i,0} - p_1 L_{i,1}\right)}_{\text{Learning Effect}}. \tag{2.5}$$

The expectation of the price change among disclosing firms will now be:

$$\begin{aligned}
\mathbb{E}\left[\Delta P_e^l \middle| discloser\right] &= \pi\left[-\left(1-p_0\right) L_{i,0} + \left(p_0 L_{\ell,0} - p_1 L_{\ell,1}\right)\right] \\
&\quad + \left(1-p_0\right)\left[-\left(1-p_{\ell,0}\right) L_{i,0} + \left(p_0 L_{h,0} - p_1 L_{h,1}\right)\right] \\
&= -\left[\pi\left(1-p_0\right) L_{\ell,0} + \left(1-\pi\right)\left(1-p_0\right) L_{h,0}\right] \\
&\quad + \pi p_0 L_{\ell,0} + \left(1-\pi\right) p_0 L_{h,0} \\
&\quad - \left[\pi p_1 L_{\ell,1} + \left(1-\pi\right) p_1 L_{h,1}\right]
\end{aligned} \tag{2.6}$$

For non-disclosing firms, the event itself will reveal their type. Denoting the portion of previously non-disclosing firms that are revealed to be low-risk as π_n, the expected response among non-disclosing firms is:

$$\begin{aligned}
\mathbb{E}\left[\Delta P_e^l \middle| non\text{-}discloser\right] &= \pi_n\left[-\left(1-p_0\right)\bar{L}_0 + \left(p_0 \bar{L}_0 - p_{\ell,1} L_{\ell,1}\right)\right] \\
&\quad + \left(1-\pi_n\right)\left[-\left(1-p_0\right)\bar{L}_0 + \left(p_0 \bar{L}_0 - p_{h,1} L_{h,1}\right)\right] \\
&= -\left(1-p_0\right)\bar{L}_0 + p_0 \bar{L}_0 - \left[\pi_n p_0 \bar{L}_0 + \left(1-\pi_n\right) p_1 L_{h,1}\right]
\end{aligned} \tag{2.7}$$

Comparing equations 2.6 and 2.7, shows that the difference between the expected price change between firms that disclose and firms that do not will still depend on how similar the two pools are in terms of risk composition, even with learning.

The same logic can be applied to determine whether the market's response to an incident will be larger or smaller had the firm decided to disclose versus not disclose. For either type of firm,

$$\mathbb{E}\left[\Delta P_e^l \middle| discloser\right] > \mathbb{E}\left[\Delta P_e^l \middle| non-discloser\right] \ \ if\ L_{i,0} > \bar{L}_0.$$

Again using the assumption that $L_\ell < \bar{L} < L_h$, high-risk firms that disclose will have a larger response relative those that do not while low-risk firms have a smaller response. Because the

expected response to the event is likely to be negative, this actually means that the response to a known high-risk firm is smaller in magnitude—i.e., less negative—than the market response for a known low-risk firm relative to a world in which they had not disclosed their risk.

2.2.2 Probability Disclosure

If firm disclosures instead center around p, the implications of the framework change slightly. Assume now that L is constant. As before, firm market value is lower with risk than without it

$$P = V_i - \tilde{p}L$$

where V_i is once again their risk-free value, but now $\tilde{p} = p_i$, firm specific risk, if they disclose and $\tilde{p} = \bar{p}$, the expected probability of the event among non-disclosers, if they do not.

With L held constant, high and low-risk firms are now distinguished by their values of p. The values of p are set so that: $p_\ell < \bar{p} < p_h$. As before, there will be an initial change in price when firms disclose their risk level:

$$\Delta P_d = -\left(p_i - \bar{p}\right) L.$$

The expectation of this value will once again be a weighted average of the expected losses:

$$\mathbb{E}\left[\Delta P_d\right] = \pi\left[-\left(p_\ell - \bar{p}\right) L\right] + \left(1 - \pi\right)\left[-\left(p_h - \bar{p}\right) L\right] \tag{2.8}$$

When the event occurs, the change in price will still be the unrealized costs

$$\Delta P_e = -\left(1 - \tilde{p}\right) L.$$

Unlike the loss disclosure case, the response to the event will actually be *smaller* for high-risk firms than low-risk firms. This is due to a greater share of the losses being priced in prior to the event. Empirically, the size of the market's response to the event will be inversely related to the size of the response to the initial filing. High-risk firms draw a more negative response upon initial disclosure, but because a larger share of the losses are priced in at that point, the secondary change is smaller. The opposite is true for low-risk firms.

With an expected loss without learning of

$$\begin{aligned} \mathbb{E}\left[\Delta P_e\right] &= \pi\left[-\left(1 - p_\ell\right) L\right] + \left(1 - \pi\right)\left[-\left(1 - p_h\right) L\right] \\ &= -\left[1 - \left(\pi p_\ell + \left(1 - \pi\right) p_h\right)\right] L \end{aligned} \tag{2.9}$$

and expected loss with learning of

$$\begin{aligned}\mathbb{E}\left[\Delta P_e^l\right] &= \pi\left[-\left(1-p_{\ell,0}\right)L_0 + \left(p_{\ell,0}L_0 - p_{\ell,1}L_1\right)\right] \\ &\quad + \left(1-\pi\right)\left[-\left(1-p_{\ell,0}\right)L_0 + \left(p_{\ell,0}L_0 - p_{\ell,1}L_1\right)\right] \\ &= -L_0 + 2L_0\left[\pi p_{\ell,0} + \left(1-\pi\right)p_{h,0}\right] - L_1\left[\pi p_{\ell,1} + \left(1-\pi\right)p_{h,1}\right],\end{aligned} \tag{2.10}$$

the expected losses of disclosures relative to non-disclosers will again depend on the relative portion of each type of firm. If they appear in a proportion such that $\pi p_{\ell,1} + (1-\pi)\, p_{h,1} = \bar{p}$, there will be no difference between them.

2.2.3 Empirical Predictions

This framework has three implications that can be tested empirically. First, the response to the initial risk disclosure will depend on whether they are high or low-risk. For high-risk firms, disclosure reduces their stock price. For low-risk firms, disclosure increases their stock price. The expected market response to disclosure will ultimately depend on which type of firm is most likely to disclose (equation 2.2). If they disclose in similar proportions, the effects will offset.

Second, whether a firm discloses their risk is predictive of the direction of the market's reaction to the event itself. However, that relationship goes in opposite directions for high and low-risk firms. As a result, the actual observed effects may also cancel each other out unless one type of firm dominates the disclosures.

Finally, the market's response to individual firm disclosures is predictive of its response to the event realizations, but depends on whether they are disclosing the losses they would face or the probability the event occurs. If the former, the response to the event will be smaller for low-risk firms than high-risk. If the latter, the response will be smaller for high-risk than low-risk firms.

2.3 Incident, Stock Price, and Company Data

Data on cybersecurity incidents come from two sources. First, events were manually collected by searching filings with the Securities and Exchange Commission (SEC) and news outlets for reports of attacks on publicly traded firms. Beginning with the SEC's EDGAR tool, I searched for form 8-K filings that included terms to suggest there had been an incident

such as "cybersecurity attack."[2][3] If the form disclosed that the firm had been a victim of cybercrime, or provided an update on a previously announced attack, I then searched for the first news articles related to the attack to determine the day that the news became public. If the 8-K filing date and original news date were different, I used the earliest of the two for the event date in my study. To find events that were not announced via an 8-K filing, I searched major news outlets such as Reuters, The Wall Street Journal, and CNBC; as well as specialty websites such as Bleeping Computer, for reports of cybersecurity incidents. I once again defined the event date as the day of the earliest public reporting on the attack.

The second source is the Privacy Rights Clearinghouse's (PRC) Data Breach Chronology Database. The PRC chronology lists cybersecurity incidents disclosed to state and federal government agencies between 2005 and 2023. The entire chronology covers breaches of businesses, healthcare providers, nonprofits, educational institutions and governments. I search among just the breaches that were classified as hacks against businesses. After this initial filtering, I further culled the data to only businesses that were publicly traded at the time of the breach, and with breaches whose announcement dates I could verify through news reports, government disclosures, or firm statements.

In total, I observe 166 events with sufficient information to identify the date the news became public information. A full list of the events used is in appendix table **??**. Table 2.1 shows the number of firms in each industry in both the event sample and among all firms registered with the SEC.

Company and stock market data were accessed via the Wharton Research Data Services (WRDS) platform. Individual stock data were retrieved from the CRSP database. I observe the daily return of each asset along with their trading volume. To estimate the market model discussed in section 2.4.1, I also use the market return, risk-free return (measured as the return on a one-month US Treasury bond), and returns on the small-minus-big and high-minus-low portfolio returns, as defined in Fama and French (1993). These data are obtained from the WRDS Fama-French Factor database.

In the event studies, I use the daily excess return as the outcome of the market model used to calculate expected returns. This is calculated by subtracting the risk-free return from the overall return for each asset on each day. The distributions of excess return on the days used to estimate the market model and during the event period are plotted in figure 2.1 and described in table 2.2. As can be seen in figure 2.2, the mean excess return during the event window is generally similar to that of the market mean, except for the day of the

[2] https://www.sec.gov/edgar/search/

[3] Form 8-K filings are used to announce significant events that would be of interest to shareholders. This would include unexpected events such as a cyberattack.

Table 2.1: Victim Firm Industry's

	Treated Firms		Overall	
	N	%	N	%
Energy	1	0.68	1,111	7.28
Materials	3	2.04	991	6.50
Industrials	28	19.05	1,766	11.58
Consumer Discretionary	27	18.37	1,907	12.50
Consumer Staples	7	4.76	643	4.22
Health Care	20	13.61	2,433	15.95
Financials	16	10.88	2,496	16.36
IT	22	14.97	2,645	17.34
Communication Services	16	10.88	651	4.27
Utilities	2	1.36	248	1.63
Real Estate	5	3.40	363	2.38

Notes: This table shows the number and fraction of firms in each industry. The first two columns are firms in the sample, the third and fourth are among all firms registered with the SEC.

event where it is noticeably lower.

Other firm characteristics, such as their total assets, industry, and number of employees are obtained through the Compustat database, also via WRDS. Summary statistics for each variable from the fiscal year prior to the event are in table 2.3. They are compared to the means of the never treated firms, also in the year prior to each event. All dollar figures are adjusted to be in 2021 dollars.

Table 2.2: Returns Above the Risk Free Rate

	Estimation Window	Event Window
Count	36,520	1,826
Mean	0.057	-0.090
St. Deviation	3.071	3.307
Min.	-43.006	-19.932
25%	-1.090	-1.247
50%	0.045	0.019
75%	1.194	1.200
Max.	67.234	32.268

Notes: This table shows the unconditional mean return on stocks above the risk-free rate in the sample for the market model estimation period prior to the event and during the event window. These are not the estimated abnormal returns, simply the average.

Figure 2.1: Distribution of Returns Above the Risk Free Rate

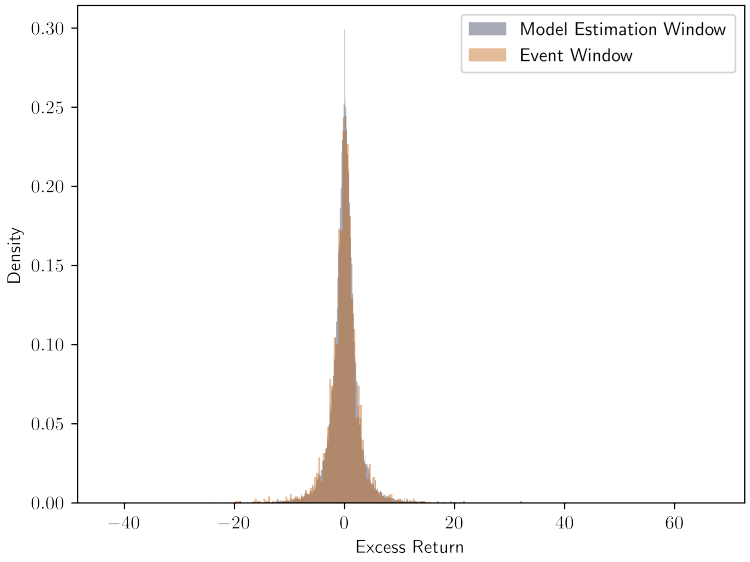
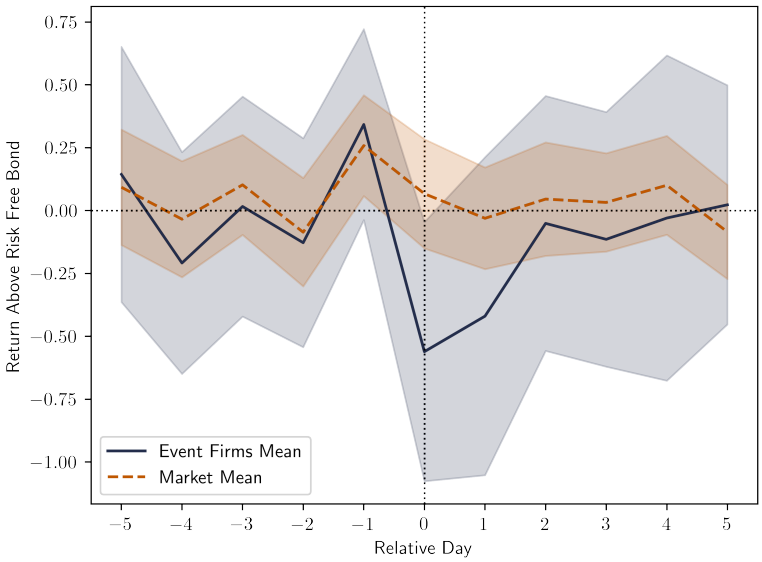


Figure 2.2: Average Daily Return Above Risk-Free Rate



Note: These are the average returns of both the market and affected firms around the event. This is not the CAR, just the average of the raw returns. The shaded areas around the lines show the 95 percent confidence interval of the mean each day.

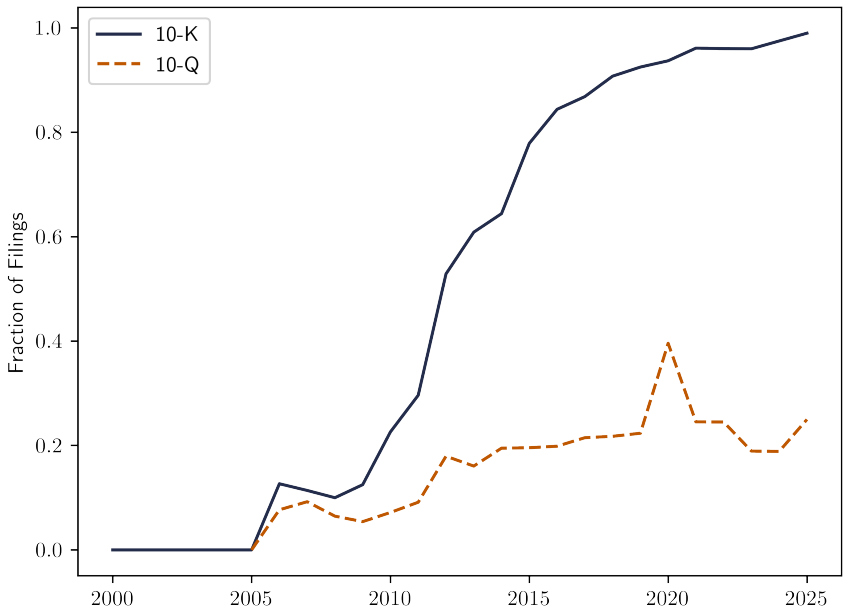
Table 2.3: Summary Statistics

	Assets	Market Value	Revenue	Tobin's Q	Intangible Ratio	ROA	Sales Growth	Employees
Obs.	166	166	166	166	166	166	141	166
Mean	56,524.13	51,983.92	20,775.61	2.48	0.27	0.02	0.10	52.33
Std. Dev.	226,219.65	142,572.20	41,727.31	2.37	0.24	0.10	0.25	83.94
Min.	20.60	19.98	15.70	0.58	0	-0.55	-0.50	0.06
25%	1,564.57	1,833.36	911.43	1.15	0.07	-0.01	0	3.29
50%	7,566.82	7,286.03	4,045.63	1.60	0.21	0.04	0.00	13.95
75%	31,105.81	37,619.15	20,673.25	2.87	0.47	0.07	0.12	65.13
Max.	2,760,475.34	1,085,179.08	276,644	14.28	0.91	0.24	1.44	400

Notes: Total assets, market value, and total revenue are in millions of 2021 dollars. Employees are measured in thousands. These statistics are taken from the annual filing in the fiscal year prior to a firm's attack.

Company risk disclosure status was determined using their annual 10-K and quarterly 10-Q SEC filings in the years prior to each event. I used the SEC's EDGAR API to download and analyze each firm's 10-K and 10-Q filings.[4] I excluded all tables, figures, and headings from the text analysis. For each company, I performed a keyword search in the risk factors section of their filing for mentions of cybersecurity risk. I marked a filing as discussing cybersecurity risk if at least one of those words was found in section 1A, the risk factors section, of the filing. A firm is designated a discloser if they discuss cyber risk in at least one filing prior to their incident. A list of the keywords used is in table B.1 of the appendix.

Figure 2.3: Fraction of Filings Mentioning Cybersecurity Risk



As shown in figure 2.3, the portion of firms that discuss cyber risk in their annual filings has increased to nearly 100 percent over time. The strategic disclosure framework in section 2.2 does result in all firms disclosing risk, however that happens immediately rather than over nearly two decades as observed in the data. This is the first evidence that in reality firms behave according to the threshold disclosure framework, rather than strategic disclosure.

Almost all firms in the sample discuss their risk in a filing prior to suffering from a cyberattack (table 2.4). There are some firms who discuss their cyber risk in one filing, then decline to do so in a subsequent filing, as seen by the differences in the prior risk mentions columns of panels A and B in table 2.4. Interestingly, while some firms who had not included the risk of a cyberattack in their filings prior to their incident begin to afterward, not all do.

[4]To access the API, I used the EDGARTools Python package developed by Dwight Gunning. https://edgartools.readthedocs.io/en/latest/

Table 2.4: Mentions of Cyber Risk

Panel A: All Forms

	Prior	Post	Switch
All Forms	0.813	0.873	0.078
10-K	0.807	0.873	0.084
10-Q	0.416	0.602	0.253

Panel B: Forms Nearest Event

	Prior	Post	Switch
All Forms	0.380	0.554	0.247
10-K	0.789	0.837	0.054
10-Q	0.253	0.404	0.193

Notes: Panel A summarizes the percentage of firms who discussed cyber risk in filings prior to and after their attack. The *Switch* column is the percentage that had not discussed risk prior to the event, but began to in any period afterward. Panel B. shows the same, but limits the sample to the forms filed closest to the event date. There are differences because some firms mention cybersecurity risk in one filing, then do not in a subsequent filing.

The number of events I observe in each year is listed in table 2.5. The second, third, and fourth columns count the number of that year's targeted firms who had disclosed cybersecurity risk in a filing prior to their incident.

One of the control variables I use when analyzing the response of the market to discussion of cyber risk in an SEC filing is the overall sentiment of the document. I measure the sentiment of the filing using a bag-of-words approach. Using the negative words list developed by Loughran and Mcdonald (2011) for sentiment analysis of financial filings, I define the sentiment of the document as the weighted count of negative words divided by the total word count. Words as weighted using term frequency-inverse document frequency (tf-idf), which attenuates the influence of common words on the overall sentiment. The weight of word i in document j is

$$w_{ij} = \begin{cases} (1 + \log(tf)) \times \left(\log\left(\frac{1+N}{1+df_i}\right) + 1\right) & \text{if } tf_{ij} > 0 \\ 0 & \text{if } tf_{ij} = 0 \end{cases}$$

where tf_{ij} is the number of times word i appears in filing j, N is the number of filings, and df_i is the number of times word i appears in all filings. Sentiment distributions for both types of filings are shown in figure 2.4.

Table 2.5: Number of Events Each Year and Disclosure Status

		Prior Cyber Risk Disclosure		
Year	Number of Events	Any Filing	10-K Filing	10-Q Filing
2006	1	0	0	0
2007	1	0	0	0
2010	2	0	0	0
2012	2	2	2	0
2013	2	2	2	2
2014	6	3	3	1
2015	4	2	2	1
2016	6	5	5	3
2017	9	7	7	2
2018	12	11	11	6
2019	33	28	28	9
2020	51	41	41	24
2021	29	26	25	16
2022	8	8	8	5

Notes: This table contains the number of events in each year, and the number of those events that discuss cyber risk in an SEC filing prior to the event.

2.4 Measuring and Examining Market Reactions

There are two parts to the empirical analysis. First, I estimate the abnormal returns in the market after SEC filings discussing cyber risk and cybersecurity incidents. Next, I analyze the influence of cybersecurity risk disclosures on each set of returns.

2.4.1 Cumulative Abnormal Returns

To measure the market's reaction to firm filings and incident disclosures, I first calculate the abnormal returns, defined as:

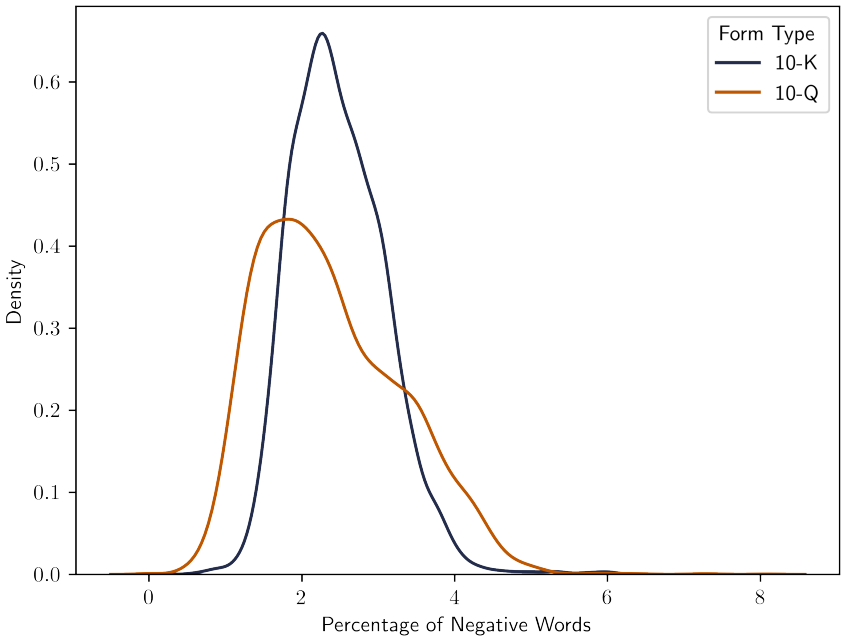
$$AR_{it} = R_{it} - \mathbb{E}\left[R_{it}\right] \tag{2.11}$$

where R_{it} is their realized return on day t and $\mathbb{E}[R_{it}]$ is their expected return. $\mathbb{E}[R_{it}]$ is estimated using the Fama-French three-factor model (Fama and French, 1993):

$$R_{it} = R_{ft} + \alpha + \beta_{i1}(R_{mt} - R_{ft}) + \beta_{i2}SMB_t + \beta_{i3}HML_t + \varepsilon_{it}. \tag{2.12}$$

where R_{ft} and R_{mt} are the return on a risk-free asset and market return, respectively. Small-minus-big, SMB_t, and high-minus-low, HML_t returns are the excess return on small-cap

Figure 2.4: Sentiment Distribution



companies and value stocks, as defined in Fama and French (1993).

Daily abnormal returns are summed over the window t_1 to t_2 to get cumulative abnormal returns:

$$CAR_i(t_1, t_2) = \sum_{t=t_1}^{t_2} AR_{it}. \tag{2.13}$$

Finally, the reported effect is the average cumulative abnormal response:

$$\overline{CAR}(t_1, t_2) = \frac{1}{N} \sum_{i=1}^{N} CAR(t_1, t_2). \tag{2.14}$$

Statistical significance is tested using the t-ratio from Kolari and Pynnönen (2010):

$$t\ statistic = \frac{\overline{CAR}\sqrt{n}}{S_{CAR}\sqrt{1 + (n-1)\bar{r}}}$$

where $\bar{r}$ is average of the cross-correlations of model residuals from the estimation period, and S_{CAR} is the standard deviation of the scaled CAR.[5] Using this test statistic has the advantage

[5]Scaled CAR are the sum of the scaled abnormal returns. The scaled abnormal return for firm i on day t is $SAR_{it} = \frac{AR_{it}}{s_i\sqrt{1+d_t}}$. Here, s_i is the standard deviation of the residual from the model fit for firm i, and d_t is a correction term equal to $x_t' \left(X'X\right)^{-1} x_t$. In this correction term, x_t is a vector of the values of the model's independent variables on day t, and X is the matrix of those values throughout the estimation period. For a detailed discussion of this test statistic, see Kolari and Pynnönen (2010).

of adjusting for any correlation present due to events happening in similar windows.

2.4.2 Risk Disclosure Effects

After estimating the mean CAR surrounding each filing and cybersecurity incident, I estimate the effects of risk disclosure.

The framework in section 2.2 presented three testable implications. First, the expected response to risk disclosure will depend on whether high or low-risk types are more likely to disclose. If the former, disclosure should result in negative abnormal returns in expectation. If the latter, disclosure should result in positive returns. When the proportions of each type are equal, the two effects cancel out. To estimate the effect of cyber risk disclosure in 10-K and 10-Q filings on abnormal returns around the filing date, I estimate the linear model:

$$CAR_i^{filing} = \alpha + \varphi First_i + \phi Risk_i + \theta Sentiment_i + \Gamma X + \mu_i + \tau + \epsilon \tag{2.15}$$

where CAR^{filing} is the cumulative abnormal return after filing i. $First$ equals one if the filing was the first time the firm discussed cyber risk. $Risk$ indicates whether the firm discussed cyber risk in the filing, regardless of whether it was the first mention. $Sentiment$ is the filing's sentiment score, as defined in section 2.3. Filing and firm specific controls are in the matrix X. These include an indicator for whether the filing is a 10-Q and firm age. μ and τ are firm and calendar year fixed effects, respectively. As a robustness check, I also estimate this model on 10-K and 10-Q filings separately, removing the 10-Q indicator from the list of covariates.

The second implication is that the market's response to the event will depend on whether they previously disclosed their risk, but whether that is detectable will depend on whether the disclosing group is predominately high or low-risk firms. The disclosure effect goes in opposite directions for each group so it is possible that the two cancel each other out. At the event level, I estimate:

$$CAR^{event} = \alpha + \psi Risk + \beta X + \gamma + \tau + \epsilon_i \tag{2.16}$$

where $Risk$ is equal to one if the firm involved had any mention of cybersecurity risk in a 10-K or 10-Q filed prior to their incident. The firm and event specific attributes in X are the logs of firm market value and total liabilities; the portion of their assets that are intangible (intangible ratio); Tobin's Q in the fiscal year prior to their event; and indicators for whether the event included ransomware and whether the firm had had a prior incident. Industry, γ, and year τ fixed effects control for attitudes towards cyber risk that are shared across

industry and time. I use industry fixed effects rather than firm fixed effects as in equation 2.15 because most firms appear in the sample only once.

Finally, I test whether the market's response to risk disclosure is predictive of its response to the event itself. Recall from section 2.2 that a firm's market value will fall if they disclose themselves to be high-risk, and increase if they reveal they're low-risk. Additionally, when the event happens, there should be a larger (in magnitude) response for high-risk firms than low-risk if the disclosures focus on loss rather than probability. While it is not possible to isolate the portion of the filing response that is attributable to the risk disclosure, the abnormal returns around that filing will incorporate it. At the incident level, I estimate

$$CAR_i^{event} = \alpha + \theta CAR_i^{filing} + \beta X_i + \gamma_i + \tau + \epsilon_i \tag{2.17}$$

CAR_i^{filing} is the abnormal return for each firm around the first filing in which they discuss cybersecurity risk. For those that never discuss cybersecurity risk, this will be zero. All other variables are defined as before. For robustness, I also estimate this equation using only firms that disclosed risk and using a dummy variable equal to one if CAR^{filing} was negative after their initial risk disclosure.

2.5 Results

Results are presented in two parts. First, I focus on the SEC filings. I show that there is no systematic response, positive or negative, to a 10-K or 10-Q release. Additionally, whether the firm discussed cyber risk for the first time in those filings is not predictive of CAR^{filing}. Second, I find the cybersecurity incidents cause a slightly less than one percent decline in firm market value. This is also not affected by whether the firm had previously discussed their cyber risk in a 10-K or 10-Q. However, CAR^{filing} is positively correlated with CAR^{event}.

2.5.1 Market Response to Filing Disclosures

I find no significant abnormal returns around SEC filing dates (table 2.6). This result holds both whether the effect is estimated among the full population of filings or estimated for 10-K and 10-Q filings separately (column 1 versus columns 2 and 3 of table 2.6).

The results of estimating equation 2.15, the regression of CAR^{filing} on risk disclosure indicators, are in table 2.7. The framework in section 2.2 shows that the market's reaction to the initial risk disclosure will vary based on whether the firm is low or high-risk. For low-risk firms, their price will increase, and it will decrease for high-risk firms. In expectation, these effects may cancel each other out. Figure 2.5 shows the distribution of CAR^{filing}.

Table 2.6: Filing Abnormal Returns

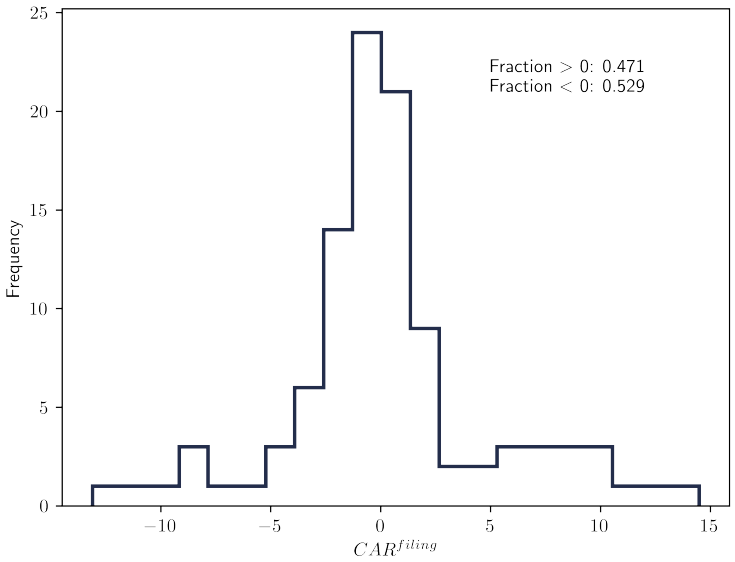
	All Filings	10-K Filings	10-Q Filings
t	(1)	(2)	(3)
-1	0.015	0.068	-0.005
	(0.041)	(0.071)	(0.049)
0	0.006	-0.054	0.028
	(0.060)	(0.097)	(0.074)
1	-0.051	-0.079	-0.041
	(0.051)	(0.078)	(0.064)
CAR	**-0.030**	**-0.064**	**-0.017**
	(0.089)	**(0.144)**	**(0.110)**
Observations	7,098	1,911	5,187
Model	Three Factor	Three Factor	Three Factor

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

Notes: This table contains abnormal returns after each filing. This is CAR^{filing}. The first column contains the mean abnormal and cumulative abnormal returns after all filings. The second two columns separately estimate the mean abnormal and cumulative abnormal returns for 10-K and 10-Q filings, respectively.

There are close to an even number of firms who had positive and negative filings. It is possible that the effects of each are offsetting.

Figure 2.5: CAR^{filing} Distribution



Interestingly, there is a positive and significant effect for just mentioning cybersecurity

Table 2.7: Disclosure Effect on the Market Response to Filings

	Dependent Variable: CAR(-1, 1)	
	(1)	(2)
Intercept	-2.733	5.186
	(215.880)	(371.233)
First Cyber Risk Mention	-0.316	-0.312
	(0.589)	(0.591)
Mentions Cyber Risk	0.901***	0.906***
	(0.335)	(0.342)
10-Q	0.522*	0.616**
	(0.266)	(0.292)
Age		-0.282
		(0.411)
Sentiment		-0.011
		(0.142)
Year Fixed Effects	Yes	Yes
Firm Fixed Effects	Yes	Yes
Observations	7,098	7,098
R^2	0.026	0.026
F Statistic	1.207**	1.207**
Model	OLS	OLS

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

Notes: This table shows the results of regressing CAR^{filing} on an indicator for whether the filing discussed cybersecurity risk. Both 10-K and 10-Q filings are included in the sample. Estimates splitting the sample by form type are in section B.3 of the appendix.

risk of 0.9 percent. This is consistent with the strategic disclosure framework where low-risk firms use their disclosures to signal that they are low-risk. Though, as previously discussed, the lack of immediate disclosure by all firms suggests that strategic disclosure is unlikely to be reality. Additionally, unless new information is being revealed in subsequent filings, the response should still be concentrated on the first disclosure, where there are no significant results. This result disappears when I limit the sample to just 10-K filings (appendix table B.5). It remains and grows larger when I limit the sample to 10-Q filings (appendix table B.6). This suggests that there is some specific information derived from disclosing risk in the 10-Q filing that may be valuable. As shown in figure 2.3, it is significantly less common to discuss these risks in 10-Qs than 10-Ks.

2.5.2 Market Response to Events

When the event is realized, there is a clear negative response in the market. The average CAR ranged from -0.880 to -1.539 percent depending on the length of event window (table 2.8). This finding is similar in magnitude to previous literature.

Table 2.8: Incident Abnormal Returns

	(-1, 1)	(-2, 2)	(-5, 5)
t	(1)	(2)	(3)
-5			-0.014
			(0.200)
-4			-0.203
			(0.182)
-3			-0.155
			(0.194)
-2		0.037	0.037
		(0.179)	(0.179)
-1	-0.126	-0.126	-0.126
	(0.153)	(0.153)	(0.153)
0	-0.507**	-0.507**	-0.507**
	(0.231)	(0.231)	(0.231)
1	-0.247	-0.247	-0.247
	(0.297)	(0.297)	(0.297)
2		-0.041	-0.041
		(0.218)	(0.218)
3			-0.176
			(0.208)
4			-0.098
			(0.295)
5			-0.010
			(0.174)
CAR	**-0.880*** **	**-0.884** **	**-1.539* **
	(0.391)	**(0.537)**	**(0.791)**
Observations	166	166	166
Model	Three Factor	Three Factor	Three Factor

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

Notes: This table contains abnormal returns after cybersecurity incidents, CAR^{event}. The numbers above each column represent the start and end date for the event windows. Each row in the table is the abnormal return on day t. CAR is the sum of the abnormal returns over the event window. The standard errors of the means are in parentheses.

Estimating equation 2.16, I find that prior risk disclosure is not predictive of CAR around a cybersecurity event (table 2.9). Along with the results in table 2.7, this suggests that high and low-risk firms are offsetting.

It is possible that my findings are influenced by there being very few firms who had not disclosed risk prior to their incidents. Only about 20 percent of firms in my sample did not discuss cybersecurity risk in a filing prior to their incident (table 2.4). This is a product of my sample time period, which contains more recent event than the previous literature. The

majority of my events occur after 2018, after it became common for firms to discuss cyber risk in the filings (figure 2.3). Additionally, by 2018 all 50 states also had a law requiring organizations to notify customers when they suffered a data breach.[6] Data breach disclosure laws may increase the likelihood of firms discussing cybersecurity risks in their SEC filings as they impose an additional cost that would be incurred in the event of an incident.

Table 2.9: Disclosure Effect on the Market Response to Incidents

	Dependent Variable: CAR(-1, 1)			
	(1)	(2)	(3)	(4)
Intercept	-0.838	-1.658	0.104	0.001
	(60.546)	(22.255)	(18.595)	(20.604)
Disclosed Risk	-1.285	-0.679		
	(1.109)	(1.260)		
CAR^{filing}			0.288**	0.290**
			(0.118)	(0.130)
Not First Event		-0.844		-2.832*
		(1.458)		(1.683)
Ransomware		0.574		-0.003
		(1.436)		(1.086)
Log(Market Value)		-0.285		-0.157
		(0.469)		(0.517)
Tobin's Q		-0.015		0.033
		(0.356)		(0.321)
Intangible Ratio		-2.451		-0.264
		(2.835)		(2.226)
Log(Liabilities)		0.286		0.147
		(0.486)		(0.503)
Year Fixed Effects	Y	Y	Y	Y
Industry Fixed Effects	Y	Y	Y	Y
Observations	166	166	135	135
R^2	0.099	0.117	0.138	0.174
F Statistic	0.950	0.860	1.422	1.969***
Model	OLS	OLS	OLS	OLS

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

Notes: The results in this table are from estimating equation 2.16. It shows the effects of prior risk disclosure on the market's response to cybersecurity incidents.

These results also contain information about how investors may process cybersecurity

[6]Alabama was the last state to adopt a data breach disclosure law with the Alabama Data Breach Notification Act of 2018 (Acts 2018-396).

risks. The average CAR^{event} in this chapter is similar to that of Kamiya et al. (2021), who show that a significant portion of the negative market response is investors updating their beliefs about the loss distribution after the event. This suggests that investors may still rely heavily on the event itself to learn about the value of the risk. That could be because the information in these filings is not meaningful, or investors are unwilling to incur the costs associated with learning about risk until it is realized. Alternatively, the losses from these events may be getting larger. This would offset any progress investors make in correctly identifying p.

The final prediction of the framework I propose is that the variance in market response will be greater among disclosing firms than non-disclosing firms. This is due to the simplifying assumptions that all firms face the same loss after and event, and that all non-disclosing firms are assigned the same prior expected loss. As a result, the market's response to an incident is the same for all non-disclosers, making the variance zero. Naturally, this will not be exactly true in reality. Well-informed investors may assign firm specific risks to those who do not disclose. Comparing observed variance in CAR^{event} for disclosers and non-disclosers, I find that the variance is higher among disclosing firms, but difference between the two, however, is not statistically significant (table 2.10).

Table 2.10: CAR^{event} Variance

	Variance			
Window	Disclosers	Non-Disclosers	Difference	P-Value
(-1, 1)	27.875	13.904	13.971	0.233
(-2, 2)	50.747	32.690	18.057	0.546
(-5, 5)	109.116	82.626	26.491	0.694

Notes: The p-value for each difference is calculated using Levene's test.

2.6 Conclusion, Implications, and Future Research

Under efficient market models, future risk should be incorporated into firm's market value if it is public knowledge. This chapter has developed a framework defining how markets may account for that risk, and how prior risk disclosure should affect the response to the realization of adverse events. I show that if the disclosure generates new, negative, information to investors, firm share price should fall. The effect of prior risk disclosure on the market's reaction to realizations of the event will be opposite for high and low-risk firms. This is true regardless of whether there is heterogeneity is losses or probability. It will also

depend on whether the firms disclose information on the losses they expect to incur or on the probability of the event happening.

Empirically, I find that there is no systematic negative response to the first time a firm discusses cybersecurity risk in their 10-K and 10-Q filings. Under my proposed framework, this implies that the responses to high and low-risk firms may be offsetting each other in expectation, depending on the relative proportion of each type that discloses.

Using a collection of cybersecurity incidents against publicly traded firms, I found an average response of -0.880 percent in the days immediately following the event, inline with previous research. I then showed that previous risk disclosure in 10-K and 10-Q filings was not predictive of the magnitude of the response, providing further evidence of offsetting effects.

Next, I find that the market's response to the first filing in which the firm disclosed their cyber risk is highly predictive of its subsequent response to the events themselves. Firms where the market responds negatively to the disclosure (i.e., high-risk firms) also suffer a more negative response in the market when they suffer a cyberattack. This is consistent with the first iteration of the framework in section 2.2 where there is heterogeneity in loss after an attack, but share the same event probability.

This chapter does not study whether the need to disclose risk and mitigation efforts has other benefits such as incentivizing firms to be more proactive in their defense. Even if the losses would be the same when actualized, increased attention on cyber risk may lower the probability of an incident in the first place, benefiting firms and customers alike. Future research could determine whether there are other such benefits to requiring firms to disclose cybersecurity risks or whether we have reached a point where this risk is ubiquitous enough that investors are able to properly calibrate expectations without much additional information from firms.

Chapter 3.

Unreliable Information in Consumer Credit Markets

In collaboration with Sarah Turner

3.1 Introduction

A person's credit score is a key signal of their creditworthiness to lenders. Formed from individual's borrowing and repayment history, these scores reveal information about their risk type, signaling to lenders how likely they are to pay back their loans.[1] Higher credit scores increase access to borrowing opportunities, and can allow for better terms on those loans because they indicate that the borrower is less risky.

As part of its response to the 2020 COVID-19 pandemic, the United States government paused activity around student loans, temporarily suspending payments and removing delinquent status from derogatory loans. Intended to last a few months, the pause ultimately stayed in place for over three years. After the pause went into effect, many beneficiaries of the policy saw large increases in their credit scores because of the student loan pauses. This was particularly true for borrowers who had previously been delinquent on their student loans. In effect, by suspending the collection of delinquent loans, the policies made these borrowers appear less risky than their credit history would suggest, adding noise and bias to the signal sent to lenders by their credit scores and potentially reducing the reliability of information in the market.

This chapter studies how that noise and bias affected consumer credit markets, focusing on auto loans and credit cards. Using a panel of consumer credit data, I estimate whether those who had a delinquency removed from their credit report because of the pause are

[1] See Arya et al. (2013) for an in-depth discussion of credit score formation.

more likely to take out new loans and go delinquent on a loan than their non-benefiting counterparts. I also estimate what characteristics of the borrower determined how large of an increase in credit score they experienced between March and September 2020.

The pause on student loan payment and collection loosened both liquidity constraints and credit score related borrowing constraints for beneficiaries. Liquidity constraints fell because they no longer needed to make payments on their loans, borrowing constraints through the aforementioned credit score increases. If these constraints were binding, then beneficiaries may respond by increasing the amount they borrow. But if the signal sent by credit scores was an accurate representation of their fundamental risk as borrowers, then distorting that signal by artificially inflating their credit scores may also result in these borrowers obtaining more credit than they are qualified for. If this is the case, they may also have higher delinquency rates than their post-pause credit score would predict.

Data for this study come from the one of the three major credit bureaus. With these data I observe information on individual loans including their scheduled payments, account status, and payment history. At the person level, I observe credit score history and demographic information. These data are recorded in March and September of each year and for this chapter I use the 2018–2023 files. I am able to identify beneficiaries of the pause by observing whether the scheduled payment on any of their loans fell to zero in September 2020 while the account balance remained positive. Among those loans flagged as being affected by the pause, I also mark whether they were delinquent in March 2020 but not in September of that same year.

These data are used to construct an unbalanced person-level panel. For the analysis, I designate the treatment group as the subset of payment pause beneficiaries who also had a delinquency resolved due to the policy. I focus on this subgroup, rather than beneficiaries as a whole, because they saw the largest credit score increase between March and September 2020. I use a propensity score matching process to construct two control groups from the non-beneficiaries. The first group is based on matches to the treated group using observations up to March 2020. This is the "but for" group. Those who are most similar to the treated leading up to the payment pause, and who, under a parallel trends assumption, are most like what they would be but for the policy. It is relative to this group that the treatment groups sees their credit constrained loosened. If the credit constraint was previously binding, beneficiaries may borrow more after the policy than this first control group is able to.

This is a partial equilibrium outcome that assumes no changes in how lenders behave. It is possible that creditors will become less willing to lend because the information environment has degraded. Narajabad (2012) proposes a model showing that improvements in information technology needed to assess risk expands credit markets as lenders can provide more tailored

terms to even risky borrowers. Reversing that logic, by reducing the quality of information available to lenders, the payment pause may cause the credit market to contract. Related, Chatterjee et al. (2023) conduct a counterfactual experiment in which lenders are unable to observe borrow history and find that interest rates will increase, potentially reducing borrowing. Each of these are examples of why information degradation may reduce the amount of available credit, and therefore reduce borrowing.

Furthermore, the control group may be affected by the policy if, due to known bias in credit scores, lenders choose to assume all credit scores are inflated. If this were the case, the control group will be harmed by their lack of score inflation, reducing their borrowing opportunities. As a result, the treatment effect I estimate on account openings will include both the benefit to the treated and the harm done to the control group. It should therefore be interpreted as an upper bound on the overall effect.

Empirically, I find evidence that borrowing does increase among the treated group. Payment pause beneficiaries were 1.07 and 4.55 percentage points more likely to open an auto loan and credit card, respectively, after March 2020 than the first control group. With auto loan opening rates fluctuating between four and six percent in most periods, and credit card opening rates between six and nine percent, these are substantial percentage increases.

The second control group is created using the same matching process, but using observations up to September 2020. This is the group that beneficiaries appear to look like, in terms of credit score, after the policy. However, because their new (higher) credit score is a result of a temporary government policy, rather than actions taken to change their credit score such as catching up on missed payments, it is possible that their underlying risk type is unchanged. As a result, they may be more likely to become delinquent on their loans than those in this post-pause match group.

As with the pre-pause matches, the post-pause matches may be harmed by the policy if lenders assume their credit scores are inflated and choose to offer them fewer loans or worse terms on loans they do offer. This could have a positive or negative effect on delinquency rates. It may decrease delinquency rates among the control group because they cannot go delinquent on loans they are not offered. This would lead to the treatment effect being overstated. Delinquency rates may increase in the control group if they are given worse terms, such as higher interest rates, that affect their ability to stay current on their payments. If lenders are not able to distinguish between treated and untreated, they should both receive similar terms on their loans, mitigating this bias in the treatment effect estimate.

I estimate that beneficiaries were 0.43 percentage points more likely to go delinquent on an auto loan and -1.08 percentage points less likely to go delinquent on a credit card than those in the post-pause match group. As with openings, those numbers are small in

magnitude, but large relative to population delinquency rates.

This chapter contributes to the literature on both the role of information in consumer credit markets and the effects of student loan relief. Seminal work on the former in Jaffee and Russell (1976) and Stiglitz and Weiss (1981) demonstrate how adverse selection and information asymmetry can erode credit markets. Chatterjee et al. (2023) show that a lending market based on credit score can reach an equilibrium equivalent to a market where a probability is assigned to each potential hidden risk type via Bayesian updating, demonstrating the importance of accurate credit scores in solving information problems. Unsurprisingly, noise in the data used to evaluate credit risk has been shown to degrade the effectiveness of many algorithms designed to classify that risk (Twala, 2013).

The effects of the payment pause on borrower behavior have been studied extensively. A consistent finding is that these payment pauses increase consumption and debt among the beneficiaries. Dinerstein et al. (2025) find that borrowers with frozen loans increased borrowing on mortgages, auto loans, and credit cards. Chava et al. (2023) also find significant increase in borrowing among those in forbearance, relative to borrowers who were not. Salman and Xie (2025) find that consumption increased in areas with higher levels of borrowers eligible for the payment pauses. Briones and Turner (2025) show that the debt payment suspension reduced hours worked as well. This chapter confirms many of the results from the previous literature using an alternative identification strategy.

The remainder of this chapter is structured as follows. Data and descriptive statistics are presented in section 3.2. I propose a matching-based empirical strategy in section 3.3, and discuss the results in section 3.4. Concluding thoughts and paths for future research are in section 3.5.

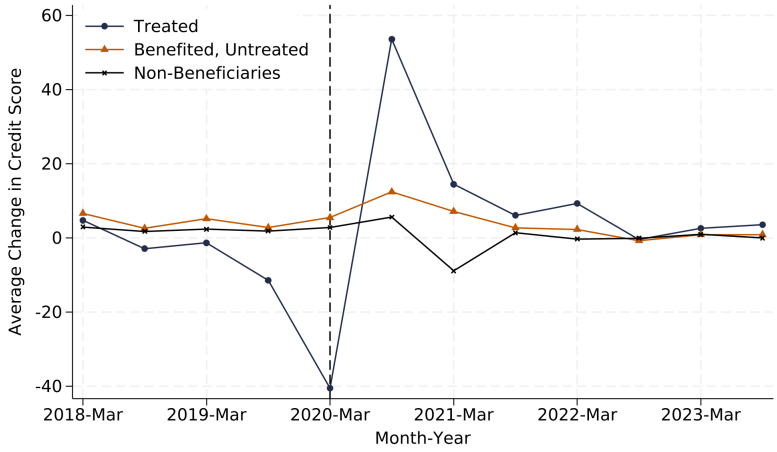
3.2 Credit Data

Data for this study are from the one of the three major credit bureaus. The data are separated into two types of datasets: attributes and trades. The attribute files contain information on the individuals in the sample, such as their credit score, marital status, and imputations of their income. The trade files contain information on individual loans and credit accounts opened by the people in the sample. This includes their scheduled and actual payments, loan status (current, past due, etc.), and account balance. An individual will appear once in an attribute file, but could have multiple entries in each trade file if they have multiple outstanding loans in a period. Observations are recorded in March and September of each year. I use the 2018–2023 files in this study.

To identify loans affected by the pause, I use the March and September 2020 trade files.

An education loan is flagged as being paused if they have a positive scheduled payment in March 2020 and a scheduled payment of zero in September 2020 without having been closed. In addition to the payment pause, the collection of delinquent loans was also suspended. To determine whether the trade benefited from this aspect of the policy, I compare the loan's delinquency status in March 2020 to its status in September 2020. Trades are flagged as benefiting from this provision if they are delinquent in March but not September, once again conditioning on not being closed.

Figure 3.1: Period-Over-Period Credit Score Change

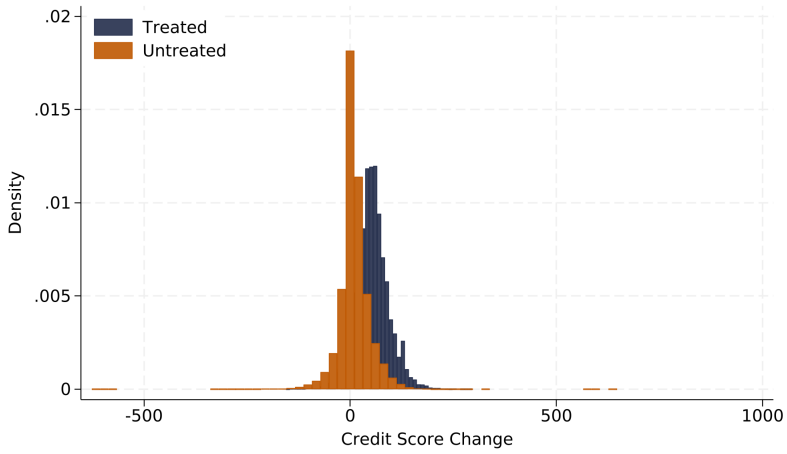


Note: These are the unconditional changes in credit score between each observation period in the study.

After determining the status of each loan, I aggregate them at the person level to set treated and untreated groups. A person is designated as treated if they have at least one student loan that had a delinquency resolved between March and September 2020 because of the pause. I choose to focus on the subset of beneficiaries who had a delinquency resolved, rather than all beneficiaries because this subgroup saw the largest credit score increases after the policy (figure 3.1). It can also be seen that credit scores in this group were falling leading into the policy change, suggesting that many of the delinquencies being resolved were relatively new. Examining the distribution of credit score changes between March and September 2020 provides further evidence that these borrowers saw disproportionately larger increases in their credit score than their peers (figure 3.2).

A comparison of the treated and untreated groups in March 2020 is in table 3.1. The two are noticeably different with the latter having much higher credit scores and balances on trades. The untreated are also significantly more likely to own homes and be married. The difference in average credit score in particular is unsurprising given that, by definition, the treated group has at least one delinquency on their credit report. The treated group does,

Figure 3.2: Credit Score Change Density

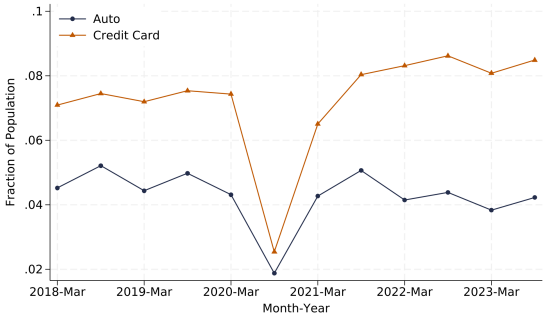


Note: Credit score change is the difference between September and March 2020 credit scores.

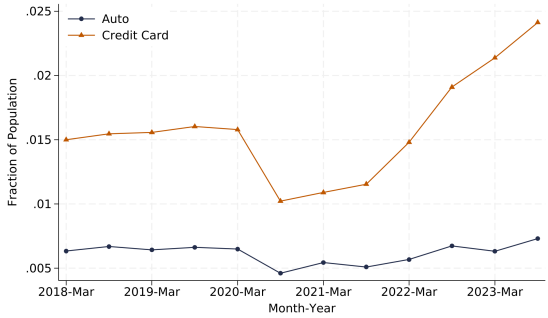
however, have much larger credit score changes on average between March and September 2020 (54 points versus 10).

Trade opening and delinquency rates are in figures 3.3a and 3.3b, respectively. Between four and six percent of the population opens an auto loan in a give period, while six to nine percent open a credit card. Less than one percent of people go delinquent on an auto loan, and between one and 2.5 percent of the population go delinquent on a credit card in a period. These delinquency rates are not conditioned on having an open loan.

Figure 3.3: Trade Opening and Delinquency Rates



(a) Trade Opening Rates



(b) Delinquency Rates

Table 3.1: Unmatched Panel Balance

	Treated	Untreated	Difference
Credit Score	504.21	677.18	-172.97***
	(66.70)	(111.25)	
Credit Score Change	53.58	9.65	43.93***
	(40.79)	(34.74)	
# Open Trades	5.99	5.09	0.90***
	(5.05)	(4.98)	
Balance On Open Trades	60,068.88	98,642.87	-38,573.99***
	(89,088.46)	(181,000.00)	
Balance on Education Trades	34,390.86	18,613.32	15,777.54***
	(47,802.24)	(38,905.38)	
# of Credit Cards	1.58	3.37	-1.78***
	(2.64)	(3.84)	
Balance on Credit Cards	2,153.30	5,252.23	-3,098.93***
	(5,720.28)	(10,652.82)	
Homeowner	0.31	0.51	-0.20***
	(0.46)	(0.50)	
Married	0.44	0.54	-0.10***
	(0.50)	(0.50)	
Observations	62,234	219,530	

Notes: This table contains the means of all the listed variables for the treated and untreated groups before matching. Each observation is of an individual in March 2020. Credit score change is the average change in credit score between March 2020 and September 2020.

3.3 Empirical Strategy

The primary focus of this chapter is to determine whether beneficiaries of the student loan payment and collection pause are more likely to open auto or credit card loans, and whether they are more likely to go delinquent on their loans. Both effects are driven by the increase in credit score caused by the pause. This change loosened credit constraints on borrowers and added noise to their risk signal, possibly making them appear to be less risky than they really are.

Before measuring these effects, I also examine the factors that determined the size of credit score changes between March and September 2020. Using just the treated group, I estimate

$$\Delta CS_i = \alpha + \beta_1 CS_i + \beta_2 Trades_i + \beta_3 Balance_i^{all} + \beta_4 Balance_i^{sl} + \beta_5 Late_i + \beta_6 N.Treated + \varepsilon_i \tag{3.1}$$

where CS_i is their credit score in March 2020; $Trades_i$ is the number of open trades they hold; $Balance_i^{all}$ is the total balance on all open trades; $Balance_i^{sl}$ is their student loan balance; $Late_i$ is their past due balance on their student loans; and $N.Treated_i$ is the number of student loan the hold that resolved a delinquency after the payment pause.

For the main analysis, I use propensity score matching to account for the large observable differences between the unmatched treated and untreated groups. Matching is conducted within groups determined by credit score bin and homeownership status. Each credit score bin is 20 points wide. Within these matching groups, I conduct one-to-one propensity score matching without replacement. Propensity scores are estimated using a logit model over a common support by removing outliers from the treated group.[2] The independent variables used to calculate propensity scores are credit score, balance on trades, number of open trades, and balance on student loans.

I create two separate control groups in this chapter. For the first, match groups are determined by credit card bin in September 2019 and March 2020, and homeownership status in March 2020. Within those groups, observations for the match variables listed above are limited to their March 2019 through March 2020 observations. This group is designed to be most like the treated group but for the collection and payment pause. After matching, they should have similar credit score patterns, including the observed drop, leading into March 2020, but will not have the benefit of the policy to boost their credit scores afterward. I refer to this as the pre-pause matches throughout the remainder of this chapter. If their credit constraint was binding, beneficiaries will likely open more loans than the control group to take advantage of their increased access to credit.

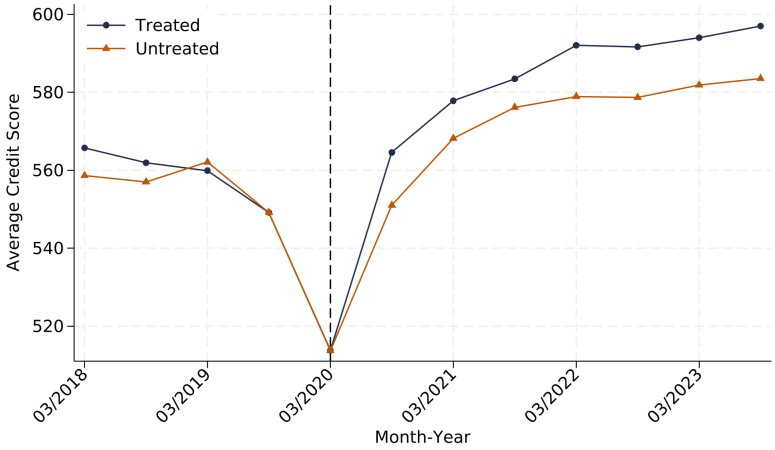
A comparison of the treatment and pre-pause matches in March 2020 can be found in table 3.2. Of the 62,234 treated individuals, 54,192 are matched.[3]. The average credit score of both groups and period-over-period change in credit score for each group are shown in figures 3.4a and 3.4b, respectively. The matching process was successful in making the two groups appear similar in credit score prior to the pause, though there are still noticeable differences in the average balance on open loans.

Creation of the second control group (the post-pause matches, hereafter) follows the same

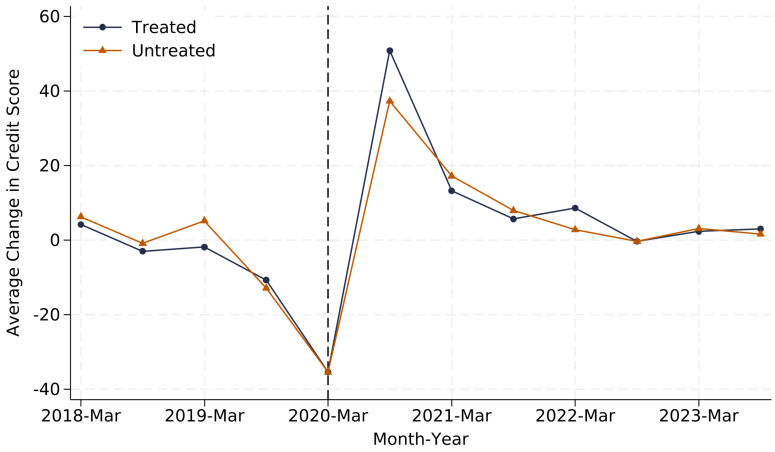
[2]An observation in the treated group is considered an outlier if its propensity score is outside the range of propensity scores seen in the control group.

[3]A discussion of which observations are matched and which ones are not is in section C.1 of the appendix.

Figure 3.4: Pre-Pause Match Credit Scores



(a) Mean Credit Score



(b) Period-Over-Period Credit Score Change

Table 3.2: Matched Panel Balance—Pre-Pause Matches

	Treated	Untreated	Difference
Credit Score	540.66	540.80	-0.13
	(51.91)	(51.92)	
Credit Score Change	41.99	27.38	14.61***
	(38.20)	(41.19)	
# Open Trades	5.98	4.40	1.58***
	(4.57)	(5.28)	
Balance On Open Trades	66,092.14	50,626.96	15,465.18***
	(89,506.38)	(100,000.00)	
Balance on Education Trades	33,914.39	28,608.45	5,305.93***
	(45,280.93)	(46,040.58)	
# of Credit Cards	2.01	2.54	-0.53***
	(2.74)	(3.99)	
Balance on Credit Cards	2,725.88	3,991.33	-1,265.45***
	(6,136.20)	(9,887.42)	
Homeowner	0.34	0.34	0.00
	(0.47)	(0.47)	
Married	0.47	0.47	0.00
	(0.50)	(0.50)	
Observations	33,394	33,394	

Notes: This table shows the mean of each variable for the treated and untreated groups after matching to create the pre-pause matches control group. As before, all values are for March 2020, except credit score change which is the change in credit score between March and September 2020.

procedure, but adds September 2020 to the list of observation periods used to create match groups and the list of variables used in propensity matching. This is the group of borrowers the beneficiaries appear most similar to after having their scores inflated. Based on credit score alone, both groups will be sending the same signal to lenders, but the treated group is potentially more risky because their "true" credit score, the score they would have if their delinquencies had not been temporarily resolved, is likely to be lower than their observed score. If this is true, the treated group should be more likely to go delinquent on a loan than the control group. Mean comparisons between the treated group and this second control group are in table 3.3, mean credit scores and credit score changes are in figures 3.5a and 3.5b. Once again the matching process resulted in close credit score matches with some differences in remaining balance on trades. There are fewer match pairs in this second group because September 2020 credit score bin was also used to create the match groups. Adding this variable created more groups to match within, spreading the data out thinner. As a

Table 3.3: Matched Panel Balance—Post-Pause Matches

	Treated	Untreated	Difference
Credit Score	543.07	543.51	-0.43
	(47.18)	(47.01)	
Credit Score Change	36.17	35.54	0.62**
	(25.11)	(25.07)	
# Open Trades	5.56	4.21	1.35***
	(4.04)	(5.14)	
Balance On Open Trades	54,244.46	43,144.29	11,100.16***
	(71,649.37)	(83,926.79)	
Balance on Education Trades	30,298.13	27,578.87	2,719.26***
	(39,154.15)	(43,229.38)	
# of Credit Cards	1.81	2.31	-0.50***
	(2.49)	(3.74)	
Balance on Credit Cards	2,235.53	3,358.75	-1,123.22***
	(5,090.23)	(8,737.68)	
Homeowner	0.26	0.26	0.00
	(0.44)	(0.44)	
Married	0.46	0.46	0.00
	(0.50)	(0.50)	
Observations	18,109	18,109	

Notes: This table shows the mean of each variable after matching to create the post-pause matches. As before, the observations are for March 2020 except credit score change which is the change in credit score between March and September 2020.

result, fewer matches were possible.

After matching, I estimated:

$$y_{it} = \alpha + \phi Treated_i + \delta D_{it} + \mu_{m(i)} + \tau_t + \epsilon_{it} \tag{3.2}$$

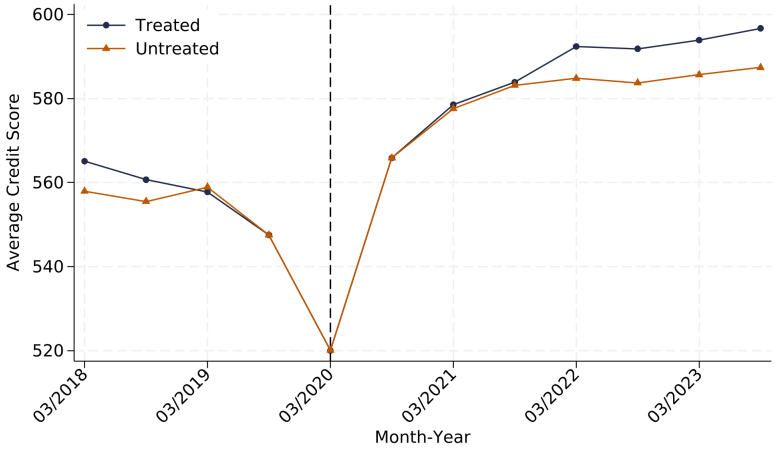
where $Treated_i$ equals one for treated individuals; D_{it} is a treatment indicator equal to one if person i is treated and the observation occurs after March 2020; $\mu_{m(i)}$ and τ_t are match pair and time period fixed effects, respectively.

To test for differential trends prior to the payment and collections pause, I estimate

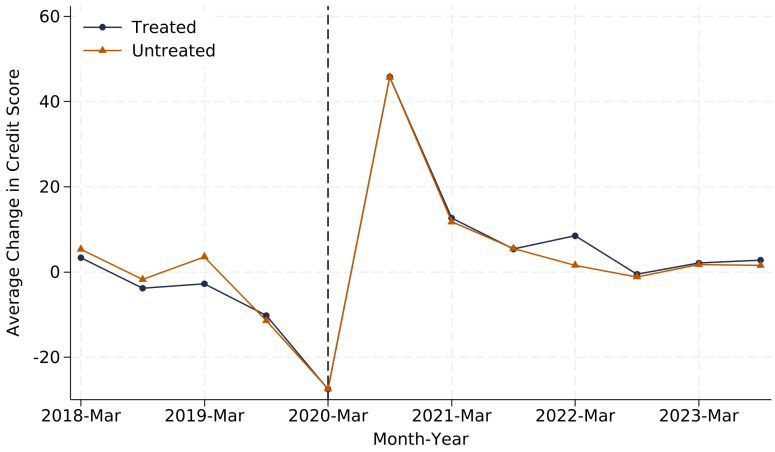
$$y_{it} = \alpha + \phi Treated_i + \sum_{\tau \neq 3/2020} \delta_\tau D_{it} + \mu_{m(i)} + \tau_t + \epsilon_{it} \tag{3.3}$$

where all parameters are defined as before.

Figure 3.5: Post-Pause Match Credit Score



(a) Mean Credit Score



(b) Period-Over-Period Credit Score Change

This empirical strategy assumes that lenders do not change their behavior after credit scores are inflated. It is possible that lenders, knowing that some credit scores are artificially high and being unable to identify which ones, treat all scores as inflated. As a result, those in either control group will have fewer opportunities to borrow, or may be offered worse terms on the loans they can get. If this is the case, the treatment effects I estimate will include both the benefit to the treatment group and the harm done to the control group. The estimated effects on auto loan and credit card openings should therefore be considered an upper bound of the true impact of the payment pause.

For delinquencies, this response by lenders could bias my estimates in either direction. First, a person cannot go delinquent on a loan they were not able to take out. The effect on unconditional delinquencies will also be an upper bound as it does not account for the decreased likelihood that the control group is able to open a new loan. The second pathway for bias is through the terms offered on new loans. If lenders offer worse terms, such as higher interest rates, because they know some credit scores have been inflated, it may be more difficult for borrowers to keep up with their payments, increasing the probability of delinquency. However, if lenders are truly unable to distinguish between the treated and control, as I am assuming, this will affect both groups as lenders offer similar terms to each.

3.4 Results

Results of estimating equation 3.1, which measures the factors that explain the size of individual credit score changes, are in table 3.4. Recall that this is only estimated using the treated group observations for March 2020 and the outcome variable is the change in credit score between March and September of that year. Most of the results are intuitive. Credit score changes are smaller the higher a borrower's initial credit score, the more open trades they have, and the further behind their payments they are. The change is larger for those with more debt. Interestingly, the number of treated trades—i.e., the number of loans where the delinquency on the loan was resolved because of the collections pause—has no effect.

Moving to the effect of the payment and collection pause on borrowing, I find that the treated are significantly more likely to open both auto and credit card trades—roughly 1.07 and 4.55 percentage points in the post-pause period, respectively, than the pre-pause matches (table 3.5). While these are small numbers in absolute terms, between four and six percent of people have historically opened auto loans and six to nine percent opened credit cards in a given period (figure 3.3a). Thus, this represents a 16 to 25 percent increase in auto loan opening rates, and a 44 to 66 percent increase in credit card opening rates. These results lend support to the hypothesis that beneficiaries take advantage of loosened borrowing constraints

Table 3.4: Credit Score Change Factor Variables

	Dependent Variable: Credit Score Change		
	(1)	(2)	(3)
Intercept	167.2***	172.5***	172.4***
	(1.1622)	(1.2117)	(1.3237)
3/2020 Credit Score	-0.225***	-0.238***	-0.238***
	(0.0023)	(0.0025)	(0.0027)
Number of Open Trades		-0.183***	-0.170***
		(0.0341)	(0.0493)
Balance on Trades		0.0000377***	0.0000290***
		(0.0000)	(0.0000)
Student Loan Balance			0.0000321***
			(0.0000)
Past Due on SL			-0.0000987***
			(0.0000)
Number of Treated Loans			-0.111
			(0.0680)
N	61906	61906	61906
R^2	0.136	0.141	0.142

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: Results from estimating equation 3.1 are in this table. The sample is limited to the treated group. There are fewer observations than in table B.1 because some members of the treated group are missing credit score observations in March and/or September 2020 and could not have their credit score change calculated.

Table 3.5: Trade Openings and Delinquencies—Pre-Pause Matches

	Openings		Delinquencies	
	Auto	Credit Card	Auto	Credit Card
	(1)	(2)	(3)	(4)
Constant	0.0636***	0.163***	0.0364***	0.0823***
	(0.0002)	(0.0004)	(0.0002)	(0.0003)
Treated	-0.0150***	-0.0782***	-0.0112***	-0.0119***
	(0.0007)	(0.0011)	(0.0005)	(0.0008)
Treated x Post	0.0107***	0.0455***	0.00222***	-0.0187***
	(0.0008)	(0.0013)	(0.0006)	(0.0010)
Match FE	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
N	1,289,496	1,289,496	1,289,496	1,289,496

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2 using the pre-pause matches control group. Standard errors are clustered at the match-pair level.

and open significantly more loans than they would have without the policy.

Beneficiaries are also more likely to experience an auto delinquency in the post-period, but less likely to experience a credit card delinquency than the pre-pause matches. While again these effects are small in absolute terms, relative to the baseline mean they are large increases. These results are not conditioned on having a loan of either type. I find similar results, though larger in magnitude, results when conditioning on having an open loan (appendix table B.7).

I find similar results using the post-pause matches. Relative to this group, beneficiaries are 0.78 and 3.5 percentage points more likely to open an auto loan and credit card, respectively (table 3.6). They are also once again more likely to go delinquent on an auto loan (0.43 percentage points) and less likely to go delinquent on a credit card (-1.08 percentage points).

While significant in their own right, comparing the effect sizes of both groups is also informative. The pre-pause matches have the same credit score positioning beneficiaries would have been in were it not for the payment pause. Their credit scores fell significantly, and while they did rebound there is still a gap between them and the treated group (figure 3.4a). The treated individuals saw their credit constraints loosened more relative to this group than relative to the post-pause matches. It is therefore unsurprising that the effect

Table 3.6: Trade Openings and Delinquencies—Post-Pause Matches

	Openings		Delinquencies	
	Auto	Credit Card	Auto	Credit Card
	(1)	(2)	(3)	(4)
Constant	0.0636***	0.163***	0.0349***	0.0798***
	(0.0003)	(0.0005)	(0.0002)	(0.0003)
Treated	-0.0115***	-0.0701***	-0.0109***	-0.0131***
	(0.0008)	(0.0012)	(0.0006)	(0.0009)
Treated x Post	0.00718***	0.0348***	0.00431***	-0.0108***
	(0.0009)	(0.0015)	(0.0007)	(0.0011)
Match FE	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
N	960,798	960,798	960,798	960,798

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2 using the post-pause matches control group. Standard errors are clustered at the match-pair level.

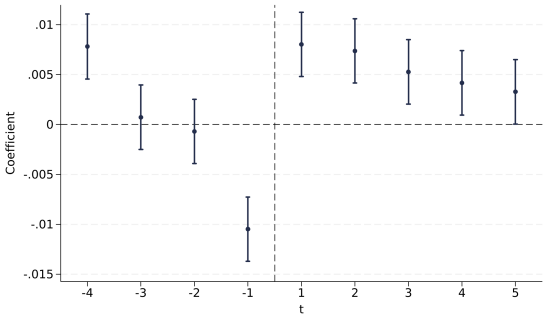
on trade openings is larger relative to the pre-pause matches than it is for the post-pause matches.

For delinquencies, the comparison between the treated and post-pause matches is most interesting. Credit score signal noise and bias introduced by the payment pause causes beneficiaries to look like this group to lenders. However, if their underlying risk is unchanged and simply masked by the higher credit scores, this group should be more likely to go delinquent than their new peers. This effect should be less prominent when using the pre-pause matches as the comparison because the treatment group is more similar to them in terms of risk profile. The results indicate this to be true, with the estimated effect on auto delinquencies estimated using the post-pause matches being close to double that of the effect estimated using the pre-pause matches. While the treated group are less likely to have a credit card delinquency than both groups, this difference is slightly smaller when compared to the post-pause matches.

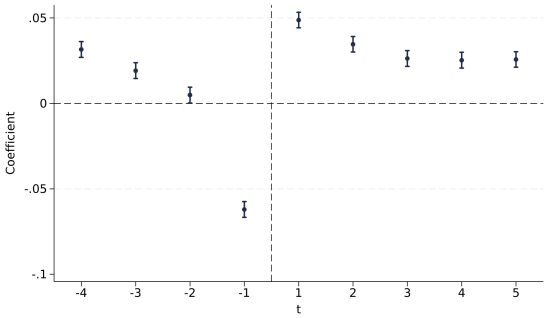
Implicit in these results in the assumption that those in the treated group only resolved their delinquencies because of the payment pause. This is not strictly true, as some likely would have caught up on their payments even without the pause. If this were accounted for, there would be less bias in the post-pause credit scores, and the estimated effects would be smaller.

Event study plots for each outcome are in figures 3.6 and 3.7. There are six months between each period. They consistently show that there are pre-payment pause differences between the treated and control group that were not resolved in the matching process.

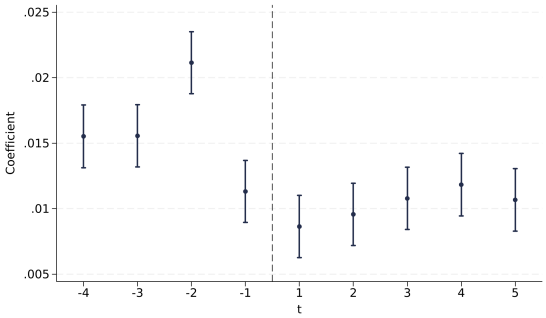
Figure 3.6: Event Studies—Pre-Pause matches



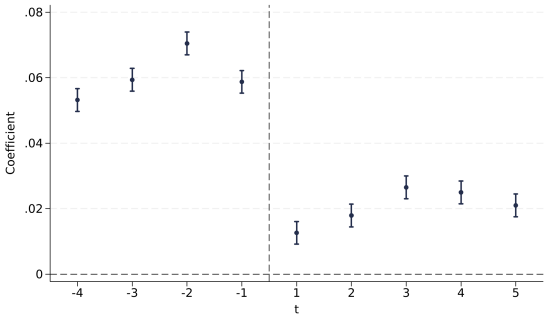
(a) Auto Trade Openings



(b) Credit Card Openings

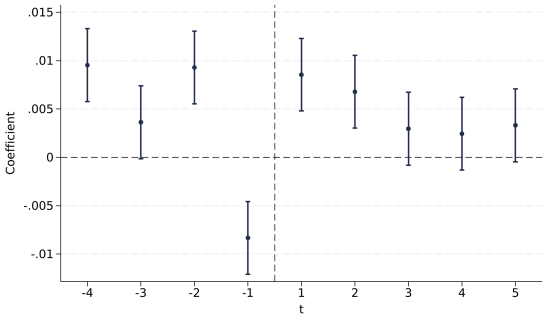


(c) Auto Trade Delinquencies

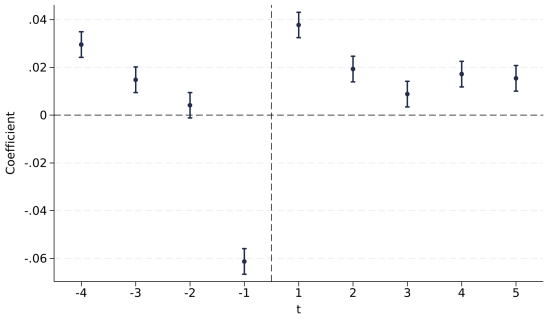


(d) Credit Card Delinquencies

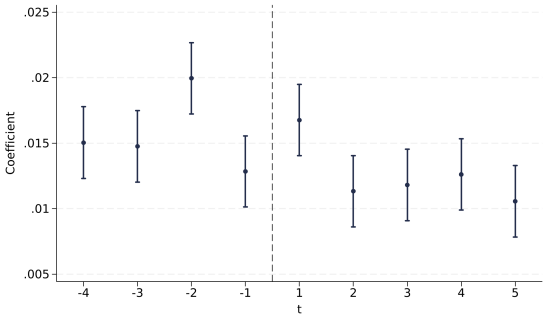
Figure 3.7: Event Studies—Post-Pause Matches



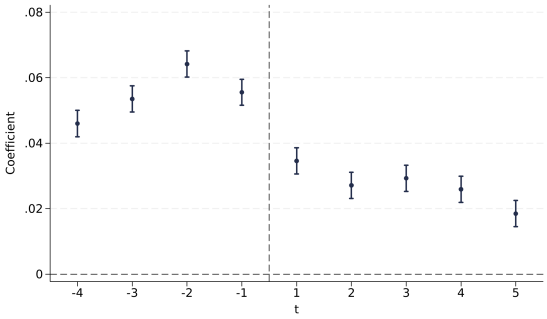
(a) Auto Trade Openings



(b) Credit Card Openings



(c) Auto Trade Delinquencies



(d) Credit Card Delinquencies

3.5 Conclusion and Future Research

Given the importance of credit scores in determining access and terms in the consumer credit market, noise that disrupts the signal scores send can have significant consequences. In this chapter I study how noise and bias introduced by the student loan payment pauses during the COVID-19 pandemic impacted borrowing and delinquency rates among beneficiaries, with a specific focus on those who were previously delinquent on a trade.

I show that these beneficiaries are more likely to open both auto loans and new credit cards than untreated individuals matched with the treated on their pre-and post-policy profiles. They are also more likely to go delinquent on auto loans than these groups, but less likely to go delinquent on credit cards. Beyond the headline results, I find that the trade openings effect is higher against the control group based on pre-policy values than the control group based on post-policy values, and the delinquency effect is larger for the post-policy value based control group than the pre-policy group.

There are a number of paths forward to build upon this research. Alternative empirical strategies, such as synthetic difference-in-differences, could be used to estimate each effect. The types of debt analyzed could be expanded to include mortgages. Delinquencies also play out overtime, meaning there could be a more dramatic effect in later years. A follow-up study conducted once more data are available would capture these long-term effects. The student loan payment pause ended in fall of 2023, after the last period of data available for this study. Payment and collection resumption would reimpose the credit constraint loosened by the payment pause, causing a second change in behavior that can be studied as those effects work their way through the economy.

A key factor in the market not discussed in this chapter is the terms lenders offer to borrowers. Higher credit scores typically result in better terms (lower interest rates, higher credit limits, etc.). However, if lenders believe there to be noise which makes it more difficult to identify underlying risk types, they may offer less favorable terms to compensate for the unaccounted for risk. This would have implications on the overall welfare impact of the policy. It would be valuable to take a general equilibrium approach to the problem in order to understand the full effects of policies that impact the consumer credit market in this way.

References

Acquisti, A., Friedman, A., and Telang, R. (2006). Is there a cost to privacy breaches? an event study. *Proceedings of the International Conference of Information Systems (ICIS)*.

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

Akyildirim, E., Conlon, T., Corbet, S., and Hou, Y. G. (2024). Hacked: Understanding the stock market response to cyberattacks. *Journal of International Financial Markets, Institutions and Money*, 97:102082.

Antman, F. M., Qu, S., Logan, T. D., and Weinberg, B. A. (2025). The long-run impacts of mentoring underrepresented minority groups in economics. (33689). DOI: 10.3386/w33689.

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.

Arya, S., Eckel, C., and Wichman, C. (2013). Anatomy of the credit score. *Journal of Economic Behavior & Organization*, 95:175–185.

Asthana, S. and Balsam, S. (2001). The effect of edgar on the market reaction to 10-k filings. *Journal of Accounting and Public Policy*, 20(4):349–372.

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 victim 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.

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13).

Blankespoor, E., deHaan, E., and Marinovic, I. (2020). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2):101344.

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

Briones, D. A. and Turner, S. (2025). Labor, loans and leisure: The impact of the student loan payment pause. (33553). DOI: 10.3386/w33553.

Chatterjee, S., Corbae, D., Dempsey, K., and Ríos-Rull, J.-V. (2023). A quantitative theory of the credit score. *Econometrica*, 91(5):1803–1840.

Chava, S., Tookes, H., and Zhang, Y. (2023). Leaving them hanging: Student loan forbearance, distressed borrowers, and their lenders. (4451747).

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.

Dinerstein, M., Earnest, S., Koustas, D. K., and Yannelis, C. (2025). Student loan forgiveness. (33462). DOI: 10.3386/w33462.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Feroz, E. H., Park, K., and Pastena, V. S. (1991). The financial and market effects of the sec's accounting and auditing enforcement releases. *Journal of Accounting Research*, 29:107–142.

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.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with numpy. *Nature*, 585(7825):357–362.

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.

Jaffee, D. M. and Russell, T. (1976). Imperfect information, uncertainty, and credit rationing*. *The Quarterly Journal of Economics*, 90(4):651–666.

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*.

Kamiya, S., Kang, J.-K., Kim, J., Milidonis, A., and Stulz, R. M. (2021). Risk management, firm reputation, and the impact of successful cyberattacks on target firms. *Journal of Financial Economics*, 139(3):719–749.

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

Kolari, J. W. and Pynnönen, S. (2010). Event study testing with cross-sectional correlation of abnormal returns. *The Review of Financial Studies*, 23(11):3996–4025.

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

Leuven, E. and Sianesi, B. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Statistical Software Components, Boston College Department of Economics.

Li, E. X. and Ramesh, K. (2009). Market reaction surrounding the filing of periodic sec reports. *The Accounting Review*, 84(4):1171–1208.

Loughran, T. and Mcdonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65.

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

Makridis, C. and Dean, B. (2018). Measuring the economic effects of data breaches on firm outcomes: Challenges and opportunities. *Journal of Economic and Social Measurement*, 43(1–2):59–83.

Makridis, C. A. (2021). Do data breaches damage reputation? evidence from 45 companies between 2002 and 2018. *Journal of Cybersecurity*, 7(1):tyab021.

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.

Narajabad, B. N. (2012). Information technology and the rise of household bankruptcy. *Review of Economic Dynamics*, 15(4):526–550.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Salman, D. and Xie, X. (2025). Relief beliefs: Effects of anticipated student loan forgiveness. (5157757).

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.

Smith, K. T., Jones, A., Johnson, L., and Smith, L. M. (2018). Examination of cybercrime and its effects on corporate stock value. *Journal of Information, Communication and Ethics in Society*, 17(1):42–60.

Spanos, G. and Angelis, L. (2016). The impact of information security events to the stock market: A systematic literature review. *Computers & Security*, 58:216–229.

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

Stice, E. K. (1991). The market reaction to 10-k and 10-q filings and to subsequent the wall street journal earnings announcements. *The Accounting Review*, 66(1):42–55.

Stiglitz, J. E. and Weiss, A. (1981). Credit rationing in markets with imperfect information. *The American Economic Review*, 71(3):393–410.

The Pandas Development Team (2024). pandas-dev/pandas: Pandas.

Tosun, O. K. (2021). Cyber-attacks and stock market activity. *International Review of Financial Analysis*, 76:101795.

Twala, B. (2013). Impact of noise on credit risk prediction: Does data quality really matter? *Intelligent Data Analysis*, 17(6):1115–1134.

Ushey, K., Allaire, J., and Tang, Y. (2024). *reticulate: Interface to 'Python'*. R package version 1.37.0.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

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.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

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

Appendices

Appendix A.

Appendix to Chapter 1

A.1 Model Derivations

A.1.1 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:

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

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

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

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

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

Equation A.2 can be rearranged to obtain the optimal S:

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

Setting the left-hand sides of equations A.3 and A.4 equal and solving for d_2 in terms of d_1 yields:

$$d_2 = \left[\frac{\alpha_2}{\omega_2 + \frac{r}{S+1}\gamma_2} \frac{\omega_1 + \frac{r}{S+1}\gamma_1}{\alpha_1}\right]^{\frac{1}{1-\rho}} d_1. \tag{A.6}$$

Which can be substituted into equation A.3:

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

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:

$$d_1^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1}\right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2}\right)^{\frac{\rho}{1-\rho}}\right]^{\frac{\nu-\rho}{\rho(1-\nu)}}. \tag{A.7}$$

which gives the optimal d_2 when inserted into A.6:

$$d_2^* = (\nu A)^{\frac{1}{1-\nu}} \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{1}{1-\rho}} \left[\alpha_1 \left(\frac{\alpha_1}{\omega_1 + \frac{r}{S^*+1}\gamma_1} \right)^{\frac{\rho}{1-\rho}} + \alpha_2 \left(\frac{\alpha_2}{\omega_2 + \frac{r}{S^*+1}\gamma_2} \right)^{\frac{\rho}{1-\rho}} \right]^{\frac{\nu-\rho}{\rho(1-\nu)}}. \tag{A.8}$$

While not a closed form solution, equations A.5, A.7, and A.8 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.1.2 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:

$$D(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3}{2} \\ 0 & \text{Otherwise} \end{cases} \tag{A.9}$$

And supply is

$$\begin{aligned} S(p) &= \mathcal{B}\mathbb{P}\left(V \leq p\right) \\ &= \mathcal{B}p^2. \end{aligned} \tag{A.10}$$

Setting equations A.9 and A.10 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

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

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

$$\begin{aligned} S^{Post}(p) &= \mathcal{B}^{Post}\mathbb{P}\left((1-\phi)V_i < p\right) \\ &= \mathcal{B}\left(\frac{p}{1-\phi}\right)^{2-\sigma} \end{aligned} \tag{A.11}$$

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

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

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

$$D^{Post}(p) = \begin{cases} \frac{Y}{p} & \text{if } \kappa \geq \frac{3-\sigma}{2-\sigma} \\ 0 & \text{Otherwise} \end{cases} \tag{A.12}$$

Setting equations A.11 and A.12 equal yields the post-GDPR equilibrium:

$$\begin{aligned} p^*_{Post} &= \left(\frac{Y}{\mathcal{B}^{Post}}\right)^{\frac{1}{3-\sigma}} (1-\phi)^{\frac{2-\sigma}{3-\sigma}} \\ Q^*_{Post} &= Y^{\frac{2-\sigma}{3-\sigma}} \left(\frac{\mathcal{B}^{Post}}{(1-\phi)^{2-\sigma}}\right)^{\frac{1}{3-\sigma}}. \end{aligned} \tag{A.13}$$

A.2 Data

A.2.1 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.

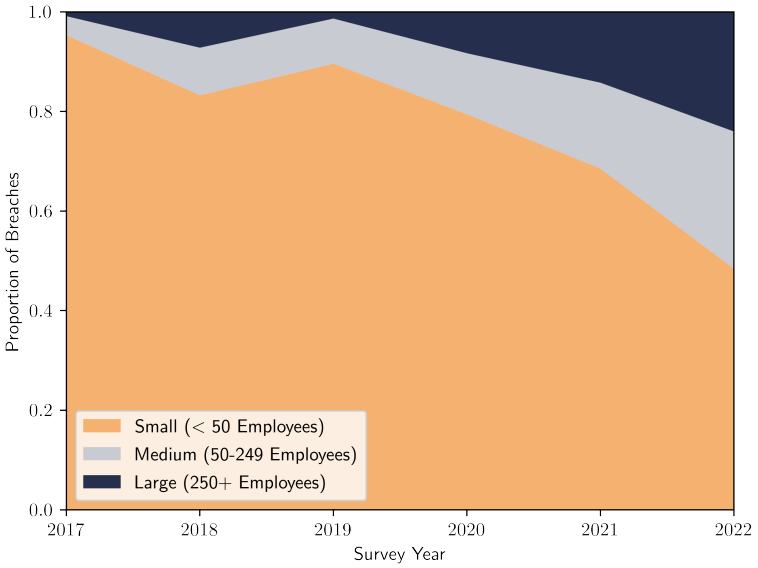
Only the 2018 and 2019 survey asked respondents whether they made any changes in response to the GDPR. The survey asked about the types of changes made as well, which I have combined into five groups: human changes (e.g., staff training and hiring), technical changes (e.g., updated system configurations 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 A.1: 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 A.1, 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.

Figure A.1: UK Data Breaches



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

A.2.2 Breach Data

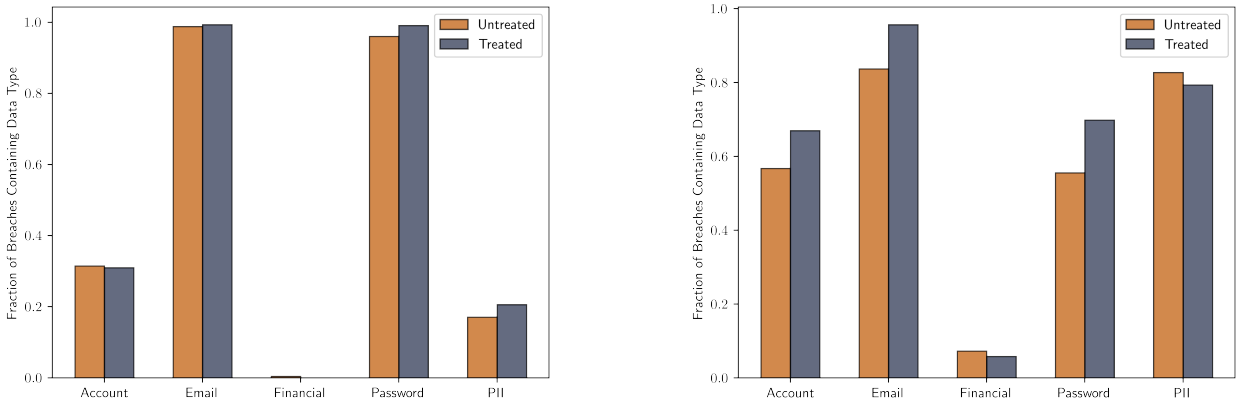
The individual breach data obtained for this study contains many more data package observations than are included in the final paper. Observations were dropped for one of three reasons. First, any breaches that could not be attributed to an organization or country were removed. Second, any data package 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 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.

A.2.3 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 1.6 and A.2), 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 A.25).

Figure A.2: Fraction of Data Packages Containing Each Data Type, Pre-and Post-GDPR



(a) Pre-GDPR

(b) Post-GDPR

A.2.4 Descriptive Information

Tables A.2-A.5 report unconditional differences in means between various data package groups.

Table A.2 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.

Table A.2: 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*

Data packages that became available before and after the GDPR are then compared in table A.3. 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 A.4 and A.5, 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 A.3: 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 A.4: 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 A.5: 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*

A.3 Results

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

A.3.1 Extensive Margin Effects

On the extensive margin, I separately estimate equation 1.4.1 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 A.6 and A.7.

A.3.2 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

Table A.6: 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 A.7: 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

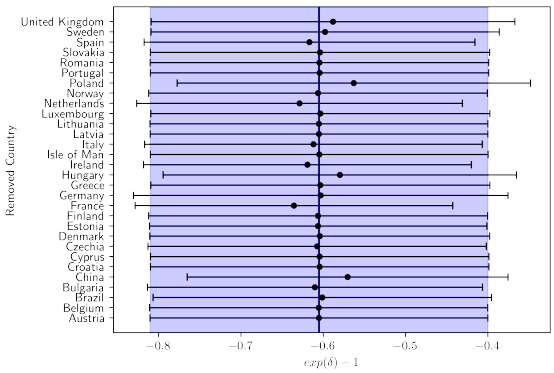
**p<0.1, **p<0.05, ***p<0.01*

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

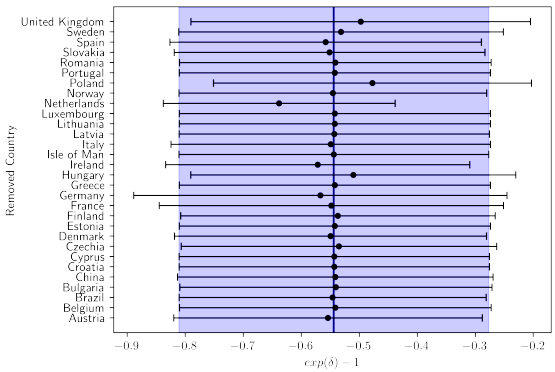
stays well within the 95 percent confidence interval of the estimate with the full sample (figures A.3 and A.4).

Next, I use different methods to construct the panel. Brazil and China each adopted

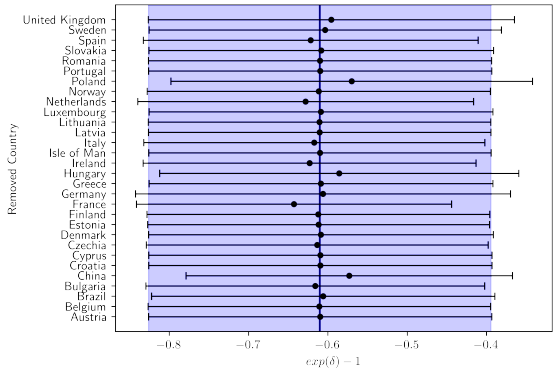
Figure A.3: Number of Data Breaches Effects Removing Countries



(a) Overall Effect



(b) Short Run Effect



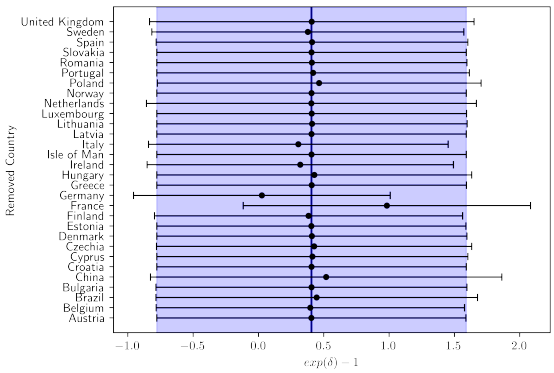
(c) Long Run Effect

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.

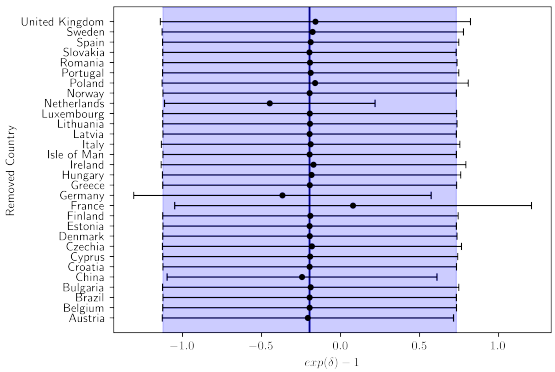
their own data privacy laws near the end of the study period. Removing them from the sample slightly reduces the estimated effect on the number of breaches, though it is still significant. As before, there are no statistically significant effects on the number of records available (table A.8).

Table A.9 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

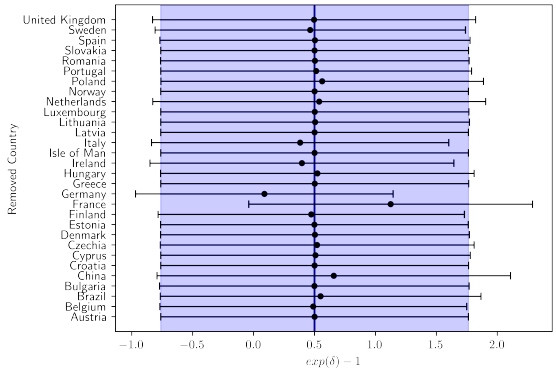
Figure A.4: Number of Records Effects Removing Countries



(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.

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 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 A.10 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

Table A.8: 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

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 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 A.11-A.16 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:

$$Y_{it} = \gamma_s + \tau_t + \delta D_i \times \textit{Post-GDPR}_t + \epsilon_{it} \tag{A.14}$$

Table A.9: 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.

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 1.9). This is likely the source of the discrepancies in effect sizes between the models and the primary reason for using the Poisson model over the linear models with a log-like transformation.

Table A.10: 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.

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 A.13 and A.14, 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 A.15 and A.16, 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 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 A.11: 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 A.12: 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 A.13: 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 A.14: 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 A.15: 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 A.16: 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 A.17, I remove the population offset. Table A.18 weights the estimates using population rather than including an offset. In table A.19, 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 A.20 and A.21, and include indicators for whether the observation is a small or large country in table A.22.[1]

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 A.23, this does not have a significant effect on my effect estimates.

Table A.17: 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

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

Table A.18: 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 A.24, 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.

A.3.3 Data Package Effects

Estimates of the change in PII fraction using a slightly different definition of PII are in table A.25. 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 A.27-A.29 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 A.27 to table

Table A.19: 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.

1.11), 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 A.30).

Table A.20: 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 A.21: 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 A.22: 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 A.23: 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 A.24: 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 A.25: 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 A.26: 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 A.27: 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 A.28: 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 A.29: 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 A.30: 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.

Appendix B.

Appendix to Chapter 2

B.1 Data

In total, I observe 12,210 filings by treated firms. There are 5,112 fewer used in both the event study and disclosure effect estimates because some filings were unable to be included in the analysis. A filing was removed for one of two reasons:

1. The filing could not be read or processed. This was typically due to an error in the encoding.

2. There were missing stock data during the event study. This could either be during the period where the risk model was estimated or during the event window itself. Some filings also happened after the last CRSP update on WRDS, meaning I could not observe any stock data around the filing.

Table B.1 lists the words used to determine whether a filing discussed cybersecurity risk.

B.2 CAR Estimates

To test the robustness of my CAR^{event} estimates, I reran the analysis using different market models and measures of market return.

Table B.2 shows the results using alternative models for estimating expected returns. Column one sets the expected returns to a constant: the mean of the returns in the estimation window. Column two uses the CAPM model, estimated as

$$R_{it} = R_{ft} + \beta_i \left(R_{mt} - R_{ft}\right) + \varepsilon_{it}$$

where R_{ft} and R_{mt} are the risk-free and market returns on day t, respectively.

Table B.1: Cybersecurity Risk Keywords

Breach of our networks or systems	Data security
Compromised data	Data-breach
Cyber attacks	Ddos
Cyber incident	Denial of service
Cyber incidents	Denial-of-service
Cyber security	Disclosure of our data
Cyber security incident	Distributed denial-of-service
Cyber security incidents	Hack
Cyber-attack	Hacker
Cyber-attacks	Hackers
Cyber-security	Hacking
Cyber-security incident	Hacks
Cyber-security incidents	Information security systems
Cyberattack	Large amounts of data
Cyberattacks	Malware
Cybersecurity	Phishing
Cybersecurity incident	Ransomware
Cybersecurity incidents	Social engineering
Data breach	Social-engineering
Data privacy	Unauthorized access to our data
Data protection	

In column three, the market model is the Fama-French three-factor model plus momentum:

$$R_{it} = R_{ft} + \alpha + \beta_{i1}(R_{mt} - R_{ft}) + \beta_{i2}SMB_t + \beta_{i3}HML_t + \beta_{i4}UMD_t + \varepsilon_{it}.$$

where R_{ft}, R_{mt}, SMB_t, and HML_t are defined in the main body of the paper. UMD_t is the momentum factor.

With each model, $\overline{CAR}$ is similar to $\overline{CAR}$ measured using the three-factor model in the main body of the paper.

For goodness of fit, table B.3 contains summary statistics for the R^2 of each model. The R^2 distributions are shown in figure B.1.

Next, I estimated cumulative abnormal return using the S&P Composite market, the value-weighted, and equal-weighted returns. These results are in table B.4. For consistency, I subtracted the daily risk-free rate from the total return. The average abnormal and cumulative abnormal returns under these return measurements are almost identical to the results using the risk-free market return from the Fama-French WRDS database.

Table B.2: CAR with Alternative Market Models

A_t	(1)	(2)	(3)	(4)
-1	0.285	0.011	-0.159	-0.126
	(0.195)	(0.194)	(0.157)	(0.153)
0	-0.618**	-0.578**	-0.499**	-0.507**
	(0.265)	(0.242)	(0.236)	(0.231)
1	-0.477**	-0.389*	-0.248	-0.247
	(0.327)	(0.293)	(0.304)	(0.297)
CAR	**-0.811****	**-0.956*****	**-0.906*****	**-0.880*****
	(0.473)	**(0.421)**	**(0.385)**	**(0.391)**
Observations	166	166	166	166
Model	Constant	CAPM	3F Momentum	Three Factor

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

Notes: Column four, the three-factor model, is the market model used in the main text.

Table B.3: Market Model R^2 Summary Statistics

	CAPM	Three-Factor + Momentum	Three-Factor
Observations	166	166	166
Mean	0.285	0.363	0.352
Std. Dev.	0.181	0.196	0.199
Min.	0.000	0.013	0.007
25%	0.150	0.219	0.207
50%	0.247	0.354	0.339
75%	0.425	0.512	0.506
Max.	0.703	0.858	0.847

B.3 Disclosure Effects

As a robustness check, I estimate the impact of cyber risk disclosure on CAR^{filing} for 10-K and 10-Q filings separately. I use the same model as described in equation 2.15, but remove the indicator for whether the filing is a 10-Q. Results are in tables B.5 and B.6. The key change is that the coefficient on mentioning cyber risk becomes insignificant for 10-K filings. However, it is still significant and larger for 10-Q filings.

Removing the year and industry fixed effects makes the effect of CAR^{filing} slightly smaller than in the main results, but it is still the only significant predictor of the market's response to cybersecurity incidents (table B.7).

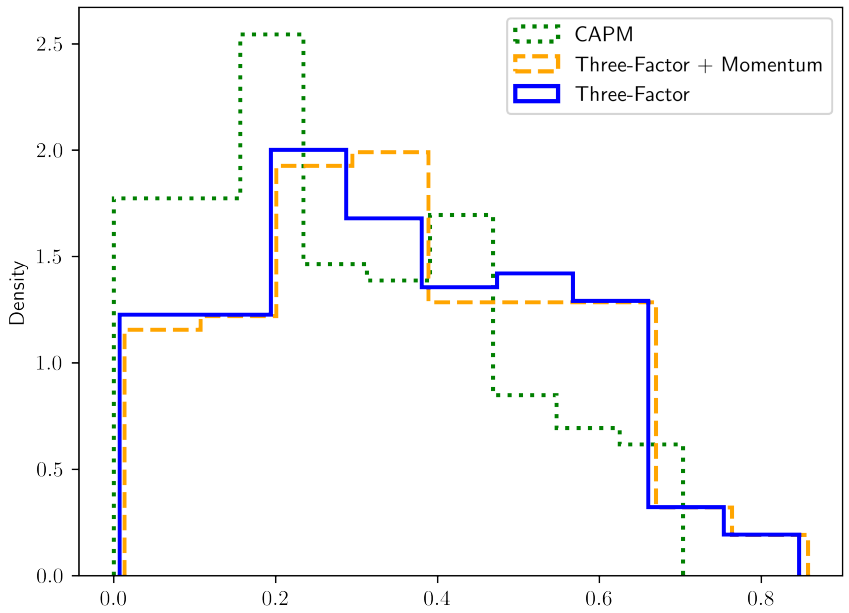


Figure B.1: R^2 Density

Table B.4: Alternative Market Return Variables

	Value-Weighted	Equal-Weighted	S&P Composite	FF-Market
t	(1)	(2)	(3)	(4)
-1	-0.135	-0.126	-0.135	-0.126
	(0.154)	(0.159)	(0.153)	(0.153)
0	-0.503**	-0.495**	-0.504**	-0.507**
	(0.231)	(0.228)	(0.232)	(0.231)
1	-0.241	-0.242	-0.232	-0.247
	(0.297)	(0.294)	(0.298)	(0.297)
CAR	**-0.879*** **	**-0.863*** **	**-0.871*** **	**-0.880*** **
	(0.390)	**(0.381)**	**(0.393)**	**(0.391)**
Observations	166	166	166	166
Model	Three Factor	Three Factor	Three Factor	Three Factor

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

Notes: Column four is the same specification in the main text.

Table B.5: Disclosure Effect on the Market Response to Filings: 10-K Only

	Dependent Variable: CAR(-1, 1)	
	(1)	(2)
Intercept	-2.983	54.775
	(17.727)	(76.068)
First Cyber Risk Mention	0.339	0.337
	(0.704)	(0.702)
Mentions Cyber Risk	0.158	0.187
	(0.646)	(0.644)
Age		-2.038
		(2.686)
Sentiment		-0.430
		(0.334)
Year Fixed Effects	Yes	Yes
Firm Fixed Effects	Yes	Yes
Observations	1,911	1,911
R^2	0.076	0.077
F Statistic	0.987	1.000
Model	OLS	OLS

$^*p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$

Notes: These are the results of estimating equation 2.15, limiting the sample to only 10-K filings.

Table B.6: Disclosure Effect on the Market Response to Filings: 10-Q Only

	Dependent Variable: CAR(-1, 1)	
	(1)	(2)
Intercept	-0.293	15.292
	(35.007)	(28.223)
First Cyber Risk Mention	-1.433	-1.397
	(1.865)	(1.873)
Mentions Cyber Risk	1.230**	1.208**
	(0.542)	(0.560)
Age		-0.462
		(0.525)
Sentiment		0.043
		(0.193)
Year Fixed Effects	Yes	Yes
Observations	5,187	5,187
R^2	0.035	0.035
F Statistic	1.176*	1.168*
Model	OLS	OLS

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

Notes: These are the results of estimating equation 2.15, limiting the sample to only 10-Q filings.

Table B.7: Disclosure Effect on the Market Response to Incidents, No Fixed Effects

	Dependent Variable: CAR(-1, 1)			
	(1)	(2)	(3)	(4)
Intercept	0.172	0.230	-0.379	1.634
	(0.681)	(2.186)	(0.340)	(1.781)
Disclosed Risk	-1.293	-0.998		
	(0.819)	(0.866)		
CAR^{filing}			0.270**	0.257**
			(0.108)	(0.115)
Not First Event		-0.899		-1.984
		(1.237)		(1.300)
Ransomware		0.205		-0.216
		(0.892)		(0.783)
Log(Market Value)		-0.358		-0.159
		(0.399)		(0.416)
Tobin's Q		0.005		-0.009
		(0.314)		(0.226)
Intangible Ratio		-1.159		-0.475
		(2.462)		(1.741)
Log(Liabilities)		0.385		-0.024
		(0.419)		(0.393)
Year Fixed Effects	N	N	N	N
Industry Fixed Effects	N	N	N	N
Observations	166	166	135	135
R^2	0.010	0.030	0.072	0.105
F Statistic	2.489	0.711	6.303**	1.370
Model	OLS	OLS	OLS	OLS

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

Notes: These results remove all fixed effects from equation 2.16.

Table B.8: Cybersecurity Incidents

Company	News	Filing	Statement	Event Date
AT&T	08-29-2006			08-29-2006
Fidelity National Information Services	07-03-2007			07-03-2007
Cisco	07-10-2010			07-10-2010
McDonalds	12-13-2010			12-13-2010
Comcast	06-19-2012			06-19-2012
Rite Aid			09-27-2012	09-27-2012
Adobe	10-03-2013			10-03-2013
Target	12-18-2013			12-18-2013
AutoNation	05-27-2014			05-27-2014
Aecom	07-16-2014			07-16-2014
Community Health Systems	08-19-2014	08-14-2014		08-14-2014
Home Depot	09-08-2014			09-08-2014
Fidelity National Financial Corportation	10-08-2014			10-08-2014
United Airlines	12-24-2014			12-24-2014
Natural Grocers	03-02-2015			03-02-2015
Sally Beauty Holdings	05-04-2015			05-04-2015
United Airlines	07-29-2015			07-29-2015
Hyatt Hotels Corporation	12-23-2015		12-23-2015	12-23-2015
Sprouts Farmers Market	03-24-2016		03-28-2016	03-24-2016
Equifax	05-06-2016		05-05-2016	05-05-2016
Noodles and Company	05-19-2016		06-28-2016	05-19-2016

Continued on next page

Company	News	Filing	Statement	Event Date
Quest Diagnostics			12-12-2016	12-12-2016
Ameriprise	12-16-2016			12-16-2016
Western Union			12-20-2016	12-20-2016
Performant Financial Corporation			04-07-2017	04-07-2017
Humana			04-18-2017	04-18-2017
Sabre Corporation	05-02-2017	05-02-2017		05-02-2017
Merck & Co., Inc.	06-27-2017	02-27-2018		06-27-2017
Mondelez International	06-27-2017	07-07-2017		06-27-2017
Fedex Corp	06-28-2017	07-17-2017	06-30-2017	06-28-2017
Equifax Inc.	09-07-2017	09-07-2017	09-07-2017	09-07-2017
Forrester Research	10-06-2017	10-10-2017	10-06-2017	10-06-2017
Insulet Corporation			10-17-2017	10-17-2017
Boeing Co	03-28-2018			03-28-2018
Under Armour Inc.	03-29-2018	03-29-2018	03-29-2018	03-29-2018
Brinker International, Inc.	05-14-2018	08-27-2018		05-14-2018
Cigna	06-04-2018			06-04-2018
T-Mobile US, Inc.	08-24-2018			08-24-2018
Orrstown Financial Services, Inc.	09-05-2018	04-18-2019		09-05-2018
Chegg, Inc.	09-26-2018	09-25-2018		09-25-2018
Chegg Corporation	09-26-2018			09-26-2018
Meta Platforms, Inc.	09-28-2018		09-28-2018	09-28-2018
Alphabet Inc	10-08-2018		10-08-2018	10-08-2018
Village Bank			11-28-2018	11-28-2018

Continued on next page

Company	News	Filing	Statement	Event Date
Marriott International, Inc.	11-30-2018			11-30-2018
Humana	01-04-2019		01-03-2019	01-03-2019
Aetna	01-07-2019			01-07-2019
CarGurus			01-18-2019	01-18-2019
Five Below			02-14-2019	02-14-2019
Toyota Motor Corp	02-20-2019		02-21-2019	02-20-2019
ABM Industries			03-12-2019	03-12-2019
Urban One Corporation	05-15-2019	05-09-2019	03-28-2019	03-28-2019
Meta Platforms, Inc.	04-03-2019			04-03-2019
Carlylye Group			04-05-2019	04-05-2019
HSBC			05-09-2019	05-09-2019
Berry Global Corporation			05-09-2019	05-09-2019
ESI			05-15-2019	05-15-2019
Ryder System Corporation			05-24-2019	05-24-2019
Xperi Corporation			05-31-2019	05-31-2019
Quest Diagnostics	06-03-2019		06-04-2019	06-03-2019
Laboratory Corporatin of America Holdings	06-04-2019	06-04-2019		06-04-2019
Natera Corporation			06-04-2019	06-04-2019
Capital One Financial Corp	07-29-2019	07-30-2019	07-29-2019	07-29-2019
AAR Corporation			08-06-2019	08-06-2019
Brixmor Property Group			08-09-2019	08-09-2019
Cable One, Inc.	08-16-2019			08-16-2019
Deluxe Corporation			08-28-2019	08-28-2019

Continued on next page

Company	News	Filing	Statement	Event Date
Cronos Group Corporation			09-13-2019	09-13-2019
Park Hotels and Resorts Corporation			09-16-2019	09-16-2019
American Express Company			09-30-2019	09-30-2019
Pitney Bowes Inc	10-14-2019	10-15-2019	10-14-2019	10-14-2019
Patrick Industries, Inc.	10-24-2019	10-24-2019	10-24-2019	10-24-2019
Dominion Energy Credit Union			10-29-2019	10-29-2019
Marriott International, Inc.			10-30-2019	10-30-2019
Golden Entertainment			11-07-2019	11-07-2019
Macy's	11-18-2019		11-14-2019	11-14-2019
T-Mobile, INC	11-21-2019		11-21-2019	11-21-2019
Avid Technology Corporation			12-24-2019	12-24-2019
Microsoft Corp	01-22-2020		12-31-2020	01-22-2020
Altice	02-05-2020			02-05-2020
MGM Resorts International	02-19-2020			02-19-2020
Dynavax Technologies Corporation			02-28-2020	02-28-2020
Carnival Corp	03-04-2020			03-04-2020
Tandem Diabetes Care Corporation			03-16-2020	03-16-2020
General Electric	03-23-2020		03-20-2020	03-20-2020
Marriott International, Inc.	03-31-2020		03-31-2020	03-31-2020
Cognizant Technology Solutions Corporation	04-18-2020	04-20-2020		04-18-2020
Stride Inc.	11-30-2020		04-18-2020	04-18-2020
Chegg Corporation	04-29-2020			04-29-2020
Pitney Bowes Inc	05-11-2020			05-11-2020

Continued on next page

Company	News	Filing	Statement	Event Date
Conduent	06-04-2020			06-04-2020
Honda Motor Co Ltd	06-09-2020		06-08-2020	06-08-2020
MaxLinear, Inc.	06-16-2020	06-16-2020	06-10-2020	06-10-2020
Xerox	06-30-2020			06-30-2020
Steel Partners Holdings			07-01-2020	07-01-2020
DXC Technology Co	07-05-2020	07-06-2020	07-05-2020	07-05-2020
FormFactor	07-09-2020		07-17-2020	07-09-2020
Twitter, Inc.	07-15-2020		07-18-2020	07-15-2020
Blackbaud, Inc.	07-16-2020	09-29-2020		07-16-2020
Orange SA	07-16-2020			07-16-2020
Telcomm Argentina	07-20-2020			07-20-2020
Garmin Ltd.	07-24-2020	07-27-2020	07-27-2020	07-24-2020
SiteOne Landscape Supply, Inc.	07-27-2020	07-27-2020	07-29-2020	07-27-2020
MGP Ingredients, Inc.	07-30-2020	02-25-2021	02-25-2021	07-30-2020
Cornerstone Building Brands, Inc.	08-11-2020	08-11-2020		08-11-2020
Salem Media Group, Inc.	08-14-2020	08-12-2020	08-14-2020	08-12-2020
R1 RCM Holdco Inc	08-14-2020			08-14-2020
Carnival Corp	08-17-2020	08-17-2020	08-17-2020	08-17-2020
Amphastar Pharmaceuticals Inc			08-27-2020	08-27-2020
Stericycle			08-31-2020	08-31-2020
Equinix Inc	09-10-2020	09-10-2020	09-09-2020	09-09-2020
IPG Photonics Corp	09-18-2020	09-21-2020		09-18-2020
Shopify, Inc.	09-23-2020		09-22-2020	09-22-2020

Continued on next page

Company	News	Filing	Statement	Event Date
Tyler Technologies, Inc.	09-23-2020	09-29-2020		09-23-2020
Universal Health Services Inc	09-28-2020	09-29-2020	09-29-2020	09-28-2020
Arthur J. Gallagher & Co.	09-29-2020	09-28-2020		09-28-2020
Barnes & Noble Education Inc	10-14-2020		10-14-2020	10-14-2020
Minerals Technologies, Inc.	10-26-2020	10-26-2020		10-26-2020
Steelcase Inc	10-27-2020	10-26-2020		10-26-2020
Mattel Inc	11-03-2020	11-03-2020		11-03-2020
The Geo Group, Inc.	11-03-2020	11-03-2020	11-03-2020	11-03-2020
Americold	11-16-2020		11-16-2020	11-16-2020
Embraer SA	12-03-2020			12-03-2020
FireEye, Inc.	12-08-2020	12-08-2020		12-08-2020
Spotify Technology SA	12-10-2020		12-09-2020	12-09-2020
Aetna	12-11-2020			12-11-2020
SolarWinds Corp	12-13-2020	12-14-2020		12-13-2020
Forward Air Corporation	12-16-2020	12-21-2020		12-16-2020
Whirpool Corp	12-28-2020			12-28-2020
Veritex Holdings Inc			01-11-2021	01-11-2021
Qualys Inc	03-04-2021	03-04-2021	03-03-2021	03-03-2021
Molson Coors Beverage Co	03-11-2021	03-11-2021		03-11-2021
Shell plc	03-22-2021		03-16-2021	03-16-2021
Insulet Corporation			03-18-2021	03-18-2021
Meta Platforms, Inc.	04-03-2021		04-06-2021	04-03-2021
The Dixie Group, Inc.	04-19-2021	04-19-2021	03-10-2022	04-19-2021

Continued on next page

Company	News	Filing	Statement	Event Date
SmileDirectClub, Inc.		05-03-2021		05-03-2021
Peloton Interactive Inc	05-05-2021			05-05-2021
Oak Valley Community Bank			05-11-2021	05-11-2021
CNA Financial Corp	05-13-2021		07-09-2021	05-13-2021
Allied Healthcare Products, Inc.	06-04-2021	06-04-2021	06-02-2021	06-02-2021
Electronic Arts	06-10-2021		06-11-2021	06-10-2021
McDonalds Corp	06-11-2021			06-11-2021
Bath and Body Works			08-10-2021	08-10-2021
P&F Industries, Inc.	08-12-2021	11-12-2021	08-12-2021	08-12-2021
T-Mobile, INC	08-15-2021		08-17-2021	08-15-2021
T-Mobile US, Inc.	08-17-2021	08-27-2021	08-17-2021	08-17-2021
ALJ Regional Holdings, Inc.	08-19-2021	08-19-2021		08-19-2021
MFA Financial Inc			09-01-2021	09-01-2021
Marcus & Millichap, Inc.	09-20-2021	09-20-2021		09-20-2021
Golden Entertainment	09-24-2021			09-24-2021
Star Group, L.P.	09-24-2021	09-24-2021		09-24-2021
Sinclair Broadcast Group, Inc.	10-18-2021	11-03-2021		10-18-2021
J.B. Hunt Transport Services			10-20-2021	10-20-2021
Kewaunee Scientific Corporation	11-10-2021	11-10-2021	03-09-2022	11-10-2021
Godaddy Inc	11-22-2021	11-22-2021	11-22-2021	11-22-2021
Radiant Logistics, Inc.	12-13-2021	01-14-2022		12-13-2021
McGrath RentCorp			12-15-2021	12-15-2021
Century Aluminum Company	02-16-2022	02-24-2022	02-24-2022	02-16-2022

Continued on next page

Company	News	Filing	Statement	Event Date
Expeditors International of Washington, Inc.	02-22-2022	05-03-2022	02-22-2022	02-22-2022
Nvidia Corp	02-25-2022		03-01-2022	02-25-2022
Aon plc	02-28-2022	02-28-2022	02-28-2022	02-28-2022
Okta Inc	03-21-2022	04-19-2022	03-23-2022	03-21-2022
Mailchimp	04-03-2022			04-03-2022
Tenet Healthcare Corp	04-26-2022	07-21-2022	04-26-2022	04-26-2022
Montrose Environmental Group, Inc.	06-14-2022	06-14-2022	06-14-2022	06-14-2022

Appendix C.

Appendix to Chapter 3

C.1 Propensity Matching

Of the 62,234 individuals identified as treated, 54,192 are matched to the first control group and 40,348 to the second control group. The treated but unmatched group are typically from the lower end of the credit score distribution where there are very few untreated individuals to be matched with. This is a result of two matching specification decisions I made. First, matches are only conducted within matching groups based on credit score bins. Second, I limit the matches to just those on a common support, meaning that any treated individual whose propensity score is not within the range of scores among the treated group will not be matched. Almost by definition, many of the treated individuals will have very low credit scores due to the fact that they have at least one delinquency on their credit report. While some untreated individuals will have non-student loan delinquencies which will also lead to low credit scores, the data show that the frequency of extremely low credit scores is higher for the treated group than untreated group (figure B.1).

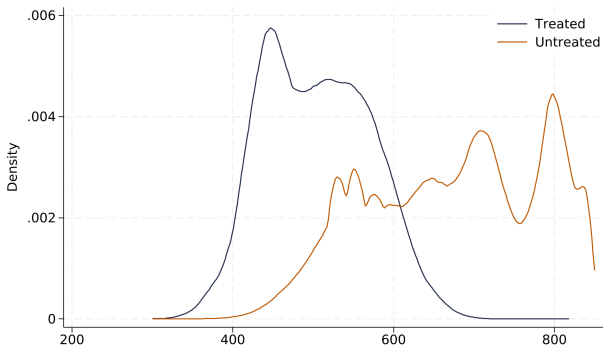
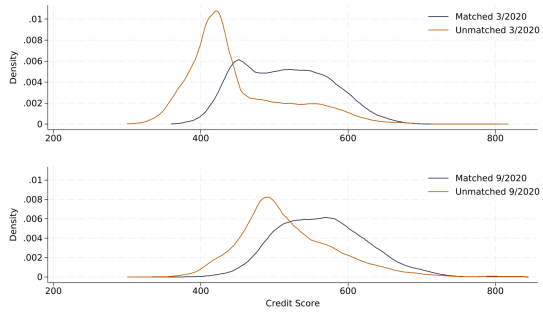


Figure B.1: Credit Score Density

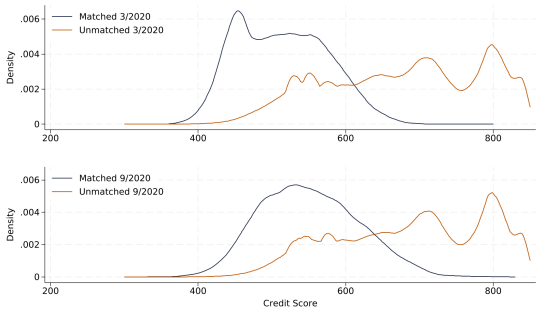
Figure B.2 shows the distribution of credit scores in March and September 2020 for the

matched and unmatched individuals, split by their treatment status. For the treated group, those who were not matched typically have a lower credit score even after the increase in credit in score from the payment pause (panels (a) and (c)). The opposite is true for the untreated group, where the matched individuals tend to be from the lower end of the credit score distribution (panels (b) and (d) of the same figure).

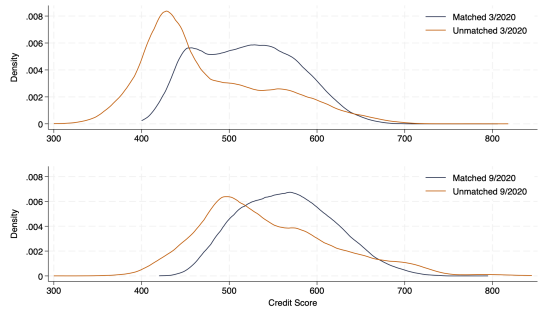
Figure B.2: Matched and Unmatched Credit Score Density



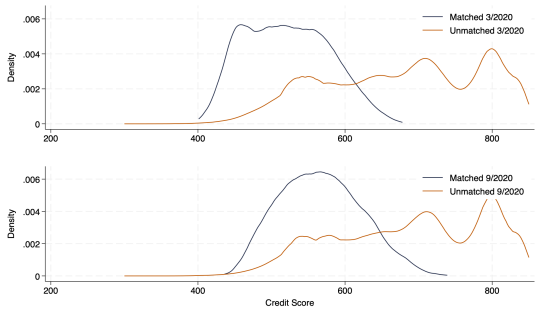
(a) Match Group One, Treated



(b) Match Group One, Untreated



(c) Match Group Two, Treated



(d) Match Group Two, Untreated

Notes: The density of credit scores for the matched and unmatched groups are presented in each figure. The first row shows the first match groups, where individuals were matched based on their observable characteristics up to March 2020. The second row shows the second match group, where individuals were matched based on their observables up to September 2020.

The density of credit score changes between March and September 2020 for the treated but unmatched group is slightly to the right of the density for the treated and matched. This is true for both match groups. As I discussed in section 3.4 and will expand on in section C.2, credit score change is decreasing in March 2020 score. Among the treated, those in the unmatched group have lower credit scores on average than their peers. It is therefore unsurprising that they saw larger increases in their credit score than those who were matched. The larger credit score increases could mean there is a greater loosening

of the credit constraint for the unmatched, but they still have lower credit scores than the matched, which could limit that benefit.

Among those who were matched, figure B.3 shows the distribution of credit score changes between March and September 2020. Panel (a) compares the treated to the first control group. Panel (b) compares the treated to the second control group.

C.2 Credit Score Change Factors

Means for the variables used in the credit score change factor regressions are in table B.1.

Table B.1: Regression Variable Means

	Summary
N	62,234
Credit Score Change	53.577 (40.792)
Credit score	504.211 (66.701)
# total of open trades	5.987 (5.054)
total balance on open trades reported in last 3 months	60,068.879 (89,088.464)
(sum) acct_balance_am	34,390.860 (47,802.237)
(sum) acct_past_due_am	2,524.613 (9,203.787)
(sum) treated	3.094 (3.273)

Notes: These are the means among the treated group for the variables in the regression in equation 3.1. Standard deviations are in the parenthesis.

In addition to the specification presented in the main text, I estimate the credit score change explanatory factors using models with non-linear relationships between score change and March 2020 credit score. Starting with a log transformation of credit score reported in table B.2, the direction of the influence of each variable is the same.

Next, I add the square of the March 2020 credit score, finding that, while higher credit scores in March 2020 still lead to lower credit score changes, this effect is diminishing. Interestingly, under this specification the number of trades that the payment pause affected now does have a statistically significant effect, with more trades causing larger credit score changes (table B.6).

Finally, I re-estimate the effects under all specifications previously mentioned, but using the full sample rather than just the treated subset. The only change I make to the model is including an indicator for whether the individual was treated. Results are in tables B.4-B.6.

Table B.2: Credit Score Change Factors — Logged

	Dependent Variable: Credit Score Change		
	(1)	(2)	(3)
Intercept	767.7***	806.7***	804.0***
	(7.1001)	(7.7084)	(8.2801)
3/2020 Credit Score Logged	-114.9***	-121.4***	-121.0***
	(1.1423)	(1.2522)	(1.3422)
Number of Open Trades		-0.161***	-0.175***
		(0.0340)	(0.0489)
Balance on Trades		0.0000370***	0.0000274***
		(0.0000)	(0.0000)
Student Loan Balance			0.0000358***
			(0.0000)
Past Due on SL			-0.0001000***
			(0.0000)
Number of Treated Loans			-0.0685
			(0.0674)
N	61906	61906	61906
R^2	0.141	0.146	0.147

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2, replacing the March 2020 credit score with it's log transformation.

C.3 Conditional Delinquency

In the main text, the delinquency effects are not conditioned on having an open auto loan or credit card. Re-running those estimates using only those who have an open loan. Results are reported in table B.7.

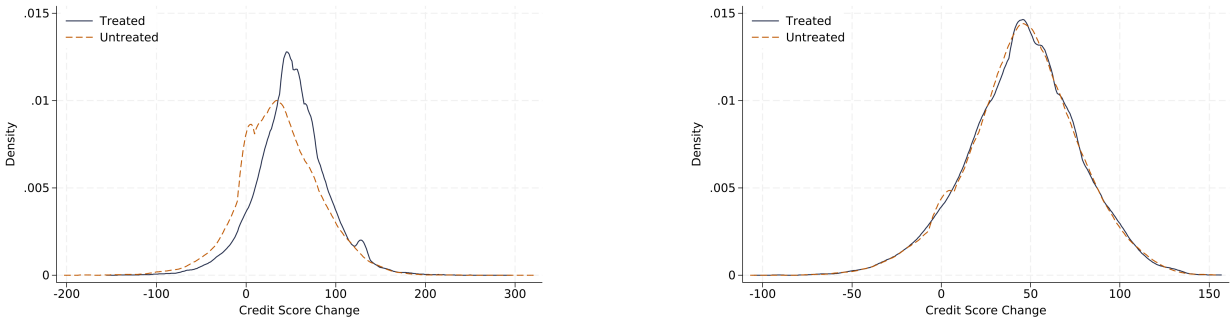
Table B.3: Credit Score Change Factors—Squared

	Dependent Variable: Credit Score Change		
	(1)	(2)	(3)
Intercept	373.6***	366.5***	376.6***
	(7.6619)	(7.6635)	(7.7343)
3/2020 Credit Score	-1.050***	-1.014***	-1.060***
	(0.0303)	(0.0304)	(0.0308)
Credit Score Squared	0.000809***	0.000762***	0.000812***
	(0.0000)	(0.0000)	(0.0000)
Number of Open Trades		-0.147***	-0.292***
		(0.0340)	(0.0492)
Balance on Trades		0.0000331***	0.0000210***
		(0.0000)	(0.0000)
Student Loan Balance			0.0000468***
			(0.0000)
Past Due on SL			-0.000107***
			(0.0000)
Number of Treated Loans			0.153*
			(0.0683)
N	61906	61906	61906
R^2	0.146	0.150	0.152

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2, adding the square of the March 2020 credit score to the right-hand side.



(a) Match Group One

(b) Match Group Two

Figure B.3: March–September 2020 Credit Score Change Distributions

Notes: This figure contains the density of credit score changes among the matched treated and untreated groups. Panel (a) compares those in the pre-pause matches. Panel (b) compares those matched with post-pause matches.

Table B.4: Credit Score Change Factors—All

	Dependent Variable: Credit Score Change		
	(1)	(2)	(3)
Intercept	80.59***	86.31***	86.75***
	(0.1285)	(0.1323)	(0.1375)
3/2020 Credit Score	-0.104***	-0.119***	-0.120***
	(0.0002)	(0.0002)	(0.0002)
Treated	25.21***	23.29***	24.90***
	(0.1412)	(0.1405)	(0.1897)
Number of Open Trades		0.580***	0.554***
		(0.0039)	(0.0040)
Balance on Trades		0.0000105***	0.0000100***
		(0.0000)	(0.0000)
Student Loan Balance			0.0000100***
			(0.0000)
Past Due on SL			-0.0000819***
			(0.0000)
Number of Treated Loans			-0.565***
			(0.0414)
N	2882961	2882961	2882266
R^2	0.122	0.134	0.135

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2 using the full sample rather than just the treated group.

Table B.5: Credit Score Change Factors — Logged, All

	Dependent Variable: Credit Score Change		
	(1)	(2)	(3)
Intercept	154.0***	158.3***	516.2***
	(0.1353)	(0.1356)	(0.8447)
3/2020 Credit Score Logged	-23.00***	-24.36***	-78.44***
	(0.0209)	(0.0213)	(0.1311)
Number of Open Trades		1.006***	0.554***
		(0.0034)	(0.0040)
Balance on Trades		0.000000884***	0.00000908***
		(0.0000)	(0.0000)
Student Loan Balance			0.0000124***
			(0.0000)
Past Due on SL			-0.0000893***
			(0.0000)
Number of Treated Loans			2.859***
			(0.0309)
N	13790629	13790629	2882266
R^2	0.081	0.087	0.130

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2 using the full sample rather than just the treated group, and using the log transformation of the March 2020 credit score instead of the untransformed values.

Table B.6: Credit Score Change Factors—Squared, All

	Dependent Variable: Credit Score Change		
	(1)	(2)	(3)
Intercept	132.3***	144.3***	148.0***
	(0.7408)	(0.7433)	(0.7458)
3/2020 Credit Score	-0.264***	-0.300***	-0.311***
	(0.0023)	(0.0023)	(0.0023)
Credit Score Squared	0.000121***	0.000137***	0.000145***
	(0.0000)	(0.0000)	(0.0000)
Number of Open Trades		0.620***	0.589***
		(0.0039)	(0.0040)
Balance on Trades		0.00000921***	0.00000844***
		(0.0000)	(0.0000)
Treated	23.08***	20.86***	22.55***
	(0.1442)	(0.1437)	(0.1915)
Student Loan Balance			0.0000134***
			(0.0000)
Past Due on SL			-0.0000893***
			(0.0000)
Number of Treated Loans			-0.632***
			(0.0414)
N	2882961	2882961	2882266
R^2	0.123	0.136	0.137

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: This table contains the results of estimating equation 3.2 using the full sample rather than just the treated group, and adding the square of March 2020 credit score to the right hand side.

Table B.7: Conditional Delinquency Effects

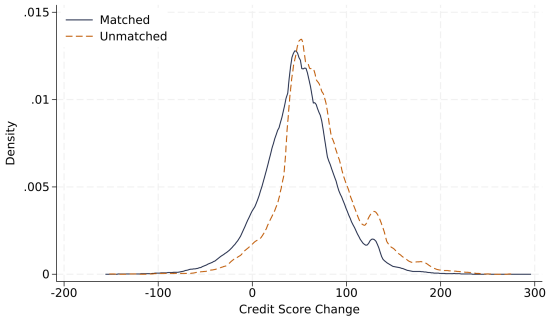
	Match Group One		Match Group Two	
	Auto	Credit Card	Auto	Credit Card
	(1)	(2)	(3)	(4)
Constant	0.0520***	0.109***	0.0494***	0.103***
	(0.0003)	(0.0003)	(0.0003)	(0.0004)
Treated	-0.0106***	-0.00969***	-0.0121***	-0.0125***
	(0.0008)	(0.0010)	(0.0009)	(0.0012)
Treated x Post	0.00364***	-0.0249***	0.00735***	-0.0120***
	(0.0010)	(0.0013)	(0.0011)	(0.0015)
Match FE	Yes	Yes	Yes	Yes
Period FE	Yes	Yes	Yes	Yes
N	847,557	939,493	647,558	718,906

Standard errors in parentheses

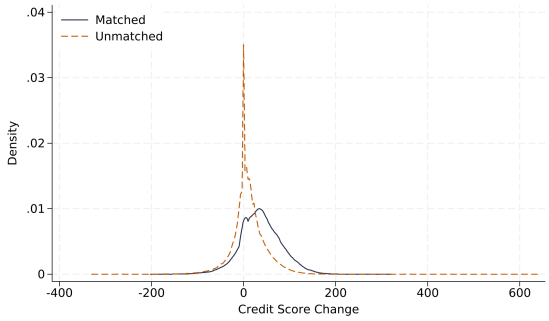
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Notes: The results in this table are from estimating equation 3.2, where the outcome variables are whether the individual went delinquent on an auto loan or credit card, conditional on having at least one trade of that type open.

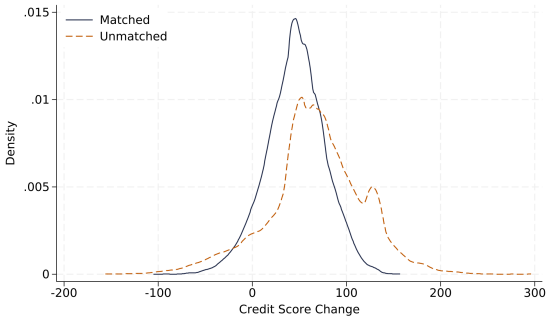
Figure B.4: Matched and Unmatched Credit Score Change Density



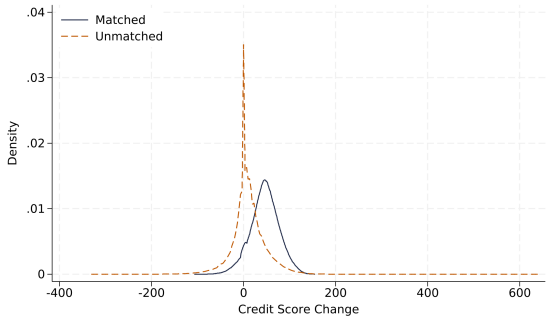
(a) Match Group One, Treated



(b) Match Group One, Untreated



(c) Match Group Two, Treated



(d) Match Group Two, Untreated

Notes: The density of credit score change between March and September 2020 for the matched and unmatched groups are presented in each figure. The first row shows the first match groups, where individuals were matched based on their observable characteristics up to March 2020. The second row shows the second match group, where individuals were matched based on their observables up to September 2020.

Appendix D.

Miscellaneous Material

I used open source software for almost all the empirical analysis in this dissertation. The software and versions used are listed in table D.1.

Table D.1: Software and Versions

Software	Version	Developer
Python	3.12	Python Software Foundation
R	4.3.3	R Core Team (2024)
Pandas	2.2.3	The Pandas Development Team (2024)
Numpy	1.26.4	Harris et al. (2020)
Seaborn	0.13.2	Waskom (2021)
edgartools	3.10.2	Dwight Gunning
Statstables	0.0.16	Anderson Frailey
Scikit-Learn	1.6.1	Pedregosa et al. (2011)
scipy	1.14.0	Virtanen et al. (2020)
psmatch2	4.0.12	Leuven and Sianesi (2003)
fixest	0.12.0	Bergé (2018)
reticulate	1.37.0	Ushey et al. (2024)