Do Data Breach Disclosure Laws Increase Firms' Investment in Securing Their Digital Infrastructure?

Raviv Murciano-Goroff, **Boston University**[*]

May 2019

Abstract

In 2002, California enacted a law requiring that companies publicly disclose data breaches. The law created additional incentives for firms to invest in information security. Since then, all other states have implemented similar laws. To evaluate the impact of data breach notification laws, I collected data on the decisions of 213,810 public and private companies regarding when to update their web server software and apply security patches during the years before and after the California legislation was signed. Comparisons are made between groups of companies within and outside of the jurisdiction of the California law using a difference-in-differences framework. I show that firms that use older server software are also more likely to suffer a successful hacking event and data breach. In addition, I find that the data breach notification law in California caused firms headquartered in that state to use web server software that was 1.8-2.8% newer. The effect of this law was most pronounced among larger firms. Ultimately, while data breach notification laws have received considerable attention in recent years, their impact on firm investment in web server security appears modest.

Keywords: Information systems, information security management, IS policy.
JEL Classification: D22, L86.

[*] E-mail: ravivmg@post.harvard.edu. Thank you to Shane Greenstein, Timothy Bresnahan, Jonathan Levin, Muriel Niederle, Richard Freeman, Matt Marx, Sonny Tambe, and Liran Einav. Thank you to the WEIS Conference and IO Seminar participants at Stanford University as well as three anonymous referees. Thank you to Kenji Nagahashi and Mark Graham from the Internet Archive.

1 Introduction

Data breaches cause tremendous economic damage to firms and consumers. According to industry reports, 1,579 data breaches occurred during 2017 with a total economic cost of $107 billion in the United States alone (Javelin Strategy, 2017; Kamiya et al., 2018; IBM, 2018). In addition, the number of people impacted by data breaches has increased over time. For example, the 2018 data breach of the credit bureau Equifax leaked the personal information of 147.9 million individuals (Fung, 2018). In the wake of such large-scale and widespread data breaches, lawmakers have sought policies that can encourage firms to invest in Information Technology (IT) security.

Data breach notification laws are one policy that has been implemented in the United States in hopes of combatting such data loss events. Enacted within all 50 states, these laws mandate that firms notify their customers if unauthorized individuals accessed sensitive company data.[1] The motivation for these laws is two fold. First, lawmakers believe that if individuals are informed about data losses then the affected consumers can take preventative measures to safeguard their financial accounts and deter identity theft. Second, proponents of these laws believe that by highlighting which firms have lax digital security they encourage companies to invest in IT security (Sullivan and Maniff, 2016).

Empirical evaluations of these laws have been narrow in scope. Regarding how these laws impacted consumers' financial losses, Romanosky et al. (2011) found that the implementation of data breach notification laws decreased the incidence of identity theft by an average of 6.1%. Similarly, Sullivan and Maniff (2016) found that particular provisions present in some states' data breach laws, such as the capping of the civil penalty possible from a data breach, are more effective for lowering identity theft.[2] Regarding how firms responded to such laws, Miller and Tucker (2011) found that firms invested more heavily in encrypting their data within states that implemented data breach notification laws with exemptions for lost encrypted data. Little empirical evidence, however, has been collected that shows if firms responded with additional investments in their management of their digital infrastructure.

In this study, I examine the potential impact of state legislated data breach notification laws on firm investment in their web server software and consciousness in keeping their web servers up-to-date. Measuring firms' investment in particular components of IT infrastructure has previously been challenging as standardized and comparable data regarding these investments across firms are not readily available. In addition, while it is easy to identify the firms that suffered data breaches, separating firms that were relatively strict versus lax in their commitment to the security of their digital infrastructure has not been done previously.

The novel dataset used for the analysis in this paper provides a window into the decision-making of

[1] No federal law currently exists that mandates all companies that experience a data breach disclose that information to consumers. For health and financial information, however, various federal laws require companies to publicly report data losses.

[2] By capping the civil penalties, firms may be more likely to promptly inform consumers of a breach. In contrast, however, this cap might lower the incentives for firms to prevent breaches.

firms regarding their investment in security related IT. I collected a panel dataset with information about the web servers used by over 200,000 public and private companies to host their website during every month between the years 2000 and 2018. The data collected contains both the vendor of the web server as well as the particular version of web server software used. By matching the software versions to their release date and known security vulnerabilities at the time, I can observe if companies used up-to-date or outdated server software as well as if they applied security patches in a timely manner.

Using this data, I am able to measure the effect of the data breach notification legislation on firms' decisions regarding the application of software updates and investments in web server upgrades. In particular, I exploit the staggered enactment of data breach notification laws across 50 states between 2002 and 2018 for this study. Starting in 2002, California established the first data breach notification law mandating that companies that suspected their data had been hacked or mishandled notify their California customers of the breach. In the subsequent years, other states enacted similar laws.[3]

A number of idiosyncratic components of these laws enable my analysis. First, these laws were implemented at the state level. Only companies that held personal information about residents within a state would be subject to a state's data breach notification law. Second, the state laws did not create new incentives for financial firms. Companies that handled financial information were already subject to a federal law requiring such firms to disclose data breaches.[4] Third, and perhaps most importantly, the timing of the enactment of these laws was plausibly exogenous to any particular firm's investment in digital security management.

I begin by focusing on the very first data breach notification law in the United States, the 2002 California law. I measure the effect of this policy change by comparing firms headquartered in California versus those headquartered in other states.[5] I examine if firms that were clearly subject to the new law by physically being located and operating within California responded by being relatively more diligent about updating their web server software and applying patches. In addition, I use financial firms and companies operating in sectors in which customer data would be unlikely to have customer data accessible on their website as a placebo test for the impact of this law.

My results show that the trailblazing California legislation did have a modest effect on firms' decisions on which version of web server software to use. On average, California based firms kept their web servers equipped with software 0.33 months newer or 1.7% younger than they would have otherwise. Companies

[3] The particular law is California Senate Bill 1386. It was signed by the Governor of California on September 25, 2002 and went into effect on July 1, 2003. The text of the bill can be found here: `https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=200120020SB1386`. This law is particularly important as it was both the first such law in the U.S. and it became the model for the data breach notification laws enacted in at least 49 states in subsequent years (Wugmeister and Lyon, eds, 2011).

[4] The Gramm-Leach Bliley Act (GLBA) is a federal law enacted in 1999 that contains similar mandates regarding data breach notification for financial companies.

[5] Ideally, one would like to identify companies that did not have California residents among their customers. In later parts of the paper, I examine firms operating in non-tradeable sectors in an effort to get a subsample unlikely to be conducting interstate commerce.

with highly trafficked websites, those in the top 1,000 most popular websites, responded by moving to servers about one month newer or 2% younger.

This paper contributes to the literature on the effectiveness of information security laws on firm investment digital security by providing empirical analysis of firm responses to these laws. Legal scholars have discussed if the framework of data breach notification laws could effectively influence firms to invest in information security and pointed to potential shortcomings in the law (Schwartz and Janger, 2007; Picanso, 2006; Needles, 2009a). The empirical studies on the impact of these laws have focused on the changes in the prevalence of identity theft because of consumers being more aware of breaches (Romanosky et al., 2011; Hoofnagle, 2007; Sullivan and Maniff, 2016). Unlike these previous works, the dataset collected and examined in this paper directly measures a firm decision regarding digital security conscientiousness.

The most closely related paper to this work isMiller and Tucker (2011) who find that hospitals in states with data breach notification laws increased their rates of encrypting their customer data. Three aspects of this paper build and add to their previous study. First, this paper examines the outcome of server software updating. Unlike encrypting data, which can be seen as mitigating the impact of a data breach, securing server software can be viewed as a preventative measure against data breaches. Second, this paper seeks to specifically identify a causal impact of data breach notifications on the outcome variable. In Miller and Tucker (2011), data breach notification laws and variations in their legal construction are used to predict the decision to encrypt as an instrument for identifying the effect of encryption on the probability of data breaches. Using the difference-in-difference approach, I identify the causal impact of these laws on the server software updating behavior of firms. Finally, while the data in this paper includes hospitals, the over 200,000 firms included in the study provide insights into the breadth of the impact of these laws across industries.

This paper also contributes to the specific literature on the effect of deterrence policies as a means to encourage information security. Previous articles have investigated the effectiveness of shaming and potential punishments in encouraging individual employees to adhere to information security protocols (D'Arcy et al., 2009; D'Arcy and Herath, 2011; Harrington, 1996). Few papers have examined how parallel policies could impact firm behavior. The insights from this paper, therefore, can help us understand how firms respond to the deterrence mechanism of publicized data breach notification when making decisions about investment in digital security.

Finally, the data collected and examined in this analysis contributes to the growing literature measuring firm investment in unpriced activities and software (Greenstein and Policy, 2014). In particular, software updates are typically unpriced, and thus not easily captured in usual measures of economic activity. This paper begins to illuminate the importance of these recurring and largely overlooked firm investments.

2 Setting

2.1 Data Security Laws

On September 25, 2002, the California Senate enacted a bill requiring any company that discovered unauthorized access of personally identifiable data to promptly disclose the security breach to their California customers.[6] This law became known as the first "data breach notification law" in the United States. The law created potentially large costs for firms that tried to hide security breaches from the public. Following the enactment of this law, customers could initiate civil action against breached firms that failed to disclose incidents and could pursue injunctive relief with no ceiling on the damages that could be assessed (Sells, 2003).[7]

The California law is notable for two reasons that are important for the analysis in this paper. First, the law changed the legal allocation of responsibility for data breaches. Instead of a hacker being solely responsible for causing a data breach, the California law placed some of the blame for the unauthorized data disclosure with the company whose digital infrastructure was compromised (Sabett, 2013). The California legislators intended the disclosure of data breaches as a way to publicly shame firms for poor information security infrastructure or policies. The increased costs of a potential data breach in turn increased the incentives for firms to invest in information security.[8]

Second, the California law only applied to companies doing business within that state and collecting data from consumers who are California residents (Perkins Coie LLP, 2018).[9] Since no other state had a data breach law between 2002 and 2005 and no federal law exists with a similar mandate for all firms across all industries, only companies with customers in California would be directly impacted by the California law.[10]

[6] The California law is known as Senate Bill (S.B.) 1386. "Personally Identifiable Information" is defined differently by different laws. In California S.B. 1386, the definition is a first name and last name along with any one of the following items: social security number, drivers license number, account number, security code, password, health insurance, or medical information. See the full text of the law: `https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=200120020SB1386`

[7] The potential costs for firms from this law were large. As one lawyer wrote, "Businesses would be ill-advised to take the new law lightly. The bill provides that an injured customer may bring a civil suit for damages and/or seek an injunction in addition to any other rights and remedies the customer may have. Although it may be difficult for an individual customer to trace an identity theft to a leak from a particular business, a business that fails to follow the law and that is exposed for not doing so could face a significant and expensive class action" (Coombs and Milner, July/August 2004).

[8] For example, many data breach notification laws incentivize firms to encrypt stored and transmitted data. Known as a "Safe Harbor" rule, many of the data breach notification laws say that if encrypted data is lost or stolen then customers do not need to be informed. By encrypting data, firms could limit their liability exposure to these new laws (Miller and Tucker, 2011; Coombs and Milner, July/August 2004).

[9] Specially, the law states, "person or business that conducts business in California." Following the enactment of the law, the precise meaning of this phrase would be debated and criticized for its ambiguity (**?**). As lawyers discussed, "Just what constitutes conducting business in California is unclear—there is no definition of 'conducts business in California.' It seems obvious that a business does not have to have a physical office in the state in order to be caught by the provisions of the law" (**?**). Another lawyer wrote, "The law applies to any person or business that does business in California, even if located out of state" (**?**).

[10] While federal laws existed for data breaches in the financial section, no other state had a similarly broad data breach notification law until 2005 (Needles, 2009a, p. 268).

In particular, firms that operated in non-tradable goods and services with highly geographically concentrated customer-bases outside of California would be unlikely to have responded to this legislation.

Finally, the California law has been the model of many subsequently enacted laws. In the 16 years following the California law, all other states have each created similar data breach notification laws. Every data breach notification law in the United States used the California law as a basic template (Sullivan and Maniff, 2016, p. 67). In Section D, I list the years in which each state's first data breach notification law went into effect. The first of these subsequent data breach notification laws did not go into effect until 2005.[11]

2.2 Web Server Technology

Websites are made accessible to users of the Internet by computer software known as web servers. Given the complexity of web server software, hackers and security experts regularly discover bugs and security vulnerabilities in server software that could enable unauthorized access to sensitive data. Therefore, the vendors of web server software frequently release updates to fix these vulnerabilities. Installing updates is an important means for keeping websites and the data behind them secure (Arora et al., 2010b).

One of the ways that firms can mitigate the risk of their websites being hacked is by being diligent in installing web server software updates. Installing updates, however, can be costly. Servers and their associated websites may need to be taken offline temporarily in order to install the updates. In addition, if new server software is incompatible with a firm's previously written website content then that firm will need to devote resources towards rewriting the code for their website. Therefore, firms face a tradeoff in keeping their web server up-to-date.

The two most popular web server software vendors between 2000 and 2005 had contrasting approaches towards software updates. Microsoft's Internet Information Services (IIS) server came packaged with Microsoft's operating systems, such as Windows NT, Windows XP, and Windows 7. The code for this server is proprietary. Microsoft made small software updates, including many security updates, available through a centralized system. Users would receive notices regarding available updates, and starting in 2001 with Windows XP, a program could automatically download and begin installing the security updates with almost no user interaction. Users would, however, often delay the installation of new software updates until they had tested those updates for conflicts with their existing content on servers dedicated for testing and development.

The Apache web server, in contrast, is developed as "open source software," meaning that the code for

[11] The California data breach notification was the more notable legal change related to data breaches and security prior to 2005. The Payment Card Industry Data Security Standard (PCI DSS), a self-organized set of standards by credit processing companies for the storage and transmission of payment, released their first version of standards in December 2004. These standards primarily affected financial institutions and payment processors, which are not included in the sample of companies that I use when analyzing the effect of the 2002 California data breach notification.

the server is posted online for others to freely use and modify. Unlike Microsoft's IIS, software updates for Apache did not automatically download or install. Administrators of Apache web servers learned about updates primarily from email lists. System administrators would then visit the Apache website, download, compile, and install the new software themselves. Alternatively, many system administrators utilize "package manager" programs that can track the release of new versions of software. Package managers can also compile and install updates, however, they require somewhat more technical sophistication than the update system provided with Microsoft IIS.[12]

Some firms outsource the operation of their server infrastructure to companies that specialize in enterprise server support. Companies, such as RedHat, manage and maintain a firm's servers, watch for bugs and security holes, and install updates when necessary. In addition, some firms host their websites on managed platforms, such as Yahoo!'s GeoCities or WordPress.com. Whether a firm installed server updates themselves or hired an enterprise support team to manage their servers, the amount of time between when software updates become available and when firms have those updates installed on their public-facing websites likely correlates with the IT security conscientiousness and investments of that firm.

Apache web server users are of particular interest because the majority of websites were served by Apache software between 1998 and 2013.[13] In addition, because Apache server updates did not have an automatic update installation feature built-in, the decision of how frequently to update this software is likely more strongly correlated with active investment decisions at a firm.[14]

3 Data

The primary data for this paper comes from a collection of meta data about the web servers used by firms to host their websites between 2000 and 2018.

When a user visits a website, the user's computer sends a request over the Internet to the server that hosts that page. The web server then responds with both the content of the webpage requested as well as meta-data about the server itself, known as "headers." These "headers" include information such as the vendor of the server as well as the version of the web server software installed on that machine.[15] The Internet Archive, a

[12] In addition, some companies outsourced the management of their web server to enterprise services. In those cases, much of the work of tracking and installing server updates fell to the enterprise solution company.

[13] Netcraft is an internet services company that has surveyed server software usage since 1995. Their report on server software market share is available here: `https://news.netcraft.com/archives/2018/09/24/september-2018-web-server-survey.html`

[14] In addition, the headers from Microsoft IIS servers only show the major versions (e.g. IIS 6 vs IIS 7). Security updates or minor version changes on Microsoft IIS are not easily detectable through parsing server headers. Minor versions are reported in the default Apache server headers.

[15] Server headers can be modified by the owner and some firms turn off displaying server information. I cross-validate that the headers align with the actual technology used at firms by comparing if observations where the website is built with Microsofts ASP.Net technology, a web framework primarily used with Microsoft's IIS server, show Microsoft IIS in the server headers. This validation and others are shown in the Appendix.

nonprofit organization, regularly makes requests to large numbers of publicly accessible websites and saves both the content of those websites and their associated "headers."

I collected information about the web servers used to host the homepages of companies located in the United States between 2000 and 2018 by examining the server headers collected by the Internet Archive. More specifically, I extracted a list of all public and private firms in the Bureau van Dyke Orbis database listed as being located in the United States and employing more than 50 workers. For each of these 271,579 companies, I downloaded the server headers collected by the Internet Archive for the company's homepage website. The Internet Archive data contained headers for 213,810 of the companies for their homepage website. I examined one set of headers per company homepage per month.[16] I then parsed the headers and inferred the vendor as well as the version of the server software being used to host the company's website (e.g. Microsoft IIS 5.0, Apache 1.3.7).[17]

This information is formed into a panel where an observation is a company's website in a month. The full dataset tracks 213,810 companies with an average of 86.68 observations per company. The variables in this panel include the vendor and the version of web server software being used by the firm in that month. Observations also include information about the associated company, such as the geographic location of the firm's operations and any subsidiaries, the sector in which the company operates based on their North American Industry Classification System (NAICS) code, the number of employees the company, the company's estimated revenue, and the company's Research & Development (R&D) expenditures that year from the Orbis database.[18] I also attach a measure of the Internet traffic to a company's website based on the position of the site in the 2010 Alexa Web ranking of websites by traffic.[19] Observations between the 2005 and 2018, I flag whether or not a hacking incident occurred on the associated website in a year. I create this indicator variable by merging the observations with the Privacy Rights Clearinghouse Database of data breaches.[20]

For each observation in the panel, I match the server vendor and version to the date in which that piece of software became publicly available for use.[21] Using this release date, I generate a variable denoted $TechAge$, which is defined as the number of months between the month of an observation in the panel and the month in which the server software being used by the firm in that observation first became available.

[16] If the Internet Archive had collected headers from a website more than once within a month, I took the first header for that site from that month.

[17] Appendix A for more details on the construction of the panel dataset and sample selection. Firms are able to change and misrepresent the server vendor and version displayed in server headers. In Appendix A.1 I show steps that I took to validate the server headers reflect the actual server technology a firm uses.

[18] In the Compustat data, the number of employees is often missing. In addition, this number is only collected annually.

[19] This data comes from the company Alexa Internet. They rank websites based one estimated hits to that site. The ranking has one as the highest traffic website and one million is the lowest trafficked site in their released dataset. The ranking is done for all websites worldwide. The ranking only goes up to 1 million. The 2010 ranking is the oldest publicly available data.

[20] The Privacy Rights Clearinghouse collects information on data breaches at firms. Their database started in 2005.

[21] I collected when versions came out by hand. I went through the "changelogs" as well as the mailing lists of the Apache server to find its release dates. For IIS, I examined the release dates of the Windows versions for which it was bundled.

The $TechAge$ of the server software used by a firm when compared with the ages of the servers used by other companies provides an indication of the firm's conscientiousness regarding digital security.

Server headers can be changed or manipulated by system administrators at firms. For example, server owners may wish to disguise the vendor and version of server software since hackers may target specific server software types.[22] About 10% of the observations in my sample display a server vendor but hide the software version number.[23] As discussed in detail in Appendix A.2, I create two additional variables for observations in which firms turn off the display of the software version they are using. In particular, I define $TechAgeMin$ to be the minimum technology age that a firm could have given the vendor of the software being used by the firm and available versions of software at that time. This variable would represent the scenario in which firms that turn off the display of their software version in the headers have switched to the latest version of available server software. I also define $TechAgeAvg$ to be the average technology age of the observations in the same month and for the same server vendor who have kept their version numbers displayed. This variable represents the scenario in which firms that hide their version number are no more or less security conscientious than those who do not. Using both of these variables in my analysis of the data breach notification law will provide bounds on the effect of the law.

This paper utilizes two datasets based on the panel data. The first dataset, which I refer to as the "Apache Users" dataset, is used to analyze the correlation of server technology age with the probability that a firm has a data breach. The Apache Users dataset contains the subset of the panel data observations in which the firm used the Apache web server software between 2005 and 2018. This dataset covers the time period in which I observe whether or not a firm experience a data breach due to a hacking incident. By comparing the $TechAge$ of firms that experience hacking incidents with those that do not, I can examine the correlation of $TechAge$ with the publicized hacking incidents.

The number of firms represented in the Apache User dataset correlates closely with government statistics, such as firms of similar sizes tracked by the Census Bureau in the Survey of U.S. Business (SUSB). In Appendix B, I show that across states there is a 0.98 correlation in the number of firms in the SUSB and the number of firms in the Apache User dataset. It should be noted that the total number of firms in this dataset are a fraction of the total number of U.S. firms of similar size. There are three reasons for this difference. First, the SUSB survey contains counts for firms with 20 employees and larger, while my sample contains the subset with 50 or more employees. Second, especially in the early 2000s, not all firms had a publicly visible website yet. Third, the Internet Archive only scans and collects headers from a subset of all public websites each month. While the total number of firms is less than the number in the SUSB, the sample in this dataset appears broadly representative of U.S. firms across states and for most industries.

[22] In Appendix A.1 I discuss my attempts to validate the server vendor information shown in the headers.

[23] There is no evidence of a differential propensity to hide the software version between California and non-California based firms.

A second dataset is used in this paper to analyze the effect of the 2002 California data breach notification law. I refer to this dataset as the "Non-tradable Firms" dataset as it contains the subset of panel data observations from the years 2000 to 2005 in which the firms in the sample have geographically concentrated demand in non-tradable sectors. In particular, I use firms that operate in the industries listed in Mian and Sufi (2014) as selling non-tradable goods and services with customers that are located within a small distance from the firm.[24] These firms include restaurants, grocery stores, clothing sporting goods, florists, shoe stores, as well as health care providers such as nursing care facilities and health care practitioners.[25] In addition, I drop any firms located in states bordering California (Oregon, Nevada, and Arizona). The identification strategy in this paper relies on being able to separate firms with California customers—that would have been impacted by the California data breach law—versus firms that were unlikely to have California customers. By using firms with geographically localized customers, this sample meets the necessary criteria for examining the effect of the data breach law but also encapsulates a large and diverse swath of economic activity.

Table 1 shows the means of various attributes of the companies in the Non-tradable Firms dataset in July 2002 prior to the California data breach notification going into effect. The means presented are from averaging over companies with non-missing values for those attributes. The means are shown separately for companies with headquarters in California and those outside of that state. The companies headquartered in California are significantly different from those based in other parts of the country along a number of dimensions. In particular, California based firms tend to have fewer employees, were founded more recently, and are less likely to be in the healthcare sector than companies based elsewhere. In addition, California based firms during the period prior to the data breach notification law going into effect used server software about 7.62% newer than non-California based firms. Because of the variation between firms located in different states and across different firms, the analysis in this paper will be done by examining changes over time within firms.

Figure 1 shows the market shares of web servers among California and non-California based firms. The Microsoft IIS server was more popular outside of California during this time period. In addition, servers other than the market leaders, Apache and IIS, had more popularity in California than other parts of the nation at the beginning of the sample time period.

Figure 3 presents the average technology age of servers used by firms in the Non-tradable Firms dataset between 2000 and 2005. The average technology age is shown separately for firms operating in California and outside of California. The average age trends upward as firms that continue using the same server

[24] See Appendix Table I, Panel B for the full list of industry NAICS codes included as non-tradable industries with geographically concentrated demand. In particular, I use all non-tradable industries listed as well as health care facilities with less than 0.8 geographical Herfindahl Index for demand.

[25] Mian and Sufi (2014) include many health care providers as being in an "Other" category of firms rather than in the "Non-tradable" segment of industries. They show, ever, that the geographic demand for this firms is equally if not more concentrated than that for many of the industries listed as "Non-tradable".

software from one period to the next would have the technology age of that server increase by a month. The red vertical line indicates when the California data breach notification law went into effect. The large structural break that occurs in July 2002 is due to the unexpected announcement of a security bug in the Apache software. Following the revelation of this security bug, many firms decided to update their software to the latest version of Apache that closed that security issue. During the period prior to the structural break as well as after the structural break, the trends in technology age appear very similar for California based firms as well as those outside of the state. I confine my analysis to the period after the structural break in order to avoid conflating the effect of the announcement of the security bug with the enactment of the data breach notification law.

4 Methodology

4.1 Correlation of Server Technology Age with Disclosed Hacking Incidents

Many factors influence the probability that a company's website is hacked. I investigate if the $TechAge$ of servers used by a firm predicts hacking incidents after controlling for other firm characteristics using the Apache Users dataset with observations from between 2005 and 2018. In particular, I estimate the following logistic regression:

$$Hacked_{jt} = \Phi\left(\alpha + \alpha_t + \beta TechAge_{jt} + \boldsymbol{\gamma' X_{jt}}\right) + \epsilon_{jt} \tag{1}$$

where α_t represents year fixed effects, and X_{jt} includes time-varying firm characteristics including the logarithm of the number of employees and an indicator for if the firm had a relatively high amount of website traffic as measured by being in the top one 500,000 websites by traffic in a given year according to Alexa Web. I estimate this regression on a panel of observations at the level of a firm in a year. These observations are created by aggregating the Apache User data to the annual level. I take the average value of the number of employees, $TechAge_{jt}$, and web traffic to get these observations. The variable $Hacked_{jt}$ is a binary variable for if the company experience a hacking event at any point within that year that was successful, publicized, and noted in the Privacy Rights Clearinghouse Database.

The variable of interest from this regression is β, which represents the change in the odds-ratio from companies using older web server software. If this coefficient is positive then running older server software is correlated with a higher probability of a successful hacking incidents. While many of the covariates in this regression are endogenous choices of the firms, and thus the coefficient on $TechAge$ should obviously not be considered a causal estimate, a positive and significant coefficient would provide evidence that even after controlling for many firm characteristics the age of server software is predictive of successful hacking incidents.

4.2 Effect of the 2002 California Data Breach Notification Law

In order to examine if the data breach notification laws incentivized firms to invest in their web server software, I use a difference-in-differences approach using the Non-tradable Firms dataset. Specifically, I estimate the following regression equation:

$$TechAge_{jt} = \alpha_j + \alpha_t + \delta LawEnacted_{jt} + \boldsymbol{\gamma' X_{jt}} + \epsilon_{jt} \tag{2}$$

where j indexes firms, t indexes months, α_j represents firm fixed effects, and α_t represents month fixed effects. An observation in this regression is a firm's website in a month. $LawEnacted_{jt}$ is an indicator for if company j was based in a state with a data breach notification law. The vector $\boldsymbol{X_{jt}}$ represents control variables and characteristics of firms, including the number of employees, the age of the firm, and the legal ownership structure. The dependent variable is the technology age of the web server used on the homepage of the company's website in a given month. Finally, ϵ represents the error term. I estimate this regression by ordinary least squares using the data in the Non-Tradable Firms dataset. As panel data is likely to have serial correlation, I cluster the standard errors by the state in which a firm is headquartered.[26] The coefficient of most interest is δ, which estimates the average effect of the data breach notification law on the technology age of firms' web server software.

The difference-in-differences method relies on the assumption that the primary driver of differences in the technology age of the server software used by firms operating inside and outside of California were caused by the implementation of the data breach laws. The empirical approach take in this paper removes the confounding impact of contemporaneous shocks impacting the secular trend in the technology age of software used. For example, when analyzing the impact of the California law, simply comparing the average technology age before and after the enactment of this law could misrepresent the actual causal effect because of events effecting the adoption of web server technology more broadly. In particular, web server vendors regularly released new versions of their software during this time period. Just comparing the technology ages of California firms' software could be confounded by the releases of new versions of software. Instead, by using companies operating outside of California as a control group, I can remove the bias from such

[26] In online supplemental material, I use a variety of other methods for estimating standard errors based on (Bertrand et al., 2004) to show the robustness of the results. In addition, because the technology age variable is bounded below in each time period by what software had been previously released and generally available, I also estimate the above model using a Tobit regression. Because Tobit models do not have an easy means for being estimated with fixed consistent fixed effects and serial correlation, I have included those estimates in the Appendix. Few firms, however, adopted technology immediately after its released, therefore, the lower-bound censoring is unlikely to be impactful.

contemporaneous events.

In order for the estimates from this difference-in-differences to be consistent, the timing of the laws being enacted cannot be correlated with investment in server technology within the state. The idiosyncratic reasons for the passage of the California law appears to provide this exogeneity. The California Senate passed their data breach notification relatively quickly following revelations of a past security incident. After the data breach notification law had been introduced in the state legislature, a California data center with sensitive information about state officials, including Senate legislators, was hacked. The data center's managers did not disclose the data breach to those affected for two months (Gaither, 2003; Richmond, 2003). Once the data center disclosed that hacking incidents had occurred and yet the data center officials had not notified affected legislators and state employees, the California Senate quickly passed the bill unanimously.[27]

For the estimates to accurately identify the causal effect of the California law, I require that the enactment of the California legislation did not impact firms outside of California. Since the California law specifically applied to firms that operated or collected personal data from California residents, firms that operated outside of California in sectors that are unlikely to have customers from other states would meet this criteria. I, therefore, investigate the impact of the California legislation using the Non-tradable firm data, which includes data on firms with geographically localized customers.

My assumption that the California and non-California based firms in the Non-tradable firms dataset constitute valid comparison groups for each other can be spot-checked by comparing the trends in the outcome variable, $TechAge$, during the period prior to the enactment of the California law. Figure 3 shows the similarity in the pre-trends between these groups. In addition, as open source web server technologies, such as Apache, are freely distributed online there is no reason why geographical boundaries should influence the adoption of particular technologies.

One limitation of the difference-in-differences methodology is if a spillover effect occurred in which firms outside of a state implementing a data breach notification law responded to the legal changes in the other states. While this is a possibility, particularly with the original 2002 California breach law, the companies in the Non-tradable Firms datasets are less likely to be impacted by this spillover effect. Because these firms have geographically concentrated consumers, firms operating outside of California would be unlikely to store personal data from California residents and thus to respond to a law targeting firms related to doing business within California. If spillovers do exist in this empirical setup then the estimated effect of the data breach notification law should be viewed as a lower-bound on the law's full impact.

While I choose to use the above described firm-level fixed effects regressions, a commonly used means for estimating the effect of state-level legal changes is developing a synthetic control for the treatment group

[27] The bill passed 78 in the affirmative, 0 in the negative, and 2 legislators did not have a vote recorded. https://leginfo.legislature.ca.gov/faces/billVotesClient.xhtml?bill_id=200120020SB1386

as described in Abadie and Gardeazabal (2003). In the synthetic control framework, the set of non-treated observations are weighted such that their attributes in the pre-treatment period closely resemble those for the treated group during the same time period. One of the limitations of this empirical method is that designing the synthetic control is computationally intensive, and thus this method is not easily applied to panel dataset with very large numbers of non-treated groups (Abadie and Gardeazabal, 2003; Abadie et al., 2010). As the panel dataset used in this paper contains over 200,000 firms, constructing synthetic controls at the firm level is computationally infeasible. Instead, I aggregate observations in the Non-tradable Firm dataset to the state-month level and estimate a synthetic control regression. The results of this estimation procedure are shown in Appendix C.

5 Results

5.1 Server Technology Age Predicts Hacking Incidents

In this section, I show that the technology age of web server software predicts hacking incidents. Table 3 shows the estimated coefficients from a logistic regression. An observation in this dataset is a company website in a year between 2005 and 2018. Only companies that used the Apache web server are used in the regression. The dependent variable is an indicator for the website being hacked within that year. The covariates in the regression include the technology age of the server software, the number of employees of the firm, indicators for the industry that the firm operates in, and fixed effects for each year. Standard errors are heteroskedasticity robust and clustered at the firm level.

The estimated coefficient on the logarithm of the technology age of the server software is 0.17. This coefficient implies that using web server software that is one month older is associated with an expected increase of 18.53% in the odds of being successfully hacked in that year after controlling for the year, the industry, number of employees, and the level of web traffic to the website of that firm. The coefficient on the indicator for the firm's website being a high-traffic site is 3.05. This implies that high traffic sites are significantly more likely to be the victim of a successful hacking attack than lower traffic sites.

Figure 2 graphically depicts an example of the predicted probability that a company website is hacked. In that figure, the vertical axis shows the probability that a company website is hacked in a year. The horizontal axis shows the technology age of the web server software. I show the predicted probability for companies operating in the Health industry (NAICS codes beginning with 62) in the year 2017 with a high level of website traffic and the average number of employees for companies in this industry (353 employees). The probability goes from around 0.010 for companies using server software that was released last month to 0.014 for those using servers with software that is one years old. While the baseline probability of a successful hacking incident in a year is low, an increase of this magnitude in the technology of server software is equivalent to a 40% higher predicted probability of a successful hacking incident occurring in the year.

5.2 California Data Breach Notification Law

The main results for the analysis of the effect of the 2002 California Data Breach Notification Law are presented in Table 4. This table displays the results of estimating fixed effects OLS model presented in Equation 2. Column (1) shows the estimates based on monthly observations of firms using Apache between 2001 and 2005 where the server version number was publicly visible. Column (2) shows estimates based on monthly observations of firms using Apache where the technology age of observations in which hidden server versions are interpolated to be based on the most recent server version available at that time. Column (3) shows similar estimates to Column (2), however, missing technology ages are interpolated to be the average technology age of the observations with visible server versions in that month.

The coefficient on data breach notification law being enacted in Column (1) is -0.330. This coefficient can be interpreted as the as the enactment of the data breach notification law in California caused firms to use server technology that was 0.330 months newer than they would have otherwise. Relative to the mean technology age of servers used during this time period, 19.42 months, the effect is an approximately 1.70% reduction. Similar estimates are found using the interpolated technology ages for missing observations. The coefficient on the law being enacted in Column (2) is -0.485, which relative to a mean 17.75 implies a 2.73% reduction, and the coefficient in Column (3) is -0.276 implying a 1.87% decrease.

These three coefficients provide relatively tight bounds on the impact of the data breach notification law. If all of the firms which switched off the display of the server version on their web servers did so because they were particularly security conscientious, we would estimate the impact of the law as lowering the average technology age by 2.73%. If, however, the firms that turned off the display of their server versions did so because they maintained older server software, we would estimate the impact at 1.87%. This small range of possibilities shows that the California data breach notification law had a small, but detectable impact on the server technologies utilized by firms at this time.

In Table 5 I explore the heterogeneity in the effect of the law across types of firms. The first three columns in that table show estimates of Equation 2 in which the implementation of a data breach notification law is interacted with the firm having more than 250 employees. These coefficients range from -0.803 to -1.041, which indicates that the California law instigated firms to use technology approximately 6.63% to 7.86% newer than they would have otherwise. The latter three columns in the table include a term interacting the California law with the firm having a high traffic website as measured by Alexa Internet's ranking of the most popular websites. These estimates range between -0.067 and -0.180. Relative to the base effect, these coefficients imply that firms with high traffic websites responded more to the California law than those with low traffic websites.

In Table 6, I show the results from running the same regression on Non-tradable firms, but examining only firms who used the Microsoft IIS server. In this regression, the coefficient on the law being enacted show an economically and statistically insignificant difference. The magnitude of this effect cannot be directly compared with the one estimated for Apache because Microsoft IIS. Whereas Apache server headers reveal the exact minor software version running on the server, the default IIS headers only show the major version number. Therefore, the set of possibility technology ages is discretized and reflects major version switches rather than the installation of minor software updates. More work will need to be done to interpret and understand Microsoft IIS users behaviors.

Overall, the results found align with the intuition that firms with more to lose by having a data breach publicized would be more responsive to the data breach notification law. The results also indicate that a firms response is more strongly correlated with the size and resources of the company rather than popularity of the associated website.

6 Conclusion

Given the gravity of data security, many laws have tried to incentivize firms to keep their digital infrastructure secure. California's 2002 data breach notification law was one of the first policies to do this and many other states have since followed suit.

Using comparisons of companies operating within California versus those based in other states, I found that the law did encourage California firms to keep their servers more secure by updating to newer versions of their server software. For companies using the open source server Apache, these gains translated to keep their servers 1.7%-2.7% more up-to-date software. Larger firms and firms with popular websites responded to the new law by decreasing the technological age of their web servers by 2%-7.8%. Ultimately, the California data breach law had a modest impact on encouraging firms to keep their digital infrastructure more up-to-date and secure.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller, "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, June 2010, *105* (490), 493–505.

_ and Javier Gardeazabal, "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, February 2003, *93* (1), 113–132.

Arora, Ashish, Chris Forman, Anand Nandkumar, and Rahul Telang, "Competition and Patching of Security Vulnerabilities: An Empirical Analysis," *Information Economics and Policy*, May 2010, *22* (2), 164–177.

_ , Ramayya Krishnan, Rahul Telang, and Yubao Yang, "An Empirical Analysis of Software Vendors' Patch Release Behavior: Impact of Vulnerability Disclosure," *Information Systems Research*, 2010, *21* (1), 115–132.

Beccaria, Cesare, *On Crimes and Punishment*, New York: Macmillan, 1963.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?," *The Quarterly Journal of Economics*, 2004, (1), 249.

Chow, Stephen Y., *Data Security and Privacy in Massachusetts*, Boston, MA: Massachusetts Continuing Legal Education, Inc., November 2015.

Coombs, Chad C. and Keenen Milner, "Practice Tips: New California Identity Theft Legislation," *Los Angeles Lawyer*, July/August 2004, *27* (21).

D'Arcy, John, Anat Hovav, and Dennis Galletta, "User Awareness of Security Countermeasures and Its Impact on Information Systems Misuse: A Deterrence Approach," *Information Systems Research*, 2009, *20* (1).

_ and Tejaswini Herath, "A Review and Analysis of Deterrence Theory in the IS Security Literature: Making Sense of the Disparate Findings," 2011, *20* (6), 643–658.

Fung, Brian, "Equifax's Massive 2017 Data Breach Keeps Getting Worse," March 2018.

Gaither, Chris, "California Law Requires That Firms Reveal Data-Security Breaches," *Knight Ridder Tribune Business News; Washington*, June 2003, p. 1.

Gatzlaff, Kevin M and Kathleen A. McCullough, "The Effect of Data Breaches on Shareholder Wealth," *Risk Management and Insurance Review*, 2010, *13* (1), 61–83.

_ and _ , "The Effect of Data Breaches on Shareholder Wealth," *Risk Management and Insurance Review*, 2010, *13* (1), 61–83.

Gavejian, Jason C. and Jackson Lewis, “New Mexico Enacts Data Breach Notification Act,” https://www.shrm.org/resourcesandtools/legal-and-compliance/state-and-local-updates/pages/new-mexico-enacts-data-breach-notification-act.aspx April 2017.

Goel, Sanjay and Hany A Shawky, “The Impact of Federal and State Notification Laws on Security Breach Announcements,” *Communications of the Association for Information Systems*, January 2014, *34*, 16.

Greenstein, S and Nagle F Policy, “Digital Dark Matter and the Economic Contribution of Apache,” *Research Policy*, 2014.

Harrington, Susan J., “The Effect of Codes of Ethics and Personal Denial of Responsibility on Computer Abuse Judgments and Intentions,” *MIS Quarterly*, 1996, *20* (3), 257–278.

Herath, Tejaswini and & H Raghav Rao, “Protection Motivation and Deterrence: A Framework for Security Policy Compliance in Organisations,” *European Journal of Information Systems*, 2009, *18* (2), 106–125.

_ and H R Rao, “Encouraging Information Security Behaviors in Organizations: Role of Penalties, Pressures and Perceived Effectiveness,” 2009.

Hoofnagle, Chris Jay, “Identity Theft: Making the Known Unknowns Known,” *Harvard Journal of Law and Technology*, 2007, *21*, 26.

Hunton Andrews Kurth LLP, “Kentucky Enacts Data Breach Notification Law,” https://www.huntonprivacyblog.com/2014/04/17/kentucky-enacts-data-breach-notification-law/ April 2014.

_ , “Alabama Becomes Final State to Enact Data Breach Notification Law,” https://www.huntonprivacyblog.com/2018/04/03/alabama-becomes-final-state-enact-data-breach-notification-law/ April 2018.

_ , “South Dakota Enacts Breach Notification Law,” https://www.huntonprivacyblog.com/2018/03/23/south-dakota-enacts-breach-notification-law/ March 2018.

IBM, “Cost of Data Breach Study — IBM Security,” 2018.

Jacobs, Bruce A., “Deterrence and Deterrability*,” *Criminology*, 2010, *48* (2), 417–441.

_ , “Deterrence and Deterrability*,” *Criminology*, 2010, *48* (2), 417–441.

Javelin Strategy, “2017 Identity Fraud: Securing the Connected Life,” Technical Report February 2017.

Jr., Detmar W. Straub, “Effective IS Security: An Empirical Study,” *Information Systems Research1*, 1990, *1* (3), 255–276.

Kamiya, Shinichi, Jun-Koo Kang, Jungmin Kim, Andreas Milidonis, and René M. Stulz, “What Is the Impact of Successful Cyberattacks on Target Firms?,” March 2018.

Mian, Atif and Amir Sufi, "What Explains the 2007-2009 Drop in Employment?: The 2007-2009 Drop in Employment," *Econometrica*, November 2014, *82* (6), 2197–2223.

Miller, Amalia R. and Catherine E. Tucker, "Encryption and the Loss of Patient Data," *Journal of Policy Analysis and Management*, June 2011, *30* (3), 534–556.

Needles, Sara A., "The Data Game: Learning to Love the State-Based Approach to Data Breach Notification Law," *North Carolina Law Review*, 2009, *88*, 267—308.

—, "The Data Game: Learning to Love the State-Based Approach to Data Breach Notification Law," *North Carolina Law Review*, 2009, *88*, 267–308.

Perkins Coie LLP, "Security Breach Notification Chart - California," https://www.perkinscoie.com/en/news-insights/security-breach-notification-chart-california.html April 2018.

Picanso, Kathryn E, "Protecting Information Security Under a Uniform Data Breach Notification Law," *Fordham Law Review*, 2006, *75* (1), 37.

Pogarsky, Greg, Alex R. Piquero, and Ray Paternoster, "Modeling Change in Perceptions about Sanction Threats: The Neglected Linkage in Deterrence Theory," *Journal of Quantitative Criminology*, December 2004, *20* (4), 343–369.

— and Alex R Piquero, "Studying the Reach of Deterrence: Can Deterrence Theory Help Explain Police Misconduct?," *Journal of Criminal Justice*, July 2004, *32* (4), 371–386.

Richmond, Riva, "E-Commerce (A Special Report) — Hacker Alert: California Will Soon Require Companies to Disclose When the State's Residents Are at Risk of Identity Theft," *Wall Street Journal, Eastern edition; New York, N.Y.*, June 2003, p. R.9.

Romanosky, Sasha, Rahul Telang, and Alessandro Acquisti, "Do Data Breach Disclosure Laws Reduce Identity Theft?," *Journal of Policy Analysis and Management*, 2011, *30* (2), 256–286.

Sabett, Randy, "Ten Years Later: The Legacy of SB 1386 Compliance on Data Privacy Laws," https://searchsecurity.techtarget.com/opinion/Ten-years-later-The-legacy-of-SB-1386-compliance-on-data-privacy-laws August 2013.

Schwartz, Paul M and Edward J Janger, "Notification of Data Security Breaches," *Source: Michigan Law Review*, 2007, *105* (5), 913–984.

Sells, Berkley D., "California Hacker Disclosure Law Will Have Wide Repercussions," *The Lawyers Weekly*, August 2003, *23* (14).

Simpson, Sally S., Nicole Leeper Piquero, and Raymond Paternoster, "Rationality and Corporate Offending Decisions," in Alex R. Piquero, ed., *Rational Choice and Criminal Behavior: Recent Research and Future Challenges*, Routledge, October 2012.

Siponen, Mikko and Anthony Vance, “Neutralization: New Insights into the Problem of Employee Information Systems Security Policy Violations,” *MIS Quarterly*, 2010, *34* (3), 487–502.

Straub, Detmar W. and Richard J. Welke, “Coping with Systems Risk: Security Planning Models for Management Decision Making,” *MIS Quarterly*, 1998, *22* (4), 441–469.

Sullivan, Richard J and Jesse Leigh Maniff, “Data Breach Notification Laws,” *Economic Review*, 2016, (First Quarter), 65—85.

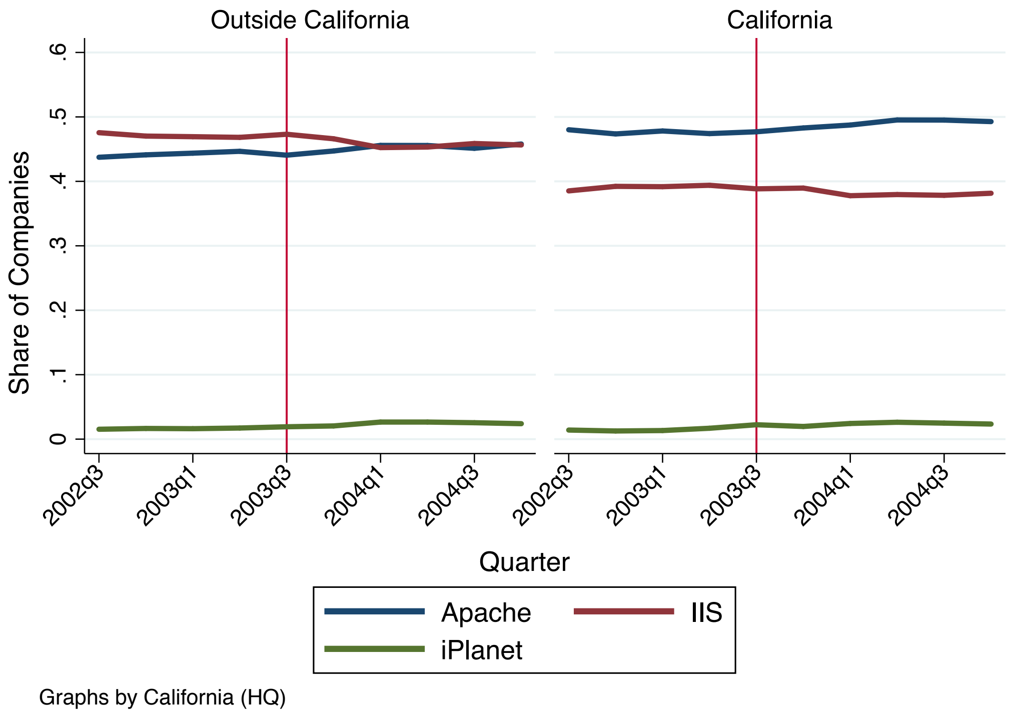
Theoharidou, Marianthi, Spyros Kokolakis, Maria Karyda, and Evangelos Kiountouzis, “The Insider Threat to Information Systems and the Effectiveness of ISO17799,” 2005, *24.*

Wang, Tawei, Karthik N. Kannan, and Jackie Rees Ulmer, “The Association Between the Disclosure and the Realization of Information Security Risk Factors,” *Information Systems Research*, June 2013, *24* (2), 201–218.

Wugmeister, Miriam H. and Christine E. Lyon, eds, *Global Employee Privacy and Data Security Law, Second Edition*, second ed., Morrison & Foerster LLP, 2011.

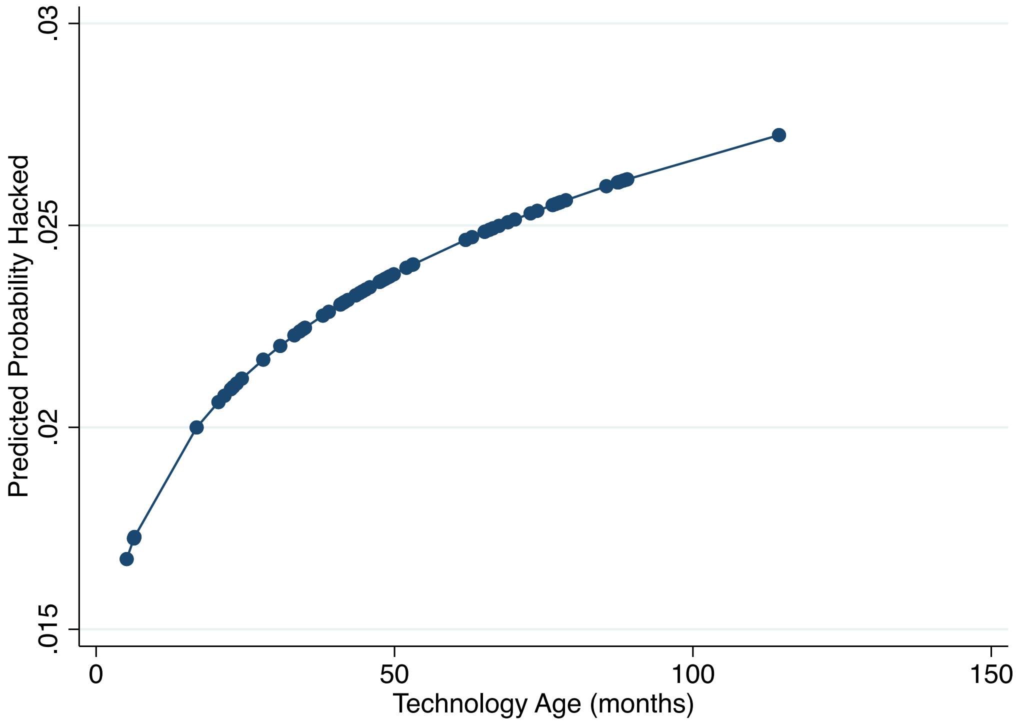
7 Tables and Figures

Figure 1: Market Shares of Servers by Quarter



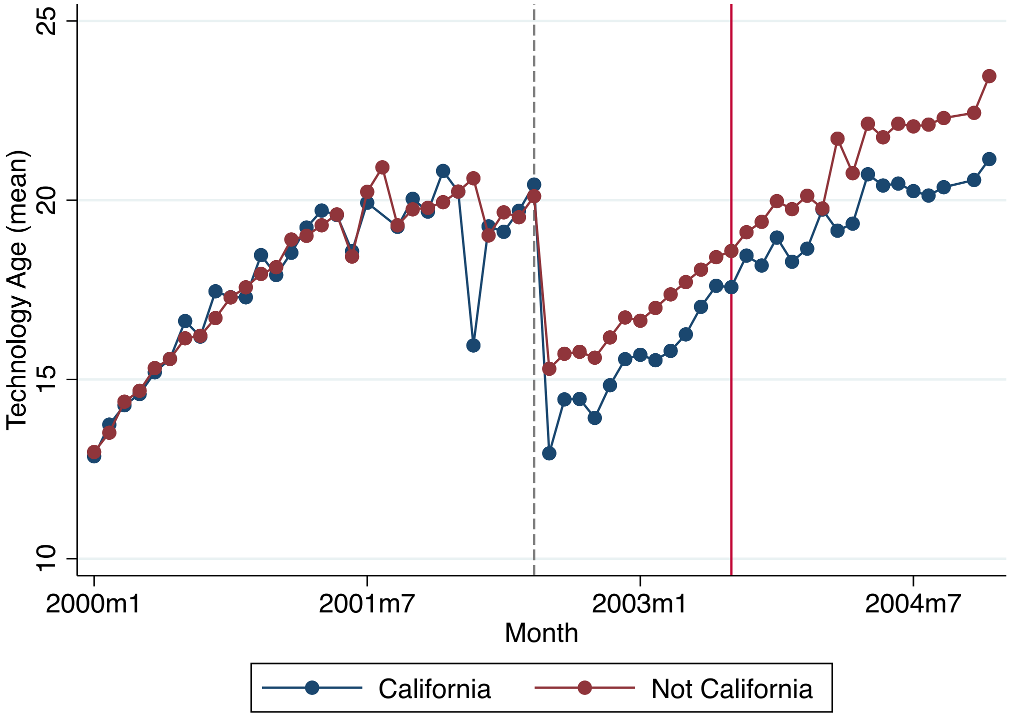
Note: The above plot shows the market share of two types of servers, Apache and IIS, for companies in the dataset. These shares are based on which server is used for serving the homepage of the company's website. A vertical line is shown for the month in which the California data breach notification law was enacted.

Figure 2: Predicted Probability of Hacking



Note: The above figure shows the predicted probability that a company's website in a year is hacked. This predicted probability is based on a logistic regression in which the dependent variable is an indicator for the site being hacked. The covariates are the log of the technology age of the server software, the year, the industry of the company, the log of the number of employees, and an indicator for the site having a high amount of traffic. The above plot shows the predicted probability for companies in the Health industry (NAICS codes starting with 62) in the year 2017 with high traffic and the average number of employees of firms operating in this industry.

Figure 3: Average Technology Age



Note: The above plot shows the average technology age of the server technologies used on the homepages of companies located in and outside of California. Technology Age is defined as the difference in the number of months between when the server is being used and when that version of the server was released. A vertical line is shown for the month in which the California data breach notification law was enacted.

Table 1: Summary Statistics of Companies

	July 2002		
	Not California	California	t
Employees (k)	372.73	219.33	5.03
– Employees: [50,100)	0.48	0.53	-11.28
– Employees: [100,500)	0.44	0.40	8.53
– Employees: [500,1000)	0.05	0.05	0.41
– Employees: [1000,)	0.03	0.02	7.88
Firm Age (Years)	302.53	256.70	7.66
Industry: Manufact.	0.32	0.35	-8.05
Industry: Accom.	0.21	0.27	-15.96
Industry: Health	0.47	0.38	20.80
Alexa Rank	409,300.30	440,527.81	-3.72
Technology Age	19.20	17.84	9.10
N	105,522		

Note: The above table shows mean attributes of companies in the dataset during two months. The top table shows the means during the month of July 2001, while the bottom table shows the means during the month of July 2004. An observation is company during that month. Alexa Rank is the ranking of the company's homepage according to Alexa Internet's estimates of their web traffic in 2010. R&D Investment is the total amount spent on Research & Development expenses in the year reported in millions of U.S. dollars. Firm age is defined as the years since the company was founded. Technology Age is the number of months between the month of the table and the month in which web server version used by that company was first released.

Table 2: Attributes of Hacked Companies

	Not Hacked	Hacked	Difference	t-stat	p-value
Tech. Age	36	40	-4.6	-3.75	0.000
Alexa Rank (2010, censored)	955,119	540,754	414,365	48.24	0.000
Alexa Rank (2010, truncated)	427,206	147,686	279,520	13.94	0.000
Employees	374	973	-600	-3.49	0.000
Firm Age	410	561	-151	-4.09	0.000
NAICS: Manufacturing	.092	.04	.052	3.72	0.000
NAICS: Wholesale	.06	.026	.034	2.98	0.003
NAICS: Retail	.053	.052	.0008	0.07	0.941
NAICS: Education	.051	.15	-.1	-9.29	0.000
NAICS: Health	.11	.18	-.071	-4.70	0.000
NAICS: Other	.64	.55	.083	3.55	0.000
State: CA	.12	.16	-.032	-2.03	0.042
State: NY	.064	.089	-.026	-2.15	0.032
State: MA	.028	.033	-.0053	-0.67	0.502
State: CT	.013	.014	-.0014	-0.26	0.792
State: FL	.058	.042	.016	1.41	0.160
State: Other	.71	.67	.049	2.22	0.026
N	172,780				

Note: The above table shows mean attributes of companies that had their website hacked versus those that did not. The data used for this table comes from the 2005-2018 dataset. It includes all firms that used the Apache web server between 2005 and 2018. A company is said to have been hacked if at any point between 2005 and 2018 that firm is listed in the Privacy Rights Clearinghouse Database as having been hacked. The NAICS variables are indicators for if the company is associated with that NAICS. The State variable is an indicator if the firm is located in that state. The average Alex Rank of web traffic is shown in two forms. The truncated version shows the average using only firms who ranked in the top one million most trafficked websites. The ranked version shows all websites based on a censored ranking at one million.

Table 3: Predicting Hacked Companies

	(1) Hacked
Log(Tech. Age Interpolated)	0.17^{**}
	(0.085)
Log(Employees)	0.05
	(0.051)
High Traffic	3.05^{***}
	(0.128)
Year FE	Yes
NAICS FE	Yes
N	1,159,536
N Domains	172,613
N Domains Hacked	423
Dep. Mean	0.00039671
Psuedo R^2	0.14

Robust standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The above table shows the estimated coefficients from a logistic regression predicting if a company's website is hacked in a given year. An observation in the data is a company homepage in a year. The observations cover from 2005 through 2018. Only firms using the Apache web server are used. The logarithm of the technology age of the server software is used. When the version of the server is hidden, I interpolate the average technology age based on firms with visible version numbers. Standard errors are clustered at the firm level and robust to heteroskedasticity.

Table 4: Impact of the 2002 California Law on Server Technology Age for Apache Users

	Tech Age		
	(1)	(2)	(3)
	Visible	Interpolated: Min	Interpolated: Avg.
Law Enacted	-0.330**	-0.485***	-0.363***
	(0.128)	(0.134)	(0.123)
Constant	13.279***	12.738***	13.318***
	(0.203)	(0.187)	(0.181)
Firm FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
N	187,145	206,848	206,848
N Domains	14,259	15,087	15,087
Dep. Mean	19.42	17.75	19.42

Standard errors clustered at state level in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The above table shows the OLS estimates of Equation 2. An observation in this regression is a company-month during the years 2000-2005 among firm websites using the Apache server software. The Law Enacted indicator represents if the observation occurred after the enactment of California S.B. 1386 and the company represented by the observation is headquartered in California. The first column shows the results using only observations from the panel in which the server version is available. The second column shows the estimated coefficients when missing version numbers are interpolated with the smallest technology age based on versions available in that month, and the third column shows the estimates when missing version numbers are interpolated with the average technology age based on observations with server version numbers in that month.

Table 5: Heterogeneity in the Impact of the 2002 California Law on Server Technology Age for Apache Users

	Tech Age					
	(1)	(2)	(3)	(4)	(5)	(6)
	Visible	Interpolated: Min	Interpolated: Avg.	Visible	Interpolated: Min	Interpolated: Avg.
Law Enacted	-0.485***	-0.633***	-0.486***	-0.320**	-0.459***	-0.341***
	(0.128)	(0.134)	(0.123)	(0.127)	(0.134)	(0.123)
Law Enacted x 250+ Employees	-1.041***	-0.970***	-0.803***			
	(0.003)	(0.003)	(0.003)			
Law Enacted x High-Traffic				-0.067***	-0.180***	-0.152***
				(0.006)	(0.004)	(0.004)
Constant	13.279***	12.738***	13.318***	13.278***	12.737***	13.317***
	(0.203)	(0.187)	(0.181)	(0.203)	(0.187)	(0.182)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
N	187,145	206,848	206,848	187,145	206,848	206,848
N Domains	14,259	15,087	15,087	14,259	15,087	15,087
Dep. Mean	19.42	17.75	19.42	19.42	17.75	19.42

Standard errors clustered at state level in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The above table shows the OLS estimates of Equation 2. An observation in this regression is a company-month during the years 2000-2005 among firm websites using the Apache server software. The Law Enacted indicator represents if the observation occurred after the enactment of California S.B. 1386 and the company represented by the observation is headquartered in California. The first column shows the results using only observations from the panel in which the server version is available. The second column shows the estimated coefficients when missing version numbers are interpolated with the smallest technology age based on versions available in that month, and the third column shows the estimates when missing version numbers are interpolated with the average technology age based on observations with server version numbers in that month.

Table 6: Impact of the 2002 California Law on Server Technology Age for Microsoft IIS Users

	Tech Age
Law Enacted	-0.111
	(0.126)
Constant	35.732***
	(0.178)
Firm FE	Yes
Month FE	Yes
N	208,374
N Domains	13,929
Dep. Mean	45.41

Standard errors clustered at state level in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The above table shows the OLS estimates of Equation 2. An observation in this regression is a company-month during the years 2000-2005 among firm websites using the Microsoft IIS server software. The Law Enacted indicator represents if the observation occurred after the enactment of California S.B. 1386 and the company represented by the observation is headquartered in California. The results use only observations from the panel in which the server version is available.

Appendix

A Data Construction

The data samples are constructed through the following process:

First, I collected the records of firms in the Bureau van Dyke Orbis Database that are located in the United States, have at least 50 employees, and have a listed website. This came to 271,579 firms. For each firm, I collected the state they operated in as well as the industry classification of their operations in the form of a North American Industry Classification System (NAICS) code. From the website addresses for the firms, I extracted the domain of the website.

Second, I took the unique domain for the firm websites (typically the homepage of the firm) and searched the Internet Archive's Wayback Machine for times when the Internet Archive had collected server headers from that website. For each website, I collected one set of server headers per month when those headers had been collected. If the Internet Archive had collected more than one set of headers, I only used the first in the month.

Third, I processed the extracted and processed the server header information. In particular, I examined the "Server" server header to see if it matched the patterns of those associated with the Apache, Microsoft IIS, and iPlanet server software. In addition, I extracted any available information about the version of server software used. For example, the server header "Apache/2.4.2 (Unix) PHP/4.2.2 MyMod/1.2" would be parsed as a server vendor of "Apache" and software version "2.4.2". I also ran the full set of servers headers through the open source computer program Wappalyzer to validate my categorization of the servers. This culminates in being able to label each of the server headers with a server vendor and version number.

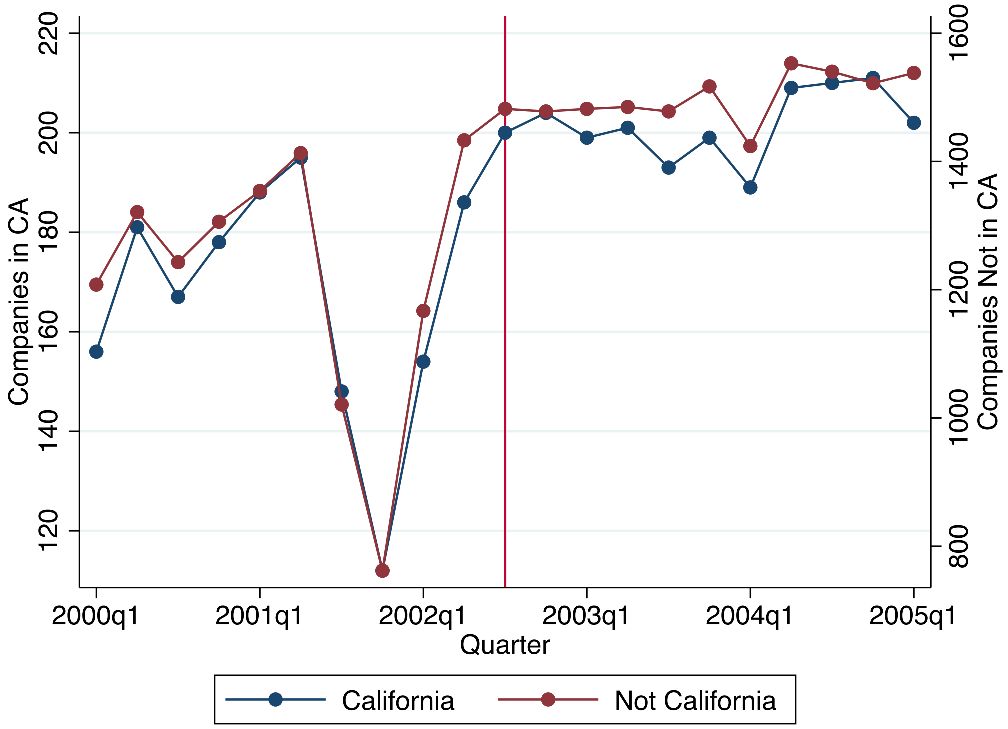
Fourth, the parsed server vendor and version numbers are merged back onto the firm data. In doing so, I create a panel of firms with observations in months when the Internet Archive collected data for their associated website homepage containing information on the type and version of the server in use.

For flagging which firms experienced a data breach, I match the firm observations with the Privacy Rights Clearinghouse database of data breach events. Specifically, I search for only recorded hacking and malware breaches. I then match these breaches to the observations in the panel dataset based on the names of the company. This matching is done by hand.

Figure 4 shows the number of company websites with server information in each month. A website could be missing either because the company did not exist at that time, because the Internet Archive did not attempt to contact that server that month, or because because the Internet Archive could not connect to the server that month. Noticeably, the Dot-Com bust caused a significant drop in the number of companies in the dataset between the end of 2001 and early 2002. This appears to be largely because the Internet Archive slowed their rate of crawling sites and focused on crawling only higher traffic websites during that period.[28]

[28] In Figure **??**, I show the average number of snapshots per year among the balanced panel companies. This shows that the number of snapshots for these firms went down only slightly during this time period.

Figure 4: Number of Active Companies per Quarter



Note: An active company is a Compustat company with a website with header information. In the above figure, I show the number of active companies per quarter. The blue line shows the number of companies located in California, while the red line shows the companies not located in California.

A.1 Server Header Inference Validation

Changing the displayed server headers for a website is relatively easy for a server administrator to do. Indeed, a number of firms turn off the display of server headers altogether. Therefore, I attempt to validate that the server vendor types that can be inferred from the server headers for their accuracy.

The first method for validating the information parsed from the server headers involves comparing the server vendor indicated from the headers with the server vendor indicated from actual content from the website and sent to the website visitors. In particular, many websites use "cookies" or small files set on website visitors' computer to save and store information. Cookies are particularly useful for storing a unique ID number for website visitors in order to track the visitors' actions over time. Websites that are built using web development frameworks name their cookies according to predictable and differentiable patterns.[29] In particular, I find that of the websites that set cookies consistent with using the ASP.Net technology—a web framework associated with the Microsoft IIS web server—99.18% of the server headers indicate that the Microsoft IIS server is being used to host the website.[30] The high correlation between the server indicated

[29] Cookies, like server headers, can be named in ways that obfuscate the framework that set them.

[30] ASP.Net can be used on servers other than Microsoft IIS although not all features are available. Therefore, while possible, it is

by the cookies and the server headers is consistent with server headers accurately displaying when IIS is being used.

The second method is to compare the types of technologies used by firms as listed in survey data with the technologies indicated from the server headers. When the server headers indicate that a firm is using Microsoft's IIS server, I check if that firm responded to the Harte-Hankes technology survey that they use Microsoft technologies. This work is still in progress.

A.2 Missing Server Version Numbers

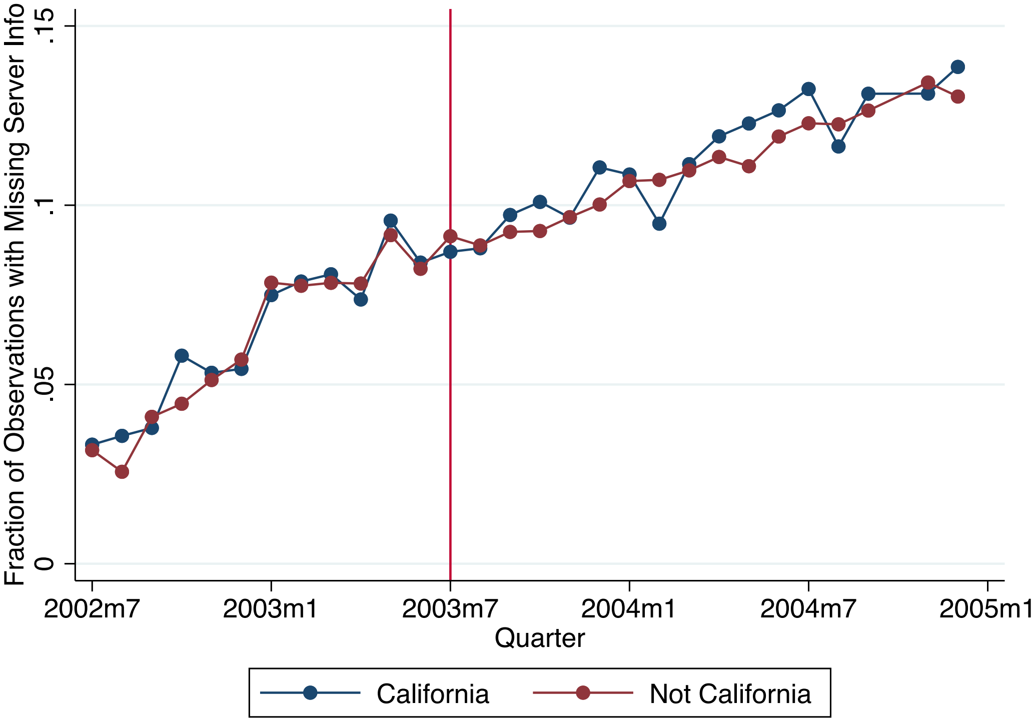
While most server software displays both the server vendor and software version number in the server headers, server administrators can hide this information from being displayed in the headers. During the time of the sample, an increasing number of firms hid the software version number from their headers. In Figure 5, I show the fraction of observations from the Non-tradable firm dataset in which the server headers displayed the vendor Apache but the version number is not displayed. The plot shows that the increasing propensity to hide the server version number is similar for both California and non-California based firms during these years.

The firms that choose to hide the server version number from their headers may be those who are more or less security conscientious. For example, server administrators who are diligent about security and wish to prevent hackers from easily targeting their servers might turn off the display of version numbers. Similarly, firms that run particularly old server software with known vulnerabilities might be more inclined to turn off the headers. In order to empirically figure out if the firms who turn off the display of server version numbers are more or less security conscientious, I examine the average technology age of server software used by firms during the time period before they turn off the display. I then compare the average technology age during the same time period with firms who keep their software version number displayed. This comparison is shown in Figure 5.

The average technology ages shown in Figure 5 indicate that firms that turn off their server version numbers are on average using server technology that was more recently released than firms that keep the display of the version numbers. Because of this, I define two additional variables to replace $TechAge$ for observations when the firm has turned off the display of software version numbers. In particular, I define one variable as the minimum $TechAge$ possible in a month given the available versions from the server software vendor. I define a second variable as the average $TechAge$ of the versions in use by the firms in a month who continue to display their server version numbers. I believe that the former interpolated variable is representative of the world in which firms that turn off their server version number are very security conscientious. I believe that the latter variable is representative of the world in which those who turn off the display of their server version numbers are no more or less security conscientious than those who do not turn off the headers.

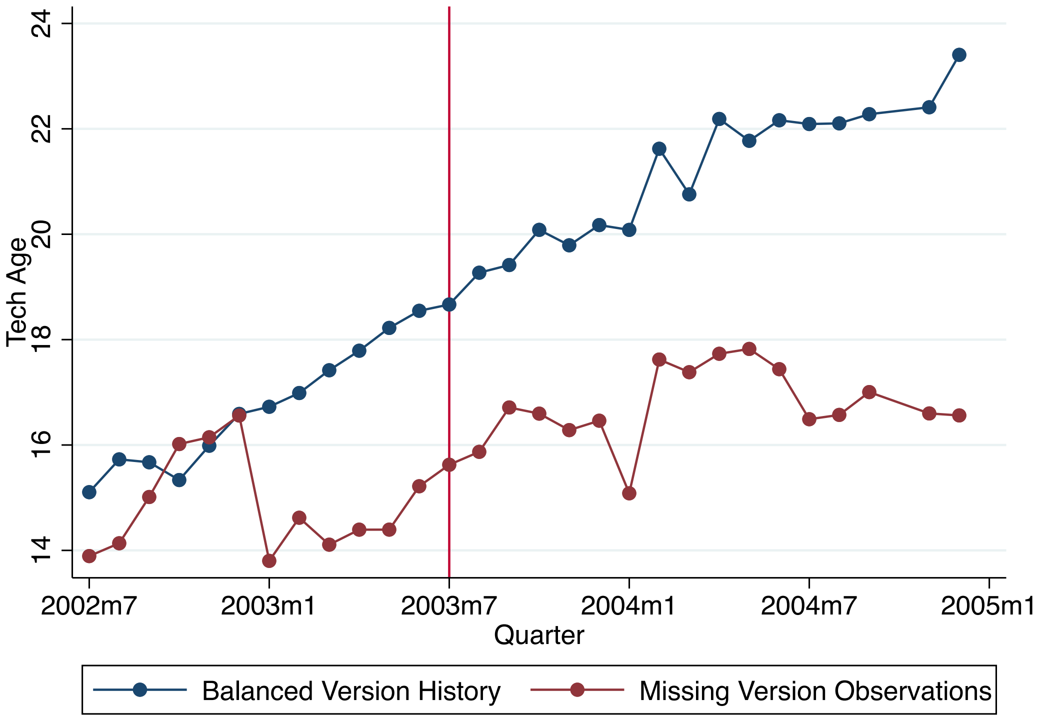
less common to use a non-Microsoft IIS server with ASP.Net based websites.

Figure 5: Number of Apache Users Hiding Firm Number



Note: The above graph shows the fraction of companies in the Non-tradables dataset where the firm's server headers showed the server software Apache but the version number was not displayed. Separate lines are plotted for California and non-California based firms.

Figure 6: Technology Age of Firms That Display or Hide Server Version Numbers

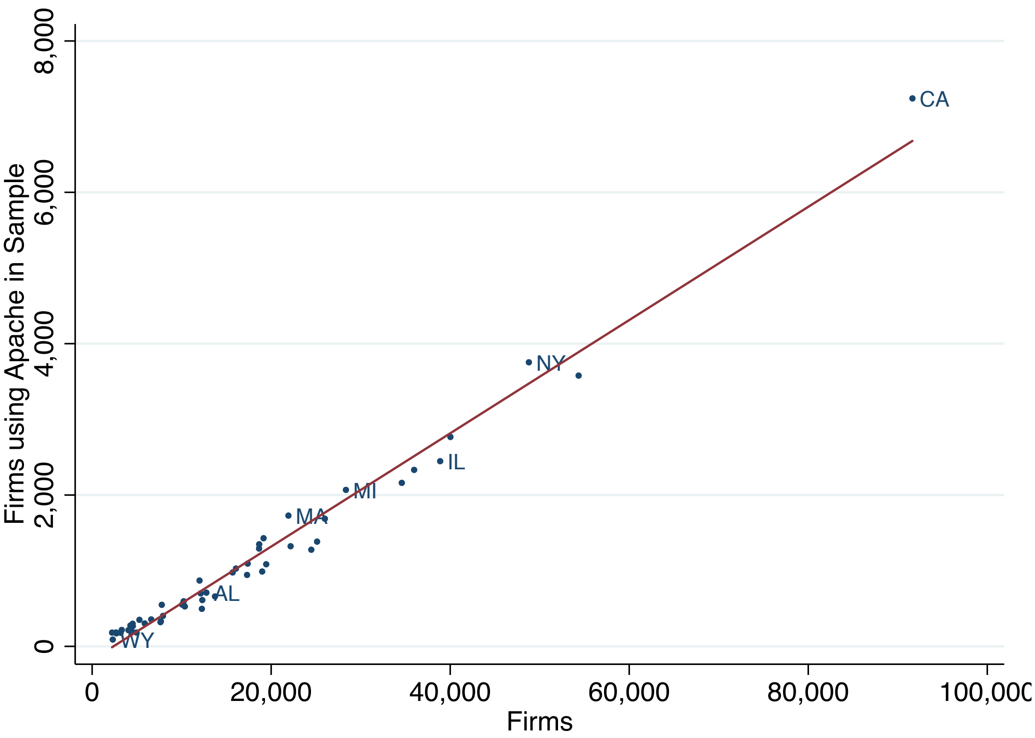


Note: The above graph shows the average technology age of observations from the Non-tradables dataset. One line shows the average technology age for firms that turn off the display of their software version numbers at some point during the sample. The other line shows the average technology age for firms that keep their software version numbers displayed throughout the sample. Note that some firms that turn off their server version number turn it back on in other months.

B Sample Representativeness

In order to assess the representativeness of the sample used in my analysis, I compare the number of firms as listed in the Census Bureau's Statistics of U.S. Business (SUSB) series with the number of firms in my sample. In Figure , I plot the number of firms with 20 or more employees in the 2003 SUSB by state on the horizontal axis and the number of distinct firms using the Apache web server at any point in 2003 in my sample on the vertical axis. I also plot a linear regression line through the unweighted data. The graph shows a positive upward slope with a correlation coefficient of 0.989.

Figure 7: Comparison of the Number of Firms in Sample versus 2003 SUSB by State



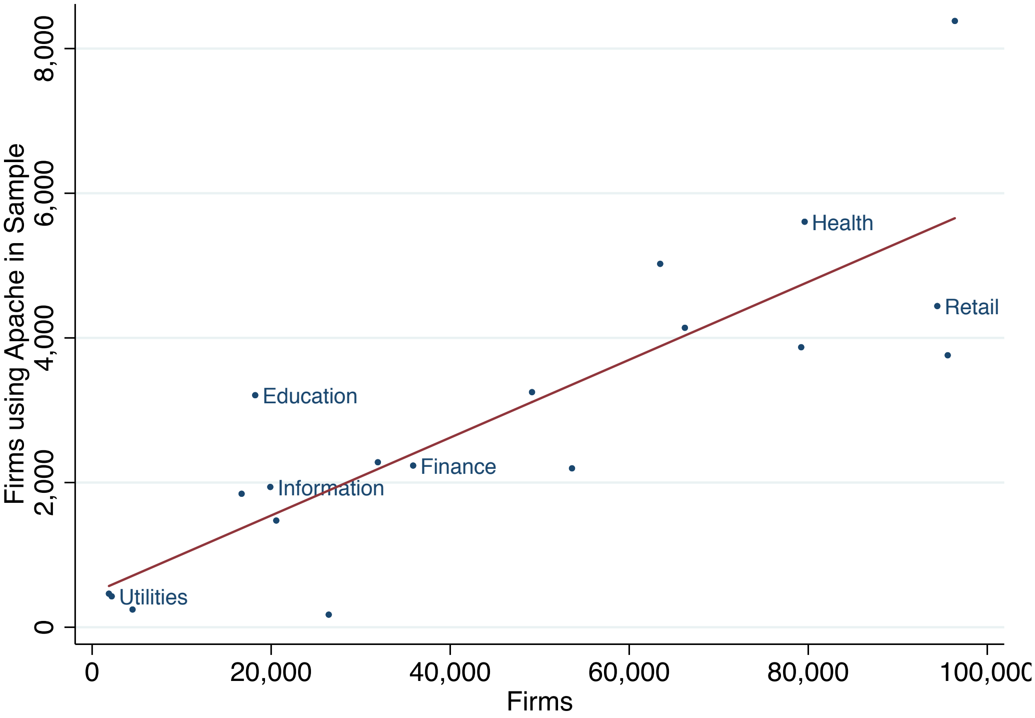
Note: The above plot shows the number of firms with at least 20 employees listed in the Census Bureau's Survey of U.S. Business for 2003 grouped by state on the horizontal axis. The vertical axis shows the number of distinct firms using the Apache web server during any month in 2003 in my sample. The points represent a state. The linear fit line is also included in the graph.

In levels, my sample appears to have around one-tenth of the number of firms as the SUSB reports. There are a number of reasons for the difference. First, the SUSB only enumerates the number of firms that are larger than 20 employees, while my sample is based on employers with 50 or more employees. Unfortunately, SUSB does not release finer bins for employer sizes for the years 2002-2005. Second, only a fraction of all employers during this time period had websites. Third, the Internet Archive only captured server headers from a subset of internet websites.

While the number of firms represented in my data may be lower than those that existed in the United States at the time, the strong positive correlation in the number of firms shows that the sample is useful for gaining insights into the larger population of firms.

In Figure , I plot the number of firms with 20 or more employees in the 2003 SUSB by two-digit NAICS code on the horizontal axis and the number of distinct firms using the Apache web server at any point in 2003 in my sample on the vertical axis. The graph shows a positive upward slope with a correlation coefficient of 0.843. While somewhat noisier than the state level graph, this figure shows that my sample has good coverage of many industries.

Figure 8: Comparison of the Number of Firms in Sample versus 2003 SUSB by NAICS



Note: The above plot shows the number of firms with at least 20 employees listed in the Census Bureau's Survey of U.S. Business for 2003 grouped by two digit NAICS on the horizontal axis. The vertical axis shows the number of distinct firms using the Apache web server during any month in 2003 in my sample. The points represent a two-digit NAICS code. NAICS codes 31-33 are combined, as are 44-45 and 48-49. The linear fit line is also included in the graph.

C Synthetic Control Estimates of the Effect of Data Breach Notification Laws

I follow the works of Abadie and Gardeazabal (2003) and Abadie et al. (2010) in estimating a synthetic control approach to investing the effect of the 2002 California data breach notification law. Beginning with the Non-tradable Firm dataset, I aggregate the observations to the state-year level. I do this by taking the average $TechAge$, firm size (number of employees), firm age, and industries over firms in a state in a year. For California, I record observations from 2004 onward as having a data breach notification law in place (as the law went into effect at July 2003).

The results of estimating this synthetic control framework are shown in Table 7. In that table, I list the estimated effect of the 2002 California data breach notification law for each month following the enactment. While the estimated effect in each month ranges, on average the effect appears to be small negative. The average of the estimated effects is -0.25 or about one-quarter of a month, however, they are not statistically significant. While the estimates show a similar pattern to the fixed effect firm-level regressions shown in the paper—that California based firms used Apache server software that was approximately one-quarter to one-third of a month newer than they would have if the law had not be in place—the synthetic control procedure at the state-level is not precise enough to be distinguishable from no effect.

Table 7: Synthetic Control Estimate of 2002 Data Breach Notification Law

	Technology Age
+1 Months	-0.289
+2 Months	1.402
+3 Months	-0.853
+4 Months	0.106
+5 Months	-0.566
+6 Months	-0.442
+7 Months	1.153
+8 Months	-0.477
+9 Months	-0.956
+10 Months	-0.504
+11 Months	-0.102
+12 Months	-1.022
+13 Months	-1.243
+14 Months	-0.644
+15 Months	-0.195
+16 Months	0.448
+17 Months	0.013

Note: The above table shows the treatment effect estimates based on a synthetic control for California. The data used for constructing the synthetic control starts with the Non-Tradable Firms dataset and aggregates to the state-month level by taking the average of all variables and observations. The synthetic control for California is constructed based on the average $TechAge$ of server software used in the period prior to the law going into effect, the average number of employees at firms, the average age of firms, and the percentage of firms in the healthcare sector within a state. The effects are then estimated and displayed for the months following the law's passage before January 2005.

D Dates of Enactment of State Data Breach Notification Laws

Information comes from Chow (2015, p. 9-13), Goel and Shawky (2014), Gavejian and Lewis (2017)and Hunton Andrews Kurth LLP (2018a), Hunton Andrews Kurth LLP (2014), Hunton Andrews Kurth LLP (2018b).

Table 8: Years When State Data Breach Notification Laws Were Enacted

State	Effective
California	7/1/03
Arkansas	3/31/05
Georgia	5/5/05
North Dakota	6/1/05
Delaware	6/28/05
Florida	7/1/05

Tennessee	7/1/05
Washington	7/24/05
Nevada	10/1/05
North Carolina	12/1/05
New York	12/7/05
Connecticut	1/1/06
Illinois	1/1/06
Louisiana	1/1/06
Minnesota	1/1/06
New Jersey	1/1/06
Maine	1/31/06
Ohio	2/17/06
Montana	3/1/06
Rhode Island	3/1/06
Wisconsin	3/31/06
Oklahoma	6/8/06
Pennsylvania	6/20/06
Idaho	7/1/06
Indiana	7/1/06
Kansas	7/1/06
Nebraska	7/14/06
Colorado	9/1/06
Arizona	12/31/06
Hawaii	1/1/07
New Hampshire	1/1/07
Utah	1/1/07
Vermont	1/1/07
District of Columbia	3/8/07
Michigan	6/29/07
Wyoming	7/1/07
Oregon	10/1/07
Massachusetts	10/31/07
Maryland	1/1/08
West Virginia	6/8/08
Iowa	7/1/08
Virginia	7/1/08
Texas	4/1/09
Alaska	7/1/09
South Carolina	7/1/09

Missouri	8/28/09
Mississippi	7/1/11
Kentucky	04/10/14
New Mexico	04/06/17
South Dakota	03/21/18
Alabama	03/28/18