Measuring the Economic Effects of Data Breaches on Firm Outcomes: Challenges and Opportunities

Christos A. Makridis and Benjamin Dean*

March 25, 2018

For Review

Abstract

We introduce a new dataset that links publicly reported data breaches and financial outcomes at the firm-level. First, we document three new facts about the incidence of data breaches: (i) heavy skewness in the distribution of the scale of data breaches, (ii) differences in data breaches by sector, and (iii) publicly traded companies exhibit much greater means and standard deviations of records breached. Second, while we find some evidence using cross-sectional variation and controlling for time-varying observable firm inputs that a 10% rise in breaches is associated with approximately a 0.2% decline in firm productivity, the result is sensitive to different specifications and datasets. Third, we show that the absence of more reliable estimates is driven by non-classical measurement error arising from sample selection problems in publicly reported breach data. We conclude by discussing the importance of developing reliable measurement approaches for answering policy questions in cyber security.

Keywords: Cyber crime; data breaches; productivity; information security.

JEL: D80, H40, L10, L86

*Stanford University, Working Paper. Corresponding Author: Christos Makridis, Department of Economics and Department of Management Science & Engineering, Huang Engineering Center, 475 Via Ortega, Stanford, CA 94305, www.christosmakridis.com. Ben Dean: Columbia University, bcd2120@columbia.edu. Acknowledgments: Christos Makridis thanks Columbia SIPA for their doctoral fellowship in cyber security studies. We also thank Jay Healey at SIPA for comments on an early version of the draft and Herb Lin and the Stanford Cybersecurity Initiative for feedback. An earlier version of the paper was titled: "The Economic Effects of Cyber Security Failures on Firms: Evidence from Publicly Reported Data Breaches".

1. Introduction

> *The loss of industrial information and intellectual property through cyber espionage constitutes the "greatest transfer of wealth in history."* – General Keith Alexander.

> *Measurement is the first step that leads to control and eventually to improvement. If you can't measure something, you can't understand it. If you can't understand it, you can't control it. If you can't control it, you can't improve it.* – H. James Harrington

Cyber security has become a pressing issue in the wake of several large-scale and high-profile data breaches in recent years. Historically, attacks were driven by financial gain, but attacks are increasingly motivated by espionage, whether for state or commercial purposes.[1] The changing nature of these attacks has generated significant concern from corporate decision-makers and policymakers. One of the most concerning elements in this landscape arises from the fact that many system vulnerabilities and their magnitudes are not known, making it tough for researchers and policymakers to bound the range of possible outcomes. For example, using an index that gauges known and unknown vulnerability threats, Figure 1 displays more than a doubling simply between 2011 and 2014, consistent with increasing concerns about cyber threats in the press [1].

[insert Figure 1 here]

Despite the increasing importance of cyber security, little is known about its effects on firm-level investment and other economic outcomes. Part of the challenge is driven by the fact that there is no ready-to-use database containing both financial firm-level outcomes and cyber security information to estimate statistical models or calibrate computational models. Unfortunately, the lack of information has made it difficult for corporate and governmental decision-makers to determine the optimal allocation of both private and public funds towards cyber security or public policies that might bring benefits that are consummate with costs.

We document and characterize two major empirical challenges to statistical inference within the realm of the economics of cyber security at the firm-level. The first challenge is unobserved firm heterogeneity. On one hand, larger and more productive companies might be more likely to be targets of cyber security incidents because attackers have more to gain through a successful attack. On the other hand, larger and more productive companies may also tend to have better

[1] See Verizon's Data Breach Investigations Report (http://www.verizonenterprise.com/DBIR/2014/reports/rp_Verizon-DBIR-2014_en_xg.pdf).

cyber security since they have more sophisticated risk management and larger information technology budgets, which are likely inputs towards information security. The magnitude of these two competing channels will determine the direction of the bias. The second challenge is sample selection. If not all cyber security incidents (including data breaches) are reported, and firms are less likely to report incidents during quarters or years when their profits and productivity are lower, then estimates of impacts of incidents will be biased towards zero.

The primary contribution of this paper is to bring new evidence to bear on the economics of cyber security by producing and analyzing the first panel database containing cyber and financial outcomes at the firm-level. Our study combines data on publicly traded firms from the available public data breach incident databases: the Privacy Rights Clearinghouse database (PRC) and the US Department of Health and Human Services (HHS).[2] In each of these databases, we identify publicly-traded firms and match them with their publicly available financial statements accessible through Compustat. Our main solution to unobserved firm heterogeneity is the inclusion of relevant financial variables (e.g., number of employees), which proxies for firm-specific characteristics that are correlated with unobserved productivity both in the cross-section and over time. We also present results using a fixed effects estimator, which removes time-invariant sources of variation across firms by exploiting only the within-firm *changes* in firm and information security outcomes. For all of our statistical analysis, we use Stata and its accompanying packages for regression analysis and basic descriptive characterizations.

Using cross-sectional variation, some of our estimates point towards a negative association between breaches and firm outcomes. For example, we find that a 10% rise in records breached is associated with a 0.2% decline in firm productivity and a 1.1% decline when we focus on the set of breaches due to thefts in the HHS data. These estimates are identified off of variation conditional on time-varying firm inputs, such as employment or capital. However, our sample sizes are small, meaning that standard assumptions needed for interpreting our estimates from a causal standpoint do not readily apply. We also show that adding panel information from PRC/HHS does not add any predictive information and, in fact, amplifies the selection problem that is already inherent in the data. While the HHS data is marginally more reliable because of the sector's regulations for reporting breaches of proprietary patient information, it too is constrained in predictive power due to its sample size and possible selection.[3]

[2] We also experimented with the VERIS Community Database, but it did not add any predictive power to our analysis. We, therefore, omit it for simplicity.

[3] http://www.forbes.com/sites/stevemorgan/2016/05/13/list-of-the-5-most-cyber-attacked-

Our paper is closely related with a broader strain of research on the economics of privacy and information security [2, 3]. While there are a number of event studies that examine the behavior of stock prices following a cyber security breach at a company, these studies have produced conflicting results, as we will discuss in the next section. There are at least two reasons. First, there is incredible heterogeneity in the type of information that might be released in a breach. For example, the release of private consumer records, such as in the case of Equifax [4], is likely to undermine consumer trust, whereas malware attacks for manufacturing companies is more likely to reduce cash flow (rather than consumer trust) because of the nature of their products and services.[4] The effects of different types of attacks may take longer to materialize for companies that rely more on intangible capital as a competitive advantage in the marketplace. Second, while many breaches take place without firm knowledge, reflected in the large uncertainty about firm vulnerabilities in Figure 1, those firms that do detect incidents may not have an incentive to report the full magnitude of a breach, as was seen in the case of Yahoo [5]. Our study comes at an especially important time as policymakers and governments are debating changes to the legal framework surrounding the reporting of breach incidents [6].

2. Literature review

2.1. Economics of cyber security

Information technology has fueled economic growth over the past three decades across industries [7], firms [8, 9], and individuals [10]. While information technology (IT) has helped promote global integration [11], greater firm-level innovation, [12], increased product variety [13], and lower costs of doing routine tasks [14], information security (or 'cyber security') has lagged. Many systemic vulnerabilities within infrastructure, logical, and software layers create risks for firms and individuals who use and rely on these technologies.

Unfortunately, there is little empirical content to ground sensible cyber security decision-making in the public and private sectors. While there is some evidence of a negative effect of breaches on firm outcomes, research thus far has produced ambiguous conclusions. Part of the reason for this ambiguity is due to the nature of the heavy-tailed distribution of data breaches (e.g.

industries/#2e911fca3954

[4] Equifax has, since then, launched a "free-for-life" service that offers consumers the ability to lock and unlock their Equifax credit report as an attempt to rebuild consumer trust (https://www.equifaxsecurity2017.com/2018/01/31/equifax-launches-lock-alert/).

number of records affected), which reduces the reliability of least squares estimators since they focus only on the conditional mean, but, more importantly, a bigger gap is simply that there is also no available linked financial and cyber security outcome data for conducting causal inference.

Nonetheless, many event studies have emerged, estimating the impact of data breaches on stock prices within narrow windows. For example, a literature review of 45 studies found that roughly a quarter of the studies do not find a statistically significant association between breach incidents and stock prices [15]. Even among those studies that do find a negative association, there is wide dispersion in their estimated conditional correlations between breaches and stock prices. One reason is the fact that the type of information accessed in a breach may vary across companies. For example, some have found a statistically significant adverse effect among breaches involving unauthorized and confidential data, but no effect among data that is not confidential [16]. Another reason is that stock prices might not fully capitalize the risk inherent in cyber security breaches. For example, others have found that, while stock prices decline following an announcement of a breach, they tend to recover almost to trend after two days [17]. Moreover, there is also evidence that breached companies face a 1.13% lower stock return the first three days following the breach, but the stock rebounds by the 14th day [18].[5]

Other research has focused on changes in the frequency and distribution of cyber security incidents over time. For example, while several event studies reveal a negative association between stock prices and breaches, it is possible that the negative correlation is spuriously driven by a decline in the costs, or at least perception of risk among investors, over time [19]. For example, information security has become a major concern among chief executives, leading to expanded data security budgets. However, others have found that neither the severity nor frequency of data breaches has increased over the past decade [20]. Instead, those incidences that have attracted attention can be explained by the heavy-tailed statistical distributions underlying the specific incident [21]. Another study found that there is a stable power-law tail distribution of personal identity losses per information security incident [22].

An implication of a heavy-tailed statistical distribution is that the empirically observed incidents only represent a sub-set of all possible incidents, which makes it harder to recover causal effects in small samples. Consistent with the view that the impacts of breaches have heavy tails,

[5] When the financial cost from data breaches are deemed ‘material’, publicly-listed companies report these costs in their quarterly financial filings to the U.S. Securities and Exchange Commission. Such past incidents have typically resulted in costs that equate to a small fraction of annual revenues (<1%), which could indicate that the direct costs of even these large-scale incidents are relatively small; see https://theconversation.com/why-companies-have-little-incentive-to-invest-in-cybersecurity-37570.

some have found that the median cost of data breaches was substantially below the mean and, moreover, represented only 0.4% of the estimated annual revenues of the companies in the sample [23]. If the majority of incidents fall beneath this threshold, as the literature would suggest, then inferences drawn from the data provide a limited view of the true cost of data breaches.

Data linking certain categories of breach, their root causes and subsequent financial/economic impact on a business may assist in better cyber risk management. Some survey evidence exists accounting the cyber security incident and cost data that are more easily collected than others, how policymakers might improve data access, why previous policy-based efforts to do so have largely failed, and what differential ignorance implies for cyber security policy and investment in cyber defenses and mitigation [24]. This paper attempts to contribute to this growing understanding of what data could be collected, how, and what the limitations of such data collection may be.

While these event studies make important strides in identifying the presence (or lack thereof) of short-run financial effects on firms, they do not focus on long-run and dynamic considerations. Our study takes a step back by creating a firm-level database containing both data breaches and financial outcomes, and examining their conditional correlations. As we will discuss shortly, our primary aim is to characterize the data and illustrate several fundamental weaknesses that undermine the ability to use these publicly available datasets on breaches for causal inference. Unfortunately, the importance of privacy in cyber security is receiving heightened attention and scrutiny, which makes storing and securing data on breaches more costly and challenging for empirical analysis [25]. Understanding whether the increasing attention towards cyber security is a function of the salience versus substance is an important issue for disciplining regulation and policy.

2.2. Methods for estimating costs

The prevailing approach for estimating the costs of cyber security incidents is to gather estimated costs and number of breaches from a cross-section of companies. Many of these analyses are implemented using either qualitative surveys (e.g. those conducted by the Ponemon Institute) or self-reported incidence reports. The survey responses provide information on the estimated total breach damage and the number of records breached such that the cost per record is obtained by taking the ratio between the two. Some of the subjective elicitation studies (e.g., [26]) have been criticized (e.g., [27]), obtaining $201 per record breached cost from the full sample. Comparisons between studies have largely relied on simple measures of R-squared to determine the quality of model fit and superiority [27]. Other research suggests that information security countermeasures might be most cost-effective with an approximate 17-21% return on investment [28].

These surveys are useful heuristics, and the efforts are laudable, but serious improvements are required to undertake robust econometric analyses relating to information security. The relationships are not micro-founded in any theories about organizational dynamics, nor are the implied cost estimates grounded in statistical inference. By averaging highly heterogeneous breaches and estimated costs together, these studies are challenged by fundamental identification problems arising from omitted variables bias. Specifically, these approaches fail to not only control for time-invariant sources of heterogeneity across companies, but also basic time-varying differences, such as employment and assets, that are correlated with the timing of data breaches and broader organizational decisions.

Deferring to simple measures of overall fit from the R-squared in order to evaluate the quality of the model is not a fruitful endeavor for at least two reasons. First, to the extent existing datasets to not contain a representative sample of data breach incidents, but are rather a selected subset due to perverse incentives for companies to report data breaches, then metrics of model fit do not have their standard interpretation [29]. Second, given that there are significant cross-sectional differences in productivity in even narrowly defined industries [30, 31], and there is still an active economics literature on the underlying sources that explain the dispersion of productivity across firms [32], reliably predicting idiosyncratic data breaches out of sample is an even more challenging endeavor.

Many studies use self-reported surveys and suffer from a variety of sampling and methodological problems [33]. For instance, self-reported incidence reports are produced annually by the Federal Bureau of Investigation's Internet Crime Complaint Center. These reports provide information on the number of incidents of various forms of Internet-related crimes, the demographics of those who reported being victim of an Internet-related crime and the losses due to the crime as reported by the person or company in question. There are many short-comings of the methodology used in these reports. Since only the companies or people who report the crime are included in the sample, then differences between the sample and population can bias out-of-sample extrapolations and aggregation exercises. T Moreover, the method for estimating the losses incurred by each person or company are not provided, and likely differ across the reporting companies, again leading to unreliable total loss estimates [34].

Other studies use an accounting approach to measure the economic costs of security incidents.

They assign values to particular activities affected by the breach and publicly report the aggregate quantities [35]. The wide variance of the estimates of the global cost of cyber-crime in these studies points to the inaccuracy of the methods used to make the estimates. One commonly cited paper by national security leaders claims global costs amounting to $445 billion while another, previously cited by the President of the United States, suggests over $1 trillion [36, 37]. In addition to the sample selection problems we pointed out above, these accounting exercises represent more about the "engineering costs" rather than the "economic costs". In reality, there are many unobserved factors and frictions that are not taken into account absent a model that describes the relevant economic forces.[6]

3. Measurement and Institutional Context

3.1. Data Sources and Software

We use a combination of the Privacy Rights Clearinghouse (PRC) and the U.S. Department of Health and Human Services (HHS) databases of cyber security data breach incidents. These datasets document the incidence of cyber security data breaches based on geographic location and organizational entity. We extracted all relevant variables and manually created common identifiers for publicly-listed firms in these datasets with their publicly-available financials through Wharton's Data Research Services Compustat ("Compustat"). In addition to providing measures of employment, capital, and revenue, among other financial outcomes, we also use Compustat to recover a time-varying measure of firm productivity by taking the residual from a regression of logged revenues on logged employment, capital, and materials.[7]

The PRC is compiled by a nonprofit corporation and contains records of breaches starting from 2005 with 5,391 separate breach incidents as of March 2017. The U.S. Department of Heath and Human Services catalogs all notifications of breaches of unsecured protected health information for incidents affecting 500 or more individuals. While the database contains 1,561 unique entries as of July 2016, only 87 are breaches for publicly traded companies that we are able to match.

[6] Engineering costs estimates have led to wildly different conclusions than economic estimates in the environmental economics literature, for example [38].

[7] While there are many ways to compute measures of productivity, this specification assumes a Cobb-Douglas production function among the three inputs, taking the residual as unexplained variation in revenue based on the inputs. There are other more recent approaches [39], but for the sake of simplicity we defer to the more transparent method.

Though these datasets are the most extensive publicly available data, they cover only a small fraction of total information security incidents.[8]

For the PRC, we observe the name of the company breached, the date that the breach was made public, the total records stolen, the ascribed cause of the breach (one of eight categories: unintended disclosure, hacking or malware, payment card fraud, insider, physical loss, portable device, stationary device, unknown or other) and the firm's industry (one of seven categories: nonprofit, healthcare and medical providers, government and military, educational institutions, businesses - retail/merchant, businesses - financial and insurance services, businesses - other).[9]

For the HHS dataset, we observe the name of the breached entity, breach submission date, the number of affected individuals, the type of breach (hacking/IT incident, improper disposal, loss, theft, unauthorized access/disclosure, unknown or other), the location of the breach (desktop computer, electronic medical record, email, laptop, network server, other portable electronic device, paper/films, or other), the type of covered entity (health plan, heath care clearing house or healthcare provider), the state in which the entity is located, whether a business associate was present, and a description of each incident. The HHS data offer an improved probability of detecting potential relationships, relative PRC, since the reporting is arguably cleaner and contains less measurement error. The healthcare sector has largely modernized and the publicly traded companies in our data generally have integrated technology into their services, providing fewer opportunities for mis-reporting. The healthcare sector has also been a major target of data breaches in recent years.[10]

While the PRC data contain many records for government departments and educational institutions, we restrict our sample to publicly listed companies so that we can link them with financial records from Compustat—between 2005 and 2016 for the PRC and between 2010 and 2016 for the HHS. Our Compustat sample contains publicly traded firms with non-zero employees and non-missing data on the number of employees, value of the capital stock, materials, and revenues at an annual frequency. We also hand-code each company with a unique identifier between the PRC/HHS and Compustat databases. These constraints produce a sample of roughly 475 unique firm observations in the PRC. Throughout our descriptive statistics and ordinary least squares

[8] There are not many studies that take a case-study approach, but there is some evidence from one study with detailed micro-data at a single firm that the number of cyber incidents is far larger than typically reported, containing over 60,000 cyber security entries in their company alone [21].

[9] There are a few entries that include multiple firms, which we omit from the analysis since there is no way to reasonably infer the fraction of the reported total accounts breached that correspond to each of the included firms.

[10] http://www.eweek.com/small-business/health-care-it-security-challenged-by-phishing-attacks.html

(OLS) regression analysis, we use `Stata MP 15.0` [40] on an Intel(R) Core(TM) i5-3230M CPU @ 2.60 GHz, 2601 Mhz, 2 cores, and 4 logical processors with 8 GB of memory on Windows 10.

A final distinction that is important to point out is that there are many years that a company may not be observed in either database even if they are observed once. For example, a firm might incur a breach that is publicly reported in 2012, but they are not observed in 2010 in the cyber databases. Does this mean that they had no cyber incidents in 2010? Or, does it mean that they did not have a large enough incident that led to public recognition? There is a potentially major selection problem here—years a firm is not observed might not be "true" zeros in the sense that they had breaches that were simply not reported. We examine the following later, but note that setting values of breaches when the company is not observed in PRC/HHS to zero is not valid given the presence of non-random reporting of breaches.

3.2. Cross-sectional Comparison with Compustat

We now turn towards a comparison between both the PRC/HHS datasets and Compustat to better understand the type of firms that we are capturing. Figure 2 plots the distribution of records affected in breaches (in logarithms of thousands of affected records) between firms that we are able to match between the Department of Health and Human Services (HHS) data and Compustat with those firms that are only in the HHS data. We also plot the distributions of logged records breached separately based on the type of breach: all companies, a breach against a business associate, a breach against a health plan, and a breach against a healthcare provider. In other words, we plot the distribution of records breached for all companies observed in the HHS data with those that match into the set of publicly traded companies through Compustat.

The data points towards three facts. First, Compustat firms have many more records affected per breach, on average, than their non-Compustat counterparts. For example, pooling all types of breaches together, Compustat firms have, on average, 716,515 records breached over the observed time series, whereas non-Compustat firms have 47,033 records breached—approximately 6% as large as their counterparts. Second, the bulk of the records breached are those involving health plans, rather than business associates or healthcare providers. For example, Compustat firms have, on average, over 2.4 million records breached. One reason for some of these stark differences is the fact that the median healthcare company is actually quite small—for example, there are many health facilities with just a few doctors. Third, publicly traded companies display considerably more dispersion in their records breached. For example, the standard deviation for the non-

Compustat sample is 440,461, whereas it is 6,876,867—over 15 times as large.

[insert Figure 2 here]

We also implement a similar exercise for firms in the PRC dataset separately by major industry, which is documented in Figure 3. We again find significant differences in the mean number of records breached covered in the Compustat versus PRC-only samples. For example, in the pooled sample, we find an average (standard deviation) of 1,442,697 (10,162,010) among Compustat firms, whereas it is 194,011 (2,301,437) among PRC-only organizations. These orders of magnitude differences again reflect the fact that publicly traded companies tend to experience cyber security attacks of a greater scale, relative to their private counterparts, except for in the "other business" sector. Companies in the finance / insurance and retail / merchant sectors are also the biggest targets by far, although the sample of medical companies in the data is limited.

[insert Figure 3 here]

We finally turn towards a more detailed look at the differences in financial outcomes between the companies covered in the Compustat dataset overall and our specific PRC and HHS datasets. We document these differences in Table 1. The sample of PRC and HHS publicly traded firms that we are able to match with Compustat are much larger than the average size of firms covered by Compustat, but are not in either the PRC or HHS data. For example, publicly traded firms in the PRC sample have roughly 87,000 employees, whereas those in the HHS sample have over 72,000, but the remaining firms in Compustat only have 10,000. We see similar differences across other financial performance measures. For example, capital stock, research and development expenditures, and revenue are all orders of magnitude larger in the linked PRC/HHS-Compustat sample versus those that remain in Compustat.

[insert Table 1 here]

Aside from making our sample more transparent for the results that follow, we view this as an important exercise because the PRC and HHS datasets are typically used as the "go-to" samples for cyber security work simply due to data availability. However, if the goal is to understand how data breaches affect major firms in the U.S. economy, then these basic descriptive results suggest that the PRC and HHS are not necessarily reliable samples. One reason we might be especially interested in the performance of these large firms is because of network externalities [3]. If, for example, a major bank is breached and credit card numbers are released, this can have profound effects on consumer confidence and spending in the aggregate economy, thereby impacting economic growth.

3.3. Descriptive Statistics

Having compared companies in the entire Compustat dataset with those observed in ours, we now provide several descriptive facts about our data. Figure 4 plots the distribution of these breaches between the two datasets. Because healthcare companies are only required to report breaches that involve a certain number of patients, the distribution for the HHS dataset is truncated. However, the PRC still produces a larger average number of records affected per breach per firm (1,442,697 versus 801,525). The distribution of records breached in the PRC adheres slightly more to a normal distribution (Shapiro-Wilk test for rejecting normality z-statistic = 2.82) than the HHS dataset (Shapiro-Wilk test for rejecting normality z-statistic = 6.174).

[insert Figure 4 here]

We turn towards plotting the number of breaches (in '0,000s) from 2005-2016 in Figure 5. One immediate observation is that the number of records breached in each of the databases is not smooth; there are large year-to-year swings that are based on, for example, one or just a few very large incidents. For example, Anthem Inc. had several large breaches in the sample: 1,023,209 records in 2010 and 839,711 records in 2014. Even though the year-to-year averages seem to represent relatively random fluctuations—partially because of sample selection issues, which we discuss later—the probability of facing a breach at a firm-level is likely endogenous.

[insert Figure 5 here]

We now turn towards differences in the type and location of breaches. There are four general categories for breach type (hacking, improper disposal, loss / theft, and miscellaneous) and there are four general categories of the location of a breach (desktop, email, laptop, and network server). Figure 6 displays these distributions. Panel A shows that malicious hacking attempts are by far the most frequent form of data breaches with an average of 488,676 records affected per breach. While records are still breached through human error, e.g., improper disposal or loss, these statistics reflect the growing concern that hacking is becoming more common. These results from the HHS are also consistent with those from the PRC in Figure 7, which again overwhelmingly shows that hacking attempts are the most common cause of data breaches. Panel B shows that breaches are most common via the network. These breaches affect an average of 512,581 records per incident with the next closest among desktops with 44,649 records breached. The fact that networks are the most common port of entry for hackers might reflect that it is easier to go unnoticed in a larger network, rather than through a particular computer.

[insert Figure 6 here]

[insert Figure 7 here]

4. Data Breaches and Firm Outcomes

4.1. Conceptual Framework

Information security arguably falls within a broader umbrella of organizational practices, which have an already well-known association with productivity at the firm-level [9, 41, 8]. We now develop a formal framework for understanding the potential importance of information security within a firm by specifying a stylized model of production at the firm-level. Suppose that firms have a production function that relates various inputs, such as capital and labor, with output

$$Y_{it} = A_{it} S_{it}^{\alpha^S} K_{it}^{\alpha^K + \sigma S_{it}} L_{it}^{\alpha^L - \sigma S_{it}} \tag{1}$$

where Y denotes output of firm i in year t, A denotes technology-neutral productivity, S denotes information security, K denotes physical capital, and L denotes labor services. Motivated by prior applications in the literature [9, 8], this production function provides a tractable framework for testing whether information security and capital are relative complements, i.e. $\sigma > 0$.[11]

Our inclusion of information security in the production function captures the increasingly important role of maintaining safe and secure data in organizations.[12] There are various reasons that information security might matter for an organization. For example, information security might

[11] We assume the usual regularity assumptions on the production function—that the factor shares are bounded by zero and one.

[12] For instance, the Bureau of Labor Statistics (BLS) forecasts that employment in cyber security will grow 18% from 2014 to 2024, which is much more rapid than the average. Starting salaries are roughly $91,000 and many expect them to continue rising (https://www.bls.gov/ooh/computer-and-information-technology/information-security-analysts.htm). According to other estimates, job postings for information security analysts have grown 74% between 2007 and 2013, which is twice the rate of other information technology jobs *https://www.tripwire.com/state-of-security/off-topic/the-top-10-highest-paying-jobs-in-information-security-part-1/). In this sense, companies are increasing their demand for information security services and view them as an integral part of their production and services processes.

matter for maintaining an organizational brand and trust with consumers. In these cases, information security is closely related with prior contributions that have modeled intangible and/or organizational capital [42, 43, 44]. Similarly, information security might matter for an organization's internal data infrastructure. In these cases, information security plays a role of coordination, especially among disparate business units [45]. However, we do not need to take an explicit stance on the mechanism.

We normalize the production function to labor. Letting lower case letters denote the transformed log variables (e.g., $x = \ln X$), Equation 1 takes the following form

$$(y-l)_{it} = \alpha^K (k-l)_{it} + (\alpha^K + \alpha^L + \alpha^S - 1) l_{it} + \sigma (k-l)_{it} s_{it} + \alpha^S s_{it} + a_i + \xi_{jt} + \lambda_t + \epsilon_{it} \quad (2)$$

where $A_{it} = \exp(a_i + \xi_{jt} + \lambda_t + \epsilon_{it})$ provides a specific functional form for characterizing firm-level efficiency in terms of an idiosyncratic firm-specific component (a_i), an industry trend (ξ_{jt}), and an aggregate trend (λ). If $\sigma > 0$, then we should also find complementarity between information security and capital through equations of the form

$$(k-l)_{it} = \gamma s_{it} + \pi X_{it} + a_i + \xi_{jt} + \epsilon_{it} \quad (3)$$

where X denotes a vector of controls. The first order-condition associate with a change in information security is given by $\partial(k-l)/\partial s = \gamma$. Complementarity implies that an exogenous increase in s will produce a rise in the factor demand for capital. In this sense, information security matters relatively more when when information technology is high, which is especially pertinent for the technology and software industry.

Practically, while we do not observe the quality of information security, we do observe the number of data breaches, which is a proxy for the opposite of information security. We would, therefore, expect to see a negative association between breaches and various measures of firm outcomes, namely various measures of productivity. Unfortunately, the size of our sample limits the amount of heterogeneity we can allow for, but we would also expect to find that data breaches are more costly for firms in industries that have greater complementarity between information security and capital.

4.2. Identification Strategy

We now consider variants of Equations 2 and 3 through regressions of the form

$$y_{it} = \beta X_{it} + \gamma b_{it} + \phi_i + \lambda_t + \epsilon_{it} \tag{4}$$

where y denotes some firm outcome (e.g., logged revenues), X denotes a vector of firm characteristics (e.g., employment), b denotes logged breaches, and ϕ and λ denote potential fixed effects on firm and year.[13] We cluster standard errors at the firm-level to allow for arbitrary degrees of autocorrelation in the error over time in the same company [46].

Consistent identification of γ in Equation 4 requires that unobserved shocks to firm outcomes are uncorrelated with changes in information security outcomes. There are at least two plausible violations. The first identification concern is a static selection problem that arises from the fact that more productive firms might also be larger targets for malicious actors—that is, because there is more to steal or gain from these companies. It is also possible, however, that less productive firms are higher priority targets for malicious actors—that is, because they may be easier to breach. These two margins represent the marginal benefits and costs to malicious information activities, requiring empirical evidence to shed light on the theoretically ambiguous effect. The inclusion of firm fixed effects addresses this concern, but requires that we have sufficient within-firm variation—that is, that we see the same company being breached multiple times. Unfortunately, since companies do not always have an incentive to necessarily report breaches, we are concerned about a major selection problem.

The second identification concern is a dynamic selection problem that arises from the fact that data breaches might respond to a rise in firm outcomes. That is, if a firm experiences a surge in sales or profitability, malicious actors may respond by increasing their attacks against the company. While in theory we can address both concerns by including firm fixed effects to remove time-invariant heterogeneity across companies, in practice we found that fixed effects estimators failed due to a lack of variation. We address this concern by controlling for relevant firm characteristics, such as employment or capital, that are likely to co-move with firm outcomes. To the extent information security incidents spike in response to firm outcomes, our identifying

[13] We also recognize our parametric assumption about the functional form between breaches and firm outcomes. Given that the distribution of breaches is skewed, we also implemented (but do not report for brevity) regressions that include higher-order terms and splines. Splines may be an effective strategy, but they require sufficient numbers of observations within each partition of the distribution. Our use of the term b is driven by the fact that we observe records breached, rather than information security s, in the data.

assumption is that they do so in a way that is captured by movements in employment or capital.

While we are well aware of these potential identification problems, and thus caution readers to interpret these as conditional correlations, we want to underscore that this is, to our knowledge, the first time that gradients between breaches and firm outcomes have been estimated for a comprehensive set of companies. In this sense, the fact that we point out major flaws with the publicly reported data breaches is a starting point for what we hope will be a more rigorous and comprehensive data gathering process in the future.

4.3. Sample Selection

There are at least two reasons that our measure of breaches could contain sample selection problems. The first is that many companies do not know that they are breached. Those that do often do not find out that they were breached until several months or years after. For example, Mandient, an incident response survey, reports that it took companies roughly 205 days in 2014 to detect data breaches. We, therefore, use annual (rather than quarterly) financial data.[14]

The second is that companies do not have a strong incentive to report all of their incidents except in jurisdictions where such reporting is mandatory. Even if there was comprehensive legislation in place, as is seen in the case of the healthcare sector where there are some mandatory disclosure laws and associated regulations, companies may still not report if the probability of getting caught and/or the fines associated with getting caught are sufficiently low.

While the first form of sample selection will tend to attenuate the estimates due to the potentially weak mapping between the time a breach is discovered and the financial outcomes of the firm, the second sample selection problem behaves as non-classical measurement error and can create bias. Anecdotal evidence suggests that larger companies may tend to have a stronger incentive not to report the full scale of their breaches. Take, for instance, the data breach(es) on Yahoo. In late 2016, Yahoo announced two incidents involving the compromise of data for over 500 million and one billion Yahoo! accounts respectively. These breaches were alleged to have occurred in 2013. Subsequent to their initial announcement of 500 million accounts, Yahoo's stock price fell by 4.4% in just a few months, totaling $1.7 billion in market value.[15] Verizon, which was in the process of acquiring Yahoo, also reduced their offer from $4.83 billion to $4.48 billion, or 7.25%. Later in 2017 it was announced that the incident previously claimed to have affected 1 billion

[14] http://ww2.cfo.com/cyber-security-technology/2015/02/fewer-companies-able-detect-cyber-breach/

[15] http://fortune.com/2016/12/15/yahoo-shares-hack/

account had in fact affected 3 billion accounts. If larger companies have a stronger incentive not to report, then we will underestimate the effect of data breaches on firm outcomes. Moreover, if it takes time to detect, assess and report incidents—and these details can change over time—then estimates on the effect of the incidents will be underestimated or overestimated accordingly.

5. Results

5.1. Cross-sectional and Panel Estimates

We begin by detailing our main estimates of Equation 4 in Table 2. We focus first on the estimates obtained from our PRC sample of firms. In the cross-section, controlling for logged employment, we find that a 10% rise in breaches is associated with a 0.2% rise in revenue per worker, a 0.2% decline in capital per worker, and a null association with total factor productivity. Once we add firm and year fixed effects, these correlations vanish even further to null and statistically insignificant associations in every case.

We now turn towards our estimates obtained from our HHS sample of firms. In the cross-section, we find that a 10% rise of breaches is associated with a 0.5% decline in revenue per worker, 0.3% decline in capital per worker, and a 0.2% decline in TFP. While the first two correlations are not statistically significant at conventional levels, the TFP estimate is significant at a 1% level. However, once we add firm and year fixed effects, our estimates all become statistically insignificant.

[insert Table 2 here]

Panel estimators are well-known to introduce bias when there is serially uncorrelated noise measurement error—that is, the signal is highly correlated over time [47, 48]. Our interpretation of the data is that a firm's incentive to report a breach is effectively serially uncorrelated noise. That is, sometimes a particularly large breach is publicly reported, other times it is not, but in both cases the noise is correlated with underlying firm fundamentals. If the incentive to misreport were simply time-invariant, then fixed effects would help. However, even in such cases, there is simply too small of a sample to do serious inference with fixed effects.

We, therefore, turn towards the HHS data, which appears more reliable (given it is based on state-level mandatory incident reporting) despite its smaller size. We specifically look for evidence of heterogeneity in the treatment effect of breaches using only the cross-section when partitioning by the type of breach. Table 3 documents these results. We find that a 10% rise in records breached is spuriously associated with a 1.5% rise in revenue per worker for hacks, but a 1.6% decline in capital per worker and 0.2% decline in TFP. We underscore, however, that our sample consists of nine observations and is, therefore, not a credible estimate. When we restrict the sample to breaches due to theft and losses, we find that increases in breaches are generally associated with declines in revenue per worker and TFP, most notably a plausible 1.1% decline in TFP associated with a 10% rise in records breached.

[insert Table 3 here]

5.2. Discussion of the Limitations

The data on which we have based our analysis is constrained in two ways. The first major constraint is the completeness of publicly available cyber security incident databases, broadly speaking, and data breach incident databases, more narrowly. Neither PRC nor HHS contain the full universe—or anywhere near it—of incidents. In fact, it is likely that the public data represents only a very small fraction of overall incidents [**?**]. To the extent that the missing observations and incomplete nature of the data is random, it will only produce larger standard errors. Normally, such measurement error is innocuous, but since our sample is so small it is likely to induce bias.

However, what concerns us even further is the fact that the companies that do not report, or those who do report simply a subset of the actual breaches that occurred, may do so systematically in ways that are correlated with unobserved heterogeneity. Our results from the the HHS sample are marginally more reliable and likely driven by the fact that companies are required to report certain types of incidents under section 13402(e)(4) of the HITECH Act. With the European Union's General Data Protection Regulation and Directive coming into effect in 2018, which include mandatory data breach notification rules, we hope this will provide a future data source with incidents across many industries.

In both cases, however, we tested our concerns for sample selection more formally by replacing instances when a given company is not observed in the PRC or HHS dataset with a value of zero for records breached. If, for example, both databases are the entire universe of breaches, it is safe to infer that there are zero breaches when a company is not observed in them. However, when implementing this diagnostic, we recovered estimates that were even more noisy and centered at zero with p-values above 0.90. Based on the evidence we have examined so far, our conclusion is that these datasets are inadequate measures of the universe of data breaches even for publicly traded companies.

We also are required to assume that records are additively separable. Our heterogeneity analysis indicates that there are substantial differences in the correlations between breaches and firm outcomes based on the nature of the breach, meaning that our assumption of homogeneous effects merely through the inclusion of logged records breached generates a form of heterogeneous treatment effects. A related issue arises from the fact that not all information lost is equally harmful. If, for example, emails from a CEO are released, that could affect firm outcomes much more than a rank-and-file worker's information might simply because the former is much more visible to shareholders and the public. We are aware of these problems, but simply due to sample size constraints are unable to resolve them.

The second major constraint is that our estimates are based solely on the set of publicly traded firms. However, large publicly traded companies often many establishments, which increases the noise-to-signal ratio in causal inference since a cyber security incident, including data breaches, might occur only at a single establishment, leaving the bulk of the firm unscathed. To the extent that data breaches affect the local establishment more than the overall firm, then focusing only on the large companies could add measurement error. Future analysis can match based on establishments using Dun and Bradstreet data, but our results more generally point towards the conclusion that neither the PRC or HHS publicly reported datasets are even useful for causal inference.[16]

As a final note, we also used data from the Veris Community Database, which contains an additional set of cyber security incidents (including a sub-set of data breaches). However, the data were less comparable to those in the PRC and HHS and did not add any predictive power in our regressions. To keep our analysis as transparent and simple as possible, we use the PRC to obtain a broad cross-section of companies and the HHS to narrow in on an industry that is most at risk by data breaches specifically.

[16] A separate concern is that the cyber security incidents tend to reflect multiple sources of security failures. For example, an attack may involve not only lost records, but also fraud and a shock to the networks that bring company operations to a halt. While the solution is to simply condition on these different measures—that is, the continuous measure of records lost and the discrete measures of the type of attack—there is simply not enough within-firm variation in the available data to separately identify the coefficients, let alone on any single coefficient.

6. Conclusion

Despite the increasing concerns and perceived damages associated with emerging cyber security incidents, there is no *causal* evidence on their potential long-run economic effects. The only available studies that speculate over the costs use either: (i) self-reported cost data where companies are asked about how much they think attacks cost them, or (ii) aggregate data where engineering cost estimates are applied to large swaths of the population. We address this empirical gap by producing the first panel dataset containing both financial and data breach incidents at a firm-level using data from the Privacy Rights Clearinghouse (PRC) and Department of Health and Human Services (HHS) from 2005 to 2016. The dataset enables us to provide, to our knowledge, the first descriptive evidence to date on the operational and financial health (e.g., cash flow and assets) of companies joint with their histories of data breaches.

The first part of our paper began by documenting several descriptive statistics about breaches—the types of breaches, their location, and incidence in publicly traded companies. The majority of breaches tend to take place either directly through an individual's desktop computer or through email, which is consistent with the anecdotal evidence about employee inattentiveness to phishing scams and other behaviors (e.g. picking up and using USB thumb drives infected with malware). We also examined the differences between the set of publicly traded companies with reported breaches and the entire universe of publicly traded companies. We find that the companies contained in our sample are roughly 10 times as larger—as measured by employees or assets—than the average company in Compustat, suggesting that only the most visible of publicly traded companies get their breach covered and reported.

The second part of our paper examined the association between breaches and firm outcomes. Using cross-sectional variation, our results suggest that a 10% rise in records breached is associated with a 0.2% decline in firm productivity. We find massive heterogeneity in the treatment effects when we partition by the type of breach, but are unable to determine whether the dispersion is spurious due to small sample sizes, particularly in the HHS data. For example, we also found that a 10% rise in records breached is associated with a 1.1% decline in firm productivity for the set of breaches due to theft. In either case, our cross-sectional estimates are non-trivial in light of other treatments that influence firm productivity [32].

Despite suggestive evidence on the negative association between breaches and firm outcomes, our data is plagued by serious sample selection problems. In particular, our results are based only off of the instances when the company is observed reporting a breach. Periods when the company does not have a reported breach does not necessarily indicate that the company indeed had zero breaches; rather, it indicates that the company simply was not caught on public record with the breach. When we instead set records breached to zero in years that the company did appear as having any publicly reported breaches, we find that there is no association between breaches and firm outcomes, indicating that there is a selection problem. Our results point towards an acute need for developing more comprehensive data on cyber security incident outcomes at the firm-level.

Decision-makers in firms and policy makers in governments require evidence in order to make well-informed decisions regarding information security investments. However, the requisite evidence is not available for econometric analysis given the data deficiencies identified in this paper. Mandatory breach notification rules, such as those coming into effect in the EU in 2018, might provide future data sources that remedy these deficiencies. Until that point, data deficiencies will persist and, in turn, effective investment decision-making for information security will continue to be hampered.

References

[1] Bailey T, Del Miglio A, Richter W. The rising strategic risk of cyberattacks. McKinsey Quarterly. 2014;.

[2] Grossklags J, Acquisti A. When 25 cents is too much: An experiment on willingnes to sell and willingness to protect information. Workshop on the Economics of Information Security. 2007;.

[3] Anderson R, Moore T. The economics of information security. Science. 2006;314(5799):610–613.

[4] Berr J. Equifax breach exposed data for 143 million consumers. CBS News. 2017;.

[5] Collins K. Yahoo and Equifax just proved that you can never trust the first number announced in the data. Quartz. 2017;.

[6] Bisogni F, Asghari H, Van Eeten MJG. Estimating the size of the iceberg from its tip: An investigation into unreported breach notifications. Workshop on the Economics of Information Security. 2017;.

[7] Stiroh KJ. Information technology and the U.S. productivity revival: What do the industry data say? American Economic Review. 2002;92(5):1559–1576.

[8] Bloom N, Sadun R, Van Reenen J. Americans do IT better: US multinationals and the productivity miracle. American Economic Review. 2012;102(1):167–201.

[9] Bresnahan T, Brynjolfsson E, Hitt L. Information technology, workplace organization and the demand for skilled labor: firm-level evidence. Quarterly Journal of Economics. 2002;117(1):339–376.

[10] Aral S, Brynjolfsson E, Van Alystne MW. Information, technology and information worker productivity. Information Systems Research. 2009;23:849–867.

[11] Van Alystne MW, Brynjolfsson E. Global village or cyberbalkans: Modeling and measuring the integration of electronic communities. Management Science. 2005;51:851–868.

[12] Tambe P, Hitt L, Brynjolfsson E. The Extroverted Firm: How External Information Practices Affect Innovation and Productivity. Management Science. 2012;58(5):843–859.

[13] Brynjolfsson E, Hu Y, Simester D. Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sale. Management Science. 2011;57:1373–1386.

[14] Autor DH, Levy F, Murnane RJ. The skill content of recent technological change: An empirical exploration. Quarterly Journal of Economics. 2003;118(4):1279–1333.

[15] Spanos G, Angelis L. The impact of information security events to the stock market: A systematic literature review. Computers & Security. 2016;58:216–229.

[16] Campbell K, Gordon LA, Loeb MP, Zhou L. The economic cost of publicly announced information security breaches: Empirical evidence from the stock market. Journal of Computer Security. 2003;1:431–448.

[17] Cavusoglu H, Mishra B, Raghunathan S. The effect of internet security breach announcements on market value: Capital market reactions for breached firms and internet security developers. International Journal of Electronic Commerice. 2004;9(1):69–104.

[18] Lange R, Burger EW. Long-term market implications of data breaches. Journal of Information Privacy and Security. 2017;forthcoming.

[19] Gordon LA, Loeb MP, Zhou L. The impact of information security breaches: Has there been a downward shift in costs? Journal of Computer Security. 2011;19(1):33–56.

[20] Edwards B, Hofmeyr S, Forrest S. Hype and heavy tails: A closer look at data breaches. Workshop on Economics of Information Security. 2015;.

[21] Kuypers MA, Maillart T, Pate-Cornell E. An empirical analysis of cyber security incidents at a large organization. Workshop on the Economics of Information Security. 2016;.

[22] Maillart T, Sornette D. Heavy-tailed distribution of cyber-risks. European Physical Journal B. 2010;75(3):357–364.

[23] Romanosky S. Examining the costs and causes of cyber incidents. Journal of Cybersecurity. 2016;2(2):121–135.

[24] Wolff J, Lehr W. Degrees of ignorance about the costs of data breaches: What policymakers cand and can't do about the lack of good empirical data. Working paper. 2017;.

[25] Wheatley S, Maillart T, Sornette D. The extreme risk of personal data breaches and the erosion of privacy. European Physical Journal B. 2016;89(7).

[26] Ponemon L. 2014 Cost of data breach study: Global analysis. Ponemon Institute. 2014;.

[27] Verizon. 2015 Data breach investigations report. VERIS Community Group. 2015;.

[28] Soo Hoo K. Return on securing investment: Calculating the security investment equation. Secure Business Quarterly. 2001;.

[29] Greene WH. Econometric analysis, seventh edition. Prentice Hall. 2012;.

[30] Syverson C. Market structure and productivity: A concrete example. Journal of Political Economy. 2004;119(2):403–456.

[31] Syverson C. Product substitutability and productivity dispersion. Review of Economics and Statistics. 2004;86(2):534–550.

[32] Syverson C. What determines productivity? Journal of Economic Literature. 2011;49:326–365.

[33] Herley C, Florencio D. Sex, lies and cyber-crime surveys. Microsoft Research. 2011;.

[34] Florencio D, Herley C. Where do all the attacks go? Workshop on Economics of Information Security. 2011;.

[35] Anderson R, Barton C, Bohme R, Clayton R, van Eeten MJG, Levi M, et al. Measuring the costs of cybercrime. Working paper. 2014;.

[36] CSIS. Net losses: Estimating the global cost of cybercrime. Center for Strategic and International Studies. 2014;.

[37] McAfee. Underground Economies: Intellectual Capital and Sensitive Corporate Data Now the Latest Cybercrime Currency. McAfee. 2011;.

[38] Fowlie M, Greenstone M, Wolfram C. Do energy efficiency investments deliver? Evidence from the Weatherization Assistance Program. Working paper. 2015;.

[39] Ackerberg DA, Caves K, Frazer G. Structural identification of production functions. Econometrica. 2015;83(6):2411–2451.

[40] StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC. 2017;.

[41] Bloom N, Van Reenen J. Measuring and explaining management practices across firms and countries. Quarterly Journal of Economics. 2007;3:1560–1689.

[42] Andrew A, Kehoe PJ. Modeling and measuring organizational capital. Journal of Political Economy. 2005;113(5):1026–1053.

[43] Eisfeldt AL, Papanikolaou D. Organizational capital and the cross-section of expected returns. Journal of Finance. 2013;68(4):1365–1406.

[44] McGrattan ER, Prescott EC. A reassessment of real business cycle theory. American Economic Review. 2014;104(5):177–182.

[45] Alonso R, Dessein W, Matouschek N. When does coordination require centralization? American Economic Review. 2008;98(1):145–179.

[46] Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? Quarterly Journal of Economics. 2004;119(1):249–275.

[47] Bound J, Krueger AB. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? Journal of Labor Economics. 1991;9(1):1–24.

[48] Bound J, Brown C, Duncan GJ, Rodgers WL. Evidence on the validity of cross sectional and longitudinal labor market data. Journal of Labor Economics. 1994;12(3):345–368.

7. Tables

Table 1: Differences in Financial Outcomes between PRC/HHS and Compustat Samples

	PRC Sample		HHS Sample		All Compustat	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
breaches, '000s	160	1337	208	884	.	.
cash, '000,000s	5307	13665	2044	3262	555	4710
employees, '000s	87	179	72	97	10	42
capital, '000,000s	17416	42649	10218	34401	3192	16216
R&D, '000,000s	1350	2548	478	1226	140	711
revenue, '000,000s	33454	51767	35693	46506	3565	16070
Observations	723		119		82759	

Notes.–Sources: Compustat (2005-2016), Privacy Rights Clearinghouse (2005-2016), and Department of Health and Human Services (2010-2016). The table reports the average and standard deviation of breaches, cash-on-hand, employees, capital, research and development, and revenue across the three major samples: PRC, HHS, and Compustat (inclusive of companies that were matched).

8. Figures

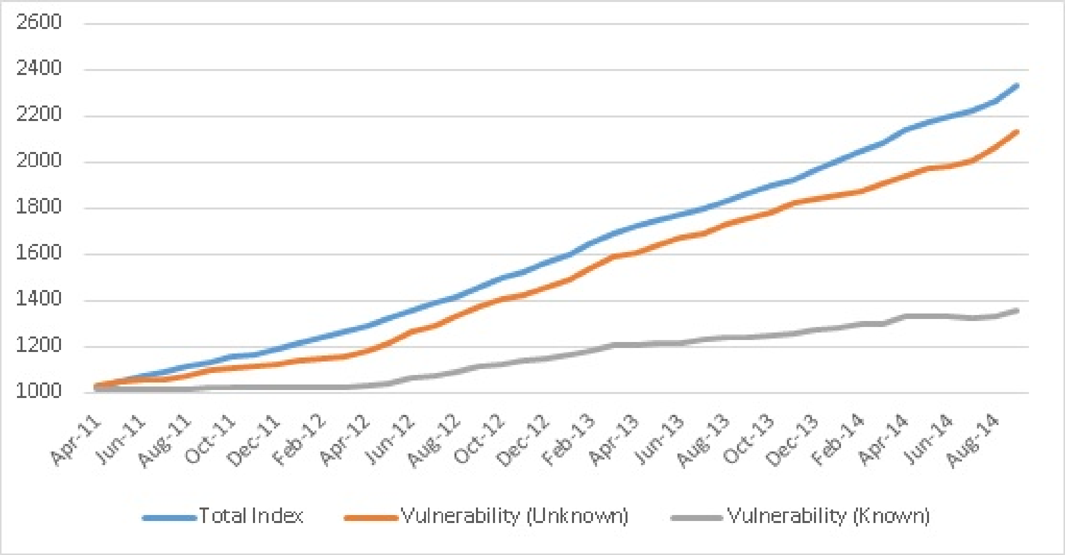

Figure 1: Index of Cyber Security Vulnerabilities, 2011-2014

Notes.—Source: Index of cyber security. The cyber security index is a sentiment-based measure of the risk to infrastructure from perceived cyber security threats (http://www.cybersecurityindex.org/). The survey is administered monthly to a cross-section of industry, government, and academic participants, and includes questions over threat levels from prospective attackers (e.g., groups), weapons (e.g., malware), targets (e.g., infrastructure), defense, and public perception.

Table 2: Estimates of Data Breaches and Firm Outcomes

Dep. var. =	ln(revenue per worker)				ln(capital per worker)				total factor productivity			
	PRC	PRC	HHS	HHS	PRC	PRC	HHS	HHS	PRC	PRC	HHS	HHS
ln(breaches)	.02*	.00	-.05	.05	-.02*	-.00	-.03	.02	.00	-.00	-.02***	.04
	[.01]	[.00]	[.07]	[.03]	[.01]	[.00]	[.05]	[.01]	[.00]	[.00]	[.01]	[.03]
ln(employment)	.12***	-.14	.37***	-.33***	.18***	-.10	.30**	-.26***	-.01	-.37***	.01	-.47***
	[.04]	[.09]	[.13]	[.07]	[.06]	[.13]	[.11]	[.06]	[.01]	[.07]	[.01]	[.08]
R-squared	.05	.99	.23	.99	.05	1.00	.18	1.00	.01	.90	.05	.84
Sample Size	546	546	83	83	465	465	83	83	465	465	83	83
Firm FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Year FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes

Notes.–Source: Privacy Rights Clearinghouse (PRC), Department of Health and Human Services (HHS), and Compustat. The table reports the coefficients associated with regressions of logged revenues per employee, logged capital per employee, and total factor productivity on logged records breached under different specifications of fixed effects and on different restrictions of the sample. TFP is computed by taking the residual from a regression of logged revenues on logged employment, logged capital, and logged materials. Standard errors are clustered at the firm-level.

Table 3: Heterogeneity in Cyber Security Outcomes by Breach Type

Dep. var. =	ln(revenue per worker)				ln(capital per worker)				total factor productivity			
	hack	theft	loss	other	hack	theft	loss	other	hack	theft	loss	other
ln(breaches)	.15***	-.62**	-.04	-.25	-.16**	-.14	.02	-.26**	-.02***	-.11*	-.01	-.00
	[.02]	[.23]	[.06]	[.24]	[.05]	[.25]	[.08]	[.09]	[.01]	[.06]	[.01]	[.02]
ln(employment)	.14	.61*	.40***	.41	.71***	.20	.26*	.25*	.05**	.06	.02	.01
	[.17]	[.31]	[.13]	[.26]	[.21]	[.32]	[.13]	[.12]	[.02]	[.07]	[.02]	[.03]
R-squared	.60	.43	.30	.33	.65	.05	.14	.66	.27	.30	.04	.02
Sample Size	9	18	50	14	9	18	50	14	9	18	50	14

Notes.–Source: Department of Health and Human Services (HHS), and Compustat. The table reports the coefficients associated with regressions of logged revenues per employee, logged capital per employee, and total factor productivity on logged records breached under different specifications of fixed effects and on different restrictions of the sample. TFP is computed by taking the residual from a regression of logged revenues on logged employment, logged capital, and logged materials. Standard errors are clustered at the firm-level.

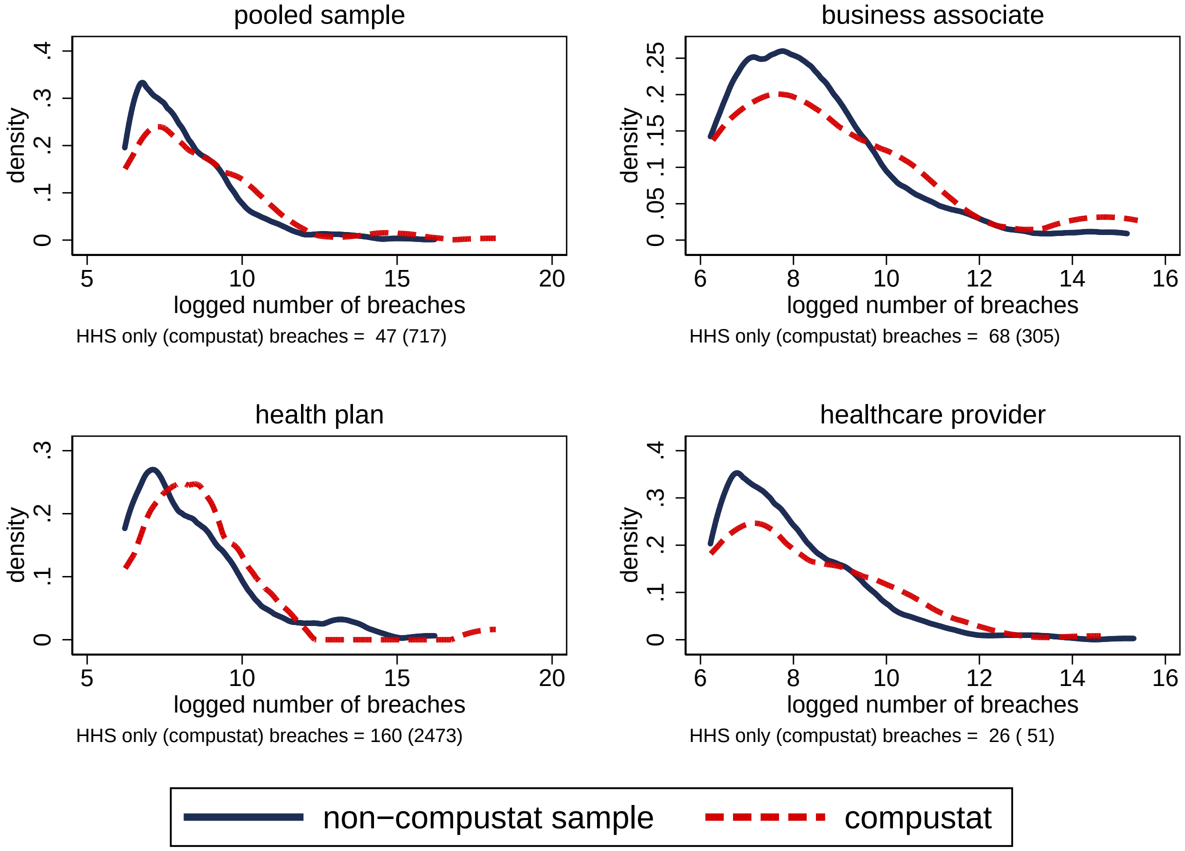

Figure 2: Distribution of Breaches, (non) Compustat and non-Compustat Healthcare Firms

Notes.–Sources: Department of Health and Human Services, 2010-2016. The figure plots the logged number of records breached (in thousands) between firms that are covered by Compustat (i.e., publicly traded firms) and firms that are not covered in Compustat (i.e., private organizations in HHS). The distributions are partitioned into four categories: the pooled sample, breaches where an business associate is hit, breaches where the health plan is hit, and breaches where the healthcare provider is hit. The figure documents big differences in the mean and standard deviation of breaches within firms covered by Compustat, relative to those only in the HHS data, which reflects the fact that many private healthcare companies are smaller practices.

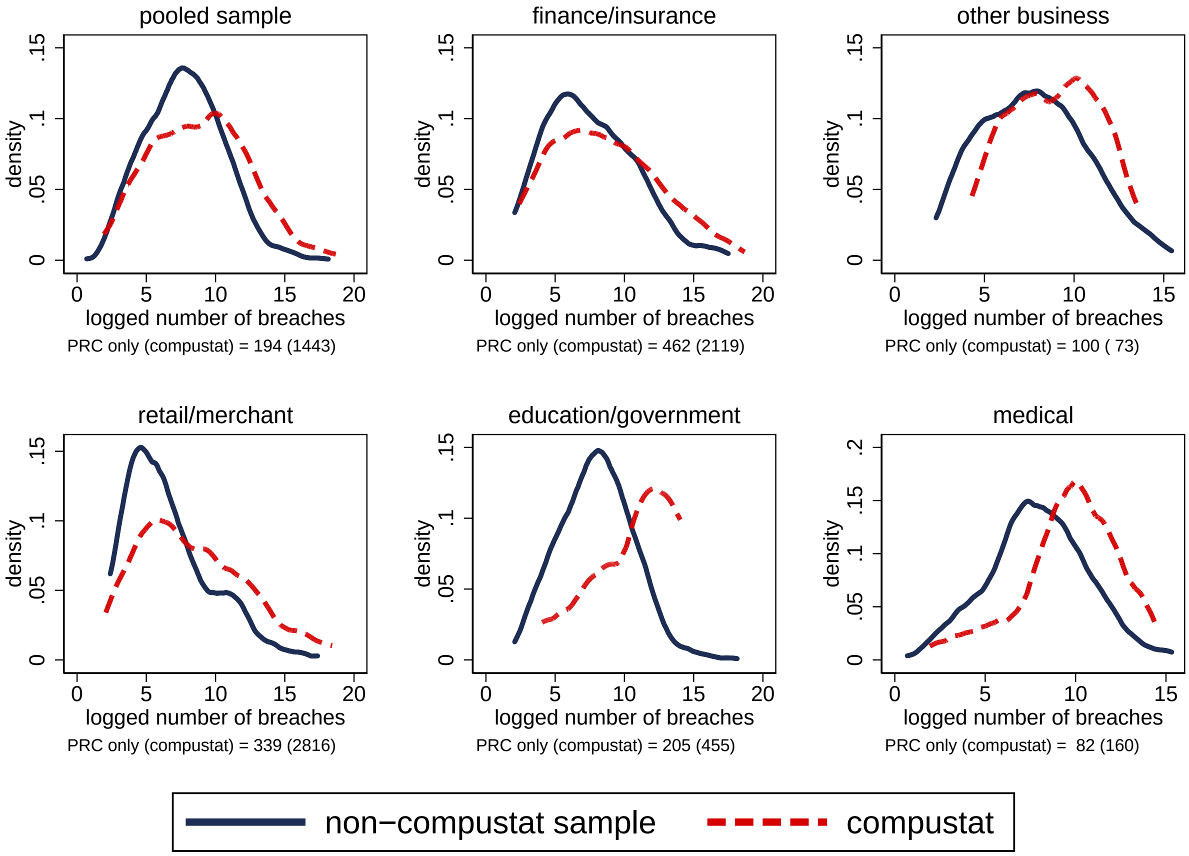

Figure 3: Distribution of Breaches, (non)Compustat and Private Rights Clearinghouse Firms

Notes.–Sources: Privacy Rights Clearinghouse, 2005-2016. The figure plots the logged number of records breached (in thousands) between firms that are covered by Compustat (i.e., publicly traded firms) and firms that are not covered in Compustat (i.e., private organizations in PRC). The distributions are partitioned into six categories: a pooled sample and separate groups by major sector. The figure documents big differences in the mean and standard deviation of breaches within firms covered by Compustat, relative to those only in the PRC data.

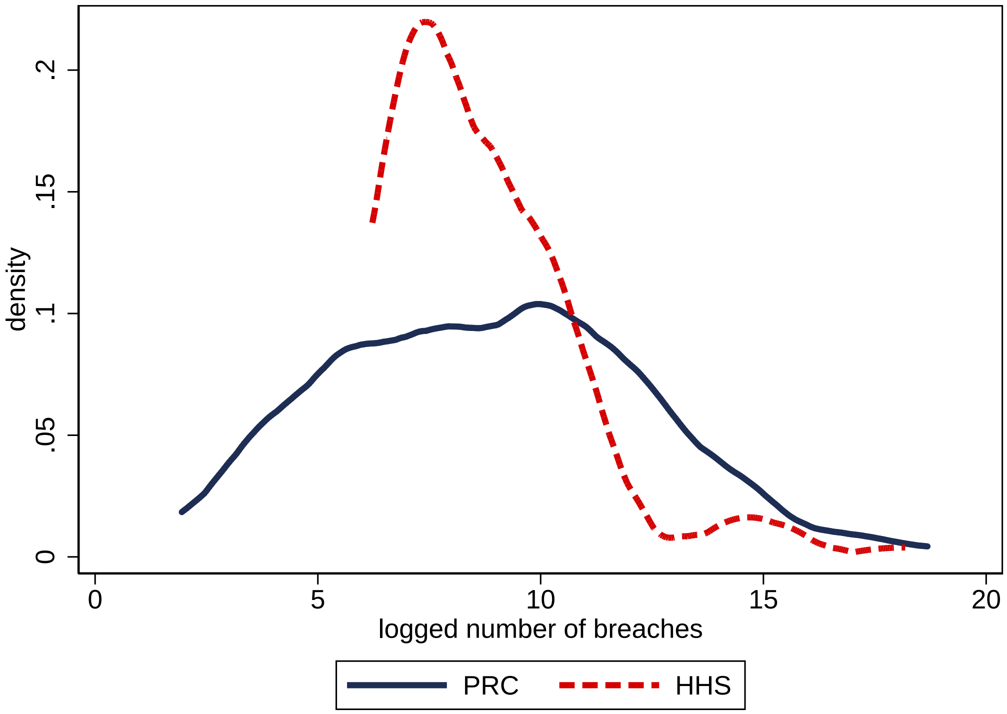

Figure 4: Distribution of Breaches Between PRC and HHS Samples

Notes.–Sources: Privacy Rights Clearinghouse Database (2005-2016) and Department of Health and Human Services (2010-2016). The figure plots the logged number of breaches in both the PRC and HHS datasets.

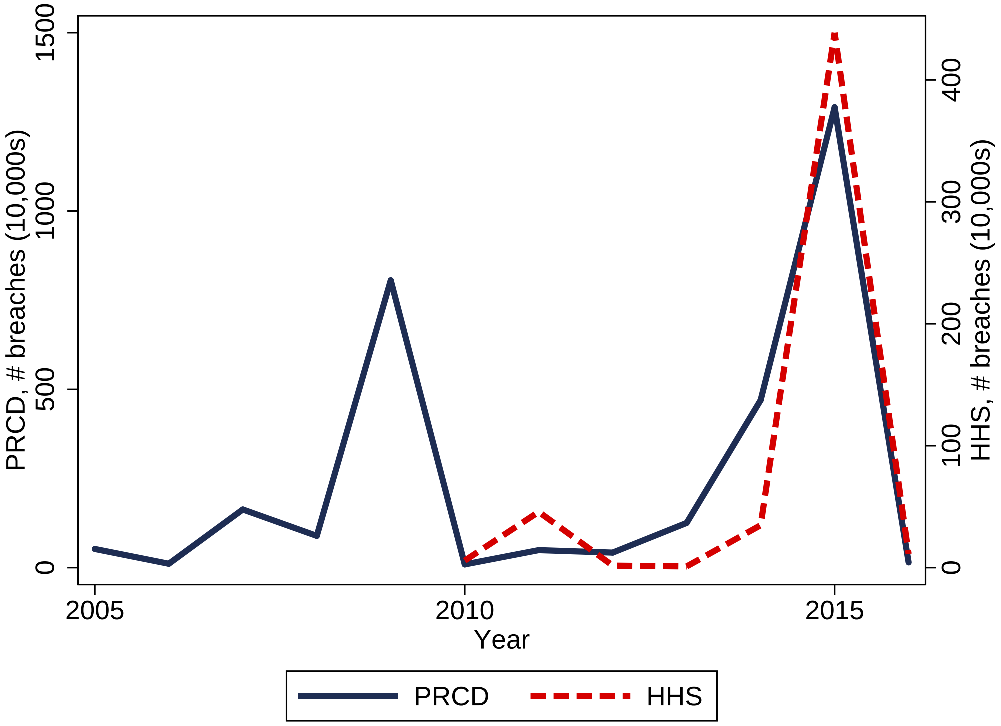

Figure 5: Records Breached (in 10,000s) and Distribution of Breaches, 2005-2015

Notes.–Sources: Privacy Rights Clearinghouse Database Community, and Department of Health and Human Services. The figure plots the mean annual number of breaches in tens of thousands restricted to the set of publicly traded companies in the corresponding databases.

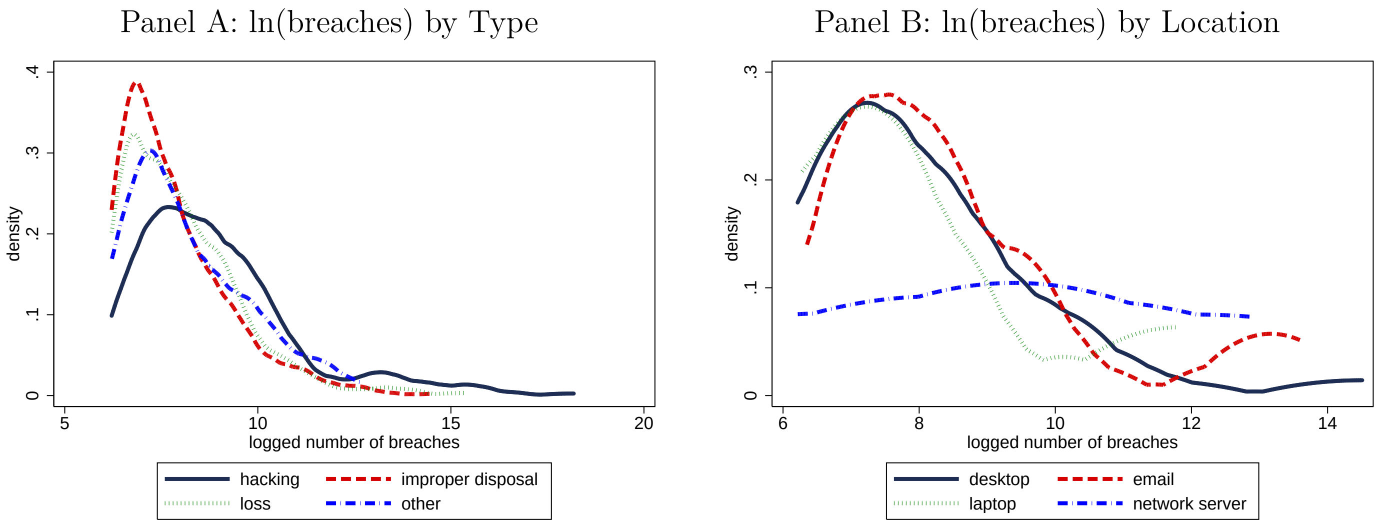

Figure 6: Cyber Security Breach Types and Location in Healthcare Companies

Notes.–Sources: Department of Health and Human Services. The figures plot the logged number of records breached by the type of data breach (i.e., how it took place) and the location of the breach (i.e., on the electronic medium it took place on).

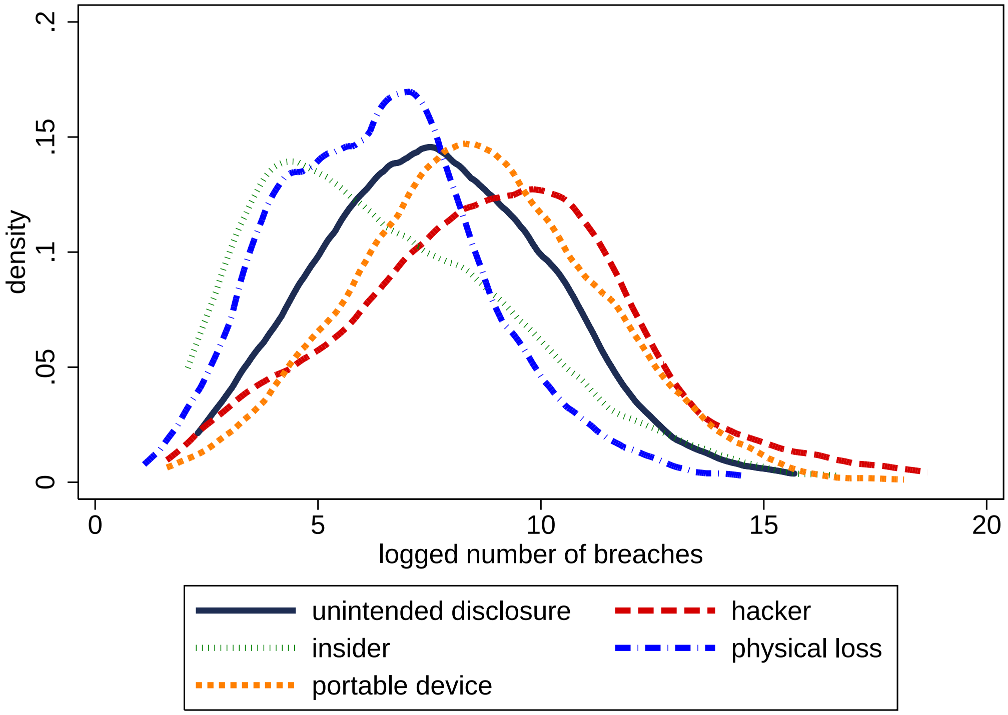

Figure 7: Cyber Security Breach Types Across Sectors

Notes.–Sources: Privacy Rights Clearinghouse. The figure plots the logged number of records breached by the type of breach.