Examining the Costs and Causes of Cyber Incidents

Sasha Romanosky
sromanos@rand.org

DRAFT
DO NOT CITE OR REDISTRIBUTE

Submitted to FTC PrivacyCon, 1.14.16

Abstract

This paper examines a sample of over 12,000 cyber events that include data breaches, security incidents, privacy violations, and phishing crimes. First, we analyze the characteristics of these breaches (such as causes, and types of information compromised). We then examine the breach and litigation rate, by industry, and we identify the industries that incur the greatest costs from cyber events. We then compare these costs to bad debts and fraud within other industries. Public concerns regarding the increasing rates of breaches and legal actions, conflict, however, with our findings that show a much smaller financial impact to firms that suffer these events. Specifically, we find that the cost of a typical cyber incident in our sample is less than $200k (about the same as the firm's annual IT security budget), and that this represents only 0.4% of their estimated annual revenues.

Keywords: cyber attack, data breach, security incident, privacy violation, data breach litigation,

Acknowledgements
I would like to thank James Anderson, Matt Crespi, Lawrence Gordon, Jim Graves, Mark Greisiger, Martin Libicki, Martin Loeb, Trey Herr, Paul Heaton, and Jamie Morikawa for their valuable comments and suggestions. I would also like to thank RAND's Institute of Civil Justice for its generous support.

Introduction

Data breaches, cyber attacks and privacy violations have become commonplace. Yet despite the body of academic literature and media stories concerning these events, there exist no rigorous research that examines a large sample of incidents in order to properly assess the risk and trends of these events.[1] Therefore using a unique dataset of over 12,000 cyber incidents recorded over the years 2004 and 2015, we conduct a thorough analysis of these incidents and examine the costs and composition of these events, by industry, and over time.

This analysis is expected to help inform three main stakeholders: private sector firms, insurance companies, and policy makers. Private sector firms will better understand the risks they face when collecting personal information and operating IT networks that are publicly accessible. Insurance companies can better estimate the risks of their insureds in order to foster a healthy cyber insurance market, and policy makers will better understand the context and impact of these cyber events across industry, and time.

Throughout this research, we distinguish between four types of cyber events: data breaches (unauthorized disclosure of personal information), security incidents (malicious attacks directed at a company), privacy violations (alleged violation of consumer privacy), and phishing/skimming incidents (individual financial crimes).

Of all cyber incidents from our dataset, we find that data breaches are by far the most common, dwarfing rates of all other cyber events. Beyond name and address, we find that credit card numbers and medical information were the most commonly compromised pieces of information. And incidents caused by malicious actions (as opposed to accidental or unintentional activities) have remained relatively constant at around 60% of all incidents. Further, of the almost 1700 resulting legal actions, over 50% continue to be private civil actions brought in federal courts, with only a 17% being criminal actions.

In order to better understand risk by industry, there are a number of potentially relevant metrics, each of which provide a useful, but singularly incomplete, insight. For example, we examine the following metrics for the incidents in our database: total number of incidents, incident rate, litigation rate, total cost, and cost per event.[2] While the Finance and Insurance industry suffers the greatest number of cyber events, Government agencies suffer the highest *incident* rate. Further, Mining and Oil & Gas firms suffer the highest *litigation* rate, while Management firms suffer the highest *cost per event*. Overall, when examining each of the metrics together, we find that the Retail, Information, Manufacturing, and Finance and Insurance industries consistently pose the greatest risk, while -- contrary to common belief – Health Care and Education services pose some of the lowest risks.

Finally, while we estimate the total costs from cyber events at approximately $10 billion annually. We find that the typical cost of a data breach is less than $200,000, far lower than the millions of dollars often cited in surveys (e.g. Ponemon, 2015). Moreover, we find that cyber incidents cost firms only a 0.4% of their annual revenues, much lower than retail shrinkage (1.3%), online fraud (0.9%), and overall rates of corruption, financial misstatements, and billing fraud (5%).

By comparing observed cyber events with the total number of firms within an industry, this research provides one of the first true estimates of firm risk, by industry type. Further, our use of cost data enables

[1] See Edwards (2015) for an excellent analysis of one data breach dataset.

[2] The incident rate represents the number of cyber incidents within a given industry, divided by the number of firms within that industry. Litigation rate is computed as the number of lawsuits within an industry, divided by the number of cyber events that occur within the industry.

us to provide a unique and novel analysis of the scope and magnitude of cyber events, as a function of firm revenues, and other forms of loss, theft, and waste.

Data

Data Source and Data Generating Process

The primary dataset used in this analysis is a dataset of cyber incidents acquired from Advisen, a for-profit organization that collects, integrates and resells many different forms of data for the commercial insurance industry. Advisen employs a dedicated team of analysts who use a variety of search strategies in order to find and classify cyber events. Specifically, Advisen collects news stories from dozens of local and national online news websites, newsfeeds, and vendor partners. It also sources information from specialized legal information services, as well as multiple online data breach clearinghouses. In addition, it collects information from state and federal governments and agencies both from publicly available websites, as well as employing the freedom of information act (FOIA) and their state analogs. While other publicly available data include 4000 - 5000 observations, Advisen manages over 15000 observations, and continues to grow. Therefore, we believe that these efforts have created one of the most comprehensive datasets of cyber events available.

While the search strategies used to collect these data (and any data, for that matter) are thorough, they are not without limitations, as shown in Figure 1.

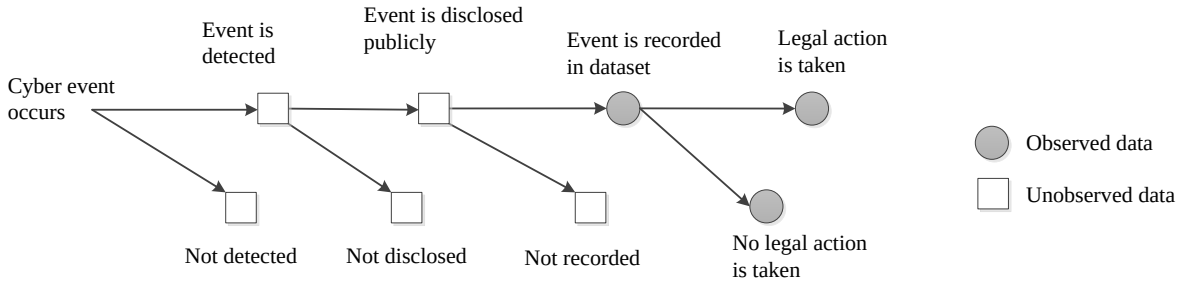


Figure 1: Data generating process

First, the methods used to collect the incidents are limited to publicly available data sources. For example, of all cyber events, only some will be detected (either by the organization suffering the event, or by a third party who, in turn notifies the firm).[3] An important factor to consider is that we do not expect all cyber events to be detected equally. For example, most states have adopted laws that require organizations to notify individuals when their information is compromised (Romanosky, 2011). However, there are no similar laws for privacy violations, or security incidents. This alone suggests a bias in available data toward data breaches, relative to other cyber events.

Conditional on detection, the event may be disclosed to the public. Certainly, firms have an incentive to at least identify attacks, successful or otherwise, against their corporate systems and network if only to stop the attack and reduce any further losses. Our data collection, therefore, is strongly reliant on the effectiveness of information security detection systems, consumer awareness (e.g. in regard to privacy violations), and law enforcement (e.g. in regard to detecting phishing and skimming crimes).

[3] In some cases, it may be law enforcement that first learns of, and notifies the firm. In other cases, it may be a security forensics firm, a security reporter, a credit card processor, or a consumer. The means by which the firm is notified, however, is not relevant for this analysis.

Of the publicly disclosed events, some (we hope most) will be captured by the data collectors and included in our data. Our data do not include events that have not been disclosed to the public, nor do they include events which have been missed by an analyst.[4] For example, very small data breaches affecting only a few individuals would likely not be captured by this (or most any) search strategy. That being said, incidents involving many thousands, millions, or tens of millions of individuals would very likely be captured. Cyber events that are detected by firms, but are willfully ignored may also not be included in these data (unless they were eventually discovered by a third party).

However, if indeed the national policy debate and legal doctrine are most influenced by these medium and large incidents, then the search strategies employed in the creation of these data (even if potentially biased toward larger incidents) would very likely produce a representative sample of the population of reported cyber events, and those in which we are most interested. Nevertheless, we recognize that the inferences made within this manuscript apply strictly to the data being evaluated.

Types of Cyber Incidents

The original dataset used in this analysis distinguished between 11 separate types of cyber events – too large for practical analysis. Given that many events shared similar fundamental characteristics, some categories were aggregated based on a combination of approaches: the type of event; whether the event was *caused by*, versus *suffered by*, the firm; industry convention as used in previous academic research and security reports; and conversations with security and privacy experts. Therefore, we have organized the data according to the following categories which are meant to be exhaustive and mutually exclusive.

Data breach: the unintentional disclosure of personally identifiable information (PII) stemming from loss or theft of digital or printed information. For example, the theft of laptop or desktop computers containing personal information of employees or customers of a firm, caused either by a hacker, or malicious employee. This category also includes the improper disposal or disclosure of personal information (i.e. to a dumpster or website).

Security incident: an incident involving the compromise or disruption of corporate IT systems (computers or networks) or its intellectual property. For example, a denial of service (DoS) attack, the theft of intellectual property, the malicious infiltration (hack) and subsequent cyber extortion of corporate information, or a disruption of business services.

Privacy violation: the unauthorized collection, use or disclosure of personal information. For example, unauthorized collection from cell phones, GPS devices, cookies, web tracking, or physical surveillance. Allegations of violations of information protection statutes such as Drivers Privacy Protection Act (DPPA), Video Privacy Protection Act (VPPA), Telephone Consumer Protection Act (TCPA), Children's Online Privacy Protection Act (COPPA), Do-Not-Call, Song-Beverly Act, and the Privacy Act. Also includes unsolicited communication from phishing emails, spam, other mass marketing communication (robocalling, texts, emails), or debt collection.

The first two categories are differentiated from the third in that the first two relate to incidents *suffered by* the firm (i.e. PII stolen from the firm, or the firm suffering a compromise of business operations because of a hack), while the third category relates to events *caused by* the firm (e.g. the firm improperly collecting or selling personal information).

[4] There are a number of reasons why a breach or cyber event would not be disclosed to the public. First, while most states have breach disclosure laws, many provide reporting exceptions such as a breach that affect only a few individuals, when the stolen information is encrypted, etc.

Phishing / Skimming: The final category relates to instances of individuals committing particular kinds of computer or electronic crimes directly against other individuals or firms. For example, these crimes would include phishing attacks (wherein criminals seek to harvest account information from users), identity theft (wherein criminals use another person's information for financial gain), or skimming attacks (where criminals install, for example, a hardware device over ATM machines in order to copy bank account and bank PIN numbers).

Descriptive Analysis of Cyber Events

Figure 2 shows the absolute number of incidents across the four groups with data breaches displayed in the left panel, and all others (privacy violations, security incidents and phishing/skimming incidents) displayed in the right panel. All figures reflect data over a 10 year period from 2005 to 2014. Note the difference in scale between the left (0-1500) and right (0-250) panels.

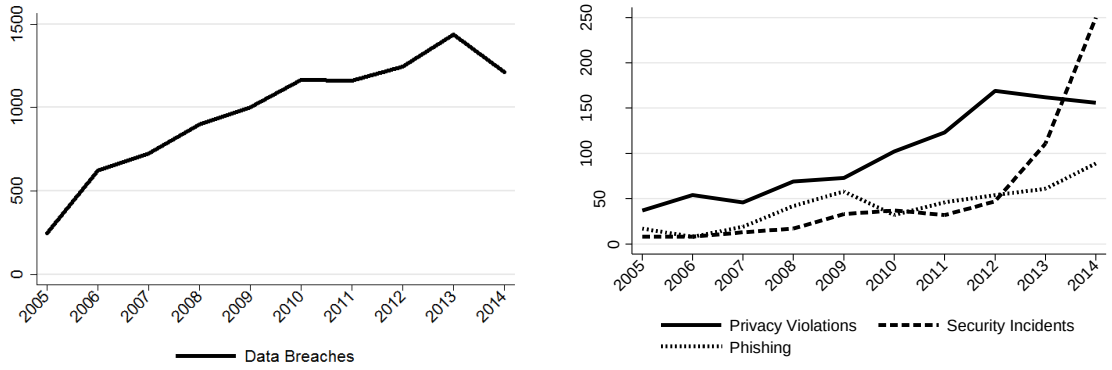


Figure 2: Four types of cyber events

These data show how some events have been increasing steadily since 2005, though at a decreasing rate. However the count of reported data breaches is in some cases an order of magnitude larger than other incidents. For example, data breach reports have seen a four-fold increase from just over 200 in 2005 to over 1200 in 2014, while privacy violations have seen only a modest increase from dozens in 2005, to around 150 by 2014. The rise in data breaches may be, in part, due to some states adopting breach disclosure laws in later years. However, for most practical purposes, the early adopting states (such as California) could well have provided incentive for most all firms to report, regardless of the state of the individuals affected by a breach. The steady increase in privacy violations is likely due to the national attention that early privacy events and alleged violations occurred (such as Google Street View, Facebook's rise in popularity, surveillance movement, behavioral advertisment tracking, etc).

Security incidents, on the other hand, have seen a very sharp rise since 2012 (64) rising to almost 250 by 2014. The underlying cause is unclear given that there has been no substantial policy or industry interventions which would drive material changes in reporting. Absent such reporting changes, one might conclude that the rise is due to an increase in actual events. i.e. that firms are simply suffering more security intrusions.

Next, we examine incidents and incident rates by NAICS industry, as shown in Figure 3.[5] (See Appendix for further descriptions of firm types by industry). The left panel identifies the most frequent incidents by

[5] Note that the NAICS industry coding, while ubiquitous, provides only one segmentation of firms by industry and may not reflect one's intuition of a firm's industry. For example Apple Inc is coded as a Manufacturing company,

industry, while the right panel illustrates the incident *rate* (i.e. number of cyber incidents divided by the total number of firms within that industry). That is, the left panel shows the total count of incidents, while the right panel shows the percentage of firms within an industry that suffer an incident.[6] See Appendix Table 4 for a full description of all industries, as well as the number of firms per industry.

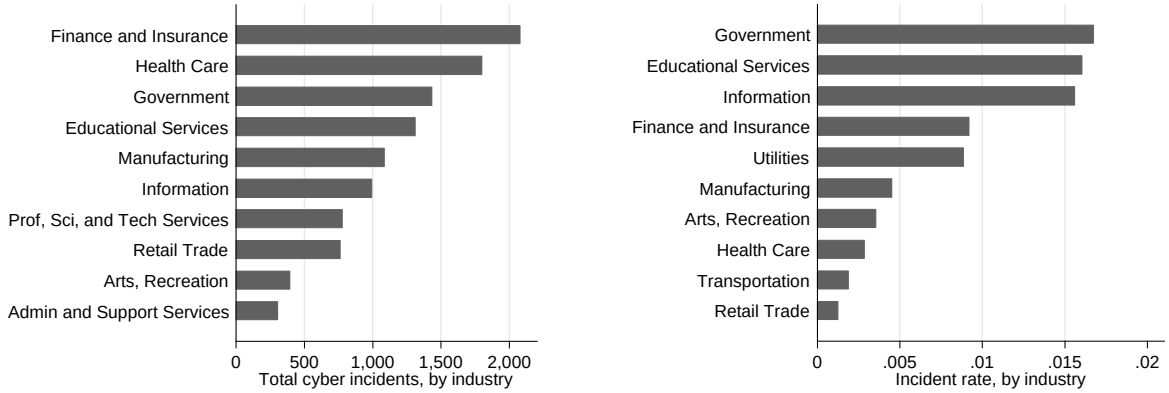


Figure 3: Cyber incidents, and rates, by industry

The results in the left panel show that Finance and Insurance (e.g. insurance carriers, credit intermediaries), Health Care (hospitals, ambulatory care), and Government entities (courts, police, administrative offices, etc) suffered the highest number of reported breaches of all industries in our dataset, followed by Educational services (schools, univesities, and supporting services), Manufacturing (e.g. computer and electronic manufacturing), and Information services (e.g. data processing and hosting services, software vendors, telecommunications companies, internet portals, etc.), respectively. The large number of incidents of Health Care and Finance and Insurance matches other survey results which found that the healthcare and financial services sectors suffered the largest percentage of breaches from their sample (NetDiligence, 2014). However, care should be taken to not immediately conclude, simply based on total incidents, that these industries pose the greatest risk of a cyber event.

For example, when we consider incident rate in our dataset (left panel), we find that Government agencies, and firms within the Education and Information services industries are affected at a much higher rate (> 1.5%) compared with all other industries. That is, about 15 out of every 1000 firms have suffered a *reported* incident, while firms within the Finance and Insurance, and Utilities industries suffered cyber incidents at a rate of about 9 out of every 1000 firms. Note that Health Care and Retail industries, however, suffer extremely low incident rates of around 0.3% or less.

It is certainly true that not all cyber events are detected, and even if detected, they may not be recorded in the dataset (see Figure 1). It is also true that there currently exists no reliable way to estimate the number of *unknown* cyber events. However, given that there are approximately 6 million US firms,[7] claims by some that "there are two kinds of companies: those who know they have been hacked, and those who

---

and health insurance companies are categorized under Finance and Insurance, rather than Health Care. SIC codes are an alternative method for segmenting industries, but NAICS codes were chosen at the discretion of the author.

[6] The US Census collects many industry-level data, including the total number of firms, establishments, employees, and payroll. The number of public agencies was collected from the dataset: Federal, state, and local government unit: http://www.census.gov/govs/cog/, for 2012.

[7] See data from the US Census at http://www.census.gov/.

don't yet know they've been hacked" are, despite being a catchy media slogan, very likely false because it would suggest that all 6 million firms had, indeed suffered a data breach or cyber incident of some kind.[8]

Additional descriptive statistics are shown below in Table 1.

Table 1: Descriptive Statistics

Variable	N	Mean	SD	Median	Min	Max
Records compromised	1201	2.39 m	19.2 m	100	1	400 m
Employees[9]	10929	20491	123 k	300	1	2.8 m
Revenues (millions)	9360	8031	30373	64	0	484 b

Type of information compromised		
Name (%)	12574	43.84
Credit card (%)	12574	24.90
Address (%)	12574	18.93
Medical (%)	12574	19.38
Financial (%)	12574	17.07
Date of birth (%)	12574	21.73
Email (%)	12574	14.81
SSN (%)	12574	3.07
Drivers License (%)	12574	11.75

Firm type	**No.**	**Percent**
Government (%)	1,762	14.0
NonProfit (%)	401	3.2
Privately Held (%)	7,072	56.2
Publicly Traded (%)	3,350	26.6

Cause			**Malicious?**
Disclosure/Disposal (%)	2,841	22.70	N
Espionage, Extortion, Fraud, Bribery (%)	504	4.03	Y
Hack/DDoS (%)	2,132	17.03	Y
Insider (%)	1,342	10.72	Y
Lost HW (%)	701	5.60	N
Phishing (%)	215	1.72	Y
Stolen HW (%)	3,541	28.29	Y
Unauthorized Use or Collection (%)	970	7.75	n/a

Next we examine the rates of incidents in our dataset according to kinds of information compromised. Specifically, Figure 4 shows that the rates of compromise of more personal information (and those that could lead to greater consumer harms) is increasing with time, and particularly, the number of cyber events that involve medical information has risen most sharply. These trends are particularly concerning given that these data are more difficult (or in the case of medical data, impossible) to change, and therefore individuals suffering a compromise of these data are argueably more at risk of financial, medical and other forms of fraud and identitytheft.

[8] This claim was self-attributed by Dmitri Alperovitch at a public symposium in Washington D.C. on September 10, 2015.

[9] i.e. employees and revenues related to the firm affected by the cyber event.

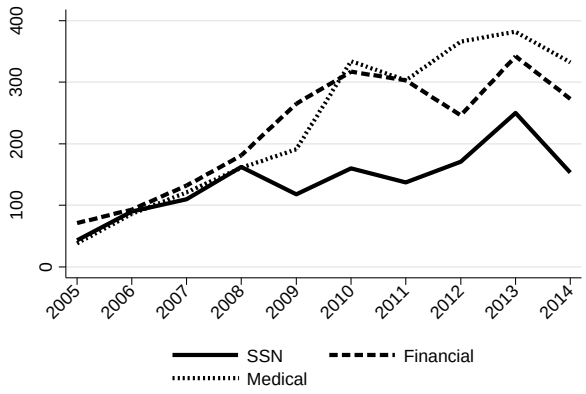


Figure 4: Cyber events by type of information compromised

In addition, contrary to other findings suggesting that the rate of malicious incidents has been increasing in recent years (relative to accidental ones), we find a relatively stable proportion of malicious events. Specifically, surveys conducted by the Ponemon Institute have shown that these rates have been increasing from 37% in 2011, to 44% in 2014 and 49% in 2015 (Ponemon, 2011; Ponemon, 2014; Ponemon, 2015). By grouping the causes of incidents according to whether the cause was malicious or not (that is, intentionally committed to cause harm, as shown in Table 1), the rates of malicious incidents from our dataset are presented in Figure 5.

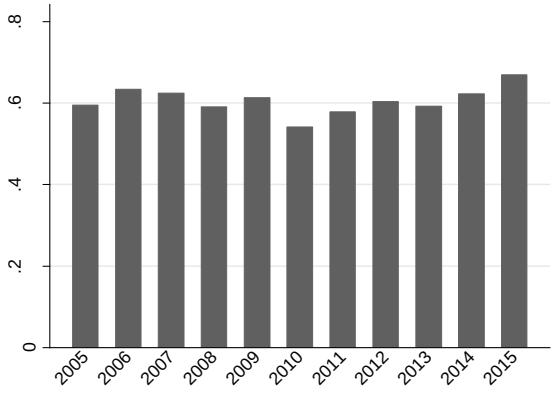


Figure 5: Rates of malicious events

Based on our data, this suggests that the rates of malicious events has remained quite steady at 60% over the past decade. That being said, the slight increase in the most recent years may, indeed, be due to the increase in security incidents and attacks against corporate systems.

Legal Actions

Figure 6 illustrates the composition of legal actions is distinguished along three dimensions: civil vs criminal, federal vs state, and private vs public actions. As fully described in Romanosky et al (2014), private actions typically reflect cases brought by individuals against firms for the unauthorized disclosure (or use) of their personal information.[10] Similarly, public actions (whether federal or state) are brought by

[10] For example many arguments in data breach cases include those similar to the ones presented in Hilary Remijas v. Neiman Marcus Group, LLC, (No. 14-3122), "(1) lost time and money resolving the fraudulent charges, 2) lost time and money protecting themselves against future identity theft, 3) the financial loss of buying items at Neiman

government agencies (e.g. FTC, state AGs) against companies for their allegedly inferior security practices or careless handling of personal information. On the other hand, the criminal actions are brought by State prosecutors against alleged perpetrators of the crimes and reflect a different category of lawsuit. The type of action is identified along with the number of observations by category, in parentheses. The weight of the line roughly corresponds to the relative number of observations from our dataset.

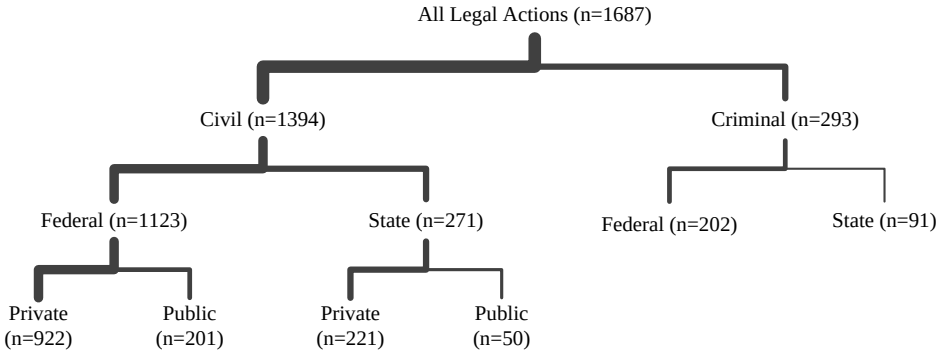


Figure 6: Composition of legal actions

Of the almost 1700 actions, about 83% are civil, 79% are federal, and 68% are private. Of just the civil actions, notice that there are about 4 times as many federal as state actions (1123 vs 271), and of those federal, almost 5 times as many private suits relative to public enforcements (922 vs 201). Of the 271 state actions, there are over 4 times as many private actions as there are public (221 vs 50). Criminal actions make up 293 of all legal actions, with the majority of them being prosecuted in federal court.

Next, we examine the frequency of actions by year of filing as shown in Figure 7. The left panel illustrates the count of *private* lawsuits brought in federal and state court, while the right panel shows the count of *public* (civil and criminal) actions brought in federal and state court from 2005-2014.[11] We limit this examination to years beginning in 2005 since this was objectively the first year that systematic reporting and recording of data breaches (and therefore subsequent litigation) began.[12]

These histograms help illustrate the change in number of lawsuits filed over time. First, note that the number of public, federal actions has been increasing since 2005 (upper histograms of the right panel), and greatly outnumbers public state actions (lower histograms of the right panel). However, the majority of all legal actions for cyber events are driven by private (civil) actions filed in federal court (left histogram of left panel), and the number of suits has been increasing steadily from a couple of dozen in 2005, to almost 200 by 2014.

Marcus that they would not have purchased had they known of the store's careless approach to cybersecurity, and 4) lost control over the value of their personal information."

[11] Note that approximately 200 suits were omitted from this analysis because of multiple suits being filed for the same cases, most of which are eventually consolidated into multi district litigation (Romanosky, 2014).

[12] Data breach disclosure laws first began in 2003, but it was not until 2005 that additional states began adopting similar laws (Romanosky, 2011). Further, it was not until 2005 that other non-profit organizations such as DatalossDB, and the Privacy Right Clearinghouse began systematically recording and reporting these events.

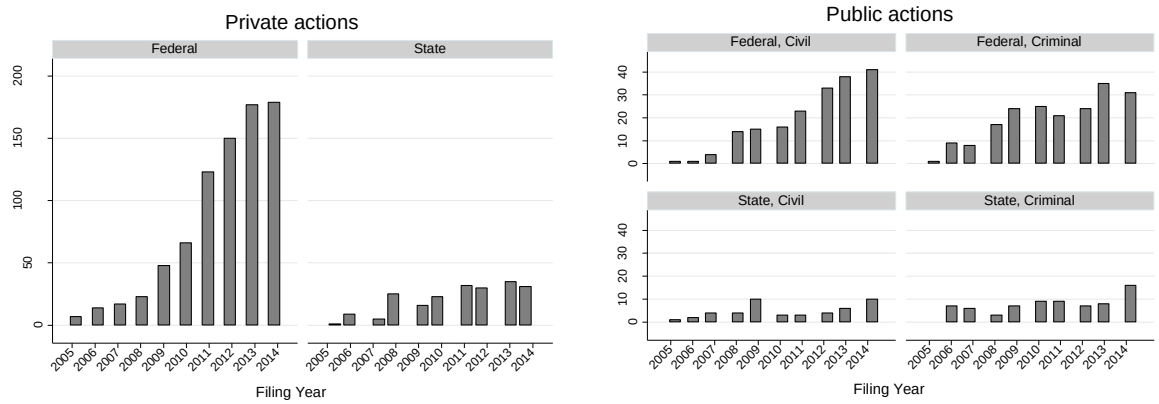


Figure 7: Count of private and public legal actions by year

An important qualifier for these data is that state lawsuits are typically much more difficult to observe and track relative to federal suits. And so it is possible that the low number of observed private state suits is largely a function of their absence from our dataset rather the absence of such cases.

Even though federal civil and criminal actions in our dataset have been increasing since 2005, they are clearly not increasing at the same rate as private actions, nor are they likely to reach such levels. The reason is simply that public agencies are much more resource constrained, suggesting that public agencies must therefore rely on suits brought by private individuals in order affect change.

Litigation rates

Next we examine the rate of legal actions stemming from the cyber events in our sample. The left panel of Figure 8 shows a count of legal actions by event type, while the right panel shows the litigation rate, also by event type. That is, the ratio of lawsuits filed in a given year to the number of events occurring in that year.[13]

The left panel, lawsuits, shows that while data breach lawsuits have remained relatively stable at just over 50 per year, lawsuits regarding privacy violations have been increasing dramatically since 2005 and especially since 2009, reaching as many as 150 suits per year in our sample. Actions for security and phishing incidents, however, have experienced very little increases since 2005.[14]

Despite sharp increase in cyber events in recent years (see Figure 2), the litigation rate for all cyber events has been generally decreasing in our sample. For example, the litigation rate for data breaches was around 20% in 2004, but has fallen to about 5% in 2014. This substantiates a similar litigation rate as found by Romanosky et al (2014).

[13] Given that lawsuits are not always filed within the same year of the event, this formulation of the litigation rate is an approximation.

[14] Note that these values may be slightly under-representative of the actual number, since we are only considering unique suits for a given incident.

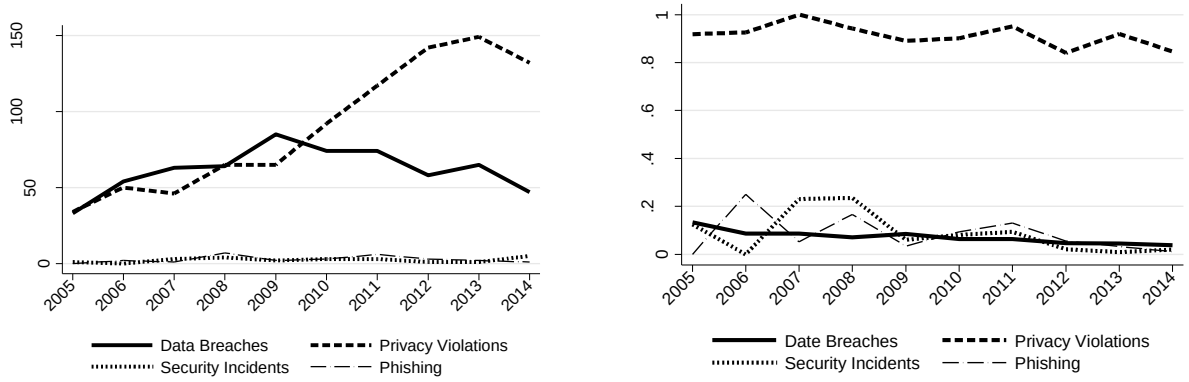


Figure 8: Legal actions and litigation rates, by type of cyber event

Litigation rates for privacy violations, however, are dramatically different, suggesting an overall average of around 85%. This is likely an overinflated result, driven by artifacts of data collection. The most reasonable explanation is that privacy violations, themselves, are only observable once a lawsuit has been filed, whereas events such as data breaches would likely not suffer from this because of the state breach disclosure laws which provide a legal obligation for firms to report these events to affected individuals (Romanosky et al., 2011)

Legal actions by organization type

While Figure 3 displayed rate of cyber events by industry, Figure 9 provides more information regarding litigation rate by entity type (left panel) and industry (right panel) in our sample. From these figures, we observe that while privately held firms experience the largest total number of breaches (approx. 7000) and lawsuits (over 1000), publicly traded companies face the largest litigation rate (643 lawsuits / 3300 breaches = 19.4% litigation rate). It is likely unsurprising to observe that government agencies and non-profit companies experience relatively small litigation rates (approx. 8% and 9%, respectively).

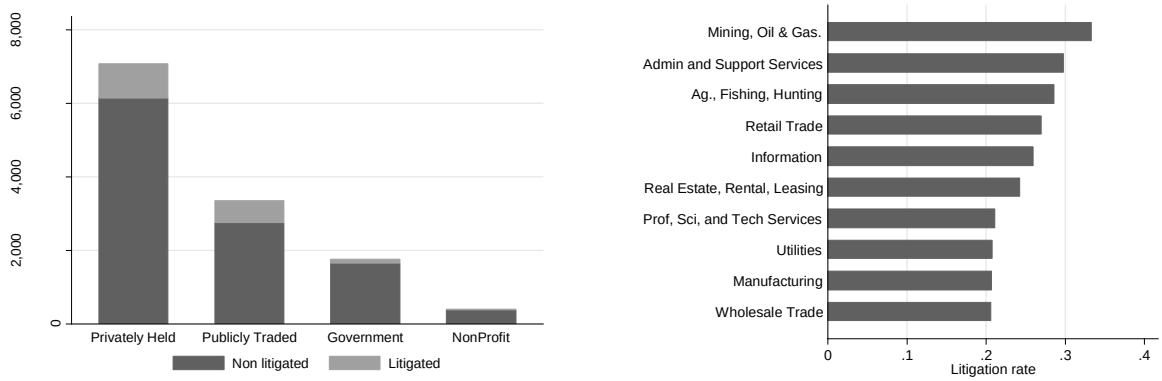


Figure 9: Litigation rate by organization type, and industry

The right panel of Figure 9 shows the top 10 litigation rates, by NAICS industry. Note that while entities in the Government and Educational sectors suffer the highest *breach* rates (as seen in the right panel of Figure 3), they suffer *litigation* rates less than 10% (not shown). Contrary to conventional thought, however, Finance and Insurance, and Health care companies also suffer relatively low (around 10%)

litigation rates. On the other hand, Administrative and Support Services (e.g. telemarketing, collection agencies and credit bureaus, etc.) Retail firms, and Information Services (e.g. software vendors, telecommunications companies, internet portals, etc.), in our dataset suffer litigation rates over 25%, while the Mining and Oil & Gas industry (mining, petroleum, gas extraction, drilling, and supporting activities) suffers the highest litigation rate of all other industries (i.e. more than 30% of all cyber events are litigated).

Cost of Cyber Events

The costs incurred by cyber events can largely be differentiated among first and third party losses. First party losses relate to expenses the firm incurred as a direct result of the incident. For example, in the case of a data breach, this would include the cost of forensic investigation in order to determine the cause, the cost of notifying affected consumers, marketing or public relations campaigns, customer support efforts, and any consumer redress in the form of credit monitoring or identity theft insurance. Third party losses, on the other hand, relate to costs incurred due to private litigation (e.g. class action lawsuits, judicial rulings, settlements or *cy près* awards[15]), or fines or fees brought by government agencies. In the case of security incidents or phishing/skimming scams, losses may include the dollar value of any financial theft.

There are a number of important qualifiers regarding the cost figures that follow. First, in some cases, they represent estimated or alleged losses, rather than actual, verifiable costs incurred. Second, they are incomplete in that they likely do not include all costs borne by firms due to these cyber events. They also only represent a small percentage of the total observations within this dataset,[16] which is in itself only a subset of all publicly reported breaches, which is a subset all known events, which is finally only a subset of all actual cyber events (See Figure 1). Third, they do not include the costs borne by consumers due to identity theft, other sorts of financial, medical, or privacy harms, out of pocket expenses. Finally, they do not include other social costs or externalities borne by other parties because of these events. Therefore, these costs represent a sample (albeit large) of estimated costs incurred by firms, and in some cases, third parties, due to a cyber event. Summary statistics are shown in Table 2.

Table 2: Cost by event type (in millions)

Event Type	N	Mean	SD	Median	Min[17]	Max
Data Breach	602	5.87	35.70	0.17	0.00	572
Security Incident	36	9.17	27.02	0.33	0.00	100
Privacy Violation	234	10.14	55.41	1.34	0.00	750
Phishing	49	19.99	105.93	0.15	0.01	710
Total	921	7.84	47.28	0.25	0.00	750

Assessing and predicting the costs of data breaches has been a struggle for many years because of the lack of quality data. And naturally, many organizations have an interest in better understanding these costs, for example, firms at risk of suffering breaches, insurance carriers, researchers, and social planners. Based on recent survey data, current estimates present the average cost of a data breach at around $5 million (or,

[15] In cases of data breaches, money from the defendant is sometimes allocated to assist with identity theft education and awareness, or to fund research in data protection or consumer privacy.
[16] Cost data are available for only 921, or 7.3% of all observations.
[17] Values are presented in millions of dollars and therefore, any zero values are artifacts of rounding functions.

$217 per record; Ponemon 2015). However, given the heavily skewed cost distribution from these data, use of the statistical mean as a measure of the cost of a data breach (or cyber event) is misleading. As shown in Table 2, while the mean loss for a data breach is almost $6 million, the median loss is only $170k. Similarly skewed values arise for phishing and security incidents. Privacy violations, however, account for a much larger median loss of $1.3 million, but is also greatly skewed.

Similar cost results for data breaches were found using a remarkable data set of actual cyber insurance claims data which finds median claim payouts of $144k, and mean payouts of almost $3 million for large companies (NetDiligence, 2014).[18] The similarity between estimates is comforting and provides some validation of the cost data used in this analysis. That is, if the insurance claims account for as much of the liability and financial loss suffered by the firm as possible, then one would welcome a strong correlation between those data, and the cost estimates from our dataset

Next we examine the change in total costs, by event type, over the 10 year period from 2004 to 2014 as shown in Figure 10. Given that the raw data exhibit extreme annual fluctuation, we therefore display smoothed lines in order to provide directional insights.[19]

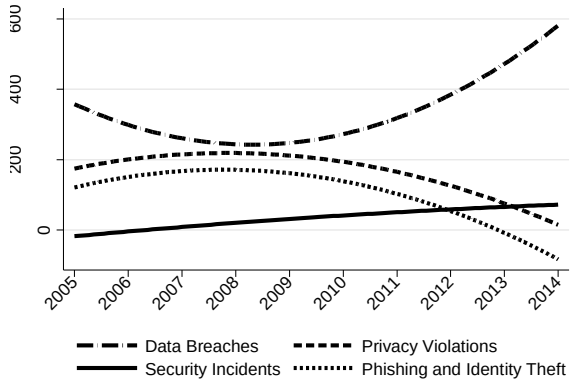


Figure 10: Total cost of cyber events over time

First, we observe that total costs for privacy violations and phishing attacks have been generally decreasing since 2008, as illustrated by the middle two dashed curves. On the other hand, costs from security incidents have been increasing steadily since 2005, though only at a moderate rate. Total costs for data breaches, however, while declining slightly from 2005 to 2008, increased steadily and dramatically after 2008. While this is in part driven by the increasing number of data breaches, it provides evidence showing how overall costs from these cyber events are, indeed, rising.

Costs by industry

Next we examine the losses by industry in our dataset, as shown in Figure 11. Note that for brevity, we only plot the top 10 industries. The left panel identifies the top 10 industries that suffer the greatest losses as a result of cyber events. However, because these data may be driven simply by the number of cyber

[18] These data are remarkable because it is extraordinarily difficult to obtain actual claims information.
[19] Note that it is this smoothing process which produces values less than zero. Clearly negative costs are not observed in our data.

events within each industry, the left panel divides the loss by the number of events (that is, it shows the total loss divided by the number of events in our dataset).[20]

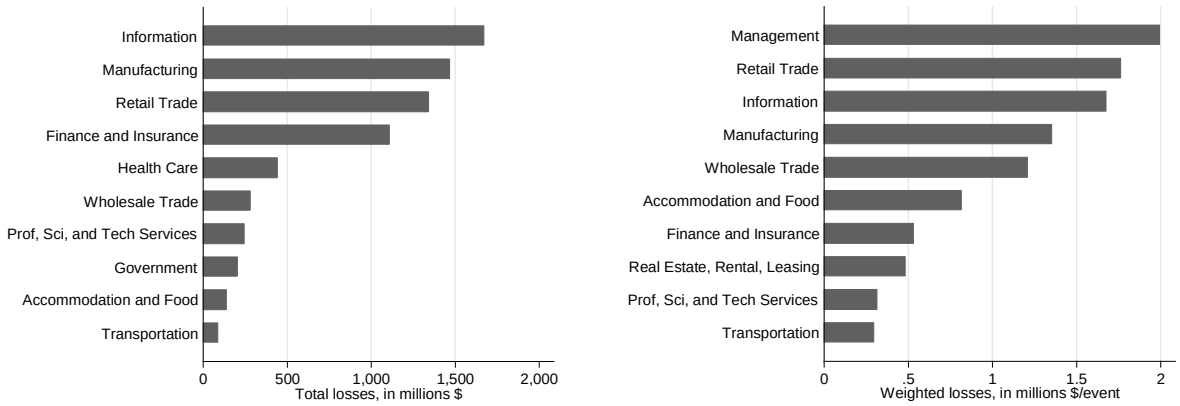


Figure 11: Losses by industry

The two panels show that overall, the Information, Manufacturing, and Retail industries suffer the greatest losses relative to any other industry, as well as the greatest loss per event. This additional measure is useful in better understanding in which industry the greatest losses exist, and therefore which industries pose the greatest risk to firms, investors, employees, and potentially consumers.

An important conclusion from this work is that while Information and Retail industries incur the greatest losses, they also seem to incur the greatest risk from legal action. Further, despite being the most heavily regulated in terms of information security controls, firms within the finance industry do not appear to be better able to prevent or mitigate losses or cyber attacks materially better relative to other industries.[21]

Cost by first and third party losses

Next we examine these same losses as a function of first and third party costs, as shown in Figure 12.[22]

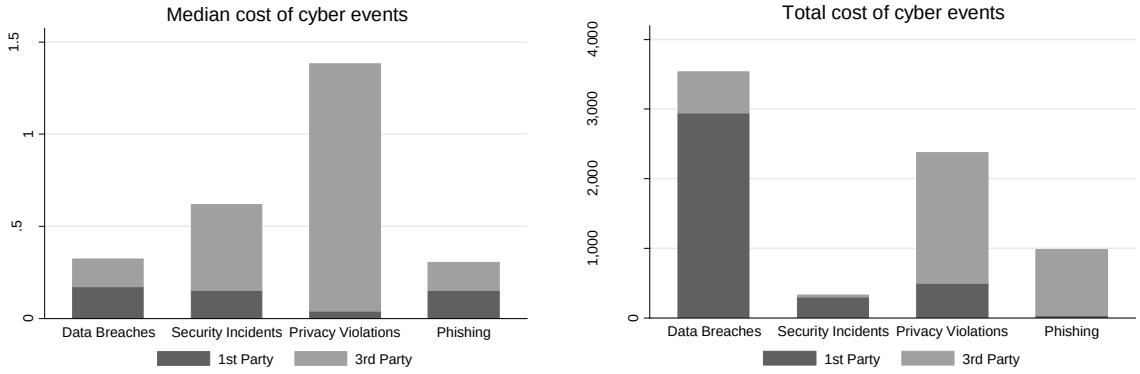


Figure 12: First and third party costs from cyber events

[20] Note that we omit 6 observations that are extreme cost outliers.
[21] Of course, we do not observe the number or severity of attacks launched against these firms.
[22] Note that we have dropped one privacy violation observation because of an extreme outlier.

The total losses are strictly a function of available data, and represent information of only 7.3% of the observations in the data (i.e. 921 out of 12,600 observations). If the available data are representative of the full population of all (reported) cyber events,[23] this suggests that the total cost of cyber events (between 2005-2015) was approximately $10 billion, or about $10 billion annually.[24] Given that the US GDP in 2013 was approximately $16.8 trillion, annual losses represents approximately 0.06% of GDP.

By comparison, Figure 13 shows other losses plotted on a cardinal scale ranging between $0 - $250 billion.[25]

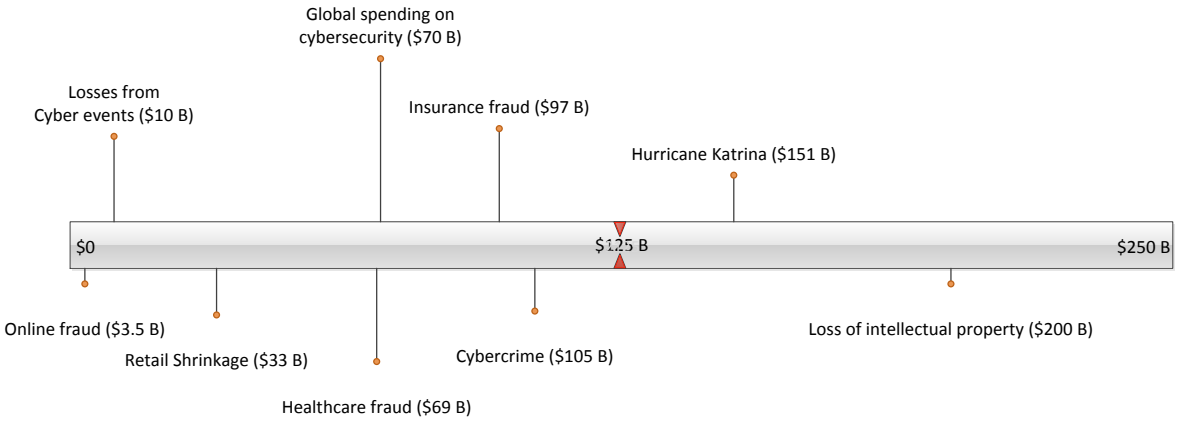


Figure 13: Relative costs and losses (annual, in billions of US dollars)

Notice that in relation to these other estimates of conventional losses, cyber events from this dataset are small.

Modeling the costs of data breaches and privacy violations

Of particular interest to insurance carriers and, indeed, firms, is to be able to develop useful predictive models concerning the costs of data breaches and other cyber events. However, very little empirical research has been conducted, and the work that does exist provides only basic insights. For example, using Ponemon data, Jacobs (2014) identifies the following model using only cost and size (i.e. number of records compromised),

$$log(impact) = 7.68 + 0.76*log(records) \qquad Eq\ 1.$$

Where *impact* refers to the cost of a data breach, and *records* refers to the number of records compromised. The interpretation of this simple regression equation is that as the number of records increases by 10 percent, the cost of a data breach would increase by 7.6 percent. While limited, this model is informative and shows an increasing linear trend between size of a breach and the incurring costs. In related work, Edwards et al. (2015) use data collected from the Privacy Rights Clearinghouse to create a

[23] This is obviously an approximation, but provides one means to estimate the annual cost.
[24] Clearly, this is an estimated value based on the observations for which cost data are available, and the approximately 12,600 total observations. Further, this is likely an underestimate of the true value given that some breaches and security incidents are not reported, and that it does not include consumer or other social losses.
[25] Sources, cost of insurance fraud and hurricane Katrina (https://www.erieinsurance.com/about-us/insurance-fraud/cost, https://www.ncdc.noaa.gov/billions/events), others (CSIS, 2014).

model associating stratified sizes of breaches with the probability of those occurrences. Unfortunately, based on data limitations, they do not model costs.

However, given our rich set of data, we are able to develop a more comprehensive model for cyber incidents that helps better understand the relevant factors driving costs. For example, we consider the following estimating model,

$$log(cost_{it}) = \beta_0 + \beta_1 * log(revenue)_{it} + \beta_2 * log(compromised)_{it} +$$

$$+ \beta_3 * malicious_{it} + \beta_4 * lawsuit_{it} + \alpha * FirmType_{it} + \lambda_t + \lambda_{ind} + \mu_{it} \qquad Eq.\ 2$$

Where *cost* is the total cost of the incident incurred and caused by firm *i* in year *t*. *revenue* is the log of the firm's revenue. *compromised* is the number of compromised records from the incident, *malicious* is a binary variable coded as 1 if the event was caused by malicious intent, and 0 otherwise. *lawsuit* is a binary variable coded as 1 if a legal action resulted, and zero otherwise. *FirmType* is a vector of binary variables describing whether the affected firm was a government agency, non profit, privately held company, or publicly traded.[26] We also include vectors of year and industry binary variables, represented by λ_t and λ_{ind}. μ_{it} is the error term, assumed to be uncorrelated with the covariates.

The results from Eq. 2 are shown in Table 3. Because of data limitations, we examine only data breaches and privacy violations. The combined event types are shown in the pooled estimates from Model 1. However, one may be concerned that results may vary systematically between data breaches and privacy violations, and so Models 2 and 3 separately estimate Eq. 1 using only data breach and privacy violation data. Note that while there exists over 10,000 data breaches and over 1000 privacy violations in the full dataset, the presence of many missing observations across multiple variables reduces the estimating dataset considerably. Therefore, the results below represent preliminary results only.

Table 3: Regression Results for Eq 2.

Dep var: log(cost)	(1) Both	(2) Data Breaches	(3) Privacy Violations
Log(revenues)	0.106**	0.133**	0.0570
	(0.0452)	(0.0524)	(0.0979)
Log(compromised)	0.255***	0.286***	0.245***
	(0.0263)	(0.0371)	(0.0386)
Malicious	-0.913***	0.0190	-4.554***
	(0.301)	(0.373)	(1.577)
Lawsuit	0.669**	0.345	-0.787
	(0.328)	(0.347)	(0.987)
Government	-0.528	-0.904	1.130
	(1.383)	(1.423)	(2.263)
Private	-0.353	-0.822	-0.278
	(1.092)	(1.110)	(0.656)
Public	0.701	0.152	

[26] Given that these represent a set of completely exhaustive and mutually exclusive categories, by convention we omit one category -- non profit firms.

	(1.110)	(1.146)	
Constant	-2.672	-4.047**	-5.451**
	(1.833)	(1.752)	(2.349)
Observations	351	265	86
R-squared	0.421	0.423	0.608
Year Controls	Yes	Yes	Yes
Industry Controls	Yes	Yes	Yes

heteroskedastic robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Results from Model 1 suggest that revenues[27] are strongly associated with incident cost. For example, a 10% increase in firm revenues is correlated with a 1.1% increase in the cost of an incident (significant at the 5% level). However, the lack of significant result from Model 3 suggests that this result is driven entirely by data breaches. That is, only for data breaches are revenues correlated statistically with the cost of a breach. This is somewhat surprising, since one would imagine that the volume of revenues of a company would be a factor that would strongly influence overall costs for any type of event.

On the other hand, the number of records compromised is found to be robust across all three models, suggesting that it is consistently and strongly correlated with incident cost (i.e. a 10% increase in the number of records compromised is associated with a 2.5%-2.9% increase in cost, significant at the 1% level). This result is unsurprising as one would expect that larger breaches impose greater cost.

Interestingly, malicious events were found to be significantly and negatively correlated with both pooled data, and privacy violations, yet not data breaches. (Note that the estimate from Model 3 is spurious as it is based on a single malicious observation.) One might expect that intentional, malicious events are more likely to cause greater damage and loss, relative to accidental ones. The lack of a significant result for malicious events and those resulting from litigation for data breaches (while simultaneously being significant for the pooled dataset) is interesting for a number of reasons. First, it suggests that, based on these preliminary analyses, there are substantial differences across data breaches and privacy violations for which pooled analysis may be deceiving. Therefore, that any rigorous empirical analysis should properly differentiate between these event types.

Placing costs in context

A critical policy question, and one that this manuscript attempts to address is: how much of a problem are cyber events? One approach to understanding the impacts is to consider these costs as a percentage of firm revenues. A similar practice is used by financial lending institutions which face the chronic problem of balancing good versus bad debt (often referred to as bad debt expenses). While they cannot avoid all bad debt, some amount is tolerable, and indeed, efficient.[28] Firms, therefore, track these bad debt expenses and no doubt apply their own threshold for determining a tolerable amount of bad debt. Firms in the retail

[27] We use firm revenues because it provides a reasonable proxy for firm size. Number of employees is a legitimate alternative. However, the correlation between revenues and employees was 0.935 and the regression results presented are robust to the substitution of revenues with employees.

[28] Efficiency comes from the realization that while some effort spent to prevent fraud and abuse is cost-effective, at some point, each additional dollar spent to prevent further waste results in a benefit less than one dollar.

industry also closely track the loss (shrinkage)[29] that they incur on an annual basis. Indeed, tracking and reporting this metric is a familiar practice for many firms. For example hospitals incurred 5.9% in bad debt due to uncompensated care in 2011 (Rodney, 2013), while Xerox and Strayer Education incurred less than 1 percent, and 3.2% of bad debt, respectively (BusinessWire, 2015; MarketWatch, 2015). Estimates from the restaurant industry suggest much higher losses around 20% (Rowe, 2011; Plotkin, undated), and annual shrinkage in the US retail industry was estimated at 1.44% of annual sales ($33.5 billion; Kays, 2010),[30]

Therefore, by dividing the total loss (sum of first and third party losses) by revenues,[31] we obtain a distribution of loss ratios that is heavily skewed with very expensive cyber events drastically driving up mean calculations. From the sample of 764 observations, 66 cyber events exhibited losses less 0.001% of revenue, and 90% of events exhibited losses less than firm revenues. Overall, we find that the median loss was just 0.4% of annual revenue.[32] That is, based on our data, we find that most cyber events cost firms less than 0.4% of their annual revenues.

Another form of comparison is to examine these losses relative to fraud. For example, the total loss from US cyber crime activities was estimated to be $105 billion annually (CSIS, 2014),[33] and US firms were estimated to have lost 0.9% of revenues to online fraud in 2013 ($3.5 billion; Cybersource, 2013).[34] Healthcare fraud was estimated at 3% ($69 billion) of total US health care expenditures (NHCAFA, 2010). Globally, losses due to fraud accounted for approximately $3.5 trillion, or 5% of annual firm revenues (ACFE, 2012),[35] while payment card fraud specifically, accounted for $11.27 billion, or 0.05% of sales volume (Nilson, 2013). These data are presented in Figure 14.

If it is true that on average, businesses lose 5% of their annual revenue to fraud, and that the cost of a cyber event represents only 0.4% of firm revenues, then one may conclude that these hacks, attacks, and careless behaviors represent a small fraction of the liabilities that firms face, and therefore only a small portion of the cost of doing business.

[29] In a retail clothing store, shrinkage represents lost or stolen merchandise. In the restaurant industry, shrinkage represents the portion of food or drink that is stolen, spoiled, or broken.
[30] Based on a survey of 100 retail clothing and supermarket stores.
[31] Note that all firm revenues are computed as 2015 dollar values.
[32] We removed 6 extreme outliers from these calculations as they were clearly anomalous events with loss/revenue ratios greater than 100
[33] This estimate is based on their estimate of cyber crime being .64% of the US GDP. Using a value of $16.77 trillion, 0.64% is approximately $107 billion.
[34] See also http://www.cardhub.com/edu/credit-debit-card-fraud-statistics/.
[35] Based on a survey conducted by the Association of Certified Fraud Examiners between 2010 and 2011 with 1388 firms. Fraud includes corruption, financial misstatements, cash theft, billing and payroll fraud, larceny.

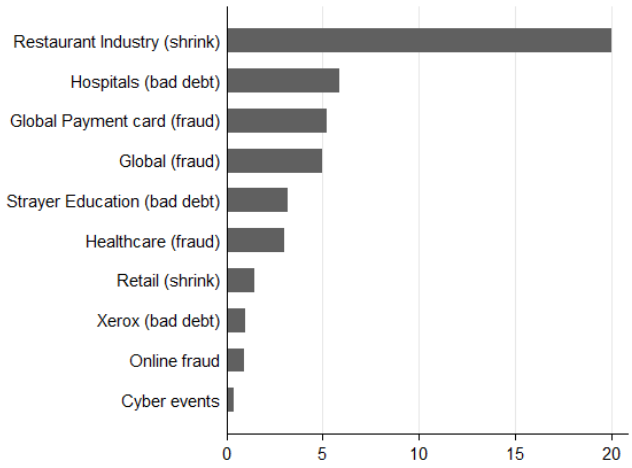
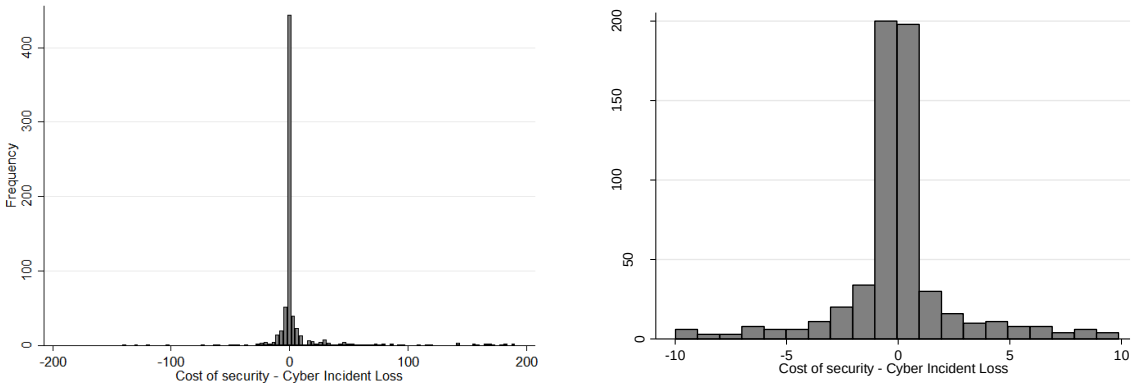


Figure 14: Loss as a percentage of revenues

How much should a firm spend on IT Security?

A question that has plagued researchers, executives and policy makers for decades is regarding the amount of money to invest in protecting a firm's corporate and customer databases, IT systems, and intellectual property. The Gordon-Loeb model (2002) is one attempt to provide such an answer. It suggests that a firm's investment in IT security should not exceed 30% of the losses it expects to incur from a data breach or cyber event.[36]

Using the data from this research, we can further understand the efficiency of IT security by comparing the losses from these events, with a firm's investment in IT security. A study from 2013 found that an average firm's IT budget was around 5% of its revenues (CIO 2013), and a Gartner survey of 1500 firms in 2010 found that firms spend an average of 5% of their IT budget on information security (Kirk 2010). Therefore, if we assume that IT security spending is, on average, 0.025% of revenues, then we can compare this with actual cyber event losses from our dataset. Figure 15 presents a histogram showing the difference between cyber incident costs and IT security spending.[37] The left panel shows observations within +/- $200 million, while the right panel displays the distribution for values within +/- $10 million.



[36] However, note that this estimate of 30% may be considered an artifact of the functional forms used in the model.
[37] Note that we omit some extreme outliers for presentation purposes.

Figure 15: Incident costs – security spending (in millions of $)

Values to the left of zero imply that incident costs are greater than IT security costs, while values to the right of zero imply that IT security costs exceed the cost of a breach. A surprising observation is the large mass right around the zero mark. There is no reason to expect, *a priori*, that IT security budgets should so closely match breach costs. Specifically, 77% of incidents cost firms between +/- $10 million of its security budget, and 50% of incidents cost firms between +/- $1 million of its security budget. That is, these data show that half of cyber events cost a firm an amount equal to its annual investment in IT security.

On one hand, this may suggest that unless a firm incurs a breach every year, it is wasting its IT security investment every year it does not suffer a breach. Alternatively, it may imply that a firm can expect to lose the equivalent of its IT security budget each time it suffers a data breach or security incident.

Limitations

Even though the data used in this analysis are likely the most comprehensive sample of cyber incidents available, they are still based on publicly available data, and would therefore not always reflect events which are unobserved by the firm. Similarly, events which are observed but not reported or discovered publicly would also not be included. Further, despite the comprehensive search strategies employed, the observations recorded may still be biased towards larger, more prominent cyber incidents. Therefore, all inferences and conclusions necessarily rely on an assumption that the observed cyber incidents in our dataset are generally representative of the universe of events.

We think this is a reasonable assumption, because of the impressively diverse and comprehensive collection of search strategies employed by Advisen that include scouring local and news sources, searching legal databases, data breach clearinghouses, and government websites. Indeed, while other leading clearinghouses contain less than 5000 incidents,[38] the full dataset includes over 12,000. Further, even if the data were biased toward larger, more severe or prominent cyber events, it is precisely these which we anticipate are used to inform and drive public policies, firm practices, and regulatory oversight.

Finally, because we are limited to the data available (which collects firm-level data), we are not able to examine consumer-level harms or costs due to these events.

Conclusion

The analysis provided in this paper is relevant to firms from all industries (including those that have yet to suffer a cyber event), policy makers, and especially insurance companies. However, it has also uncovered an unsettling paradox. On one hand, the analysis on breach and litigation rates suggest similar patterns – that cyber incidents and resulting legal actions are becoming more frequent and therefore potentially more expensive to the firms collecting personal information, and suffering these events. In addition, the kinds of information being compromised (SSN, medical and financial), are those that could well lead to more severe and longer lasting forms of consumer identity theft and fraud.

[38] As of early 2015, the Privacy Rights Clearinghouse, a non-profit organization dedicated to issues concerning consumer privacy, was cited as having less than 4500 incidents (Edwards et al, 2015).

On the other hand, as we examine the actual costs of these events in our dataset (clearly one of the most important outcome measures), we find that most events cost firms only a fraction of the millions of dollars that is commonly cited as the cost of a data breach. In addition, we find that cyber events cost firms only 0.4% of revenues, far less than any other loss due to fraud, theft, corruption, or bad debt.[39] Further, based on consumer surveys, other research has shown that consumers are generally very satisfied with firm responses to data breaches, and that only a small percentage (11%) of customers are lost due to attrition (Ablon et al., unpublished). If true, this suggests that perhaps the media attention and concern surrounding cyber events (and data breaches in particular), are unnecessarily exaggerated relative to the actual impact they cause.

Appendix

NAIC Industry Descriptions

Table 4: NAIC industry descriptions

Industry (NAICS Code); number of firms	Description and Examples
Accommodation and Food Services (72); 495,347	Hotels, inns, etc, food services, etc.
Admin and Support Services (56); 327,214	Telemarketing, collection agencies, credit bureaus, travel agencies, armored car services, hazardous waste removal, etc.
Agriculture, Forestry, Fishing and Hunting (11); 21,351	Eg farming, orchards, cattle, chicken, logging, farm management, etc.
Arts, Recreation and Entertainment (71); 114,969	Theatre, dance, sports, museums, casinos, amusement parks, golf, fitness centers
Construction (23); 640,951	Residential, commercial construction, highway/street bridge construction, etc.
Educational Services (61); 84,503	Schools, univesities, training and trade schools
Finance and Insurance (52); 234,841	Commercial banking, savings, credit card issuers, mortgage brokers, investment banking, insurance carriers and brokers.
Government (Public Administration) (92); 90,107	Courts, police, fire protection, administrative offices, national security, international affairs
Health Care and social assistance (62); 640,724	Hospitals, dentists, doctors, medical centers, ambulance services, psychiatry and nursing, day care services.
Information (51); 71,108	e.g. news/book publishing, motion picture and music publishing, radio, television, software vendors, telecommunications companies, internet portals, etc.
Management of Companies and Enterprises (55); 26,819	Management of companies and enterprises, offices of holding companies , corporate, subsidiary, and regional managing offices
Manufacturing (31); 256,363	Food product manufacturing, breweries, wineries, fabrics/clothing, construction materials, newsprint/books, pharmaceuticals, plastics, rubber, iron, steel, computer and computing products, semiconductor, audio/video, truck, car manufacturing, medical equipment

[39] Clearly, however, in some cases, data breaches and other cyber attacks have caused massive losses to firms, as well as some cases of identity theft do cause extreme harms to individuals. Note, however, that these discussions relate to average or median outcomes.

Mining, Quarrying, and Oil and Gas Extraction (21); 22,149	Mining, petroleum, gas extraction, drilling, and supporting activities
Other Services (81); 667,176	Automotive and computer repair, civic organizations, religious organizations, salons and other personal services
Prof, Sci, and Tech Services (54); 772,685	Legal, tax, engineering services, computer programming, management consulting, advertising and public relations companies, direct mail
Real Estate and Rental and Leasing (53); 270,034	Residential and commercial property and equipment leasing, car and truck rental
Retail Trade (44); 650,749	Automotive, furniture, home centers, food markets, clothing, electronics
Transportation (and warehousing) (48); 168,057	Air, train transportation, trucking, taxi, limousine, postal service
Utilities (22); 5,973	Power generation and distribution (solar, hydro, nuclear, wind), water and sewage treatment facilities
Wholesale Trade (42); 315,031	Furniture, automotive, lumber, commercial and industrial equipment, farm machinery, etc.

References

Ablon, L., Heaton, P., Lavery, D., Romanosky, S. Consumer Attitudes Toward Data Breach Notifications and Loss of Personal Information, RAND Corporation.

Association of Certified Fraud Examiners (ACFE), 2012 Report to the Nations on Occupational Fraud & Abuse, 2012.

BusinessWire, 2015, Xerox Reports Second-Quarter 2015 Earnings, http://www.businesswire.com/news/home/20150724005172/en/Xerox-Reports-Second-Quarter-2015-Earnings#.Vbk4yWfbK01.

CSIS, Net Losses: Estimating the Global Cost of Cybercrime, Center for Strategic and International Studies, 2014.

Cybersource, 2013, 2013 Online Fraud Report, Online Payment Fraud Trends, Merchant Practices, and Benchmarks.

Edwards, B., Hofmeyr, S., Forrest, S., 2015, Hype and Heavy Tails: A Closer Look at Data Breaches, Workshop on the Economics of Information Security, 2015

Gordon, L.A. and M.P. Loeb, 2002, The Economics of Information Security Investment, ACM Transactions on Information and System Security, pp. 438-457.

Jacobs, J., 2014, Analyzing Ponemon Cost of Data Breach , Available at http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/.

Kays, Joseph, 2010, Business Expense, Fall 2010.

MarketWatch, 2015, Strayer Education, Inc. Reports Second Quarter Revenues and Earnings; and Summer Term 2015 Enrollments. http://www.marketwatch.com/story/strayer-education-inc-reports-second-quarter-revenues-and-earnings-and-summer-term-2015-enrollments-2015-07-29.

Moore, Rodney, 2013, U.S. hospitals provided $41.1 billion in uncompensated care in 2011, representing 5.9 percent of annual hospital expenses. Available at http://www.healthcarefinancenews.com/news/stopping-rise-hospital-bad-debt-0

National Health Care Anti-Fraud Association (NHCAFA), 2010, Testimony of the National Health Care Anti-Fraud Association, to the House Insurance Committee, House of Representatives, Commonwealth of Pennsylvania.

NetDiligence, 2014, Cyber Claims Study, NetDiligence.

Nilson Report, 2013, Global Credit, Debit, and Prepaid Card Fraud Losses Reach $11.27 Billion in 2012.

Plotkin, Robert, undated, Sticky Fingers — Bartender Theft in the New Economy. Available at http://www.barprofits.com/pages/newsletter/vol2-issue4/page06.php.

Ponemon, 2011, 2011 Cost of Data Breach Study: United States.

Ponemon, 2013, 2013 Cost of Data Breach Study: Global Analysis.

Ponemon, 2014, 2014 Cost of Data Breach Study: United States.

Ponemon, 2015, 2015 Cost of Data Breach Study: United States.

Romanosky, S., Telang, R. & Acquisti, A. (2011). Do Data Breach Disclosure Laws Reduce Identity Theft? Journal of Policy Analysis and Management, 30(2), 256-286.

Romanosky, S., Hoffman, D., and Acquisti, 2014, A. Empirical Analysis of Data Breach Litigation, Journal of Empirical Legal Studies, 11(1), 74-104.

Rowe, Megan, 2011, 7 Ways to Stem Bar Shrinkage, Restaurant-hospitality.com. Available at http://restaurant-hospitality.com/trends/seven-ways-stem-bar-shrinkage0911.