Empirical Analysis of Data Breach Litigation

Sasha Romanosky,
Heinz College of Public Policy and Information Systems,
Carnegie Mellon University
sromanos@cmu.edu

David Hoffman,
Beasley School of Law, Temple University
hoffmand@temple.edu

Alessandro Acquisti,
Heinz College of Public Policy and Information Systems,
Carnegie Mellon University
acquisti@andrew.cmu.edu

June 1, 2012

This research was supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office, and by Temple Law School's Conwell Corps Program. We would like to thank Antima Chakraborty, Carol Anne Donohoe, Ian Everhart, Caitlin Jones, Kevin Leary and Jake Oresick for their research assistance. We would also like to thank Paul Bond, Aaron Burnstein, Fainna Kagan, Amelia Haviland, Mark Melodia, Kristen Matthews, Peter Oh, Barrie Nault, David Navetta, Mohammad Rahman, Theresa Romanosky, Boris Segalis, Brendon Tavelli and 7 anonymous attorneys for their valuable insights and suggestions.

Published Version:
Journal of Empirical Legal Studies
Volume 11, Issue 1, 74–104, March 2014

Abstract

In recent years, many individuals have sought legal redress for harms caused by the loss or theft of their personal information. However, very little is known about the drivers, mechanics, and outcomes of these lawsuits, making it difficult to assess the effectiveness of litigation at balancing organizations' usage of personal data with individual privacy rights. Using a unique database of manually collected lawsuits, we analyze court dockets for over 230 federal data breach lawsuits from 2000 to 2010. We investigate two research questions: Which data breaches are being litigated? Which data breach lawsuits are settling? Our results suggest that the odds of a firm being sued are 3.5 times greater when individuals suffer financial harm, but 6 times lower when the firm provides free credit monitoring. Moreover, defendants settle 30% more often when plaintiffs allege financial loss, or when faced with a certified class action suit. By providing the first comprehensive empirical analysis of data breach litigation, these findings offer insights in the debate over privacy litigation versus privacy regulation.

Empirical Analysis of Data Breach Litigation

1. INTRODUCTION

The surge in popularity of social media, e-commerce, and mobile services is proof of the benefits consumers are enjoying from information and communication technologies. However, these same technologies can create harm when personal consumer information is lost or stolen, causing emotional distress or monetary damage from fraud and identity theft.[1] Since 2005, an estimated 543 million records have been lost from over 2,800 data breaches,[2] and identity theft caused $13.3 billion in consumer financial loss in 2010 (BJS, 2011). In response, federal legislators have introduced numerous bills that define appropriate business practices regarding the collection and protection of consumer information,[3] and federal regulators have drafted privacy frameworks for consumer data protection (Department of Commerce, 2010; FTC, 2010). A significant concern for policy makers, therefore, is balancing *ex ante* regulation with *ex post* litigation to protect both consumer and commercial interests. For instance, the Department of Commerce inquired: "should baseline commercial data privacy legislation include a private right of action?" (Department of Commerce, 2010, 30). At issue is the degree to which federal consumer litigation deters privacy harms, or whether a new federal privacy statute is required.

On one hand, a weak litigation regime would be ineffective at deterring a firm's harmful or negligent behavior. Lawsuits that are inappropriately disposed of eliminate a plaintiff's ability to obtain appropriate relief for legitimate harms. For example, a case was successfully brought against Rite Aide for carelessly disposing pharmacy labels and employment applications in a public trash dumpster.[4] In the settlement, Ride Aide agreed to "a comprehensive information security program that is reasonably designed to protect the security, confidentiality, and integrity of personal information collected from or about consumers."[5] Without legal action, such careless practices may have never been corrected.

On the other hand, a heavy-handed litigation regime could impose excessive legal fees and damage awards, over-deter firms, and – according to some – stifle innovation. Online movie-rental site, Netflix, offered a $1 million prize to anyone who could sufficiently improve its movie recommendation algorithm. To facilitate the contest, Netflix published (what was believed to be) anonymized rental information for a sample of its users. Due to lawsuits stemming from the re-identification of these data, Netflix cancelled a subsequent contest. While the total social value of such innovation may be limited, the Netflix case provides one example of how litigation can impact firms' product development.

[1] See Solove (2007) for a description of the potential harms associated with breaches of personal information.

[2] See Privacy Rights Clearinghouse, http://www.privacyrights.org/data-breach. Last accessed Jan 22, 2012.

[3] For example, the Cyber Security and American Cyber Competitiveness Act of 2011 (S.21), the Data Security and Breach Notification Act of 2011 (S.1207), the Commercial Privacy Bill of Rights Act 2011 (S.799), the Personal Data Privacy and Security Act of 2011 (S.1151), the Data Breach Notification Act (S.1408), the Personal Data Protection and Breach Accountability Act of 2011 (S.1535), the Secure and Fortify Electronic Data Act of 2011 (H.R.2577), the Cybersecurity Enhancement Act of 2011 (H.R. 2096).

[4] See In re Rite Aid Corp., FTC File No. 072-3121 (July 27, 2010).

[5] *Id.*

Our manuscript attempts to offer novel insights over this debate by providing the first comprehensive empirical analysis of data breach litigation, and investigate the characteristics, drivers, and outcomes of data breach litigation.

Determining whether current US privacy laws are too weak or too strong is not easy. It is difficult (and perhaps impossible) to assess the aggregate costs and benefits for both consumers and firms of different privacy regimes in purely monetary terms (Romanosky and Acquisti, 2010). However, even just understanding the landscape is a problem. Little is known about the trends in data breach litigation – which breaches are litigated and which are not, and with what outcomes. While there exists some legal scholarship regarding data breach litigation (Citron, 2007, 2011; Rice, 2007; Serwin, 2009), it typically examines a narrow subset of lawsuits, usually focusing on high-profile cases or those with published opinions. Unfortunately, given that as few as 15% of all lawsuits produce reported opinions (Hoffman et al., 2007), any conclusions reached from examining particular, high-profile cases are likely unrepresentative of the full population of data breach lawsuits. Consequently, it is still unknown what characteristics these lawsuits actually possess, and how "successful" they have been.

To our knowledge, no empirical research involving data breach lawsuits has been conducted. The purpose of this manuscript is to address this research and policy gap by empirically investigating a representative collection of federal data breach lawsuits and their outcomes. We overcome common sample selection issues by searching Westlaw and acquiring data directly from court dockets (PACER), in combination with other publicly available data sources.[6]

In addition to presenting rich descriptive information about these suits, we explore two questions. First, what kinds of data breaches are being litigated in federal court, and why? Second, what kinds of data breach lawsuits are settling, and why? Our first question examines federal lawsuits resulting from reported data breaches, while the second question includes all known federal lawsuits related to the unauthorized disclosure of personal information. By providing the first comprehensive empirical analysis of data breach litigation, these findings offer insights in the debate over privacy litigation versus privacy regulation. Specifically, we believe that answering these questions will help inform firms, consumers, and policy makers regarding the risks associated with the collection and use of personal information, and the characteristics and outcomes of federal data breach litigation.

Our analysis reveals that federal data breach lawsuits typically exhibit the following characteristics. First, plaintiffs seek relief for one or more of: actual loss from identity theft (e.g. financial or medical fraud), emotional distress, cost of preventing future losses (e.g. credit monitoring and identity theft insurance), and the increased risk of future harm. Second, the lawsuits are usually private class actions, though some are brought by public entities such as the Federal Trade Commission or state attorneys general. Third, defendants are typically large firms such as banks, medical/insurance entities, retailers, or other private businesses. Fourth, complaints allege a staggering range of both common law (tort, breach of contract) and statutory causes of action. And fifth, cases generally either settle, or are dismissed, either as a matter of law, or because the plaintiff was unable to demonstrate actual harm.

In addition, we find that that the odds of a firm being sued are 3.5 times greater when individuals suffered financial harm, but over 6 times lower when the firm provides free credit monitoring to those affected by the breach. Moreover, the odds of a firm being sued as a result of improperly disposing data are 3 times greater relative to breaches caused by lost/stolen data, and 6 times

[6] We discuss the consequences of limiting our search to Federal Lawsuits *infra* at Section 6.

greater when the data breach involved the loss of financial information. Our analysis suggests that defendants settle 30% more often when plaintiffs allege financial loss from a data breach, or when faced with a certified class action suit. The odds of a settlement are found to be 10 times greater when the breach is caused by a cyber-attack, relative to lost or stolen hardware, and the compromise of medical data increases the probability of settlement by 31%.

The next section provides background literature related to data breaches, docket analysis and litigation. We then examine which breaches are litigated, and, conditional on suit, which cases settle. Discussions of limitations and final conclusions complete the manuscript.

2. RELATED WORK

In recent years, economists have researched a number of empirical and theoretical aspects of data breaches, such as the effect of breaches on a firm's stock market price (Campbell et al., 2003; Cavusoglu et al., 2004; Acquisti et al., 2006; Kannan et al., 2007; Gordon et al., 2011), the effect of data breach disclosure laws on identity theft (Romanosky et al., 2011), and the conditions under which disclosure laws may reduce the social costs of these breaches (Romanosky et al., 2010). This work shows that while disclosure of a breach does appear to reduce identity theft, conclusive evidence of the impact on stock market price is unsettled. In addition, a growing body of legal scholarship relates to data breaches. For example, policy and legal scholars have discussed the outcomes of data breach litigation (Citron, 2007; Hutchins, 2008; Lesemann, 2009; Solove, 2005); they have summarized the legal theories that plaintiffs allege when trying to recover damages from data breaches (Citron, 2007, 2011; Rice, 2007; Serwin, 2009); and they have examined alternative policy mechanisms that can be used to reduce the harm from data breaches (Romanosky and Acquisti, 2009).

An emerging body of legal scholarship analyzes court dockets. This form of empirical research makes very practical use of publicly available -- and generally very detailed -- collection of pleadings, motions, rulings and administrative record keeping that compose a legal dispute. For example, Kim et al. (2009) use docket analysis to compare judicial decisions between district and appellate judges. Hoffman et al. (2007) use dockets to examine the incentives for judges to justify their legal decisions (i.e. orders versus opinions) and to publish these decisions in court reporters. Boyd and Hoffman (2010) use dockets to examine federal veil piercing litigation and examine the characteristics that lead to a plaintiff's greater 'success' rate.

In addition, there are efforts to construct repositories for domain-specific lawsuits, making them available for public analysis and research. Such efforts include the Securities Class Action Clearinghouse,[7] Intellectual Property Litigation Clearinghouse,[8] Civil Rights Litigation Clearinghouse, [9] and the Equal Employment Opportunity Commission (Kim et al., 2009).

Intuitively, economic analysis of litigation suggests that individuals are more likely to file suit when their expected rewards exceed their expected costs (Cooter & Ulen, 2008, 414-484; Cooter and Rubinfeld, 1989). This hypothesis has been supported by some empirical work (Clermont and Eisenberg, 2002), especially in the area of financial patent litigation (Lerner, 2010). For instance, Dunbar and Sabry (2007) examine plaintiff demographics, injury severity, and economic factors in the propensity for victims of work, car, or product-related injuries to sue. They find that severity of injury is significantly correlated with litigation. Viscusi (1986, 326) provides evidence

[7] See http://securities.stanford.edu/, last accessed June 10, 2011.
[8] See http://www.law.stanford.edu/program/centers/iplc/, last accessed September 29, 2011.
[9] See http://www.clearinghouse.net/, last accessed June 10, 2011.

that case outcomes are correlated with the defendant’s alleged violations of government regulations.

Priest and Klein (1984) propose a theoretical model of plaintiff win rate, which holds under general conditions and is robust to multiple types of liability regimes, judicial biases, and distribution of disputes (see Wittman (1988) for empirical validation). Alternatively, Shavell (1996) presents a brief but competing model in which he argues, under other conditions, that any frequency of plaintiff victory is possible. A find from this literature which holds particular relevance is that is that statistical models studying outcomes often suffer from omitted variable and sample selection biases when the collection of suits reaching judicial ruling (or settlement) is not representative of the larger set of cases that begin a dispute (Clermont and Eisenberg, 1998; Clermont and Eisenberg, 2002; Boyd and Hoffman, 2010).

As a whole, these bodies of research help inform this manuscript in a number of ways. First, the economic analysis of litigation provides the foundational theories upon which we develop our hypotheses. In addition, we leverage the existing research on docketology to help inform our data collection, and we leverage the research on settlement and dispute resolution to overcome chronic forms of bias. However, while economic and legal scholarship has examined various aspects of data breaches, their harms, and the legal theories brought by plaintiffs, to our knowledge this is the first paper to empirically examine privacy litigation generally, and data breach litigation specifically.

3. DATA

This manuscript combines a number of datasets. We first obtained a list of publicly reported data breaches. We then used Westlaw (an online legal research service) to identify federal data breach lawsuits. Finally, we used PACER (“Public Access to Court Electronic Records”) to obtain docket filings. For the purpose of this manuscript, a data breach is defined broadly as the unauthorized disclosure of personal information by an organization.

3.1. Data Collection

To address our first research question (“Which breaches are being litigated?”), we first gathered a list of reported US data breaches from the Open Security Foundation (“Datalossdb”), a non-profit organization devoted to collecting and recording data breaches and IT vulnerabilities, and which is one of the most comprehensive collections of reported data breaches.[10] This dataset contains the name and industry of the breached entity, the number of records compromised, the date of breach, the cause of breach,[11] and the types of information lost. Then, we used Westlaw to identify which of these reported breaches resulted in federal litigation.

To address our second research question (“Which data breach lawsuits settle?”), we used Westlaw to perform a systematic search for all federal lawsuits in which plaintiffs alleged an unauthorized disclosure of their personal information.[12] (The lawsuit observations previously

[10] These data are used per the OSF license agreement which states: “permission is granted to use this database in non-profit works and research.”

[11] The causes of the breach, as coded by Datalossdb, included 51 unique types. However, many categories are variations which can easily be reduced to the following three: “loss or theft” (i.e. accidental loss or theft of computing hardware that happened to contain personal information), “disclosure” (i.e. personal data carelessly made publicly available), and “hack” (i.e. the deliberate theft of personal information through cyberattack).

[12] Certainly, the ideal dataset would include all state and federal suits. However, per Section 1, we focus on federal suits only in this manuscript.

used are, of course, a subset of the results from this search.) Specifically, we searched Westlaw's Pleadings database using the following search strings: "personally identifiable information," "personal information," and either "data breach," "security breach," or "privacy breach." These search terms balance specificity without biasing search results to specific causes or types of data breach lawsuits. We then manually examined the results and extracted those cases relating to unauthorized disclosure of personal information.[13] We believe this is an appropriate combination of methods for identifying all lawsuits either filed in, or removed to, federal court and therefore represents the most complete collection of federal data breach lawsuits. We address issues related to collecting state actions later in this manuscript.

We then used PACER to retrieve the court docket for each case. From the docket itself we coded the following information: presiding judge, date filed, date terminated, forum, the law firms involved in the suit and number of docket filings. We then purchased the complaint (or amended complaint where appropriate) and coded information relating to the breach such as the date of breach, size, and cause of the breach, types of information compromised, and all causes of action. We also identified whether any dispositive motions were filed, and coded the disposition of the case. Settlement information (such as actual confirmation of a settlement, and amounts of any damage awards) was obtained either from the docket filings, or from directly contacting the litigating attorneys.[14]

3.2. Data Generating Process

Data breach and lawsuit data are generated from the processes shown in Figure 1.

<Insert Figure 1 here>

Stage 1: Reported and unreported breaches

As mentioned, for the purpose of this manuscript, a "data breach" is defined as the unauthorized disclosure of personal information. From this population of events only a subset will become public knowledge and "reported" by the Datalossdb clearinghouse. Specifically, the only breaches that are included in this clearinghouse are those relating to social security numbers, financial/banking information, credit card numbers, or medical information, and where the number of records compromised exceeds 10.[15] This collection of reported breaches originates from a community of dedicated security professionals, who obtain data breach information from news sources across the country, from Freedom of Information Act (FOIA) requests to state agencies, and from many individual contributing members across the country. This group has been systematically collecting data breach information since at least 2005.[16]

Awareness of breaches also stems from US state laws requiring companies to notify individuals when their personal information is lost or stolen. California first adopted this type of law in 2003,

[13] For consistency in analysis, we omitted cases relating to, for example, a breached entity suing an individual alleged to have stolen data (21 instances), individuals suing entities for unauthorized collection or use of personal information (114), or known state actions (79).

[14] Class action settlements were sometimes publicly available, and in some cases we were able to obtain settlement details for individual actions. Many times, however, only confirmation of settlement was available, with all other details being privileged.

[15] Note that the sample of "unreported breaches" (the dotted line from Stage 1 to Stage 2) also contains observations which would be non-litigated, federally-litigated, or state-litigated. However, when addressing our first research question ("Which data breaches are litigated?"), we do not include these observations.

[16] See http:// datalossdb.org/about, last accessed Jan 25, 2012.

with other states following (by the end of 2011, at least 46 states had adopted similar laws).[17] Two characteristic of state disclosure laws can affect the proportion of all breaches that are reported. First, there is heterogeneity among the state laws regarding the threshold of disclosure; about half of the laws require notification only if there is a reasonable risk of malicious use of the data (high threshold), as opposed to simple loss of the data (low threshold). Second, it is the residence of the individual that drives disclosure, not the location of the breach. That is, disclosure to an individual is only required if the state in which the individual is a citizen has adopted a disclosure law. These properties suggest that breaches are less likely to be systematically reported if they affect citizens of states with higher thresholds for notification, or affect citizens of a state without a disclosure law.

There are, however, a number of mitigating factors that should reduce this systematic non-reporting. First, conversations with defense attorneys suggest that, because it is quite costly for firms to separate disclosure requirements among differing states' citizens, it is easier for firms to simply notify all individuals. Indeed, firms also choose to notify all individuals independently of any particular state law, as a means of managing public relations or due to pressure from states' attorneys general. For example, Choicepoint notified consumers of all affected states from its breach in 2005 even though only California had a disclosure law at that time (Ryan, n.d.). Similarly, firms, confronted with legal requirements from disparate states, may simply choose to follow the strictest law – which would require notification regardless of any threshold.

In addition, one may be concerned that organizations weigh the costs and benefits of disclosure and rationally choose not to notify consumers. However, conversations with privacy attorneys suggest that firms find this practice too risky and obey the law. Together, these effects should therefore minimize systematic non-reporting of data breaches.

Stage 2: Non-litigated, state-litigated, and federally-litigated data breaches
Stage 2 describes three separate outcomes from the sample of reported breaches: non-litigated, federally-litigated, or state-litigated.[18] Because our key research questions relate to federal policy solutions to resolving the externalities caused by data breaches, our empirical focus compares *federally-litigated breaches* with *non-federally-litigated breaches* (i.e. both state- and non-litigated breaches). It is important to note that by pooling state- and non-litigated breaches we are still able to obtain unbiased estimates of *federal* lawsuits resulting from *reported* data breaches.[19] We discuss data limitations from unobserved state lawsuits in Section 6.

Stage 3: Federal lawsuits observed from Westlaw
For Stage 3, we obtained a sample of federal lawsuits through Westlaw using a systematic search strategy designed to identify the largest collection of data breach lawsuits practical, and then manually edited the list of suits matching our research question. Investigations by researchers have concluded that the Westlaw Pleadings database (used in this analysis), "covers or nearly

[17] See http://www.ncsl.org/issues-research/telecommunications-information-technology/security-breach-legislation-2011.aspx, last accessed Jan 25, 2012.

[18] Arbitration is one further category of outcome that may exist. In these cases, plaintiffs, as a result of enjoying a firm's good or service, are contractually bound to resolve any legal dispute through arbitration, rather than civil court. However, we are unaware of any arbitrations in which privacy rights have been adjudicated.

[19] Alternatively, had we complete data on all three outcomes, one might choose to estimate a multinomial logit model in order to separately estimate marginal effects on federal-versus state-litigated breaches. Or, one might pool state and federal suits together in order to draw inferences about all litigated breaches. However, because our topic of interest is primarily federal policy matters, we pool all non-federally litigated outcomes (that is, state and non-litigated breaches).

covers the universe of federal claims [as it related to veil piercing lawsuits]" and that it "was designed to collect all federal complaints since 2000 that lawyers litigating commercial cases would have a plausible interest in learning about. Thus, Pleadings may exclude civil rights cases, or habeas petitions, or family disputes, but attempts to collect every tort, contract, or federal statutory claim brought against corporate defendants" (Boyd and Hoffman, 2010). Therefore, we do not believe that the use of Westlaw would pose any significant selection bias for our analysis.

It is relevant to also mention that the sample of unreported breaches may result in no federal or state litigation, although - for clarity - only the path to federally-litigated breaches is drawn in Figure 1 (these data are included for the purpose of our second research question: "Which data breach lawsuits settle?").

4. WHICH DATA BREACHES ARE BEING LITIGATED IN FEDERAL COURT?

4.1. Hypotheses

Cooter and Rubinfeld (1989) examine prior theoretical models of litigation to create a unified framework for legal disputes. They present an analytical foundation describing the tensions faced by injurer and victim (defendant and plaintiff) at each stage of a dispute. First, when deciding whether or not to prevent an accident, an injurer balances the (marginal) cost of care with the (marginal) cost of an accident. Then, when deciding whether or not to sue, a plaintiff compares the cost of litigation with the expected benefit from an award. Finally, when deciding whether to settle or proceed to trial, both plaintiff and defendant balance their expected costs of litigation with the outcome from trial. This section is concerned with the second stage (the alleged victim's decision to file suit), which is increasing in both the probability of success and magnitude of award (her expected gain). Below, we adapt these conditions to data breach litigation to construct appropriate hypotheses.

First, we consider the magnitude of a potential award. Given that most data breach lawsuits are class actions, the magnitude of a plaintiff's award becomes a function of the size of the class, which is proportional by the number of records compromised in the data breach. If it is true that class action lawsuits are, in general, driven by class action plaintiffs' attorneys, it follows that the larger the data breach, the greater the potential fee award to the attorney, and the greater the incentive to bring and litigate the suit.[20] Therefore, *the probability of a lawsuit is positively correlated with the number of records lost (H1a).*

Next, the probability of a favorable outcome is multifaceted. Among other things, it is a function of whether an alleged harm can be attributed directly to the breach, the cause of the breach, and the types of information lost.

Plaintiffs in many data breach lawsuits seek relief for harms such as actual financial loss from identity theft, emotional distress, costs of credit monitoring, and anticipated future losses. However, a critical factor affecting the success of a lawsuit is the presence of a cognizable harm for which the law could provide a remedy. In the context of data breach litigation, this is manifested by whether or not the plaintiff can allege (though would not yet have to prove)

[20] It is not the purpose of this research to address the motivations of attorneys, but merely to understand and apply relevant behavior in forming reasonable hypotheses. Conversations with class action plaintiffs attorneys confirm that while it is true that attorneys do seek plaintiffs, plaintiffs also seek attorneys for class action litigation.

financial harm. Moreover, plaintiff harm (loss) is also a function of whether the breached firm provided any initial compensation immediately following the breach and before litigation. This redress is commonly offered in the form of credit monitoring or identity theft insurance. Full compensation for any loss will decrease plaintiffs' legal remedies. Therefore, *the probability of a lawsuit is positively correlated with the presence of actual harm, and negatively correlated with credit monitoring (H1b).*

The legal merits matter. In the context of data breaches, a plaintiff's case is strengthened by her ability to prove that the defendant had a legal duty to protect their personal information, and somehow failed in that duty. This could occur in two different ways.

The first manner relates to the cause of the breach, which typically occurs in one of three ways: improper disclosure or disposal of personal information (e.g. tossing tax records in a dumpster); a computer hack (e.g. computer-based theft of information); loss or theft of hardware (e.g. petty theft of computer hardware that happens to contain personal information). Of these methods, we consider that the first cause (the careless handling of personal information) may provide the strongest legal argument, because it involves the negligent behavior on the part of the data custodian, as opposed to the misfortune of petty theft. Therefore, *lawsuits are more likely to occur from breaches caused by improper disclosure of information, relative to the computer hack, or loss of hardware (H1c).*

The second manner relates to the types of information compromised. It is reasonable to consider that the greater the legal duty to protect certain information (typically enforced through statute), the greater the probability of a favorable outcome. For instance, organizations using medical and financial data are governed by a regulatory environment requiring the enhanced protection of such data. The Health Information Portability and Accounting Act (HIPAA) requires patient consent before the disclosure of medical information between health agencies. The Gramm-Leach-Bliley Act (GLBA) and Fair Credit Reporting Act (FCRA) require greater security controls protecting an individual's credit data. In addition, many state and federal laws require the proper disposal of social security numbers (Dickey et al., 2011) and the storage and transmission of credit card data is also protected through contractual agreements by the credit card companies under the Payment Card Industry Data Security Standard (PCI-DSS). Therefore, *the probability of a lawsuit is positively correlated with the compromise of personal information requiring a heightened level of protection, such as social security numbers, financial, credit card and medical data (H1d).* [21]

4.2. Descriptive Statistics

The entire Datalossdb clearinghouse consists of almost 3,000 data breaches. However, since the primary research question of this section focuses on estimating the probability of a federal lawsuit conditional on covariates, we must prune the dataset in a number of ways. While the first recorded data breach occurred in 1903, systematic collection did not begin until 2005 (after the first breach disclosure law was adopted). Therefore, we limit the duration of our analysis from 2005 to 2010. Observations with missing or ambiguous data are also omitted,[22] though the descriptive analyses presented below are robust to their inclusion. The resulting dataset consists of 1,772 US data breach observations, of which only 65 (3.7%) were litigated in federal court.

[21] Note that we employ the general categories used in the Dataloss clearinghouse and that these categories are not mutually exclusive: a data breach can compromise one or more types of data.

[22] For example, the number of records compromised in some breaches is not known.

Figure 2 compares the number of reported data breaches with the number of federally-litigated breaches during the period 2005 to 2010. In the left panel, lawsuits are scaled according to the left axis (0-16), while reported breaches are scaled according to the right axis (0-600). The right panel shows the ratio of filed lawsuits to the number of breaches reported in that year (i.e., the portion of federally-litigated breaches over time). The right panel shows that, in 2005, the proportion of federal lawsuits was about 10%. However, since 2005, the proportion of federal lawsuits appears to be declining slightly, reaching around 3% in 2010.

<Insert Figure 2 here>

Notice that the number of reported breaches generally increased from 2005 to 2008, and decreased thereafter. Federal lawsuits, on the other hand, fluctuated slightly until 2008, after which they also declined. The rise in *reported* breaches is likely a result of state data breach disclosure laws, which became most popular beginning in 2005 (Romanosky et. al, 2010, figure 6). But why have they since declined? If it were true that data breach incidents were primarily collected from news articles, then this decline might be caused by the erosion of media or consumer interest.[23] Another possible explanation is that US data breach disclosure laws have, indeed, forced firms to internalize more of the cost of a breach, inducing them to invest more to protect personal information, and reducing the number of actual data breaches. This claim is partially substantiated by Verizon (2010, 7) and Romanosky et al. (2011) showing a reduction in data breaches observed on their computing networks.

As one might expect, the number of compromised records is highly correlated with breaches litigated in federal courts (i.e., federally-litigated breaches). The mean number of records compromised by non-federally-litigated breaches (n=1,708) is just over 98,000, while the mean number of records compromised in federally-litigated breaches (n=103) is over 5.3 million, providing suggestive support for H1a.

Figure 3 compares federally-litigated and non-federally litigated breaches as a function of the presence of actual harm (left panel), and the causes of breach (right panel).

<Insert Figure 3 here>

Note that the percentages displayed sum to 100% across categories. For example, as shown in the left panel, 78% of federally-litigated breaches did not result in financial loss, while 22% did result in financial loss.[24] However, breaches appear less likely to be litigated in federal court absent financial harm, providing suggestive support for H1b. The right panel of Figure 3 shows that breaches resulting from the unauthorized disclosure (or disposal) of personal information and computer hack (cyberattack) are *more* likely to be litigated in federal court, while breaches due to lost/stolen hardware are *less* likely to be litigated in federal court, providing suggestive support for H1c. Note that these figures reflect data from all years, but that the patterns presented in both panels are robust when examining individual years.

[23] However, note that while any changes in reporting may affect the proportion of breaches "reported" by Dataloss, this will not bias our regression estimates from our first research question ("Which breaches are litigated") because our inferences consider the Dataloss clearinghouse data as exogenously provided. Further, we do not use this data when examining our second research question ("Which lawsuits settle?").

[24] The presence of actual harm is coded as follows: for known lawsuits, we code 1 if the complaint includes some allegation of financial loss as a direct result of the breach. For data breaches not resulting in lawsuit, we refer to news articles associated with the breach, and similarly code a 1 if the article mentions financial loss resulting from the breach. Given that it is extraordinarily difficult to obtain full information about all possible financial losses, our results very likely provide a lower threshold of loss.

Figure 4 compares breaches that were and were not federally-litigated as a function of the types of personal information compromised. Note that a single breach may result in the compromise of multiple types of personal information.

<Insert Figure 4 here>

Breaches involving financial data (FIN) and credit card numbers (CCN) are more likely to be litigated in federal court, which provides some support for H1c. Social security numbers (SSN), on the other hand, compromised about 78% of non-litigated breaches, though only 58% of litigated breaches. Medical data (MED) appear to be equally represented in federally-litigated and non-federally-litigated breaches.

4.3. Estimating Model

To test hypotheses H1a-H1d, we estimate a binary outcome model predicting the probability that a reported data breach will result in a federal lawsuit,[25]

$$lawsuit_i = \alpha_0 + ActualHarm_i + CreditMonitoring_i + BreachSize_i + Cause_i + ProtectedPII_i + OtherPII_i + Industry_i + Year_i + \varepsilon_i \qquad (1)$$

where *lawsuit* is a binary variable that takes the value 1 if a reported breach, *i*, results in a federal lawsuit, and 0 otherwise.[26] Although we cannot determine with absolute certainty whether financial loss had occurred following a data breach, we can proxy for this by observing any evidence from news reports following the breach. Therefore, *ActualHarm* is coded as 1 if we observe any evidence of financial loss due to the breach, and 0 otherwise.[27] *CreditMonitoring* is a dummy variable coded as 1 if there was any evidence that the breached firm provided any sort of credit monitoring or identity theft insurance to the individuals following the breach.[28] *BreachSize* is a continuous variable representing the log of number of records compromised. *Cause* is a vector of mutually exclusive and completely exhaustive dummies reflecting the cause of the data breach: improper disclosure or disposal, computer hack or lost/stolen hardware.[29] *ProtectedPII* is a vector of dummies representing types of personally identifiable information (PII) should require

[25] Eq. 1 is shown as a linear probability model for clarity only. Actual regressions are estimated using logit. Also note that we limit inferences to predictions of the probability of a *known federal* lawsuit conditional on a *reported* data breach.

[26] Note again that this coding inherently pools state-litigated and non-litigated breaches, thereby ensuring that estimates of federal lawsuits from reported breaches are unbiased.

[27] Of the 1772 data breaches, we were unable to find news reports for 83 of them. In the absence of evidence, we took the most conservative approach and coded these breaches as not causing actual harm. We then performed a robustness check by considering that all 83 observations did cause actual harm. All estimates maintain qualitative magnitude and significance except for ActualHarm which reduces in magnitude by one third and therefore loses statistical significance. One may also be concerned that plaintiffs may wait many years following a breach before filing suit, however we do not find evidence of this. In a sample of 146 single-suit breaches, 78% were filed within one year, and 87% were filed within two years of public notification.

[28] This information was obtained from breach disclosure notices obtained by the Datalossdb clearinghouse, or through news reports, when available. Given that perfect information is not always available, we code this variable equal to 1 only when there is actual evidence of redress. As a result, this variable is likely an under-estimate of the true frequency.

[29] As is customary with categorical variables, we will omit one of these from the regression analysis. Given that the selection is arbitrary, we omit "lost/stolen."

a heightened level of protection, as described in the hypothesis: social security number, medical, financial, credit card). *OtherPII* controls for all other data types (email address, name/address, date of birth and miscellaneous). *Industry* is a vector of dummies representing the industry of the breached firm, whether the firm was a non-profit or publicly traded. *Year* is a vector of year dummies (2005 to 2010) reflecting the year of the data breach, and ε_i is the random error term, assumed to be independent of the observed covariates. Identification of the variables of interest comes from the portion of federally-litigated breaches. Descriptive statistics for the variables used in Eq. 1 are shown in Table 3.

4.4. Results

The results of Eq. 1 are presented in Table 1 and reflect the average marginal effects of the explanatory variables on the probability of lawsuit estimated using a logit regression.[30] Model 1 presents just the variables of interest from H1a-H1d and includes only *Year* controls, whereas Model 2 includes all data types. Models 3a and 3b control for industry variables; they are based on the same estimating equation, but Model 3b presents the results as odds ratios.

<Insert Table 1 here>

The results are robust across all models, with the third model – which controls for all variables - providing the better fit for the data and generally more conservative estimates. Though not shown, results are also robust to the exclusion of individual years 2005-2010, and to probit models. Further discussions therefore focus on results from Model 3a.

In regard to the effect of the size of the breach on probability of lawsuit, our results suggest that a 10-fold increase in the number of compromised records increases the average probability of lawsuit by 8% (from 3.7% to 11.7%), a statistically significant amount (at the 1% level), which supports H1a.[31]

Supporting H1b, the presence of actual (financial) loss is associated with a 2.5% increase in the probability of litigation (though, only significant at the 10% level), while the presence of credit monitoring is associated with a 3.7% decrease in probability of litigation (significant at the 1% level). Described in terms of odds-ratios (Model 3b), these results suggest that the odds of a firm being sued are 3.5 times *greater* when individuals suffer actual (financial) harm, but 6 times *lower* (1/0.152) when they provide free credit monitoring following a breach. While credit monitoring is widely touted by as a best practice following a data breach and, indeed, is included as part of a recent federal data security bill (HR2221), we provide the first statistical evidence to substantiate the practice's value in reducing an organization's *ex post* liability costs.

Next, we examine the relative odds of a lawsuit occurring given the different cause of the data breach (unauthorized disclosure, hack, or lost/stolen). Our results suggest that the odds of a firm being sued due to the unauthorized disclosure/disposal of consumer information are 3 times

[30] Note that the marginal effects for logit models are nonlinear functions of the parameter estimates, and so the effect of a regressor on the probability of lawsuit can either be presented as the effect for the "average observation" (i.e. marginal effect computed at the sample mean of the regressors) or, the "average effect" (i.e. computing the marginal effect for all observations and taking the average). We believe the second approach is more appropriate for our model because: 1) we avoid the confusion of subjectively determining the value of the regressor at which to compute the marginal effect, as in the case of the logged regressor, and 2) given that most explanatory variables are dummies, we do not need to justify having to calculate the marginal effect at a sample mean of a binary regressor.

[31] A 10 fold increase represents a change of 900%, or 0.009*9 = 0.081 or 8.1%.

greater, relative to breaches caused by lost/stolen data (significant at the 5% level), supporting H1c. Breaches caused by cyberattack, however, are not statistically more likely to result in a suit. These results suggest that individuals are much more likely to punish (alternatively, attorneys are more confident in filing suits against) firms when the firm is thought to have behaved negligently with consumer information, relative to the firm being the unfortunate victim of computer hardware theft.

Among all types of personally identifiable information (PII) requiring greater protection, we find that only the compromise of financial data is significantly correlated with the probability of lawsuit: the compromise of financial data increased the probability of lawsuit 5.1% (significant at the 1% level), which provides only partial support for H1d. That is, the odds of a firm being sued are 6 times greater when the breach involved the loss of financial information.

Surprisingly, however, not all forms of data were found to be positively correlated with litigation. Indeed, breaches involving the compromise of medical or credit card data produce no significant effect. The cause for this could be that plaintiffs (and attorneys) believe that loss of financial information may more easily lead to financial harm, thereby elevating their subjective belief of a successful lawsuit. That is, they may feel that it is easier to justify bringing a claim for the breach of financial information because of the increased risk of harm.

Overall, we find that our hypotheses support theoretical models of litigation. In this arena, dominated by class-action practice, parties appear to behave in a rational and wealth-maximizing manner. In the context of data breaches, this translates to a higher probability of a federal lawsuit given evidence of actual financial loss, stronger claims of negligence (unauthorized disposal of information), and heightened protection of personal financial information. However, notwithstanding the statistically significant results, none were large in magnitude. That is, no marginal effect was larger than 5%. It is yet unclear whether the magnitude of these findings is, in itself, unexpected, though it does warrant further consideration.

Next, we examine the characteristics of data breach lawsuits leading to settlement.

5. WHICH DATA BREACH LAWSUITS SETTLE?

5.1. Hypotheses

Section 4 leveraged the theoretical analysis of dispute litigation to develop hypotheses explaining the probability of a federal data breach lawsuit. We continue that process to develop hypotheses regarding the probability of settlement once a suit has been filed.

Cooter and Rubinfeld (1989) consider that a plaintiff (and her attorney) will decide to settle when the expected gains from settlement exceed the expected gains from trial. However, the vast majority of data breach lawsuits terminate before trial, either through dismissal (motion to dismiss or summary judgment) or by settlement. Indeed, of over 230 suits in our dataset, we observe only two instances of a plaintiff prevailing on a favorable ruling by a judge or jury.[32] Therefore, we can simplify the theoretical model by stating that a plaintiff (and her attorney) will settle when the expected benefits from a settlement award exceed the cost of further litigation.

[32] Conner v. Tate, 130 F. Supp.2d 1370 (ND Ga. 2001) in which a woman allegedly disclosed an illegally wiretapped conversation to local police, and Beaven et al. v. US Department Of Justice, 5:03-cv-00084 (ED La. 2007) in which the plaintiffs alleged a violation of the Privacy Act (1974).

We now adapt this theory to data breach litigation by examining conditions that would increase either the probability or magnitude of settlement.

The recognition of the legal merits or "case strength" of a lawsuit has been the topic of much analysis in legal scholarship (see, generally, Boyd and Hoffman, 2010, and Eisenberg and Lanvers, 2009; and see Johnson et al., 2007, Cox et al., 2008, and Choi, 2007, in regard to securities class action litigation). Data breach lawsuits are often dismissed because of lack of identity theft following the breach (GAO, 2007). For example, judges often determine that the plaintiff could not show that she suffered harm in a sufficiently concrete way to justify her proceeding in a lawsuit.[33] However, there are cases when plaintiffs *do* suffer actual harm and *are* therefore able to overcome this procedural obstacle and obtain settlement.[34] Hence, we consider that in the context of data breach lawsuits the presence of "actual harm" represents an appropriate measure of a meritorious legal claim that should affect the probability of settlement. Therefore, *the probability of settlement is positively correlated with lawsuits in which the plaintiff is able to demonstrate actual harm (H2a).*

A second factor which may affect the magnitude of the settlement award is whether, in class action lawsuits, the class achieves certification. Class certification represents the difference between damages potentially awarded to only a few named plaintiffs, versus thousands or millions of plaintiffs. Indeed, "class certification stands not as a mere judicial byway on the road toward full-fledged trial on the merits but, almost invariably, as the last significant judicial checkpoint on the road toward settlement" (Nagareda, 2010, p152). Therefore, *the probability of settlement is positively correlated with achieving class certification (H2b).*

A final driver potentially affecting the magnitude of settlement is statutory damages. Plaintiffs bring many kinds of common law claims (e.g. negligence, breach of contract) and statutory causes of action. For example, the Computer Fraud and Abuse Act (CFAA), the Fair Credit Reporting Act, and Electronic Communications Privacy Act. A defining characteristic of these Acts is their mere violation can justify plaintiff relief through statutory damages. For example, the Wiretap Act allows recovery up to $100 per day or $1000, whichever is greater;[35] the CFAA allows statutory damages of $5000 per incident (record compromised). Hence, we consider that defendants may be more likely to settle when complaints include causes of action with statutory damages. The reasons are twofold. First, these allegations shift the burden from the plaintiff having to demonstrate harm to the defendant having to prove that they did not violate the law, increasing the defendant's cost of litigation. Indeed, "the only real significant liability threat to those companies sustaining a data breach is the advent of statutory damages – damages that would ensue with or without any showing of real harm to a plaintiff" (Paray, 2011). Second, there may be a saliency effect when the defendant is forced to consider the potentially massive damage award that is the product of the statutory damages and the size of the class. Therefore, *the probability of settlement is positively correlated with lawsuits in which the plaintiff seeks statutory damages (H2c).*

[33] In Shafran v. Harley-Davidson, Inc., 1:07-cv-01365 (SD N.Y. 2008), "Plaintiff has failed to show an actual resulting injury that might support a claim for damages. As damages are an essential element of each of plaintiff's claims, plaintiff's claims fail as a matter of law."

[34] In Stollenwerk et al. v. Tri–West Healthcare Alliance 05-16990 (9th Cir. 2007), "[Plaintiff's] personal data was used on six occasions to open or to attempt to open unauthorized credit accounts in [plaintiff's] name. Unknown individuals successfully opened at least two credit accounts and generated more than $7,000 in unauthorized charges to these accounts."

[35] 18 U.S.C.A § 2520(c)(2).

5.2. Descriptive Statistics

To address our second research question, we relax the restrictions imposed in Section 4 and employ our full set of federal data breach lawsuits. Note that this dataset is more comprehensive than that used in Section 4, in that it includes all federally-litigated breaches. Therefore, our sample now consists of 231 lawsuits filed in 50 unique district courts from 2000 to 2011.[36] Because we seek to compare settled and dismissed cases, we omit 23 pending lawsuits and 2 cases which ended in trial decision. We also drop 42 public actions, because they are perfect predictors of settlement (i.e., the government never loses). The resulting dataset of 164 observations consists of lawsuits that terminated either by settlement (n=86) or dismissal (n=78). The left panel in Figure 5 illustrates the number of such cases sorted by year of disposition, while the right panel shows the settlement rates, by year of disposition.

<Insert Figure 5 here>

An interesting finding from this analysis is that overall settlement rate in our dataset (86/164 = 52%) is higher than legal privacy scholarship would suggest, but also lower than the approximately 80% settlement rate expected from most tort cases (Eisenberg and Lanvers, 2009, table 4).[37] The right panel of Figure 5 shows an early, erratic trend followed by a fairly constant settlement rate of around 50% after 2004.[38]

Figure 6 examines the proportion of cases in which plaintiffs were able to show actual damage (H2a), where the case achieved class certification (H2b), and where the plaintiff sought statutory damages (H2c). Note that in the following figures, percentages sum to 100% in each adjacent column pair.

<Insert Figure 6 here>

The top two pair-wise comparisons illustrate a similar result: the majority of cases that allege actual harm or achieved class certification, settled.[39] That is, of the cases that alleged actual harm (n=28), 71% of them settled, whereas only 49% of them without actual harm (n=135) settled. Similarly, of the cases that achieved class certification, 85% settled, whereas when the class was not certified, only 48% settled.[40] The bottom panel, on the other hand, is more balanced. Of the cases that include causes of action with statutory damages, 59% settled, and only about 45% otherwise. Again, note that these figures reflect data from all years, and that the patterns presented in both panels are robust across individual years.

Interestingly, however, the top panels show another consistent result: data breach lawsuits *lacking* actual harm or class certification are almost as equally likely to reach settlement as dismissal.

[36] Note that this section employs observations only related to lawsuits, and no longer the Dataloss clearinghouse.

[37] Although Eisenberg and Lanvers (2009) do find settlement rates as low as 50% for constitutional cases.

[38] It is likely only a coincidence that this 50% settlement rate matches the theoretical settlement rate of Priest and Klein (1984) since that rate defines a trial settlement rate, whereas we observe no trials in this sample.

[39] Note that these are not the same cases. In fact, there are only two cases which both included class certification and actual harm.

[40] Note that for the purpose of this figure, we include 4 cases which were certified only for the purpose of settlement (i.e. not litigation). When estimating Eq. 2, however, class certification for these 4 cases is coded as 0. We thank Paul Bond for bringing this distinction to our attention.

That is, in cases without these characteristics, the plaintiff faces approximately a 50/50 chance of obtaining a settlement. We further discuss this observation in the following section.

5.3. Estimating Model

We again employ a discrete outcome model to estimate the probability of settlement,[41]

$$settlement_i = \alpha_0 + ActualHarm_i + CreditMonitoring_i + ClassCertified_i + StatutoryDamages_i + Breach_i + Industry_i + Forum_i + Year_i + \varepsilon_i \qquad (2)$$

where *settlement* is a binary outcome variable coded as 1 if the lawsuit, *ie,* terminated in settlement and 0 otherwise.[42] *ActualHarm* is coded as 1 if the plaintiff's complaint alleges an actual loss due to the breach (for instance, if the plaintiff alleges fraudulent charges on a credit card, stolen money from a checking or savings account, or other such costs incurred from criminal activity). Other forms of alleged harm such as preventive costs from credit monitoring, emotional distress, invasion of privacy, embarrassment are coded as 0.[43] *CreditMonitoring* is coded as per Eq. 1. *ClassCertified* is coded as 1 of the suit achieved class certification. *StatutoryDamages* is coded as 1 if the complaint alleged violation of a federal statute allowing for statutory damages. *Breach* is a vector of controls for the size and cause of the data breach and types of information lost (PII). *Industry* is a vector of controls for the firm's industry, whether the firm is non-profit or publicly traded (as in Eq. 1). *Forum* is a vector of controls for the circuit court region in which the case was heard, whether the case was removed from state court, and the sex of the judge (Boyd and Hoffman, 2010). As a measure of the complexity of the case, we also control for the number of causes of action and number of times the complaint was amended. *Circuit* controls for the circuit region where the case was litigated. We may also be concerned with forum shopping (litigants filing cases in more favorable districts), and the possibility of a successful outcome in one case affecting the outcome in another case (formally referred to as the stable unit treatment value assumption, or SUTVA; Rubin, 1990). We partially control for both of these effects by coding a variable, *Standing*, equal to 1 if a suit is filed in a district which had granted standing to a plaintiff in a previous data breach lawsuit. We also include a proxy for the size of the firm with the log of the total number of employees of the defendant. In cases with multiple defendants, we consider only the first named defendant. We also code a variable, *Multisuit*, as 1 if a data breach resulted in more than one consolidated lawsuit. *Year* is a vector of dummies representing the year when the case was disposed and ε_i is the random error term, assumed to be independent of observed covariates. Descriptive statistics for the variables used in Eq. 2 are shown in Table 3.

Given that our analysis examines the binary outcome of lawsuits (settlement versus dismissal) estimation of Eq. 2 should not suffer from the familiar issue of case-selection (Priest and Klein, 1984; Clermont and Eisenberg, 1998). Also, recall that we are implicitly examining the determinants of lawsuit outcome, conditional on a complaint having been filed. Therefore, we are careful to interpret any parameter estimates as the marginal impact on the probability of a federal lawsuit, *not a data breach*, being settled. This is because such a model would suffer from

[41] Again, we present a linear probability model for readability, though the estimating model is nonlinear.

[42] We determine settlement either by contacting attorneys directly, or from case dockets. Also, recall that data breach lawsuits may terminate for a number of reasons and we therefore focus our analysis to address the probability of a lawsuit resulting in settlement, versus being dismissed as a matter of law, summary judgment, or other technical legal reasons, such as subject matter jurisdiction.

[43] Note that we take no normative position in regard to whether such harms *should* be considered as actual harm. We merely make this distinction for the purpose of hypothesis testing.

selection bias if the set of breaches that resulted in settlement were systematically different from breaches otherwise disposed.

5.4. Results

Table 2 presents the results of Eq. 2, reporting the average marginal effects of the explanatory variables on the probability of settlement.[44] Model 1 includes just the variables of interest and year fixed effects, while Model 2 includes subsequent controls for *Breach* and *Industry* characteristics. Model 3a and 3b include the full set of controls and estimate the same equation, with Model 3b presenting the results as odds-ratios.

<Insert Table 2 here>

Note again that the results are generally robust across all models, with Model 3a providing the best fit for the data and including all covariates. Though not shown, results are also robust to the exclusion of individual years 2005-2010. Robustness checks using alternative specifications of the year of disposition (i.e. the year of the breach, and the year the complaint was filed) reveal qualitatively similar results. Further discussions therefore focus on estimations from Model 3a.

These results suggest that, after controlling for all variables, plaintiff allegations of financial harm are correlated with a 30% increase in the probability of settlement (from 52% to 68%, significant at the 1% level), supporting H2a. Similarly, the certification of a class action, as Nagareda (2010) theorizes, increases the probability of settlement by 30% (significant at the 1% level), supporting H2b. In addition to each being highly statistically significant, these estimates are also large in magnitude and therefore of strong practical significance.

On the other hand, we find that causes of action asserting a violation of a federal statute with statutory damages were not positively correlated with settlement, lending no support for H2c. This finding is somewhat surprising given that this hypothesis had a strong theoretical and practical justification: these claims can help shift the burden of proof from the plaintiff in having to demonstrate actual harm to the defendant in having to prove it did not violate the law. A possible explanation for this result could be that the novelty of federal-statute based privacy litigation made it harder for the parties to arrive upon a shared understanding of the merits.

Of the breach characteristics, only breaches caused by cyber attacks were found to be positively and significantly correlated with settlement (29%, significant at the 1% level), relative to lost/stolen hardware. That is, the odds settling for a litigated breach caused by cyber attack are almost 10 times greater relative to a litigated breach caused by lost or stolen hardware. The size of the breach was again not found to be positively correlated with the outcome. This is also somewhat surprising, as one might expect that defendants would be strongly induced to settle due to potentially greater publicity from larger breaches.

Of the types of information compromised, breaches relating to financial and credit card information were found to be negatively correlated with settlement (though not statistically significantly so). It is therefore interesting that while the compromise of financial information appears to lead to more litigation, it does not appear to increase a plaintiff’s chance of a settlement. Instead, loss or theft of medical information is most strongly correlated with settlement (31%, significant at the 1% level).

[44] Note that some observations were dropped because of perfect prediction while some covariates were dropped because of colinearity, resulting in fewer than the full set of 164 observations being estimated.

Overall, despite our relatively small sample size, we are still able to show statistically significant results. Interestingly, while the compromise of financial data and breaches caused by improper disposal/disclosure appeared to drive litigation (Table 1), the compromise of medical data and breaches caused by cyber attack appear to drive settlement (Table 2). Moreover, Figure 6 demonstrates that even without actual harm or class certification, lawsuits still tend to settle about half of the time. That is, cases with merit were much more likely to settle - yet, cases without merit still settle about half of the time.

A possible explanation could be that defendants choose to settle for reasons entirely unrelated to the merits of a case. For example, they may be rationally choosing to settle to avoid further litigation costs, publicity, or distraction. Specifically, defendants may be balancing between the costs of an immediate and "certain" settlement, versus a future "uncertain" amount (that includes a settlement award with some probability in addition to legal fees). Nevertheless, a full explanation, we believe, warrants more consideration.

6. LIMITATIONS

The first limitation of this work stems from the lack of observed state data breach lawsuits, which limits our interferences to federal data breach actions. However, under the Class Action Fairness Act (CAFA, 2005), we are relatively confident that all large class actions (and certainly multi-state actions) would, indeed, be either filed in, or removed to, federal court. Conversations with defense attorneys strongly support this intuition. Moreover, the absence of these suits would not bias our regression estimates, as we are, in effect, drawing inferences by pooling non-litigated and state-litigated breaches. Further, we partially controlled for this effect by estimating Eq. 1 using only data breaches compromising greater than 1000 records and found that our results were robust to even just these breaches.

Another possible limitation concerns the relatively small sample of "success" observations of Eq. 1 and Eq. 2. An often-cited rule of thumb is that one should observe 10 "success" observations for each parameter being estimated (Peduzzi, 1996), which may otherwise lead to biased or inefficient parameter estimates. We partially address this concern by first estimating Eq. 1 and Eq. 2 using only year controls (Model 1 in both Table 1 and Table 2) and thereby fulfilling this rule of thumb. We find that, in general, parameter estimates are robust to subsequent models that control for additional effects. Moreover, any concerns of efficiency would produce a lower bound on any statistical significance. Therefore, our results as presented reflect conservative significance levels.

When estimating the probability that a lawsuit will result in settlement, it may be the case that individuals (through their lawyers) contact the breached firm on their own in hopes of achieving a 'backdoor' settlement without first filing a formal legal complaint. However, discussions with privacy attorneys suggest that this does not happen with any practical frequency.

Finally, from our use of Westlaw and PACER, we are unable to obtain a complete record of the discovery process. Hence, it is possible that some pretrial activity occurs which may influence the outcome of a lawsuit and that which we are unable to control. For example, discussions of the defendant's liability insurance coverage or IT security practices may occur which potentially drive a defendant to settle, whereas otherwise they would not. We argue, however, that the vast majority of (if not all) defendants in our sample will have basic general liability insurance policies - information about which would automatically be discoverable according to the Federal Rules of Civil Procedure 26(a)1(A)(iv), thereby reducing any variation across observations.

7. DISCUSSION

Recent events concerning breaches of consumer personal information have prompted a flurry of lawsuits by alleged victims of identity theft. These disputes have generated considerable Congressional activity concerning the collection, use, and dissemination of personally identifiable consumer health, financial and behavioral information. But is litigation an effective solution?

Consider both the probability of data breach litigation and settlement. On one hand, the overall federal litigation rate for reported data breaches is only about 4%, which may provide comfort to firms (potential defendants) that collect personal information. On the other hand, the settlement rate for all known federally litigated breaches is much higher than one might expect (50%), which would alternatively be encouraging to plaintiffs. Moreover, if actual harm (as defined within this manuscript) is indeed an appropriate measure of case merit, then the results presented in Figure 6 and Table 2 may provide some assurance that data breach lawsuits are being appropriately disposed of, on average. That is, those cases that *should* settle (because of the presence of actual harm), *do* settle. In fact, the top left panel of Figure 6 suggests that defendants settle perhaps too often (i.e. in absence of actual financial harm, and therefore case merit).

In regard to settlement awards, we naturally find great variation. After contacting litigating attorneys for the 86 settlements, award details were acquired for 28 of them, with details regarding the remaining cases either held privileged (10 cases) or unknown entirely (48). The mean value of settlements awarded to plaintiffs was about $2,500 per plaintiff (min = $500, max = $15k , n=19) with most awards being a nominal amount of around $500 and often awarded to named plaintiffs only. Attorney fees, on the other hand were substantially larger, with a mean sum of $1.2m (min = $8k, max, $6.5m, n=15). Importantly, however, settlements may also provide individual redress for identity theft losses and expenses, and cy pres awards to research, non-profits, and charities which have ranged from $50k, to $9.5m.

A final observation from this work lies with the diversity of the legal claims brought by plaintiffs. From our data, we identified over 86 unique causes of action (from only 231 cases) for essentially the same event: the unauthorized disclosure of personal information. We found 34 different kinds of tort causes of action, 15 contract, 4 violations of state statutes, and 33 violations of federal statutes. The 20 most common are shown in Figure 7.

<Insert Figure 7 here>

What does this suggest about how well equipped the current legal system is in efficiently resolving modern data breach harms? Generally speaking, common law is well suited to address many sorts of injuries in the presence of clear physical, property or economic loss. But privacy harms differ in important ways: there are many means by which personal information can be collected, used and distributed, and there are equally as many manners in which consumers may feel emotionally, statistically, or financially harmed. This is no more clearly reflected than in Figure 7. Perhaps precisely because of this complexity, it is unfair (and perhaps unrealistic) to expect a single privacy statute – a single private right of action – to provide redress for all privacy harms.

8. CONCLUSION

The unauthorized disclosure of personal information by firms imposes externalities on consumers from medical and financial fraud, and other forms of identity theft. A great deal of recent federal Congressional and Agency activity has been devoted to reducing the risk of data breaches, and crafting legislation to empower consumers to bring federal actions. However, very little is known about the characteristics of data breach litigation and the outcomes of these cases.

This manuscript hopes to address and inform this policy debate by examining two main research questions: which data breaches are being litigated?; and which data breach lawsuits settle? Our first research question examines federally litigated breaches resulting from reported data breaches, while the second question includes all known federal lawsuits related to the unauthorized disclosure of personal information.

Our results suggest that the odds of a firm being sued in federal court was 3.5 times *greater* when individuals suffered financial harm, but over 6 times *lower* when they provided free credit monitoring. Moreover, the odds of a firm being sued from improperly disposing data were 3 times greater, relative to breaches caused by lost/stolen data, and 6 times greater when the data breach involved the loss of financial information.

Turning to data breach settlements, our results suggest that defendants settle 30% more often when plaintiffs allege financial loss from a data breach, or when faced with a certified class action suit. Surprisingly, however, plaintiffs seeking statutory damages were not more likely achieve a settlement. The odds of a settlement were 10 times greater when the breach was caused by a cyber-attack, relative to lost or stolen hardware. While the compromise of financial information lead to more litigation, it did not appear to increase a plaintiff's chance of a settlement. Instead, compromise of medical information was most strongly correlated with settlement.

In addition to the regression analysis presented, we performed a number of robustness checks. We verified that the proportions of federally-litigated breaches resulting in actual harm were generally consistent across 2005-2010, as were the causes of litigated breaches. Similarly, we find consistent results when examining the portion of settled cases in which actual harm was alleged, which achieved class certification, and in which plaintiffs sought statutory damages. Further, regression results were robust to probit analysis, and to alternative model specifications, as shown in Table 1 and Table 2.

We also uncovered some novel descriptive data. For example, it seems that only about 4% of reported breaches resulted in federal litigation, and that contrary to conventional legal scholarship, the overall settlement rate of known federal lawsuits was around 50%. Moreover, it is perhaps staggering that of the federal actions coded, we found over 86 different causes of actions brought by plaintiffs for essentially the same kind of event.

Overall, we believe this research can be of use to various parties. First, it can help provide firms with prescriptive guidance regarding the relative chances of being sued, and having to settle. This research could also be useful to insurance markets as a means for assessing and pricing cyber-insurance policies. Moreover, we believe that this work can help inform both plaintiff and defense attorneys in better understanding overall trends of data breach litigation. Finally, we hope that our research can inform the policy debate and help create a balanced privacy framework protecting both the interests of consumers who provide personal information, and organizations that collect and innovate using this information.

9. REFERENCES

Acquisti, A., Friedman, A. & Telang, R. 2006. Is There a Cost to Privacy Breaches? An Event Study. Fifth Workshop on the Economics of Information Security. Jun. 26.

Boyd, C., & Hoffman, D. 2009. Disputing Limited Liability. Northwestern University Law Review, 104; Temple University Legal Studies Research Paper No. 2009-38.

Boyd, C., and Hoffman, D. 2010. Litigating Toward Settlement. Temple University Legal Studies Research Paper No. 2011-8. Available at SSRN: http://ssrn.com/abstract=1649643.

Bureau of Justice Statistics, 2011. Identity Theft Reported by Households, 2005-2010. U.S. Department of Justice, Bureau of Justice Statistics.

Campbell, K., Gordon, L., Loeb, L., & Zhou, L. 2003. The Economic Cost of Publicly Announced Information Security Breaches: Empirical Evidence from the stock market. Journal of Computer Security, 11(3) 431-448.

Cavusoglu, H., Mishra, B. & Raghunathan, S. 2004. The Effect of Internet Security Breach Announcements on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers. International J. of Electronic Commerce 9(1).

Clermont, K. and Eisenberg, T. 1998. Do Case Outcomes Really Reveal Anything About the Legal System? Win Rates and Removal Jurisdiction, 83 Cornell Law Review, 581.

Clermont, K. and Eisenberg, T. 2002. Litigation Realities. 88 Cornell Law Review, 119-154.

Choi, S. 2007. Do the Merits Matter Less after the Private Securities Litigation Reform Act? Journal of Law, Economics, & Organization, 23(3), 598-626.

Citron, D. 2007. Reservoirs of Danger: The Evolution of Public and Private Law at the Dawn of the Information Age. 80 S. Cal. L. Rev. 241-248.

Citron, D. 2011. Mainstreaming Privacy Torts. 99 Cal. L. Rev. 101-189.

Cooter, R., and Ulen, T. 2008. Law & Economics. Pearson Education, Inc, 5th ed.

Cooter, Robert D., Rubinfeld, Daniel L. 1989. Economic Analysis of Legal Disputes and Their Resolution. Journal of Economic Literature, 27(3).

Cox, J., Thomas, R. and Bai, L. 2008. There are Plaintiffs and... There are Plaintiffs: An Empirical Analysis of Securities Class Action Settlements, 61 Vanderbilt Law Review, 355.

Department of Commerce, 2010. Commercial Data Privacy and Innovation in the Internet Economy: A Dynamic Policy Framework, US Department of Commerce.

Dickey, T., Ganz, D. and Lever, J. 2011. Privacy Protection and Data Breaches: HR Tip of the Month, The Blog of the National Law Review. Available at http://nationallawforum.com/2011/04/24/privacy-protection-and-data-breaches-hr-tip-of-the-month/. Last accessed July 24, 2011. April 24.

Dunbar, F & Sabry, F. 2007. The propensity to sue: why do people seek legal actions? Journal of the National Association for Business Economics 42(2).

Eisenberg, T., & Lanvers, C. 2009. What is The Settlement Rate and Why Should We Care? Journal of Empirical Legal Studies, 6(1), 111-146.

Federal Trade Commission. 2010. Protecting consumer privacy in an era of rapid change. Federal Trade Commission.

Felstiner, W., Abel, R., & Sarat, A. 1980. The Emergence and Transformation of Disputes: Naming, Blaming, Claiming. Law and Society Review. 15(3/4), 631-654.

Government Accountability Office. 2007. Data Breaches Are Frequent, but Evidence of Resulting Identity Theft Is Limited; However, the Full Extent Is Unknown. GAO publication GAO-07-737.

Gordon, L.A., Loeb, M., & Zhou, L. 2011. The Impact of Information Security Breaches: Has There Been a Downward Shift in Costs? Journal of Computer Security.

Heckman, J. 1979. Sample selection bias as a specification error. Econometrica. 47(1), 153-161.

Hoffman, D., Izenman, A., & Lidicker, J. R. 2007. Docketology, District Courts, and Doctrine, 85 Wash. U. L. Rev. 681.

Hutchins, J. 2008. A New Frontier in Privacy Litigation: The Advent of Private Lawsuits Over Data Security Breaches. 2008 ABA Annual Meeting, Section of Litigation, New York, New York. August 8.

Johnson, M., Nelson, K. and Prichard., A. 2007. Do the Merits Matter More? The Impact of the Private Securities Litigation Reform Act, Journal of Law, Economics, & Organization, 23 (3), 627-652.

Kannan, K., Rees, J. & Sridhar, S. 2007. Market Reactions To Information Security Breach Announcements. International Journal of Electronic Commerce 12(1) 69-91.

Kim, P., Schlanger, M., Boyd, C., & Martin, A. 2009. How Should We Study District Judge Decision-Making? Journal of Law and Policy, 29(83).

Lerner, J. 2010. The Litigation of Financial Innovations. Journal of Law and Economics, 53(4), 807-831.

Lesemann, D. J. 2009. Once More unto the Breach: An Analysis of Legal, Technological and Policy Issues Involving Data Breach Notification Statutes.

Nagareda, R. A., 2010 Common Answers for Class Certification. Vanderbilt Law Review En Banc, 63, 149-170.

Paray, Paul, E. 2011. The Elephant in the Room: The Potential for Privacy Breach Statutory Damages. Digital Risk Strategies. February 18.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. 1996. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology 49:1373-1379.

Priest, G., and Klein, B. 1984. The Selection of Disputes for Litigation, 13 Journal of Legal Studies, 1-55.

Proskauer Rose LLP. 2009. California District Court Closes the Gap Left by Ruiz. Proskauer Rose LLP. Available from http://privacylaw.proskauer.com/2009/04/articles/data-breaches/california-district-court-closes-the-gap-left-by-ruiz/. Last accessed July 24, 2011. April 9.

Rice, D. 2007. Civil Actions For Privacy Violations 2007: Where Are We? Howard Rice Nemerovski Canady Falk & Rabkin.

Romanosky, S., & Acquisti, A. 2009. Privacy Costs and Personal Data Protection: Economic and Legal Perspectives of Ex Ante Regulation, Ex Post Liability and Information Disclosure. Berkeley Technology Law Journal, 24(3).

Romanosky, S., Sharp, R., & Acquisti, A. 2010. Data breaches and Identity Theft: When is Mandatory Disclosure Optimal? Paper presented at the Ninth Workshop on the Economics of Information Security (WEIS 2010), Harvard University, Cambridge, MA.

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.

Rubin, D, B. 1990. Formal Modes of Statistical Inference For Causal Effects. Journal of Statistical Planning and Inference, 25, 279-292.

Ryan, Ellen, “States offer data breach protection,” National Association of Attorneys General, no date. See http://www.naag.org/states-offer-data-breach-protection.php, last accessed Jan 25, 2012.

Serwin, A. 2009. Poised on the Precipice: A Critical Examination of Privacy Litigation Santa Clara Computer and High Technology Law Journal, 25(4).

Shavell, S. 1996. Any Frequency of Plaintiff Victory at Trial Is Possible, 25 Journal of Legal Studies, 493-501.

Solove, D. J., 2005. The New Vulnerability: Data Security and Personal Information. Securing Privacy In The Internet Age, Radin & Chander, eds., Stanford University Press, 2005; GWU Law School Public Law Research Paper No. 102.

Solove, D. 2007. The New Vulnerability: Data Security and Personal Information, in Securing Privacy In The Internet Age (Anupam Chander et al., eds.). 111, 115-116.

Solove, D. 2010. Are People Really Harmed By a Data Security Breach? Concurring Opinions Legal Blog. Available at http://www.concurringopinions.com/archives/2010/09/are-people-really-harmed-by-a-data-security-breach.html.

Telang, R. & Wattal, S. 2007. An Empirical Analysis of The Impact Of Software Vulnerability Announcements On Firm Stock Price. IEEE Transactions on Software Engineering paper 33(8) 544-557.

Verizon Communications Inc. 2010. 2010 Data breach investigations report. Verizon Communications Inc.

Viscusi, W. K. 1986. The Determinants of the Disposition of Product Liability Claims and Compensation for Bodily Injury. Journal of Legal Studies, 15(2), 321-346.

Wittman, D. 1988. Dispute Resolution, Bargaining, and the Selection of Cases for Trial: A Study of the Generation of Biased and Unbiased Data. Journal of Legal Studies, 17(2), 313-352.

10. APPENDIX

10.1. Figures

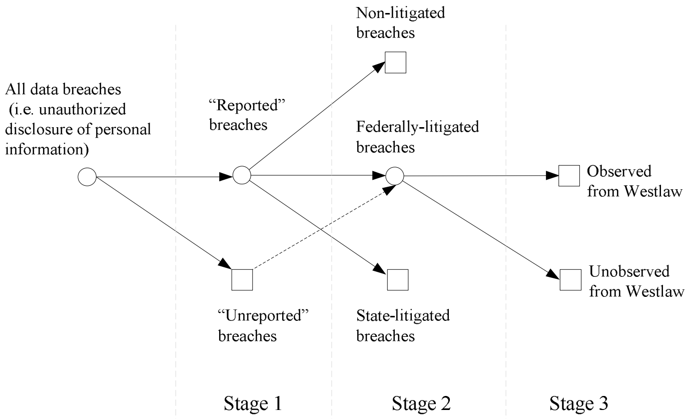

Figure 1: Data generating process

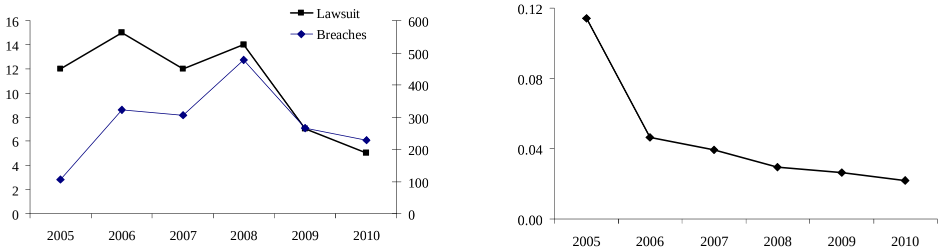

Figure 2: Reported breaches vs. known lawsuits

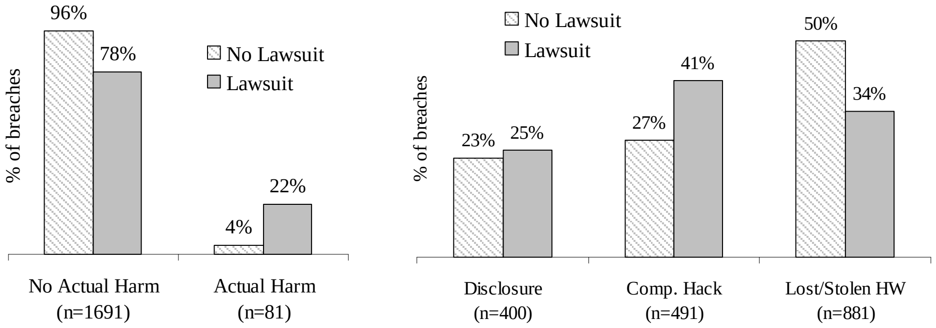

Figure 3: Presence of harm, and cause of breach

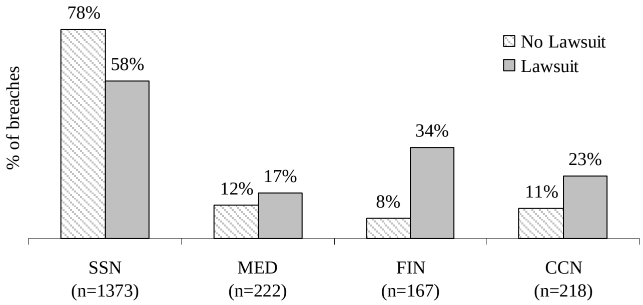

Figure 4: Types of personal information compromised

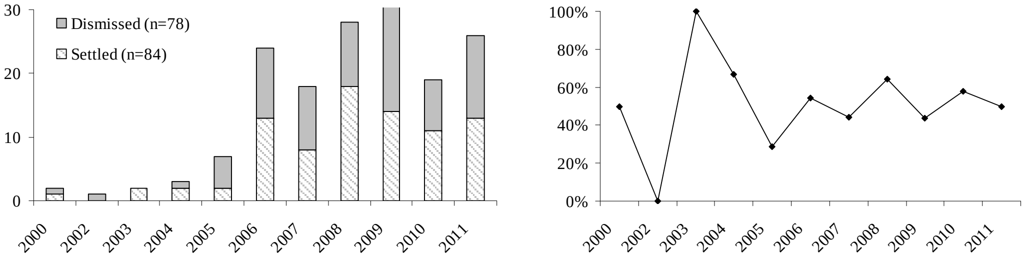

Figure 5: Settlements and dismissals

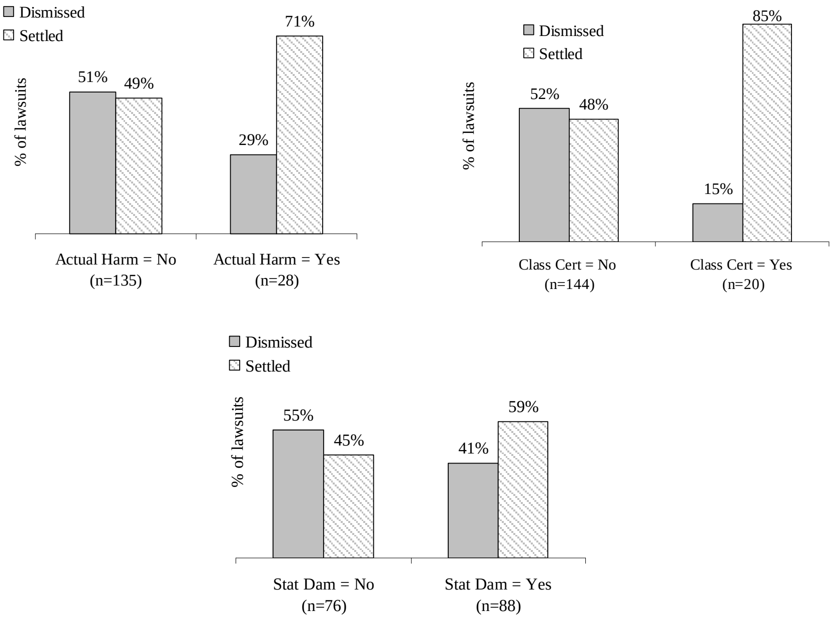

Figure 6: Pair-wise comparisons by settlement

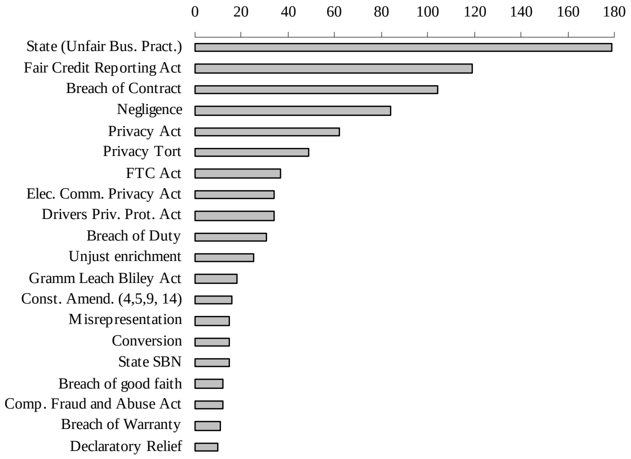

Figure 7: 20 most common causes of action

10.2. Tables

Table 1: Regression results of Eq. (1)

Dep var: lawsuit	Basic model (1)	All data types (2)	Full model (3a)	Full model (odds ratio; 3b)
Log(records)	0.013***	0.012***	0.009***	1.592
	(0.002)	(0.002)	(0.001)	
Actual Harm	0.046***	0.045***	0.025*	3.557
	(0.014)	(0.014)	(0.014)	
Credit Monitoring	-0.017*	-0.017*	-0.037***	0.152
	(0.009)	(0.009)	(0.010)	
Cause_Disclosure	0.025*	0.013	0.027**	3.122
	(0.013)	(0.011)	(0.013)	
Cause_Hack	0.004	-0.001	0.016	2.085
	(0.010)	(0.009)	(0.012)	
PII_SSN	-0.006	-0.001	0.010	1.729
	(0.009)	(0.009)	(0.007)	
PII_Medical	0.034**	0.025*	0.010	1.619
	(0.016)	(0.014)	(0.014)	
PII_Financial	0.094***	0.079***	0.051***	5.875
	(0.025)	(0.023)	(0.016)	
PII_Credit Card	0.019	0.018	0.005	1.263
	(0.014)	(0.013)	(0.010)	
Year Controls	Y	Y	Y	
PII Controls		Y	Y	
Industry Controls			Y	
Observations	1772	1772	1772	
Log likelihood	-174.63145	-165.70501	-131.40823	
Pseudo R2	0.3733	0.4053	0.5284	

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Regression results Eq. 2

Dep var: settled	Basic model (1)	With breach and industry controls (2)	Full model (3a)	Full model (odds-ratios, 3b)
Actual Harm	0.275***	0.310***	0.302**	9.19
	(0.095)	(0.106)	(0.119)	
Credit Monitoring	-0.041	-0.008	0.102	2.11
	(0.101)	(0.130)	(0.145)	
Class Certification	0.407***	0.327**	0.304***	9.31
	(0.140)	(0.143)	(0.117)	
Statutory Damages	0.163**	0.192*	0.097	2.04
	(0.078)	(0.103)	(0.096)	
Log(records)		0.003	-0.006	0.959
		(0.009)	(0.009)	
Breach_Disclosure		0.085	0.170	3.63
		(0.138)	(0.135)	
Breach_Hack		0.243**	0.290***	9.59
		(0.122)	(0.111)	
PII_SSN		0.113	0.078	1.79
		(0.101)	(0.108)	
PII_Medical		0.310**	0.312***	15.00
		(0.142)	(0.094)	
PII_Financial		-0.123	-0.072	0.589
		(0.114)	(0.096)	
PII_Credit Card		-0.083	-0.045	0.715
		(0.118)	(0.109)	
Year Controls	Y	Y	Y	
Circuit Court Region Controls		Y	Y	
PII Controls		Y	Y	
Industry Controls		Y	Y	
Forum Controls			Y	
Observations	158	156	154	
Log Likelihood	-93.475653	-78.888117	-64.067586	
Pseudo R^2	0.1456	0.2701	0.3991	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Summary Statistics for Eq. 1 and Eq. 2

	Eq. 1, n = 1772		**Eq. 2, n = 164**	
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
Log(records compromised)	7.91	2.87	9.58	5.46
Actual Harm	0.05	0.21	0.17	0.38
Breach_Disclosure	0.23	0.42	0.58	0.50
Breach_Hack	0.28	0.45	0.23	0.42
PII_SSN	0.77	0.42	0.37	0.48
PII_Medical	0.12	0.33	0.09	0.29
PII_Financial	0.09	0.28	0.27	0.45

PII_Credit Card	0.12	0.32	0.26	0.44
PII_Email	0.03	0.16	0.04	0.19
PII_NameAddress	0.77	0.42	0.34	0.47
PII_DateofBirth	0.16	0.37	0.15	0.35
Ind_Business	0.27	0.44	0.49	0.50
Ind_Education	0.28	0.45	0.02	0.15
Ind_Financial	0.12	0.33	0.28	0.45
Ind_Government	0.18	0.38	0.12	0.32
Non-profit	0.03	0.16	0.18	0.38
Publicly traded	0.12	0.32	0.41	0.49
Class certification			0.12	0.33
Statutory Damages			0.54	0.50
Multisuit			0.18	0.38
Removed			0.14	0.35
Female Judge			0.24	0.43
Settled			0.52	0.50
Standing			0.08	0.27
Log(employees)			8.73	2.80