Empirical Analysis of Data Breach Litigation

Sasha Romanosky
Heinz College
Carnegie Mellon University
sromanos@cmu.edu

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

Alessandro Acquisti
Heinz College
Carnegie Mellon University
acquisti@andrew.cmu.edu

Abstract

The surge in popularity of social media, cloud computing, and mobile services has created an unprecedented opportunity for the collection, use and sale of personal consumer information. While these services provide many benefits, individuals suffer when this information is lost, stolen, or improperly accessed, causing them to seek legal redress. 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? By providing the first comprehensive empirical analysis of data breach litigation, our findings offer insights in the debate over privacy litigation versus privacy regulation.

Keywords: data breach, identity theft, privacy, litigation

Introduction

The surge in popularity of social media, cloud computing, and mobile services has created an unprecedented opportunity for the collection, use and sale of personal consumer information. While these services clearly provide many benefits to producers and consumers, individuals suffer harm when their personal information is lost, stolen, or improperly accessed, 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 (Bureau of Justice, 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; Federal Trade Commission, 2010). 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.

[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).

Acknowledgements: 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, by Temple Law School's Conwell Corps Program, and by the Information Law Institute at New York University School of Law. 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.

Romanosky, S., Hoffman, D., Acquisti, A. (2013). Empirical analysis of data breach litigation. *iConference 2013 Proceedings* (pp. 124-137). doi:10.9776/13162

However, little is known about the trends in data breach litigation – which breaches are litigated and which are not, and with what outcomes. Current scholarship examines only a narrow subset of lawsuits, usually focusing on high-profile cases or those with published opinions. And so, to our knowledge, no empirical research involving data breach lawsuits has been conducted. The purpose of this manuscript is to 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?

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.

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). An emerging body of legal scholarship also 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 (Kim et al., 2009; Hoffman et al., 2007; and Boyd and Hoffman, ND). 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).

Data

This manuscript combines a number of datasets described below. For the purpose of this manuscript, a data breach is defined broadly as the unauthorized disclosure of personal information by an organization.

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.[4] 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. (The lawsuit observations previously 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

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

and extracted those cases relating to unauthorized disclosure of personal information. 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 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.

Data Generating Process

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

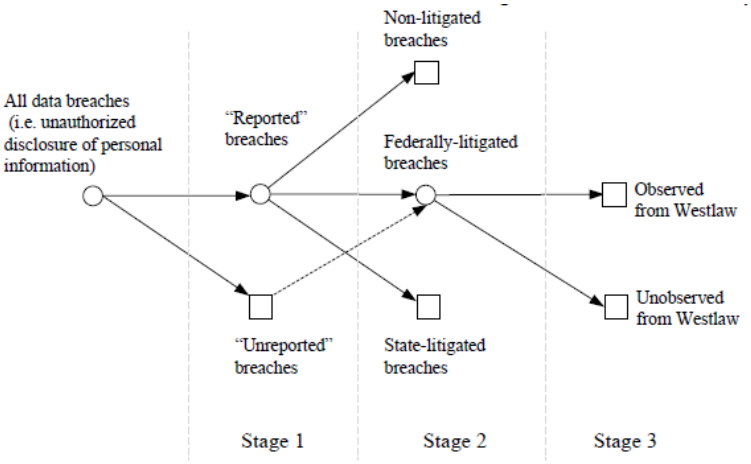

Figure 1. Data generating process.

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.[5] This group has been systematically collecting data breach information since 2005.[6]

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.[7] 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

[5] 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.

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

[7] 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.

pooling state- and non-litigated breaches we are still able to obtain unbiased estimates of federal lawsuits resulting from reported data breaches.[8]

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 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." (Boyd and Hoffman, ND). 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?").

Which Data Breaches are Litigated in Federal Court?

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.[9] *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) 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*

[8] 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).

[9] 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.

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).* [10]

Descriptive Statistics

Our final sample of Datalossdb data consists of 1,772 US data breach observations, of which only 65 (3.7%) were litigated in federal court. 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.

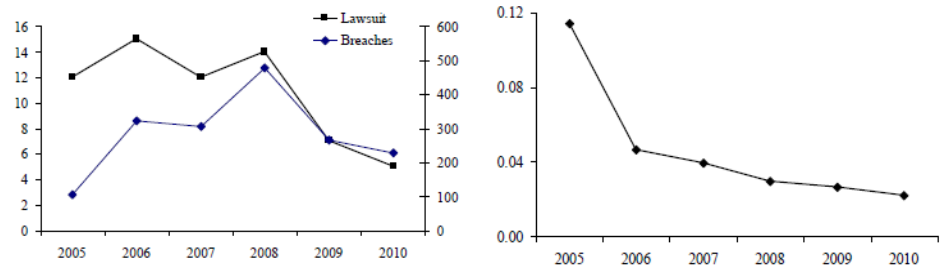

Figure 2. Reported breaches vs. known lawsuits.

[10] 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.

Figure 3 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.

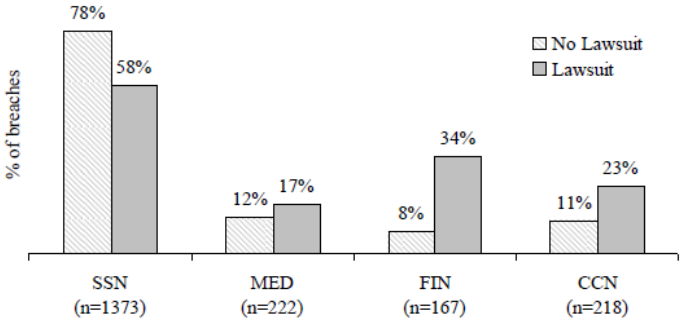

Figure 3. Types of personal information compromised.

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.

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,[11]

$$lawsuit_i = \alpha_0 + ActualHarm_i + CreditMonitoring_i + Cause_i + ProtectedPII_i + Controls_i + \varepsilon_i \quad (1)$$

where *lawsuit* is a binary variable that takes the value 1 if a reported breach, i, results in a federal lawsuit, and 0 otherwise.[12] 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.[13] *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.[14] *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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

lost/stolen hardware.[15] *ProtectedPII* is a vector of dummies representing types of personally identifiable information (PII) should require a heightened level of protection, as described in the hypothesis: social security number, medical, financial, credit card). *Controls* represents a vector of controls for all other data types (email address, name/address, etc), industry of the breached firm, whether the firm was a non-profit or publicly traded, and year dummies (2005 to 2010). ε_i is the random error term, assumed to be independent of the observed covariates. Descriptive statistics for the variables used in Eq. 1 are shown in Table 3.

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.[16] 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.

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.[17]

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 greater, relative to breaches caused by lost/stolen data (significant at the 5% level), supporting H1c. These results suggest that individuals are much more likely to punish 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.

[15] 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."

[16] 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.

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

Table 1
Regression results 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*

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.

Which Data Breach Lawsuits Settle?

Hypotheses

The previous section 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 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. 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. 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, ND, 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). However, there are cases when plaintiffs do suffer actual harm and are therefore able to overcome this procedural obstacle and obtain settlement. 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; 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).

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 (though we omit pending and public action suits). The resulting dataset of 164 observations consists of lawsuits that terminated either by settlement (n=86) or dismissal (n=78).

Figure 4 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.

The top two pair-wise comparisons illustrate a similar result: the majority of cases that allege actual harm or achieved class certification, settled. 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. 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.

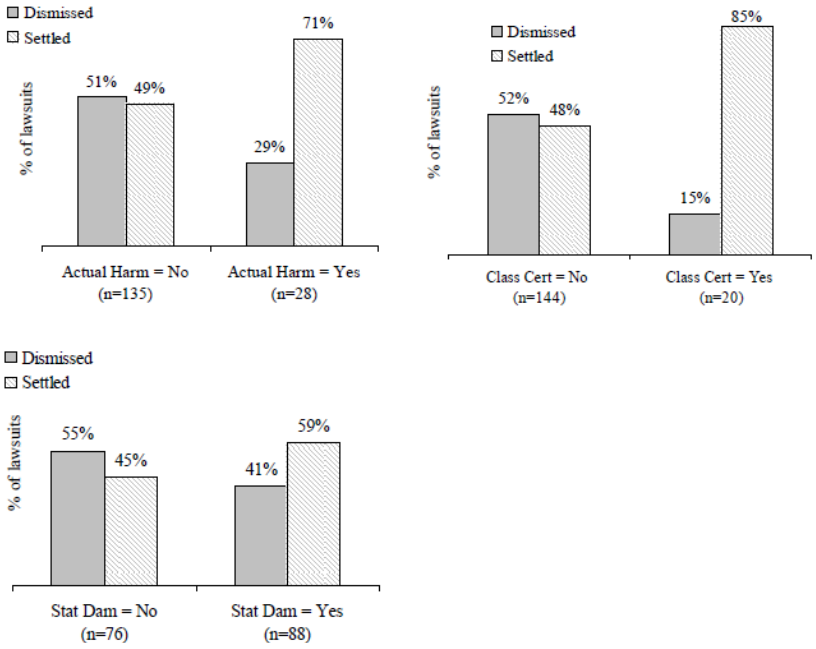

Figure 4. Pair-wise comparisons by settlement.

Estimating Model

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

$$settlement_i = \alpha 0 + ActualHarm_i + ClassCertified_i + StatutoryDamages_i + Controls_i + \varepsilon_i \quad (2)$$

where *settlement* is a binary outcome variable coded as 1 if the lawsuit terminated in settlement and 0 otherwise. *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). *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. *Controls* is a vector of explanatory variables that includes size and cause of the breach, types of information lost, industry and circuit court controls, number of causes of action and number of times the complaint was amended, and year when the case was disposed. 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.

Results

Table 2 presents the results of Eq. 2, reporting the average marginal effects of the explanatory variables on the probability of settlement. 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. Further discussions therefore focus on estimations from Model 3a.

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		Y	Y	
Region Controls				
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*

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.

Interestingly, while the compromise of financial data and breaches caused by improper disposal/disclosure appeared to drive litigation, the compromise of medical data and breaches caused by cyber attack appear to drive settlement. Moreover, 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.

Discussion and Conclusion

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 here 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 4 suggests that defendants settle perhaps too often (i.e. in absence of actual financial harm, and therefore case merit).

References

Acquisti, A., Friedman, A. and 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. ND. Litigating Toward Settlement. Journal of Law, Economics, and Organization. Forthcoming.

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., and 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., and 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. 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.

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

Cooter, R. and Rubinfeld, D. 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.

Eisenberg, T., and 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.

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., and Zhou, L. 2011. The Impact of Information Security Breaches: Has There Been a Downward Shift in Costs? Journal of Computer Security.

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

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. and 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., and 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.

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

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

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

Appendix

Table 1.
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_creditcard	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 action suits			0.76	0.43
Class certification			0.12	0.33
Statutory damages			0.54	0.50
Multisuit cases			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