Estimating the size of the iceberg from its tip

An investigation into unreported data breach notifications

Fabio Bisogni[1,2], Hadi Asghari[1], Michel J.G. Van Eeten[1]

[1]Delft University of Technology, Faculty of Technology, Policy and Management

[2]Fondazione FORMIT

Introduction

A decade has passed since the enactment of data breach notification laws (DBNLs) in numerous U.S. states. These laws mandate companies that have suffered a data breach to inform the customers whose data might have been exposed. The intent of DBNLs can perhaps be best summed up in the phrase: "sunlight is the best disinfectant". Whether the goal of incentivizing better security practices has been realized is the subject of an ongoing debate (e.g., Romanosky et al. 2011, Bisogni 2016). What is clear, however, is that they have offered more visibility into the state of data breach events in the United States.

That being said, it is also clear that an unknown number of breaches are hidden from view. The Identity Theft Resource Center's (ITRC) Breach Report, and similar databases, only contain breaches that have become public knowledge. As Figure 1 illustrates, a breach first needs to be detected by the affected organization (move from 4 to 3), then one or more relevant parties need to be notified (move from 3 to 2), before it can become publicly reported (move from 2 to 1). A simple statistic highlights that many of the breaches never make it past the last hurdle. The notification letters that are made public by the Attorney General in four U.S. states account for approximately 40% of all reported breaches in ITRC in 2014, while these states host only 14% of U.S. firms and 15% of the population. The organization maintaining the ITRC also acknowledges the issue: "We are certain that our ITRC breach list underreports the problem" (ITRC 2017).

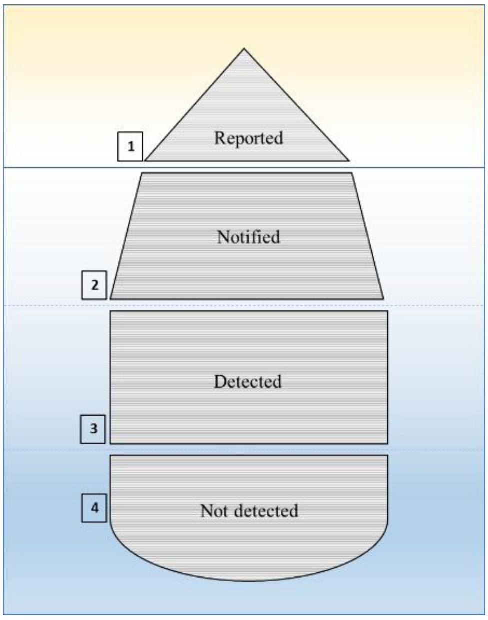

Figure 1 – Data Breach Iceberg

This paper sets out to provide an enhanced understanding of the submerged part of the iceberg. We first leverage differences among DBNLs in different U.S. states to estimate the impact of certain provisions on how many breaches have triggered notifications, yet did not been become publicly reported. In other words, we can estimate level 2 of the iceberg. Data breach statistics highlight significant differences among U.S. states (see Figure 2). We model the number of reported breaches as a function of the different DBNL provisions across the states, while controlling for the size of different sectors in each state and other factors.

Our model also includes the impact of the "risk of harm" exemption in some DBNLs, which allows breach organizations to not notify affected consumers, if after a reasonable investigation they determine that there is no reasonable likelihood of harm to customers stemming from the breach. States with this exemption report fewer breaches. This means that affected organizations never notified anyone in the first place. By modelling the impact of the risk of harm exemption on the number of reported breaches, we can estimate how breaches are detected but not notified because of this exemption – a portion of level 3 of the iceberg.

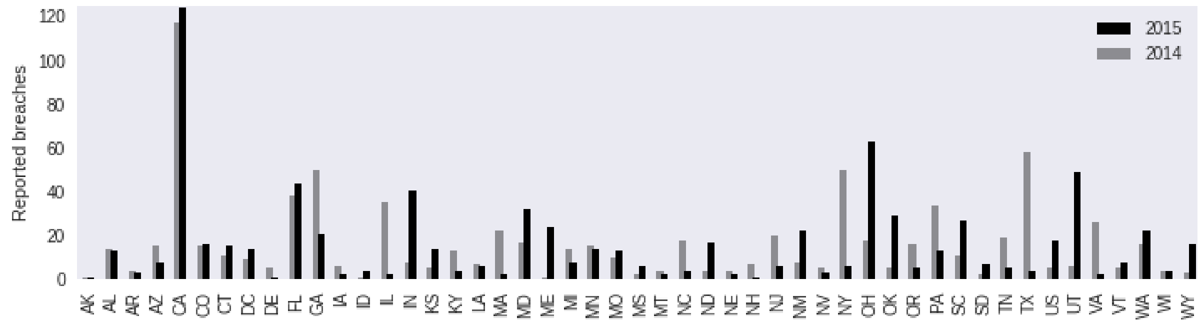

Figure 2 - U.S. Data breach Statistics 2014-2015 by State

Finally, we catch a glimpse of the deepest part of the iceberg – level 4 – through analyzing the notification letters in four states. In those states, the Attorney General publicly reports all notifications. We have coded all breach causes mentioned in those letters. Interestingly enough, the sector with the lowest breach rate ('retail and other business') is also the one with the highest ratio of breaches caused by 'hacking' and lowest ratio of 'unintended disclosure'. This suggests that security practices in this sector do not detect a significant number of breaches, contributing to a breach rate that is between 2 and 12 times lower than other sectors. The notification letters also allow us to look at notification and detection times by modeling the time span between the notification and, respectively, the breach discovery by the organization and the breach event. By doing so we managed to identify those breach causes that more than others require notification times not in line with the individuals' need to defend themselves promptly against potential harm.

Our analysis shows that there is quite a lot that is not known about U.S. data breaches. That being said, the security community knows much less about breaches in Europe. This is evident by browsing public databases that have gathered known data breaches, such as the ITRC, which contains only breaches affecting U.S. residents. The European Union (E.U.) has recently introduced its own industry-wide DBNLs: a directive[1] and regulation[2] will extend the weaker and sector-specific security breach notification laws that applied to the telecom sector. Our analysis helps the E.U. to learn from the results of almost 15 years of regulations in US, since the enactment of the first DBNL in California[3], giving relevant insights in view of the adoption of the Data Protection Package[4].

In short, the contributions of this paper are as follows: (i) to model the impact of DBNL provisions on the number of known data breaches and breach notification times, while controlling for sector and state differences; (ii) to estimate the number of breaches about which notifications have been issued but that are not publicly reported; and (iii) to discuss key elements of DBNL that make those laws effective in view of the implementation of the European regulation on security and data breaches.

[1] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016
[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016
[3] California Civil Code § 1729.98 enacted in 2003.
[4] Consisting of the General Data Protection Regulation (GDPR) and the Directive for data processing by law enforcement for the purposes of prevention, investigation, detection or prosecution of criminal offences. The Directive is to be implemented by 6 May 2018, and the Regulation will apply from 25 May 2018.

Objectives of Data Breach Notification Laws

Data breach notification laws are typically justified with two objectives. The first is that customers have the right to know when their personal information has been stolen or compromised. As Schwartz and Janger (2007) describe it, informing customers allows them to protect themselves – by changing their passwords, for instance, or by monitoring their credit card statements for signs of abuse.

Prior work has found little evidence that this objective is being realized. Bisogni (2016), for example, shows that it takes several months after a breach occurs for it to be detected and notified. By the time the consumer is informed, the attackers have had plenty of time to do damage. Romanosky, Telang, and Acquisti (2011) suggest that the adoption of state-level data breach disclosure laws might have reduced identity theft from these breaches by 6.1%, on average.

A second objective is to create incentives for organizations to take adequate steps to secure the personal information they have stored. The reputation damage resulting from a reported breach would activate 'the sunlight as disinfectant' principle, leading to companies to invest more in cybersecurity, and disinfect organizations of shoddy security practices (Ranger 2007).

Researchers have assessed reputation damage mostly through the effects of a breach on stock market prices. Acquisti, Friedman, and Telang (2006) investigated the impact of a privacy breach on stock market prices. They found a reduction of 0.6% on the day of the breach disclosure. Campbell et al. (2003) identified a significant and negative effect on stock price for data breaches caused by "unauthorized access to confidential information". Cavusoglu, Mishra, and Raghunathan (2004) reported that the disclosure of a security breach results in the loss of 2.1% percent of the breached company market value within two days of the announcement. Ko and Dorantes (2006) reported a mixed effect: although a breached firms' overall performances are lowered (relative to firms that incurred no breach), their sales increased significantly. Sinanaj and Zafar (2016) present another mixed result: breaches have a negative and immediate impact on social media and corporate reputation, while they do not have a significant effect on stock market valuations. Kwon and Johnson (2015) used a propensity score matching technique to investigate how data breaches affect subsequent outpatient visits and admissions in the United States, finding that the cumulative effect of breach events (and also of number of breached records) over a three-year period significantly decreases the number of outpatient visits and admissions. This shows that the effect of a data breach has a significant impact on the consequent consumer decisions.

The incentives fostering investments in internal security were studied by Gordon, Loeb and Lucyshyn (2003). They found that expenditures to prevent information security breaches have been growing rapidly in recent years. The empirical evidence provided in their paper supports the argument that one key driver of actual expenditures on information security activities is the occurrence of actual security breaches. This is also confirmed by Moore, Dynes and Chang (2016), who found that most firms indicated that cybersecurity was becoming a major focus, either as a result of their own data breach experience or those of other firms. Having similar events clearly changed thinking in most firm's senior management about cyber-risk management.

The pursued final result coming from the striving towards these two objectives is summarized in the Federal S.177 - Data Security and Breach Notification Act of 2015: "to protect consumers by requiring reasonable security policies and procedures to protect data containing personal information, and to provide for nationwide notice in the event of a breach of security". DBNLs also contribute to improving the security of the overall Internet ecosystem by increasing transparency for the security community, policy makers and citizens. In this respect, ENISA (2011) believes that the introduction of a data breach notification requirements is an important development with the potential to increase the level of data security and foster reassurance amongst citizens on how their personal data is being secured and protected. The introduction of DBNLs acts in an environment where two opposing elements cohabit, an increase breach risk due a. o. to higher digitalization and an increase investment in security due to better awareness of the risk itself. Edwards et al. (2015) developed Bayesian Generalized Linear Models applied to a public dataset to investigate trends in data breaches in the United States, showing that neither size nor frequency of data breaches has increased over the past decade, possibly highlighting how the existing forces are compensated.

We build on prior work in terms of looking at the degree in which DBNLs help public visibility of data breaches. Earlier research has studied specific sectors (e.g., medical sector investigated by Kwon and Johnson 2015) or state-level differences in the overall number of reported breaches (e.g. Faulkner 2007), but no work has looked at both simultaneously, as far as we know. This combined focus is important, as there are key differences among the sectors in terms of the use of information technology and the presence of more specific laws on how to deal with personal data, as is the case for finance and health.

Elements of Data Breach Notification Laws

Differences in the provisions of DBNLs in the different U.S. states can be grouped into 4 categories:

1. Definitions and scope: This includes definitions used for personal information and security breach; what is covered in terms of data owned, licensed or maintained; and finally, notification triggers generated by data acquisition and data access. The significant elements are: *broader definition of personal information*, *notification by access*, and *limited coverage*.

2. Safe Harbors and exemptions: DBNLs can offer various exemptions. The 'no risk safe harbor' applies when an entity's risk assessment concludes that no reasonable harm will be done to individuals whose personally information was breached. Organization may also be exempted from notification if they have used encryption, or if they have implemented another notification policy, or if they have complied with other laws such as the Gramm-Leach-Bliley Act or the HIPAA. The significant elements are: *encryption*, *risk of harm analysis*, *other laws applicable,* and *own notification policy.*

3. Notification flow: In addition to notifying the affected individuals, many states foresee a notification of the authorities. These typically include the Attorney General and Customer Reporting Agency, and in Idaho and Illinois, the Office of the CIO of the Department of Administration and General Assembly. Additionally, some DBNLs strictly specify what should go into the notice content for residents. It is assumed that such an obligation will contribute to the ability of the affected individuals to act on this information and to enhance the functioning of the notification as a reputational incentive. The key elements are thus: *notification to credit reporting agency*, *notification to attorney general*, *notice requirement,* and *specific time frame*.

4. Penalties: In all states, Attorney Generals can impose penalties on organizations for not notifying them about a breach. In some states, the DBNL contains a limit to the financial penalty that can be imposed by the AG. The cap is defined either in terms of an amount per breach or per breached record. Some states have another relevant DBNL provision that explicitly allows affected individuals to bring a *private cause of action* against a person or entity that violates the law, in order to recover any damages suffered. The key elements are this: *penalties limit,* and *private right of action*.

The key differences across the different DBNLs are summarized in Table 1.

Table 1 – Key DBNL provisions

Law element	Explanation	Present in n. states
Broader Definition of Personal Information	It indicates whether the statute covers more information than meets the standard definition of personal information (PI)[5]. An expanded definition of PI includes other pieces of data, most notably health and medical information.	30
Notification by access	The breach of security is defined by data acquisition. However, in some cases definition is extended to unauthorized access.	3
Limited Coverage	In some cases, the laws do not apply to organizations that own, license or maintain data that includes PI, but they regulated only those case where the data including PI are owned and/or licensed.	35
Type of data	Acquisition includes acquisition by photocopying, facsimile, or other paper-based method.	8
Encryption	This provision describes the requirements for receiving an exemption from a state notification law. States in which this exemption is easiest to attain have laws exempting notification if breached data were encrypted or redacted.	30
Risk of harm analysis	It refers to whether a statute requires a breached organization to notify only if the organization determines that the breach constitutes a reasonable likelihood of harm to the customer.	39
Other Laws applicable	When compliant with other laws, the Gramm-Leach-Bliley Act, the HIPAA or Primary Regulator organizations are exempted from Data Breach Notification Law provisions.	43
Own Notification Policy	Such exemption exists when a state allows an organization that maintains its own notification procedures as part of its information security policy to be deemed in compliance with the state notification law, so long as the organization does, in fact, disclose breaches.	14
Notification to credit reporting agency	In the event an Entity provides notice to more than a certain number of persons (it varies from 1,000 to 10,000 according to the state) at one time pursuant to the general security breach section, the Entity shall notify, without unreasonable delay, all consumer reporting agencies that compile and maintain files on consumers on a nationwide.	31
Notification to Attorney General	In the event a business provides notice to an affected person pursuant to this section, the business shall notify without unreasonable delay the state AG's office of the nature of the breach, the number of consumers affected by the breach, steps taken to investigate the breach, steps taken to prevent a similar breach in the future, and information regarding the timing, distribution, and content of the notice. The AG's website contains a form to be used for notification.	22
Notice requirement	in some cases, certain elements of the notice are mandatory and identified in the law. Such elements include the type of personal information subject to an unauthorized access or acquisition, the specification of the reporting entity's name and contact information so that affected individuals can obtain additional information, specific information on what has happened (a general description of the breach incident).	15
Specific time frame	This provision specifies that notification must occur within a given number of days (usually 30 or 45). Notification laws without a specific time limit require notification as quickly as possible and without unreasonable delay.	7
Private right of action	This provision gives customers the ability to sue organizations for failure to comply with the data breach notification statute.	16
Penalties limit	It defines a limit to the financial civil penalty imposed on an organization found in violation of the statute	26

Data

To recap, our main research objective is to study the impact of key provisions of DBNLs on the number of that move from being detected to notification to being public reported, while controlling for sector

[5] An individual first name or initial in combination with a last name and a social security number, driver's license number, state ID card number, or financial account number (see Baker 2014)

and state differences. We also study notification letters for additional insights into underreporting and timing of breach notifications. In this section we describe the datasets we used.

Breach Datasets. There are numerous initiatives aimed at providing details on data breaches—such as Privacy Rights Clearinghouse (PRC) database, the Identity Theft Resource Center (ITRC) Breach Report, and the Veris Community Database.[6] Given the current coverage of data breach notification laws in U.S., one might expect a joint institutional repository for data breaches, but this is not the case. As of today, 22 states require notifications to the Attorney General, but only 7 states publish details of the events and the notification letters.[7]

We use two datasets of breach incidents: the ITRC list, and a self-compiled dataset based on the notification letters made available by the A.G. in four states (see Appendix I). We identified ITRC as the most comprehensive source of breaches in US: in 2014 PRC reported 330 breaches versus 783 reported by ITRC. From the ITRC dataset, we collected the date of the breach, the sector and state of the breached firm, and the number of records breached.

From the notification letters, we have manually extracted the date of the letter and the breached company. We then determine the company's sector. Other details include the type of incident, number of affected records, and the incident date, which can be compared to the notification date—i.e., when the letter was sent. Table 2 lists the summary statistics from these datasets by five primary sectors. The data we extracted from the notification letters covers all of 2014, while from the ITRC dataset we selected the 2014 and 2015 events.

Table 2- Datasets

	ITRC dataset (for breach count model)	**Notification letters (for breach time model)**
Educational Institutions (EDU)	108	21
Financial and Insurance Services (FIN)	106	97
Retail and Other Business (BSO) [8]	561	222
Medical & Healthcare Providers (MED)	590	71
Government and Military (GOV)	137	19
Total breaches	1502	430[9]
Median records breached[10]	2500	6 (only NH/MD customers)
States Covered	47[11]	CA, MD, NH, VT
Dates Covered	1-Jan-2014 to 31-Dec-2015	1-Jan-2014 to 31-Dec-2014

[6] Appendix I provides links to these breach datasets, as well as the data sources used by ITRC

[7] California, Maryland, New Hampshire, Vermont (included in our analysis), Washington, Oregon and Montana. Washington and Oregon started, respectively, from mid-2015, beginning of 2016 to give such visibility, after law revision. Montana from mid- 2015.

[8] We merged BSR (Retail/Merchant) with BSO (other business)

[9] The majority of these breaches, that is 311, are also in the ITRC. The missing records are due to ITRC grouping smaller breaches together, and ITRC recording some breaches in 2013, while the letters were sent in 2014. However, we use the datasets in separate models, so the overlap, or lack of, doesn't matter.

[10] ITRC reports breached records, or customers, for 55% of the incidents. For our dataset, it is stated in 39% of incidents, and only in letters to the Attorneys General of New Hampshire and Maryland.

[11] The ITRC dataset includes all U.S. states except West Virginia, plus District of Columbia. We further remove the records for the states of Alabama, New Mexico, and South Dakota, as do not have DBNL.

Sector Size*.* We build a breach rate for each *state and sector*, by dividing the number of breaches per state and sector by the number of firms active in that state and sector. Firm data is extracted from the 2012 U.S. Census. Since the census data excludes governmental offices, we use the number of medical centers also as the denominator for breaches in the governmental sector. Our assumption is that the number of medical centers are driven by the number and size of cities in a state, which similarly influence the number of governmental offices.[12]

Control variables. When modelling the relationships, we control for the size of various sectors in different states, and attribute the remaining variation to differences among the DBN laws. However, there might be other systematic reasons that lead to less, or more, data breaches occurring, or being reported, in a state. We use a number of variables to control for such differences, such as crime rates, household income, or the concentration of firms per population. In the end, these controls had no decisive impact on our models (see Appendix IV).

DBNL Provisions. We select the DBNL provisions to include as variables in our models. We do not include all provisions for a substantive and statistical reason. The substantive reason is that some provisions could not be codified clearly among the states - for example, the definition of personal information has too many variations across the states. The statistical reason is that some of the categories were too sparse and included only a few states, which would bias the regression results. The selected provisions are as follows:

inform_credit, inform_ag_np & inform_ag_p: All DBNLs require affected consumers to be notified. In order to identify the effect of additional notification flows on the number of reported breaches, we code two variables for whether the law also requires informing credit agencies and informing Attorneys General. In the latter case, we distinguish between *inform_ag_np* and *inform_ag_p,* where the difference is whether the Attorney General publishes the notification letter on their website (ag_p) or not (ag_np). These provisions affect the probability that a specific event can be known also by additional actors such as banks and the media. Given the fact that more actors are aware of the breach, it becomes more likely that such event will end up in the public domain. Breaches included in the ITRC list are not only those reported by Attorneys General, but also breaches that media reported, with or without the AG being notified or reporting about them.

penalty_cap & priv_cause: all DBNLs include penalties for not notifying about a breach. Some however include a cap to the financial penalties. This cap can be fixed per breach (e.g. Oklahoma) or per single violation (e.g. District of Columbia) or both (e.g. Utah). The existence of a cap related to the penalty for not complying with the law, defines a priori the risk for not notifying. Some laws include a so-called "private right of action": the possibility for consumers to sue entities for failing to comply with the data breach notification statute. This increases the potential penalty for non-compliance in terms of breach notification.

risk_harm: The safe harbor provisions in different DBNLs are difficult to bring into a common set of categories.[13] We focus solely on the presence of a risk of harm exemption, which states that a breached organization only has to notify if the organization determines that the breach constitutes a reasonable likelihood of harm to the customer.

[12] Except for the District of Columbia, which we exclude due to concentration of governmental offices.
[13] Several statues include encryption as a safe harbor provision, but some do this with a definition of encryption, while other leave it undefined. The exemption due to the application of sectoral specific regulation, i.e. Financial and Medical sectors is already pictured by the sectoral analysis we performed.

Limitations. A major limitation stemming from the ITRC and other breach datasets, is that the breach is reported in the location of the headquarters of the company. This might not be where the breach actually occurred in case of companies active in multiple states. Additionally, breach notification procedures are tied to the residency of the affected customers (in such cases a company active in several states might follow the strictest DBNL in all the states to simplify its processes). This limitation is common to all papers doing similar analysis. One solution presented in Appendix II, is to rerun our models with a dataset that excludes the financial and business sectors that contain the most multi-state companies. The direction of the coefficients does not change, indicating robust results.

Explaining the number of reported breaches per state

We can now model the impact of the different DBNL provisions on the number of reported breaches. Our approach assumes that the probability of a breach in a specific sector (i.e., the number of breaches per organization in that sector) is the same across different states. In other words, we assume that differences in number of reported breaches per state and sector are caused by differences in the DBNLs and the control variables, rather than by systematic differences in security practices among states or by attacker preferences for companies in certain states over others.

Figure 3 plots the breach count versus firm count, for each combination of state and sector (color coded by sector). Given the distribution of the data, we use a *negative binomial regression* to model *breach rates*—the number of breaches per state-sector, offset by the number of organizations in that state-sector. This is the recommended way to model rates (Hilbe, 2011). Using a negative binomial distribution is also consistent with prior work on breaches (Edwards et al. 2015).

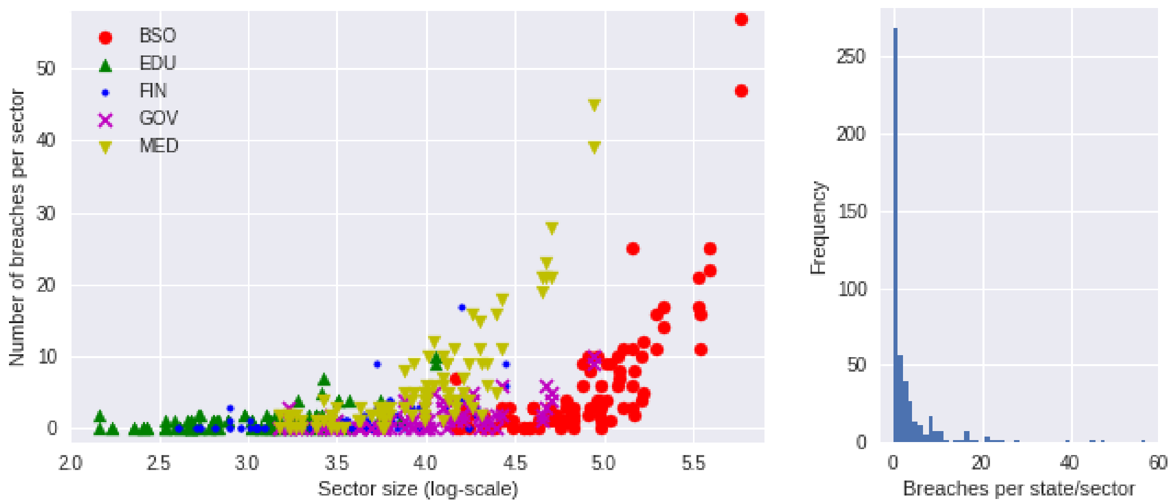

Figure 3 –Left: Breach count versus organization count per sector, Right: histogram of breaches per sector/state/year

The regression results are presented in Table 3. In general, regression coefficients in a negative binomial distribution can be interpreted as 'incident rate' ratios, and they are also 'multiplicative'. That is, they tell us how much more or less likely an incident (here, a data breach) is likely to be counted (here, detected and reported).

We see that the sectoral differences are much stronger than DBNL provisions. And among the DBNL provisions, we see that attorney generals who publicly report notifications cause, on average, a 43% increase in reported breaches in that state. This effect was to be expected, though not perhaps its magnitude. More surprising is the fact that the requirement to report to credit agencies leads to a 34% increase, all other things being equal. Allowing the risk of harm exemption significantly decreases reports, by 21%. The penalty variables have no significant effect. We now discuss these main findings in more detail.

Table 3 - Breach Count Regression Model

Variable	Coefficient (Std Err)	Incident Rate Ratio (95% CI)
inform_ag_p	0.361 (0.149) *	1.43 (1.07 -1.93)
inform_ag_np	0.044 (0.101)	1.05 (0.86 - 1.28)
inform_credit	0.289 (0.094) **	1.34 (1.11 - 1.61)
penalty_cap	-0.053 (0.085)	0.95 (0.80 - 1.12)
priv_cause	0.119 (0.094)	1.13 (0.94 -1.35)
risk_harm	-0.231 (0.104) *	0.79 (0.65 – 0.97)
fin	1.376 (0.136) ***	3.96 (3.03 -5.15)
med	2.186 (0.098) ***	8.90 (7.33 -10.80)
edu	2.482 (0.135) ***	11.97 (9.13 -15.58)
gov	0.745 (0.127) ***	2.11 (1.64 - 2.70)
bso	NA—base sector	NA
sector_size	*Offset*	*Offset*
(Intercept)	-9.963 (0.135) ***	NA

Negative binomial disp: 5.179. N= 478. Deviance null/residual: 1179/563. McFadden pseudo R^2: 0.23. (Full diagnosis available in Appendix II)

Effects of Notification Flows. Reported data breaches increase by more than one third when notification letters are published by Attorneys Generals or when credit agencies must be notified by breached organizations. In the first case, the contribution of Attorneys General in improving the level of visibility of breaches from notified to reported is clear and direct, as they themselves publish the received notification letters.

For credit agencies, the mechanism is less clear. One explanation is that agencies contribute to increase the number of reported breaches by informing other actors outside the communication flow dictated by the state DBNL, who then contribute to make the breach public. Another explanation is that these agencies provide an additional notification to consumers, in addition to the one they receive directly from the breached organization. This might cause them to take the breach more seriously and might increase the probability that consumers report to Media. This interpretation is consistent with the findings of Ablon et al. (2016), who found that a surprising 44% of consumers learned of the breach from other sources, before receiving an official breach notification. The most common method that participants recalled was through media reports (28%), followed by notifications from a third party, such as a bank (16 %).

Consumers can play an important role in informing the media, in addition to the Attorney General and the credit agencies. To explore the role of consumers in making breaches public, we compare the size of reported breaches across the states with different notification authorities.

So far the level of our analysis has allowed us to consider all data breaches equal not including in the investigation the number of records affected. We however know that data breaches come in all shapes and sizes, from the breach enabling the access to a few records containing Personal Information to the impactful mega breaches.[14] The number of accessed records defines therefore the size of the breach.

Figure 4 presents cumulative distribution functions (also known as CDFs, or cumulative histograms) of *the number of records per reported breach*[15]—a proxy for breach size—in six possible combinations covering all options related to the notification flow. The x-axis is the number of records in a breach (cropped at 6000 for readability), and the y-axis is the cumulative percentage of all breaches with that number of breached records or lower. The combinations of notification flows include the Attorney General not being informed, being informed but not publishing, or being informed and publishing notification letters for the rows, and Credit agencies being or not being informed for the columns.

In two combinations, consumers are the main, if not only, actor that can make the media aware of the breach: *not_inform_ag/not_inform_credit* (no authority informed) and *inform_ag_np/not_inform_credit* (AG is informed but doesn't publish the notifications). In these states, the median number of records affected by the breach is higher than all the other combinations, that is 2929 and 6000 respectively. This is consistent with consumers being the main source for reporting breaches: having larger breaches means more affected consumers, which increases the probability that one or more of them makes the breach public.[16] [17]

We can make two additional observations. The states in the top row, where the AG is notified and publishes the notifications, have the smallest median records. This means in these states we know about both small and larger breaches. Similarly, if we compare the two columns, the column where the credit agency is notified consistently has smaller median records. This again points to some public reporting mechanism after the credit agencies are notified.

---

[14] A mega-breach is commonly defined as a breach of more than 10 million records.

[15] As stated earlier, the ITRC has the number of records for only 55% of the breaches.

[16] Given that our analysis is based on medians, the impact of Mega breaches are limited. In the ITRC database only six breaches have more than 10 million records in the timeframe 2014-2015.

[17] It's important to note that the reputational effects of (missing) notifications may also depend on the nature and significance of the PI breached, in addition to the size of a breach. However, we cannot say much about the nature of the PI from the data, and assume breaches to be similar in this regards.

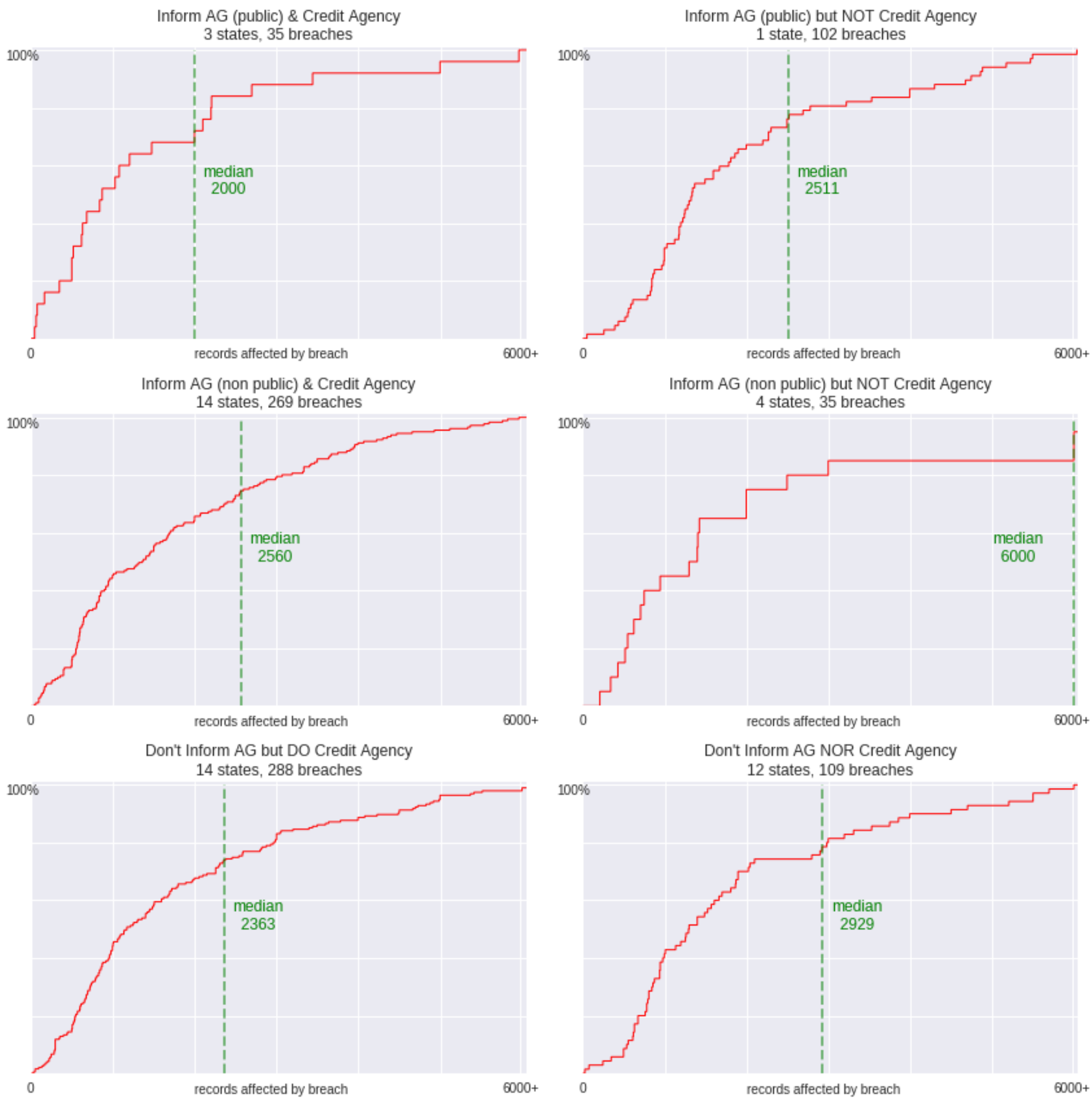

Figure 4 – Cumulative histogram (or CDF) of records affected by breach (the x-axis cut at 6000 for readability; the y-axis is the percentage of breaches with x or less records affected). Each plot represents one notification-flow combination

Effects of Penalties. When companies become aware of a data breach they face two options:

1) They decide not to notify and bear the risk of penalties, of a private right of action, if present, and of potential reputational damage, should the breach become public at a later stage. Such behavior will generate immediate savings in terms of not having to cover the notification process, customer services operations, customer redress, etc.
2) They decide to notify and accept the consequence of such a disclosure, such as the cost of notification, call centers, customer support, identity theft insurance or credit monitoring, legal fees, regulatory fines, and the potential loss of market value or lost business, while avoiding the penalties and reputational risk related to a breach becoming public at a later stage.

When assessing the two options, organizations must be aware that data breaches are not just breaches of security, but can have an important impact on the trust between companies and their customers, and can result in not only negative publicity, but lost business, lawsuits, and fines that can threaten the viability of the business.

The breach of the trust of consumers that might result from the first option - when organization intentionally hide a breach that then becomes known to the community - can be remediated less than if companies follow option 2. In the latter case organizations can in fact manage to communicate that data breaches are a common phenomenon in the sector and that are not necessarily dependent on the company security investment and practices.

Our model findings suggest that companies evaluate whether to notify mostly based on the reputational damage that results from the missing notification, more than the tangible consequences of not notifying as included in the DBNL. The direct financial consequences are checked via the private cause of action and penalty cap provision, which were both insignificant. On the contrary, the coefficient for credit reporting agencies being notified (inform_credit) suggests that companies fear the reputational consequences of missed notification. If a hidden breach becomes public (thorough other channels), organizations will have misled not only consumers, but also other organizations.

***Risk of Harm Exemption*.** The negative impact of the risk of harm exemption on the number of reported breaches confirms that, where the option is given, companies tend to use it in one out of five breaches. Consequently, 21% fewer data breaches are notified.

Sectoral Differences. The differences among sector show a much stronger impact on reported breaches that the different DBNL provisions. Finance, medical, education and government have respectively 4, 9, 12 and 2 times higher rate of reported breaches per firm than 'retail and other business'.

Sectoral differences are to some extent expected, as some are covered by additional federal laws that have breach notification aspects (e.g. the Gramm–Leach–Bliley Act for financial institutes, or the Health and Accountability Act). The financial sector specifically is also known to have a higher level of security than other sectors (Security Scorecard 2016). But what explains the lower breach rates among retail and other business? Possible explanations include being targeted less for a breach (suggesting it would be less attractive than the financial as well as the health sector), not detecting breaches (due to underinvestment in security), or the breaches not becoming public (due to their size). An additional piece of information, the typical causes of breaches in each sector, can shine some light. This information is available from the AG notification letters (our database). Table 4 presents the difference between the *observed* and *expected* causes of breaches in each sector.[18] The expected value is calculated by multiplying the row total and column total for a cell, and dividing it by the grand total. We can observe the following patterns:

- In Retail and other business, hacking makes up a larger proportion of all breaches than in other sectors. This might indicate a lower levels of network security. Unintended disclosure is lower than in other sectors, which points to either underreporting or less vigilant monitoring and process controls in place to identify these kinds of events.
- In contrast, unintended disclosures cause a very high proportion of breaches in the Governmental sector, highlighting either weak personal data handling processes in place or effective monitoring on the processes themselves or a combination of both. Insider attacks

[18] We use the breach causes from privacyrights.org, which are as follows: unintended disclosure (sensitive information posted publicly on a website, mishandled, or sent to the wrong party via e-mail, fax, or mail), physical loss (lost, discarded, or stolen nonelectronic records, or portable or stationary devices), insider (someone with legitimate access intentionally breaches information—such as an employee or contractor), hacking and malware (electronic entry by an outside party, malware, or spyware), and unknown or other (all other cases, including payment card fraud).

make up the lowest proportion, where compared to other sectors more security background checks are performed.
- In the Medical sector, physical losses have the highest value, possibly reflecting the peculiarity of Health services, where data travels more physically during service delivery than elsewhere. This might explain the high proportion of insider theft as well, as many professionals need to have access to the data.
- In the Finance sector, physical loss shows the lowest values, possibly due to them using digital capabilities more than the rest of the economy (McKinsey Global Institute 2015).

Table 4 – Contingency table with the difference between observed and expected breaches by cause and sector.

	Hacking	Insider	PhysicalLoss	Unintended Disclosure
Sector				
BSO	+24.3% (133 vs 107)	-8.7% (21 vs 23)	-5.3% (36 vs 38)	-43.1% (29 vs 51)
EDU	-10.0% (9 vs 10)	-50.0% (1 vs 2)	0% (4 vs 4)	+40.0% (7 vs 5)
FIN	-2.1% (46 vs 47)	0% (10 vs 10)	-43.7% (9 vs 16)	+36.4% (30 vs 22)
GOV	-66.7% (3 vs 9)	-100.0% (0 vs 2)	-33.3% (2 vs 3)	+250.0% (14 vs 4)
MED	-50.0% (17 vs 34)	+85.7% (13 vs 7)	+83.3% (22 vs 12)	+12.5% (18 vs 16)

Alternative Model Specifications. In addition to the presented breach count regression model, we attempted a number of different model specifications, namely including interaction terms (between laws and sectors), limiting the dataset to the three local sectors, and adding control variables. These more complex models however do not better based on the Akaike information criterion (AIC) than our parsimonious model. They are presented in the Appendices for interested readers.

Estimating the total number of data breaches

Since the model identifies the effect of DBNL provisions on the number reported breaches, we can now estimate how many breaches would be notified and reported across the U.S., if all DBNLs–or a federal law—would require the credit agencies to be notified, the Attorney General would make all notifications public and the risk of harm exemption is removed. This means the underreporting would be limited to those cases where the breached companies do not detect the breach or where they do not disclose detected breaches.

The results of the estimation are represented in Figure 5, with one sub-figure for each sector. The green dots represent the observed breach counts from (each dot is one state/year). The blue line is the fitted model, showing the breach count that the model predicts for that combination of independent variables (inform_credit, inform_ag_p, risk_harm). The red pluses are the predicted counts, if the laws in all states were stricter—i.e., require the credit agencies to be notified, the Attorney General would make all notifications public and the risk of harm exemption is removed. (The Pearson correlation between the predicted and observed values ranges from 0.53 for finance to 0.92 for BSO and medical).

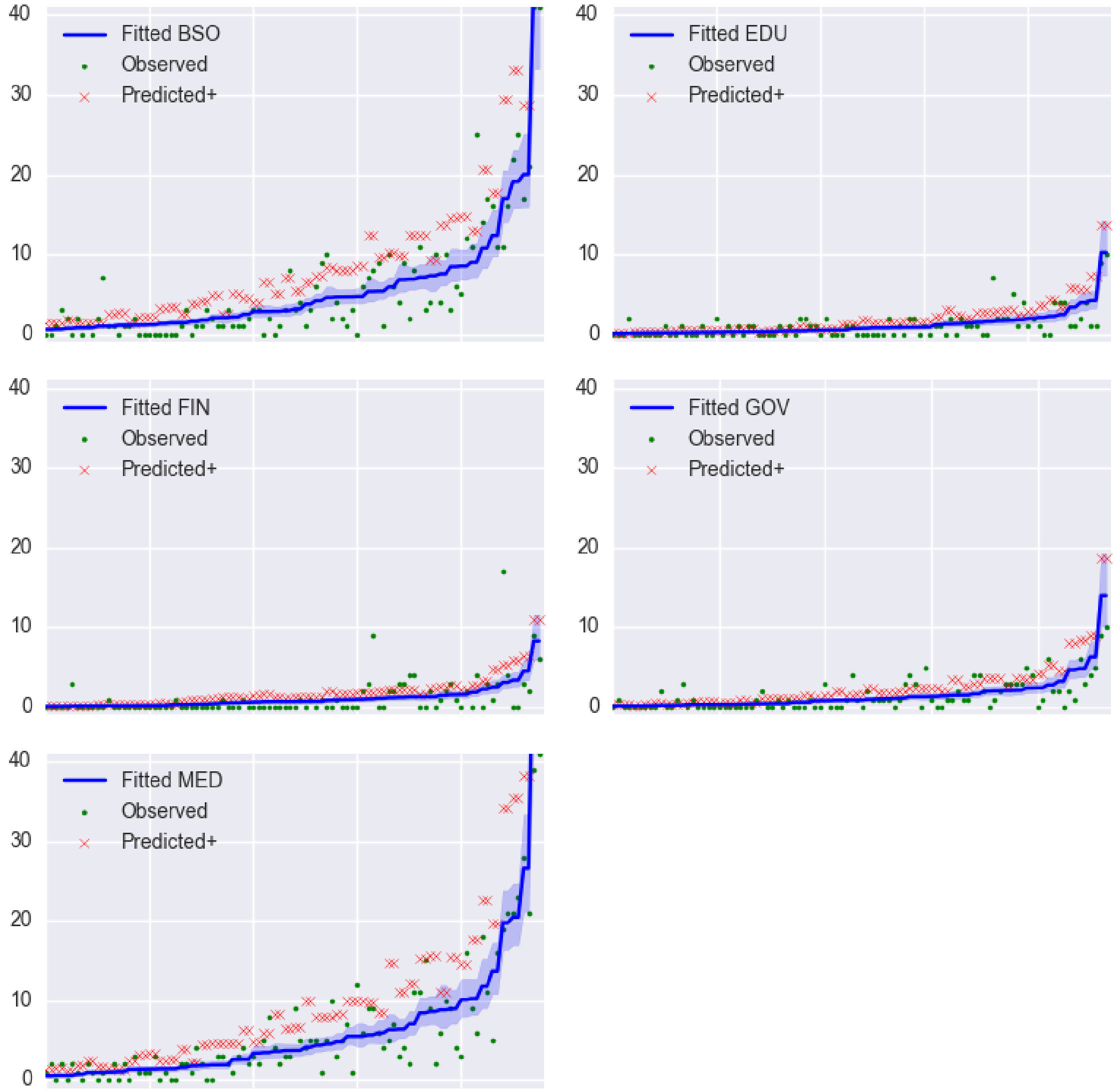

Figure 5 – Prediction results by sector (each point is a state; the x value represents the observed/fitted/or predicted+ number of breaches in that state; states are sorted by the number of breaches)

The prediction result indicates that 1,264 data breaches would have been publicly reported for 2015, if the notification to credit reporting agencies had been mandatory, all Attorneys General had published notification letters on their website, and the risk of harm analysis exemption didn't exist.[19] An additional 46% of data breaches is generated by the first two provisions applied to those States not having them, while 17% by the exclusion of the risk of harm analysis exemption. Overall this is 483 over the actual 781 data breaches reported. In other words, the current patchwork of data breach notification laws in place in US hides from the public more than 500 data breaches per year.

[19] If we instead set all states to allow the risk of harm analysis exemption, the number of breaches will be 1005 more severe breaches.

Modeling Notification Time Regression

Given informing customers faster is another aim of DBNLs, we also model the *uninformed exposure time* (the time between a security breach and the firm notifying about it), and the *notification time* (the time the organization needs to assess the situation after breach detection, to finalize the letter, and to inform the customer and relevant parties). Note that during both periods, customers are not aware of the risk they are exposed to and cannot undertake any defensive action.

This data is available in the 2014 notification letters retrieved published by AG websites in four states. The histograms for both variables are presented in Figure 6, with an average of 44 days for the notification time and 102 days for the uninformed exposure time. We use a negative binomial regression to model how sector, state, and breach cause influence these times.

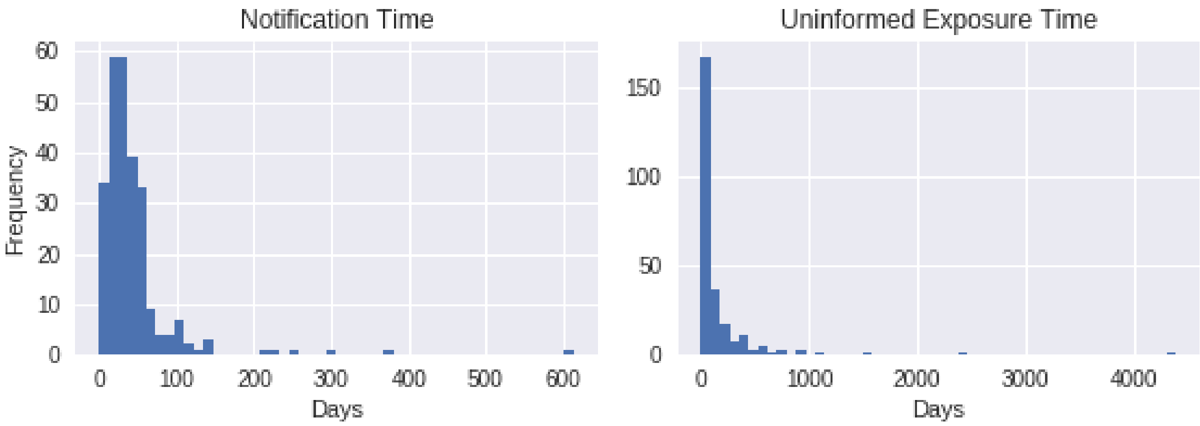

Figure 6 Histogram of notification (n) and uninformed exposure (ue) time, public AG dataset

The results are reported in Table 5 for the uninformed exposure time. The financial sector detects breaches in about half the average time (of 102 days across all industries and breach events), which might reflect the maturity of the sector in terms of information security (e.g., see Security Scorecard 2016). Compared to the baseline unintended disclosure, hacking events take 71% longer to detect, and insider events, five times more. And when the affected consumers are residents in more than one state in our dataset, breaches take 30% longer to detect, possibly highlighting the organizational complexity. In general, given its strong link with the industry and breach event, uninformed exposure time seems to represent competency.

For the notification time, no variable - state, sector, or breach type - adds to the intercept-only model based on the Akaike Information Criterion (AIC) or pseudo R^2. All we can say is that it takes on average 44 days for the notification process (see appendix V for details)[20]. We can however conclude that no DBNL element in the four states of California, Maryland, New Hampshire, and Vermont, causes a notification to be sent faster.

[20] Interestingly enough, five of the seven states that indicate a time frame for notification in US indicate a limit of 45 days, in line with the notification time average. Specifically, Ohio, Rhode Island, Vermont, Washington and Wisconsin.

Table 5 – Uninformed-Exposure Time Regression Model

Variable	Coef. (Std Error)	Incident Rate (95% CI)
(Intercept)	4.629 (0.298) ***	**102 days** (59.9-195.9)
Hacking	0.537 (0.207) ***	**1.71** (1.19-2.44)
Physical	-0.588 (0.207) **	**0.65** (0.37-0.83)
Insider	1.594 (0.252) ***	**6.00** (3.62-8.21)
Unintended	*baseline*	--
BSO	-0.342 (0.340)	0.71 (0.35-1.32)
FIN	-0.747 (0.352) **	**0.47** (0.23-0.92)
EDU	-0.268 (0.467)	0.77 (0.30-2.03)
MED	0.318 (0.351)	1.38 (0.67-2.60)
GOV	*Baseline*	--
Multistate	0.258 (0.142) *	**1.30** (0.97-1.74)

Negative binomial disp: 0.948. N=257. Deviance null/residual: 440/296. McFadden pseudo R^2: 0.04.

Discussion and Conclusion

We modeled the impact of DBNL provisions on the number of known data breaches and breach notification times, while controlling for sector and state differences. We concluded that the data breaches that are publicly known are just the tip of the iceberg. The dimensions of what is visible and what is hidden below the surface has proven to be dependent on how DBNLs are designed. We managed to unveil the number of breaches that could be reported in case certain provisions would be adopted uniformly across the U.S.

Breaking down the iceberg structure, we estimated that: (i) 46% more breaches would be reported because of the *inform credit agency* provision and the provision *notification publication by informed Attorneys General,* moving part of the breaches in block 2 to block 1; (ii) 17% more breaches would become known from block 3 (detected, but not notified) as the effect of the elimination of the *risk of harm analysis provision*; and (iii) an undefined percentage of undetected breaches that can be identified from the sectoral results of the regression model.

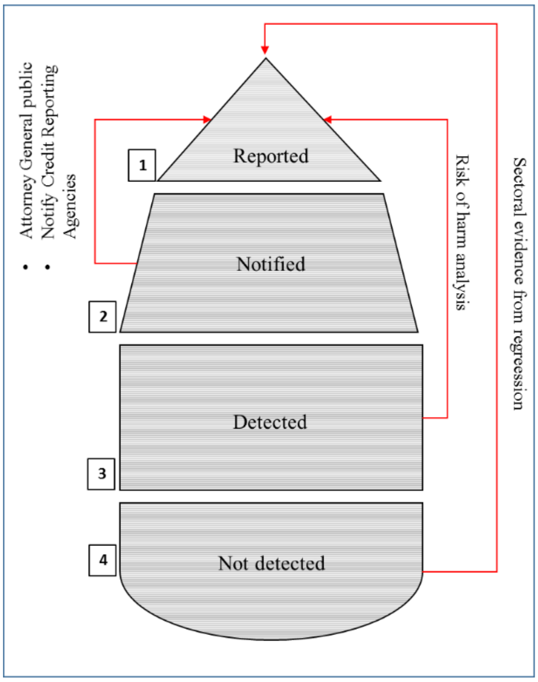

Figure 7 - Increased reported data breaches

By looking at the uninformed exposure time, we also managed to identify those breach causes that more than others represent an obstacle for a safer environment. Particularly for breaches in the category of 'Hacking' and 'Insider', the detection and notification timing is poor, greatly eroding the utility of the notification for helping individuals to defend themselves promptly against potential harm such as identity theft.

In short: two core elements play an important role in bringing detected and notified breaches into the light, namely the inclusion of more actors in the notification flow and, related to this, the publication from their side of the received notifications. While the first one requires a revision of the law and therefore lengthy processes, the second one can be easily implemented in those states that already include Attorneys General in the notification flow. Such authorities are in fact in the position to foster the visibility of the received data breach notifications by publishing them and consequently supporting both DBNL objectives (sunlight as disinfectant and consumers' right to know).

We also noticed key sectoral difference in how breaches lead to public breach notifications, or not. Certain sectors lag behind in terms of detection time, possibly of security measures, and consequently of the notification itself. Organizations in the 'retail and other business' sector report fewer breaches overall and more breaches caused by hacking. This strongly suggests shortfalls in detection and notification for breaches with other causes.

The findings of our analysis might impact other studies that have relied on datasets on publicly reported breaches, in two ways: (1) the breach frequency has been underestimated, given the relevant impact of AG disclosure, and (2) comparisons of breach counts and breach magnitudes across different are likely to be biased by systematic and substantial sector differences in whether and how breach events are reported.

Of course, uncertainty about actual breach frequency was already visible from different existing datasets. Many studies relied on PRC, while we used ITRC. By analyzing both databases for the year 2014, we have already noticed a relevant discrepancy in total figures: 330 breaches reported by PRC versus 783 reported by ITRC.

From both sources statements we can appreciate the use of similar rational when intercepting data breaches[21]. ITRC however specified also single used data sources as highlighted in annex 1. We relied therefore on ITRC, which is able to identify a larger number of breaches including those published by Attorneys General.

Not all research is impacted, of course. Romanosky et al. (2011), for example, focused solely on the date of implementation of DBNLs and its relation with identity thefts. Event studies on the effect of a single data breach on affected organization's market values is also clearly not impacted (Gordon et al. 2011, Acquisti et al. 2006, Cavusoglu et al. 2004, Campbel et al.2003). Effects are also limited for studies that are confined to a single sector regulated by federal law (e.g., Kwon et al. 2015) where the data are less impacted by sectoral differences.

Other work would benefit from taking our findings into account, specifically the systematic bias in under-reporting in sectors (e.g., for BSO) and states (e.g., for states where Attorneys General do not publish breach notifications). Particularly vulnerable are studies that look at trends and use the PRC dataset—e.g. Garrison et al. (2011), who present a longitudinal analysis of data breaches that focus on the analysis of time series of data breaches, and Edwards et al. (2015), whose research models breach frequency. Finally, Romanosky et al. (2014), who used DatalossDB[22] to identify the subset of data breaches that became public knowledge and "reported" and to then analyze which of those generated litigation. The use of the current dataset might enlarge the sample of reported breaches that are subsequently classified under non-litigated, federally litigated, or state litigated. Additionally,

[21] ITRC: Each selected incident is required to have been reported to a state Attorney General's office or published by a credible media source, such as TV, radio, press, etc. The item will not be included at all if ITRC is not certain that the source is real and credible.
PRC: PRC's Chronology includes breaches reported through either government agencies or verifiable media sources.

[22] DataLossDB.org operated until mid-2015

the fact that a large number of breaches would be known because Attorneys General publish them, could result in a different distribution among those three classes.

Our analysis also draws attention to the effects of core elements of DBNLs. Such effects could be relevant for European policy makers and authorities, in light of the European Union adoption of the Data Protection reform package. This means notification requirements have gained in importance in Europe as well. In particular, it is envisaged that the data controller must notify about the personal data breach without undue delay, not later than 72 hours after having become aware of it, to the responsible national supervisory authorities. When the breach is likely to result in a high risk to the rights and freedoms of individuals, the data controller must notify the breach to the individual (data subject) without undue delay. From our analysis of notification time we know that companies from four States in US require in average 44 days to notify consumers. The European limit of three days seems very challenging for organizations, if not outright unrealistic.

The setup of the European package creates an additional challenge. The three day limit only exists for reporting to national authorities, not the affected individuals. But since the national authorities do not (yet) make the notifications public, this incentivizes companies to “over inform” the authority as a matter of caution, also knowing there will be little reputation damage. To illustrate the point, the number of breaches reported in the Netherlands alone in the first quarter of 2017 is 2,300 (Dutch DPA, 2017). This count is higher than our upper bound estimate for the whole of the U.S. These extra notifications impose an administrative burden on the notifying firms and the regulator, without any clear security benefits. As a solution, we suggest that the national supervisory authorities publish breach notifications after a grace period, similar to some US Attorneys Generals. This way, the private administrative burden at least also generates some social benefits.

Our research could also help E.U. Member States to increase the positive effects of the Data Protection reform package in other ways. One lesson would be to include other actors in the notification flow to support better visibility of the breaches. Another lesson is to adopt a sectoral approach to help balance the discrepancies in detection timing across sectors, which we would predict will occur in the E.U. as well. Soft law initiatives such as codes of conduct and codes of practice, implemented at Member State level, could support managing such aspects and foster the appointment of sectoral bodies as industry reference points, collecting and analyzing information on notified data breaches and advising on existing security risks and on detection measures available. Such initiatives would supplement and support the implementation of the Data Protection reform package.

To conclude, we should stress that getting more breaches reported is not a goal in itself. It is a mechanism to improve the security of the Internet ecosystem by making the state of affairs transparent to the security community, policy makers and citizens. In the long run, careful monitoring is needed to see whether these outcomes are indeed achieved. Will reported breaches become background noise, the inevitable consequence of a digitizing society; will they generate increasing tangible negative consequences in terms of “naming and shaming” or companies going after the customers of breached competitors, or will other hitherto unanticipated consequences emerge? The instrument of DBNLs as a way to improve security will remain an important topic of study for the foreseeable future.

Appendix I – Data sources

ITRC current data sources (as of 28th February 2017)

California Attorney General's Office	*letters already available in 2014*
Maryland Attorney General's Office	*letters already available in 2014*
New Hampshire Department of Justice	*letters already available in 2014*
Vermont Attorney General's Office	*letters already available in 2014*
Health & Human Services (HHS.gov)	*sectoral DB*
HIPAA Journal	*sectoral DB*
www.databreaches.net	*multisectoral DB*
Maine Attorney General's Office	*No letters available only list of breaches*
Indiana Attorney General's Office	*No letters available only list of breaches*
Montana Attorney General's Office	*from mid-2015 letters available*
Oregon Attorney General's Office	*from 2016 letters available*
Washington Attorney General's Office	*from mid-2015 letters available*

Data breach databases websites

http://veriscommunity.net/vcdb.html

http://www.idtheftcenter.org

https://www.privacyrights.org

Attorney General websites accessed for notification downloads

https://oag.ca.gov/ecrime/databreach/list

http://www.oag.state.md.us/idtheft/businessGL.htm

http://doj.nh.gov/consumer/security-breaches/

http://www.atg.state.vt.us/issues/consumer-protection/privacy-and-data-security/vermont-security-breaches.php

Appendix II. Regression Diagnostics

The residuals versus predicted, and observed versus predicted plots for the breach-count model are as follows. (As a reminder, the McFadden pseudo R-square is 0.23).

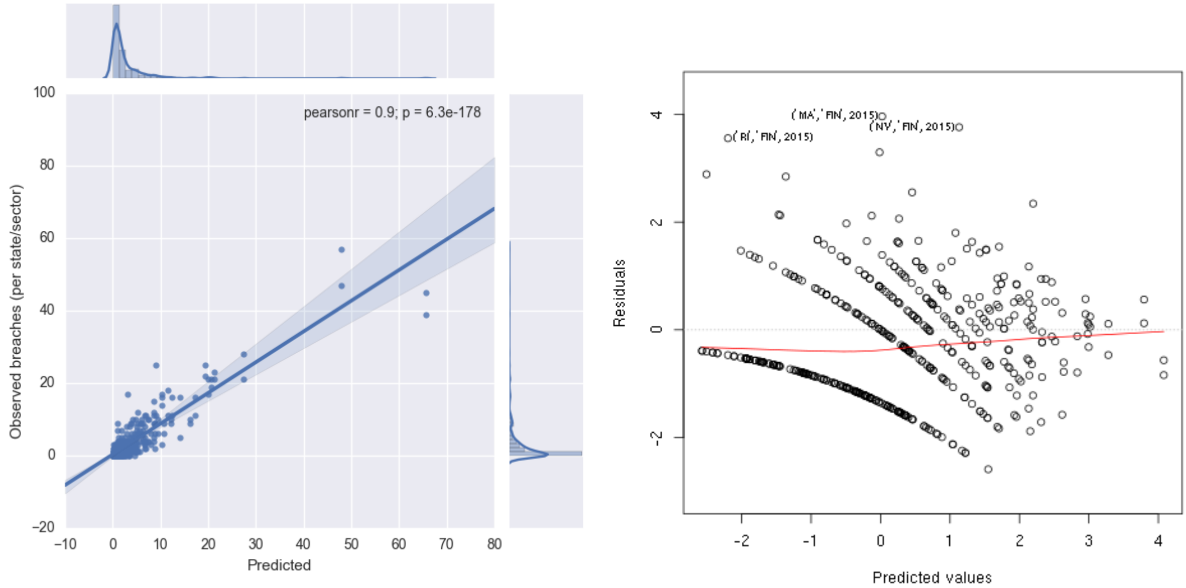

Figure 8 – Breach count model diagnostic plots

The variance inflation factor is between 1 and 2 for all the variables, showing no multicollinearity.

Outliers and datasets with reduced sectors. The top outliers are from the finance sector. This reflects the fact that financial firms are mostly headquartered in specific U.S. states due to tax laws and such, and that ITRC records the place of breach for large companies to the headquarters. This is a systematic bias, not a measurement error that might be corrected by removing outliers. We can however run the regression on a reduced dataset that excludes the multistate firms—e.g. without the financial sector, and the business sector. The results are presented below. The direction of the coefficients remain the same, indicating robust results.

	Dependent variable:		
		breaches	
	(1)	(2)	(3)
inform_agp	0.361**	0.292**	0.247
	(0.149)	(0.142)	(0.164)
inform_agnp	0.044	-0.040	-0.124
	(0.101)	(0.098)	(0.115)
inform_credit	0.289***	0.250***	0.215**
	(0.094)	(0.091)	(0.107)
penalty_cap	-0.053	-0.063	-0.077
	(0.085)	(0.082)	(0.097)
priv_cause	0.119	0.174*	0.165
	(0.094)	(0.091)	(0.108)
risk_harm	-0.231**	-0.234**	0.017
	(0.104)	(0.099)	(0.117)
fin	-8.587***		
	(0.161)		
med	-7.777***	-7.726***	-7.842***
	(0.134)	(0.126)	(0.142)
edu	-7.481***	-7.439***	-7.536***
	(0.162)	(0.154)	(0.167)

gov	-9.218***	-9.179***	-9.287***
	(0.154)	(0.146)	(0.159)
bso	-9.963***	-9.901***	
	(0.135)	(0.127)	
Observations	478	382	286
Log Likelihood	-802.491	-677.914	-463.575
theta	5.171*** (1.196)	8.250*** (2.417)	12.175** (6.001)
Akaike Inf. Crit.	1,626.981	1,375.828	945.150

Note: Negative binomial regression, with sector size as offset, in datasets with all sectors, and all excluding finance, and business. AICs cannot be compared as the datasets differ.
*p<0.1; **p<0.05; ***p<0.01

Appendix III. Alternative breach count models

We present two alternative model specifications: adding a dummy variable for the breach year, and adding an interaction term. Comparing the Akaike Information Criteria of these models against our simple model, both are slightly worse, suggesting that the year dummy or the interaction terms do not add to the preferred parsimonious model.

	Dependent variable:		
		breaches	
	(1)	(2)	(3)
inform_agp	0.361**	0.360**	0.371
	(0.149)	(0.149)	(0.257)
inform_agnp	0.044	0.044	0.102
	(0.101)	(0.101)	(0.174)
inform_credit	0.289***	0.289***	0.316*
	(0.094)	(0.094)	(0.161)
penalty_cap	-0.053	-0.054	-0.001
	(0.085)	(0.085)	(0.145)
priv_cause	0.119	0.119	0.193
	(0.094)	(0.094)	(0.157)
risk_harm	-0.231**	-0.232**	-0.675***
	(0.104)	(0.104)	(0.175)
fin	1.376***	1.376***	-9.118***
	(0.136)	(0.136)	(0.386)
med	2.186***	2.187***	-7.793***
	(0.098)	(0.098)	(0.201)
edu	2.482***	2.483***	-7.148***
	(0.135)	(0.135)	(0.342)
gov	0.745***	0.746***	-9.651***
	(0.126)	(0.126)	(0.313)
y2015		0.027	
		(0.078)	
bso			-9.747***
			(0.203)
inform_agp:fin			0.466
			(0.528)
inform_agp:med			-0.121
			(0.363)
inform_agp:edu			-0.054
			(0.476)
inform_agp:gov			-0.129
			(0.437)
inform_agp:bso			
inform_agnp:fin			0.559
			(0.350)
inform_agnp:med			-0.136
			(0.239)
inform_agnp:edu			-0.207
			(0.344)
inform_agnp:gov			-0.447
			(0.315)
inform_agnp:bso			
inform_credit:fin			0.346
			(0.363)
inform_credit:med			-0.091
			(0.221)
inform_credit:edu			-0.461
			(0.313)
inform_credit:gov			0.133
			(0.299)
inform_credit:bso			
penalty_cap:fin			0.027
			(0.297)
penalty_cap:med			-0.029
			(0.201)
penalty_cap:edu			-0.082
			(0.286)
penalty_cap:gov			-0.239
			(0.266)

penalty_cap:bso			
priv_cause:fin			-0.418
			(0.329)
priv_cause:med			-0.158
			(0.221)
priv_cause:edu			-0.186
			(0.324)
priv_cause:gov			0.359
			(0.290)
priv_cause:bso			
risk_harm:fin			0.427
			(0.365)
risk_harm:med			0.600**
			(0.246)
risk_harm:edu			0.588*
			(0.344)
risk_harm:gov			0.988***
			(0.318)
risk_harm:bso			
Constant	-9.963***	-9.976***	
	(0.135)	(0.140)	
Observations	478	478	478
Log Likelihood	-802.491	-802.433	-785.665
theta	5.171*** (1.196)	5.179*** (1.197)	6.639*** (1.734)
Akaike Inf. Crit.	1,626.981	1,628.867	1,641.330

Note: Negative binomial regression, with sector size as offset.
*p<0.1; **p<0.05; ***p<0.01

Appendix IV. Models with additional state-level controls

We have tested a number of alternative models that include state-level controls. The idea behind them was to control for differences among states that might affect the number of breaches reported (other than the DBNL provisions and sector sizes that we included). We tested the following:

- *Household income* (from the U.S. Census) as a proxy for states' wealth. Most directly, richer individuals might be more interesting targets for identity theft. Additionally, household income is highly correlated with gdp-per-capita, which also reflects wealthier companies, that might have more resources to invest in cybersecurity, and generally, better overall infrastructure.
- We look at *crime rates* (from the Internet Crime Compliant Center report) in the categories of crimes that could also be causes of data-breaches, and all categories of crimes. This could reflect the prevalence of crime in a state, leading to insider breaches or physical theft. It can alternatively also reflect how often citizens report crimes in a state.
- We look at the *centralization of sectors* in various states, e.g. the number of banks per capita, by dividing the number of firms in each sector by the state's population (from the U.S. Census).

We did not find general attributes on cybersecurity investment in states, or business attitudes to risk across U.S. states, that could be interesting controls. Another common choice is to add one dummy variable for each state, but adding 48 dummies does not make sense in a dataset with 480 observations.

The results are provided in the table below. Overall, they offer little improvement over our base model in terms of AIC, and also do not change the sign of the coefficients, indicating our existing models are robust.

	Dependent variable:			
	breaches			
	(1)	(2)	(3)	(4)
inform_agp	0.128	0.543***	0.528***	0.335**
	(0.160)	(0.184)	(0.174)	(0.149)
inform_agnp	-0.062	0.123	0.110	0.019
	(0.104)	(0.112)	(0.107)	(0.102)
inform_credit	0.258***	0.262***	0.270***	0.304***
	(0.093)	(0.094)	(0.093)	(0.094)
penalty_cap	-0.042	-0.044	-0.032	-0.055
	(0.084)	(0.084)	(0.085)	(0.085)
priv_cause	0.108	0.072	0.063	0.133
	(0.094)	(0.098)	(0.098)	(0.094)
risk_harm	-0.191*	-0.345***	-0.367***	-0.236**
	(0.103)	(0.125)	(0.128)	(0.103)
fin	1.370***	1.390***	1.392***	2.021***
	(0.135)	(0.135)	(0.135)	(0.398)
med	2.193***	2.185***	2.186***	2.779***
	(0.097)	(0.097)	(0.097)	(0.359)
edu	2.471***	2.489***	2.490***	3.147***
	(0.134)	(0.134)	(0.134)	(0.410)
gov	0.754***	0.753***	0.755***	1.337***
	(0.125)	(0.126)	(0.126)	(0.368)
house_income	0.00002***			
	(0.00001)			
victim_cause		-0.0001*		
		(0.0001)		
pop2012			-0.000*	
			(0.000)	
orgs_p100				0.406*
				(0.235)
Constant	-10.934***	-9.812***	-9.779***	-10.639***
	(0.299)	(0.162)	(0.167)	(0.419)
Observations	478	478	478	478

Log Likelihood	-796.066	-801.131	-800.855	-801.033
theta	5.441*** (1.264)	5.302*** (1.247)	5.301*** (1.245)	5.226*** (1.206)
Akaike Inf. Crit.	1,616.133	1,626.263	1,625.709	1,626.067

Note: Negative binomial regression, with sector size (organizations) as offset.
AIC of model without controls: 1627.0. *p<0.1; **p<0.05; ***p<0.01

Appendix V. Alternate time models

	Dependent variable: notification_time			
	(1)	(2)	(3)	(4)
hacking	0.130		0.241*	0.254*
	(0.126)		(0.134)	(0.134)
physical	0.068		0.020	0.017
	(0.158)		(0.162)	(0.161)
insider	0.124		0.116	0.137
	(0.194)		(0.196)	(0.198)
unintended				
bso		0.058	0.006	0.127
		(0.230)	(0.238)	(0.241)
fin		-0.082	-0.080	0.022
		(0.263)	(0.262)	(0.262)
edu		0.115	0.087	0.146
		(0.295)	(0.298)	(0.296)
med		0.419*	0.467*	0.448*
		(0.243)	(0.248)	(0.245)
gov				
CA				-0.070
				(0.125)
MD				-0.374**
				(0.149)
NH				-0.308*
				(0.173)
VT				-0.342**
				(0.144)
multi			0.017	0.552**
			(0.113)	(0.234)
Constant	3.701***	3.649***	3.529***	3.728***
	(0.101)	(0.219)	(0.223)	(0.240)
Observations	260	260	260	260
Log Likelihood	-1,237.540	-1,232.446	-1,230.251	-1,225.652
theta	1.498*** (0.126)	1.552*** (0.131)	1.575*** (0.133)	1.626*** (0.138)
Akaike Inf. Crit.	2,483.081	2,474.891	2,478.502	2,477.305

Note: Negative binomial regression. AIC/LL of null model is 2478.2/-1237.1.
p<0.1; **p<0.05; ***p<0.01

	Dependent variable:			
	uninformed_exposure_time			
	(1)	(2)	(3)	(4)
hacking	0.412**		0.537***	0.623***
	(0.170)		(0.183)	(0.183)
physical	-0.660***		-0.588***	-0.430**
	(0.201)		(0.207)	(0.207)
insider	1.689***		1.594***	1.792***
	(0.247)		(0.252)	(0.251)
unintended				
bso		0.238	-0.342	-0.552
		(0.357)	(0.340)	(0.343)
fin		-0.570	-0.747**	-0.855**
		(0.392)	(0.352)	(0.353)
edu		0.216	-0.268	-0.493
		(0.524)	(0.467)	(0.467)
med		0.709*	0.318	0.189
		(0.380)	(0.351)	(0.349)
gov				
CA				0.277*
				(0.167)
MD				0.278
				(0.190)
NH				0.587***
				(0.200)
VT				-0.019
				(0.187)
multi			0.258*	-0.230
			(0.142)	(0.277)
Constant	4.595***	4.795***	4.629***	4.381***
	(0.137)	(0.343)	(0.298)	(0.322)
Observations	257	257	257	257
Log Likelihood	-1,496.118	-1,531.416	-1,483.143	-1,478.444
theta	0.877*** (0.068)	0.713*** (0.054)	0.948*** (0.074)	0.976*** (0.077)
Akaike Inf. Crit.	3,000.235	3,072.833	2,984.286	2,982.888

Note: Negative binomial regression. AIC/LL of null model is 3088/-1542.
*p<0.1; **p<0.05; ***p<0.01

Reference List

Ablon, Lillian, Paul Heaton, Diana Lavery, Sasha Romanosky. “Consumer Attitudes Toward Data Breach Notifications and Loss of Personal Information.” Rand. 2016.

Acquisti, Alessandro, Allan Friedman, and Rahul Telang. “Is There a Cost to Privacy Breaches? An Event Study.” Paper presented at the fifth workshop on the Economics of Information Security, University of Cambridge, England, June 2006.

BakerHostetler, “State Data Breach,” 2014.

Bisogni, Fabio. "Proving Limits of State Data Breach Notification Laws: Is a Federal Law the Most Adequate Solution?" Journal of Information Policy 6 (2016): 154-205.

Campbell, Katherine, Lawrence A. Gordon, Martin P. Loeb., and Lei Zhou. “The Economic Cost of Publicly Announced Information Security Breaches: Empirical Evidence from the Stock Market.” Journal of Computer Security 11 (2003): 431–48.

Cavusoglu, Huseyin, Birendra Mishra, and Srinivasan Raghunathan. “The Effect of Internet Security Breach Announcements on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers.” International Journal of Electronic Commerce 9 (2004): 70–104.

Chlotia Posey Garrison, Matoteng Ncube, "A longitudinal analysis of data breaches", Information Management & Computer Security, Vol. 19 Issue: 4 (2011): 216-230.

Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016.

Dutch DPA, 2017. “Overzicht meldingen datalekken eerste kwartaal 2017”, Available online at: https://autoriteitpersoonsgegevens.nl/nl/nieuws/overzicht-meldingen-datalekken-eerste-kwartaal-2017

Economic Census (U.S. Government). 2012. Online at: http://www.census.gov/econ/census/. Accessed 01-Feb-2017.

Edwards, Benjamin, Steven Hofmeyr, and Stephanie Forrest. “Hype and Heavy Tails: A Closer Look at Data Breaches.” WEIS 2015: 14th Workshop on the Economics of Information Security, June 2015.

ENISA. “Data Breach Notifications in Europe”. 2011.

Faulkner Brandon, 'Hacking into Data Breach Notification Laws' (2007) 59(5) Florida Law Review. 1097, 1104.

Federal S.177 - Data Security and Breach Notification Act of 2015.

Gordon, Lawrence A. and Loeb, Martin P. and Lucyshyn, William, Information Security Expenditures and Real Options: A Wait-and-See Approach (May 31, 2003). Computer Security Journal, Vol. XIX, No. 2, Spring, 2003.

Hilbe, Joseph M. “Negative Binomial Regression” Cambridge University Press, 2nd Edition (2011).

Internet Crime Compliant Center. “2015 Internet Crime Report”. US Department of Justice, Federal Bureau of Investigation. 2016.

ITRC 2017. Identity Theft Resource Center. Online at: http://www.idtheftcenter.org/data-breaches.html. Accessed 28-Feb-2017.

Ko, Myung, and Carlos Dorantes. “The Impact of Information Security Breaches on Financial Performance of the Breached Firms: An Empirical Investigation.” Journal of Information Technology Management 17 (2006): 13–22.

Kwon, Juhee, and Eric Johnson. “The Market Effect of Healthcare Security: Do Patients Care about Data Breaches?” WEIS 2015: 14th Workshop on the Economics of Information Security, June 2015.

McKinsey Global Institute. “Digital America: a tale of the haves and have-mores”. December 2015.

Moore, Tyler, Scott Dynes, and Frederick Chang. Identifying how firms manage cybersecurity investment. In 15th Workshop on the Economics of Information Security (WEIS), 2016.

Ranger, Steve. “Data Breach Laws Make Companies Serious about Security.” September 3, 2007. Silicon.com. http://management.silicon.com/itdirector/0,39024673,39168303,00. htm?r=1.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016.

Romanosky, Sasha, Rahul Telang, and Alessandro Acquisti. “Do Data Breach Disclosure Laws Reduce Identity Theft?” Journal of Policy Analysis and Management 30, no. 2 (2011): 256–86.

Romanosky, Sasha, David Hoffman and Alessandro Acquisti, A. “Empirical Analysis of Data Breach Litigation.” Journal of Empirical Legal Studies, 11 (2014): 74–104.

Schwartz, Paul, and Edward Janger. “Notification of Data Security Breaches.” Michigan Law Review 105, no. 913 (2007).

Security Scorecard. “2016 Financial Industry Cybersecurity Report”. Security Scorecard R&D Department. August 2016.

Sinanaj, Griselda and Zafar, Humayun, "Who wins in a Data Breach? - A comparative study on the Intangible Costs of Data Breach Incidents" (2016). PACIS 2016 Proceedings. Paper 60.

Telang, Rahul, and Sunil Wattal. “An Empirical Analysis of the Impact of Software Vulnerability Announcements on Firm Stock Price.” IEEE Transactions on Software Engineering 33 (2007): 544–57.