More Than a Suspect: An Investigation into the Connection Between Data Breaches, Identity Theft, and Data Breach Notification Laws

Bisogni, F.; Asghari, H.

DOI
10.5325/JINFOPOLI.10.2020.0045

Publication date
2020

Document Version
Final published version

Published in
Journal of Information Policy

Citation (APA)
Bisogni, F., & Asghari, H. (2020). More Than a Suspect: An Investigation into the Connection Between Data Breaches, Identity Theft, and Data Breach Notification Laws. *Journal of Information Policy*, *10*, 45-82. https://doi.org/10.5325/JINFOPOLI.10.2020.0045

More Than a Suspect: An Investigation into the Connection Between Data Breaches, Identity Theft, and Data Breach Notification Laws

Author(s): Fabio Bisogni and Hadi Asghari

Source: *Journal of Information Policy*, 2020, Vol. 10 (2020), pp. 45-82

Published by: Penn State University Press

Stable URL: https://www.jstor.org/stable/10.5325/jinfopoli.10.2020.0045

MORE THAN A SUSPECT

An Investigation into the Connection Between Data Breaches, Identity Theft, and Data Breach Notification Laws

Fabio Bisogni and Hadi Asghari

ABSTRACT

This article investigates the relationship between data breaches and identity theft, including the impact of Data Breach Notification Laws (DBNL) on these incidents (using empirical data and Bayesian modeling). We collected incident data on breaches and identity thefts over a 13-year timespan (2005–2017) in the United States. Our analysis shows that the correlation is driven by the size of a state. Enacting a DBNL still slightly reduces rates of identity theft; while publishing breaches notifications by Attorney Generals helps the broader security community learning about them. We conclude with an in-depth discussion on what the European Union can learn from the US experience.

Keywords: data breach notification laws; identity theft; data breaches

Information technology enables the collection and storage of large amounts of personal data. While these activities provide unquestionable economic benefits, it has also proven impossible to keep personal data fully secure against criminal misuse. Surveys report that in 2017, identity thieves fraudulently obtained approximately $16.8 billion from 16.7 million American consumers.[1] According to the same study, in the past 6 years, identity thieves have stolen over $106 billion from their victims. Having access to

1. Javelin.

Fabio Bisogni: Delft University of Technology, Faculty of Technology, Policy and Management, Delft, the Netherlands; Fondazione FORMIT, Rome Italy
Hadi Asghari: Delft University of Technology, Faculty of Technology, Policy and Management, Delft, the Netherlands
DOI: 10.5325/jinfopoli.10.2020.0045

personally identifiable information[2] is a prerequisite for perpetrating identity crime. Data breaches are a key source for this access.[3]

California was the first state to enact a data breach notification law[4] (hereafter also DBNL), emphasizing the potential criminal harm of identity theft as their main rationale for the *duty to notify*.[5] Other US states have since enacted DBNLs. In Europe, the General Data Protection Regulation (GDPR) similarly recognizes (in its preamble) that identity theft is a major risk when a data breach is not addressed in an appropriate and timely manner.

Despite these legal rationales, little research exists to date on the relationship between data breaches, identity theft, and the impact of DBNLs on related trends over time. It is clear that data breaches are numerous and increasing: the Identity Theft Resource Center (ITRC) reported 1,579 data breaches in the United States in 2017, an increase from 1,091 in 2016 and only 421 in 2011. Yet there is no definitive estimate of how many cases of identity theft have resulted from data breaches. In a small-scale effort, the US Government Accountability Office (2007) examined 24 large data breaches between 2000 and 2005, and conclusively linked four of them to subsequent outbreaks of fraud. Romanosky et al.[6] have done one of the few studies on the impact of DBNL on identity theft, measuring and estimating this effect using panel data (from 2002 to 2009) from the US Federal Trade Commission.

---

2. The U.S. government defined the term *"personally identifiable information"* in 2007 in a memorandum from the Executive Office of the President, Office of Management and Budget (OMB), [M-07-16 SUBJECT: Safeguarding Against and Responding to the Breach of Personally Identifiable Information FROM: Clay Johnson III, Deputy Director for Management (2007/05/22)] and that usage now appears in U.S. standards such as the NIST Guide to Protecting the Confidentiality of Personally Identifiable Information (SP 800-122).["Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)" (PDF). Special Publication 800-122. NIST.]

The European Regulation (EU) 2016/679 (General Data Protection Regulation—GDPR) in its Article 4 defines personal data as any information relating to an identified or identifiable natural person; an identifiable natural person being one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person.

3. Garrison and Ncube; Roberds and Schreft.

4. California Civil Code § 1729.98 enacted in 2003.

5. Skinner; Draper; Bisogni.

6. Romanosky et al.

This paper addresses this research gap by investigating the relationship between data breaches and identity theft in more depth, including the impact of DBNL enactments (and revisions) on these incidents (using empirical data and Bayesian modeling). We collected incident data on breaches and identity theft over a 13-year timespan (2005–2017) in the United States. The databases we used included those of the ITRC (data breaches), Privacy Rights Clearinghouse (data breaches), Consumer Sentinel Network (identity theft), and Perkins Coie (DBNLs).
Our analysis reveals that the correlation between data breaches and identity theft is driven in large part by the size of the state. Enacting a DBNL still slightly reduces rates of identity theft in the enacting states, while Attorney Generals publishing breach notifications helps the broader security community learn about them. We conclude the paper with an in-depth discussion on what the European Union (EU) can learn from the results of 15 years of regulations in the United States (since the enactment of the first DBNL in California), including insights that are relevant for the governance and monitoring of the GDPR and other statutes of the Data Protection Package.[7]

Background

In this section, we explore general aspects related to identity theft, data breaches and the laws adopted in the United States and Europe to control these two issues. It is generally acknowledged that identity theft can take many forms. The US Government Accountability Office, in its report to congressional requesters dated July 2007,[8] divided identity theft into two categories: existing-account fraud and unauthorized creation of new accounts. Examples of these categories are, respectively, the misuse of credit card numbers (credit card information is stolen) and opening a credit card account in someone else's name (personal information is stolen). Another

7. This includes also the Directive (EU) 2016/680 for data processing by law enforcement for the purposes of prevention, investigation, detection or prosecution of criminal offences (implemented 6 May 2018). See https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en (last visited 13/1/2019).

8. GAO-07-737, a report to congressional requesters.

classification is provided by the ITRC; ITRC identifies five categories of identity theft:

- Financial identity theft: when the imposter uses another individual's personal identifying information, primarily a Social Security number, to establish new credit lines;
- Criminal identity theft: when a criminal gives another person's personal identifying information, in place of his or her own, to law enforcement;
- Identity cloning: when the imposter uses the victim's information to establish a new life. He or she actually lives and works in the victim's identity;
- Medical identity theft: use of someone else's data in order to obtain medical services or goods;
- Commercial identity theft: similar to financial identity theft except the victim is a commercial entity.[9]

In all of the abovementioned cases, data, mostly personal identifiable information, in the hands of thieves is a prerequisite to perpetrate the crime, and therefore the means to access this information plays a central role. Data breaches appear to be the primary source for accessing personal information and thereby the primary source of identity theft.[10]

However, we do not have a definitive estimate of how many cases of identity theft have resulted from data breaches. This type of estimation was the goal of the US Government Accountability Office (2007) in examining 24 data breaches between 2000 and 2005 in which large amounts of data were compromised. The GAO conclusively linked four large breaches to subsequent outbreaks of fraud. However, the sample was very limited; thus its findings cannot be generalized. An additional study providing evidence that a significant proportion of identity theft can be attributed to inadequately secured commercial data is the one conducted by Gordon et al.[11] The study examined 274 cases of identity theft prosecuted by the Secret Service from 2000 to 2006 and found that 50% of the cases resulted from compromised data at a business.

---

9. Di Ciccio indicates also Synthetic Identity Theft (use of different subjects' personal data combined in order to create a new identity), Ghosting (creation of a new identity, different from the original one by exploiting the data of a deceased person), Cyber Bullying (Impersonation: impersonation in a different person, by means of cellular phones or web services, with the purpose of sending messages with objectionable contents).

10. Garrison and Ncube.

11. Gordon et al.

The nature of a causal connection between security breaches and concrete harm suffered by consumers is not always easy to determine. In fact, a data breach does not necessarily result in identity theft, as data may be stolen without being used for fraudulent purposes. Moreover, identity theft can occur without a data breach. In consumer surveys, victims of identity theft who know how their information was stolen commonly attribute their loss to channels that are not linked to technology, such as lost or stolen wallets (43% of cases reported in Javelin),[12] fraud by acquaintances (13%) or stolen mail (3%). Only 11% of cases are reported to be linked to data breaches and 11% to online methods.

It is evident from the existing literature that most of the analysis performed on data breaches and identity theft have been carried out in the United States, which is a pioneer country in terms of DBNLs. DBNLs in the United States were promulgated under the main objective of reducing identity theft. Yet the measurement of this specific effect has been the subject of limited research. The work of Romanosky et al.[13] is the only empirical study measuring this effect; using panel data from the US Federal Trade Commission,[14] the researchers estimated the impact of data breach disclosure laws on identity theft from 2002 to 2009. They found that the adoption of data breach disclosure laws reduces identity theft caused by data breaches by 6.1% on average. Our study not only updates this analysis with a wider time span, but also extends it to the effect of specific law provisions and legal revisions not only on identity theft but also more directly on data breaches. Moreover, we tested different statistical models to identify the strongest model for such estimation concluding that Bayesian model is more adequate.

To lay proper foundations for a United States–Europe comparison, it is important to highlight that DBNLs not only attempt to fulfill a specific purpose, the mitigation of identity theft, but also confront conflicting goals of consumer protection and corporate compliance-cost minimization. In contrast, comprehensive information privacy legal frameworks, such as that of Europe, have an expansive purpose to ensure legal protections related to the protection of personal information.[15] Information

12. Javelin Strategy & Research, *2009 Identity Fraud Survey Report.*

13. Romanosky, Telang, and Acquisti.

14. The same source we used for Identity Theft data.

15. Information privacy law is based on the notion that individuals have rights relating to control over their personal information (Kang), or at least, have rights pertaining to who can access their personal information (Gavinson) or a combination of both (Moor).

privacy laws set minimum standards that relate to fair information practices and provide individuals with a series of limited rights of involvement in the process of personal information exchange.[16] The relation between laws protecting privacy and laws addressing concerns about identity theft is complex and sometimes antagonistic. For example, Towle[17] described the dilemma as follows: customers argue both for and against more privacy, creating tension under identity theft statutes and attribution procedures. Vendors and organizations generally find themselves between a rock and a hard place. They are asked to increasingly respect more privacy in not forcing customers to provide extensive identification data before entering into a transaction, but also less privacy in ensuring that no one is violating their customer's identities.

Identity theft and data breaches have become a relevant issue in the EU not only for individual member states, but also in the broader EU agenda. The main result is the GDPR 2016/679, which entered into force on May 24, 2016 and applied, after a 2-year transition period, from May 25, 2018. Contrary to its predecessor, Directive 95/46/EC,[18] the GDPR equally applies directly to every citizen and organization falling within the scope of EU law. Hence, the GDPR is well placed to become a significant piece of legislation. The connection between identity theft and data breaches is clearly defined in the preamble of the GDPR (EU) 2016/679, point (85):

> *A personal data breach may, if not addressed in an appropriate and timely manner, result in physical, material or non-material damage to natural persons such as loss of control over their personal data or limitation of their rights, discrimination,* ***identity theft*** *or fraud, financial loss, unauthorised reversal of pseudonymisation, damage to reputation, loss of confidentiality of personal data protected by professional secrecy or any other significant economic or social disadvantage to the natural person concerned.*

In the European context, the situation is partially different from the United States, where consumers are not protected by a general right of information privacy. Indeed, the breach notice is not associated with any

16. See Privacy Rights Clearinghouse, Why Privacy, https://www.privacyrights.org/why-privacy-0 (last visited January 13, 2019).

17. Towle, 261–264.

18. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281/31 (Data Protection Directive).

right to compensation.[19] In addition, the GDPR extends the notification duty to all data controllers. Apart from these factors, the GDPR mainly follows the approach of the United States DBNLs, with one important difference: the regulatory environment that it creates includes a much-improved enforcement mechanism for data protection violations compared to the US scheme. This also means that companies reporting a breach may face substantive fines by the regulators, in addition to any possible action by the individuals affected.[20] In the European context, while class actions are not normally found in European jurisdictions, fines can be levied by the data protection authorities without a need to show a concrete loss for individuals.

The GDPR will therefore not only reaffirm the general right to information privacy, but also provide an enforcement mechanism following the evolution of privacy regulations. (This issue is further analyzed in the Discussion).

Research Method

The remainder of this paper investigates the relationship between data breaches, identity theft, and DBNL enactment and revisions. We start by investigating the causal connection between data breaches and identity theft, with the aim of identifying the strength of correlation between the two variables. We then move to the effects that DBNL enactment and revision have on both, also considering the level of notification publicity that these laws may introduce. As illustrated in Figure 1, for this analysis we take into account other important predictors related to state wealth and infrastructure, to digital threats (for data breaches), and to crime and breached records (for identity theft).

19. Winn: "Attempts to establish a right to damages following receipt of a security breach notice through class action lawsuits have generally only succeeded in clarifying the degree to which no such right exists, although many businesses suffering breaches have chosen on a voluntary basis to provide their customers with credit monitoring services to reduce the risk of harm from identity theft."

20. In the United States, there is no general tort of privacy violation, however, individuals affected by a data breach can sue if they can prove that they suffered economic harm through the negligence of the breached entity. The availability of class actions in the US legal system gives this opportunity.

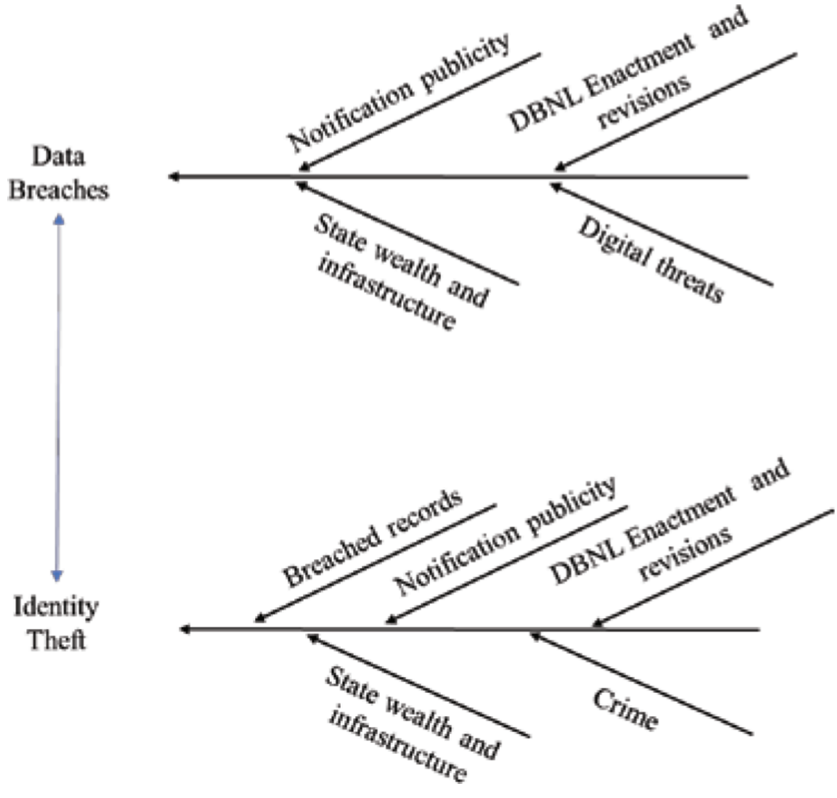

FIGURE 1 Causal Diagram.

The effects of DBNLs can be studied thanks to specific datasets. The introduction of legislation to address the threat of data breaches in the United States has indeed fostered a number of databases which gather information about data breaches and identity theft at state level.

We combine the following sources as summarized in Table 1:

TABLE 1 Summary Statistics

Years	**2005–2017**
Total DB	8,171
Average DB per year	626
Max. value (2017)	1,557
Min. value (2005)	133
Total IDT	3,879,919
Average IDT per year	295,455
Max. value (2015)	440,068
Min. value (2010)	238,107

- The data breaches come from the ITRC[21] and Privacy Rights Clearinghouse[22] databases. The Privacy Rights Clearinghouse database provided the only data for the time span 2005 to 2012. For the timespan 2013 to 2017 we used ITRC data.[23]
- The identity thefts data come from the Consumer Sentinel Network database.[24] These statistics are consumer reported and collected by the Federal Trade Commission for each state. (They reflect reports by individuals once they discover a theft, and not an automated check and balance by other agencies such as consumer credit bureaus).

Our dataset comprises 650 total records, with each record containing the number of data breaches and identity theft in one of the 50 states from 2005 to 2017. We add a number of common predictor variables to this dataset, including the population of states,[25] number of firms per state[26] and GDP per capita.[27] (These variables are used to normalize, as predictors and as controls; further explanations are provided for each use.)

Finally, we added the date when DBNL laws came into effect for each state, and dates of their subsequent revision/amendment, based on data from Perkins Coie.[28] California was the first US state to enact a DBNL (in 2003); 33 states enacted their DBNLs before 2015; Alabama and South Dakota were among the last (enacting in 2018). The majority of states have revised or amended their DBNLs a number of times since enactment, as provided in Table 2.[29]

---

21. https://www.idtheftcenter.org/data-breaches/ reporting 9,774 data breaches in the time span 2005–2018.

22. https://www.privacyrights.org/data-breaches/ reporting 9,002 data breaches in the time span 2005–2018.

23. ITRC included a higher number of data breaches for the analyzed period (4.851 vs. 3.546). For the time span 2005-2012 ITRC data were only available at aggregated level, so we used PRC data. The similar number of data breaches collected by the two sources in this time span (only c. 5% of difference) suggests that potential data heterogeneity between datasets is very limited.

24. https://www.ftc.gov/enforcement/consumer-sentinel-network/reports.

25. https://www.ftc.gov/enforcement/consumer-sentinel-network/reports.

26. US Census Bureau. Number of Firms, Number of Establishments, Employment, and Annual Payroll by Enterprise Employment Size for the United States and States, Totals: 2016. https://www.census.gov/data/tables/2016/econ/susb/2016-susb-annual.html

27. US Bureau of Economic Analysis. Last updated: May 1, 2019— new statistics for 2018; revised statistics for 2010–2017.

28. https://www.perkinscoie.com/. Given that we consolidated our dataset at yearly level, we considered enactments and revisions in the last quarter of a year for the subsequent year.

29. The average time for the first revision (or amendment) is 6 years and 2 months. Among the states, 10 went through a second revision, with an average time (from the previous change) of 3 years and 3 months; 4 went through a third revision (within 2 years and 2 months); and 2 through a fourth one.

TABLE 2 States Enacting DBNLs and Subsequent Revisions. For example, as of 31 December 2018, of the 16 DBNLs Enacted in 2006, 5 had No Revision, 9 had One Revision, and 2 had Two Revisions

Year of Enactment	Number of Enacting States	States With 0 Revision	States With 1 Revision	States With 2 Revisions	States With 3 Revisions	States With 4 Revisions
2003	1	–	–	–	–	1
2005	10	1	5	3		1
2006	16	5	9	2	–	–
2007	9	3	4	–	2	–
2008	5	2	2	1	–	–
2009	4	2	1	1	–	–
2011	1	1	–	–	–	–
2014	1	–	1	–	–	–
2017	1	1	–	–	–	–
2018	2	2	–	–	–	–
Total	**50**	**17**	**22**	**7**	**2**	**2**

We began our analysis with descriptive statistics, followed by a difference-in-differences (DiD) analysis for the effects of DBNLs.[30] DiD models are, in short, not suitable for our analysis, as they generate high standard error, and cannot reliably estimate the enactment effect. DiD requires assuming parallel trends for states before (or only after) enactment, which does not hold upon inspection. A further problem is that DiD treats the year and state intercepts (dummies, or fixed-effects) as completely independent of each other. This conceptualization ignores the fact that external events may impact data breaches across all states.[31] Nevertheless, as DiD models are used by a number of prior studies involving data breaches, we included them as a baseline.[32]

Our main analysis employed multilevel (also known as hierarchical or random-effects) Bayesian regression models.[33] The detailed model specifications are presented in the Findings section (where we also explain the variables and interpret the results). The motivation for using Bayesian multilevel

30. Angrist and Pischke.

31. Some prior work has attempted to resolve the fact that the year and state dummies are not completely independent in this instance using *robust and cluster-corrected* error terms (e.g., Romanosky, Telang, and Acquisti). However, the Bayesian multilevel method that we present next is a more flexible and robust approach.

32. The DiD models are estimated using non-Bayesian MLE methods.

33. The multilevel refers to stacking of distributions in the model definitions, due to the pooling of the intercepts.

modeling is that it can more precisely estimate the effects of a common intervention (DBNL) while allowing for differences among states by pooling together the varying intercepts[34] for each state (and similarly pooling the varying intercepts for each year). The two classic approaches to modeling interventions across multiple states, which are opposites on a spectrum, are to specify the model with only the intervention variable and no additional dummies for the states, or to specify the model with the intervention and add a unique dummy (or intercept) for each state. On the one hand, the first option (no unique intercepts) ignores structural differences among states that may affect the observation, and results in very poor model fit and estimates. The second option (per state intercept), on the other hand, assumes that the states are completely independent from each other, and may lead to the intercept capturing too much of the variance (and noise) in the data. Pooling the intercepts is the more realistic and accurate middle ground that Bayesian multilevel modeling allows: the states have unique intercepts, but those intercepts are kept as close to each other as possible in the fitting process. Given the increase in computational power, Bayesian models are increasingly recommended for problems with inherent clusters.

We follow Bayesian inference and reporting procedures as recommended by McElreath[35] and Kruschke.[36] Our *Jupyter notebooks*, which make use of *PyStan*[37] and *ArviZ*[38] packages, in addition to the classic Python analysis toolkits, are available upon request.[39]

Findings

Correlations Driven by Size

Figure 2 plots data breach and identity theft trends from 2005 to 2017.[40] The figure depicts clear and parallel growing trends until 2015.

34. see McElreath, chap. 12.

35. McElreath.

36. Kruschke.

37. PyStan (https://github.com/stan-dev/pystan) provides a Python interface to Stan, a package for Bayesian inference using the No-U-Turn sampler, a variant of Hamiltonian Monte Carlo.

38. ArviZ (https://arviz-devs.github.io/arviz/) is a Python package for exploratory analysis of Bayesian models.

39. We also contemplated the use of Bayesian multilevel ARMAX models. However, given the fact that data breaches and identity thefts have clear time trends, and given that the random variation is less important than the overall magnitude of these incidents, ARMAX models are not informative for our study.

40. See Appendix I for a comparison with breached record counts.

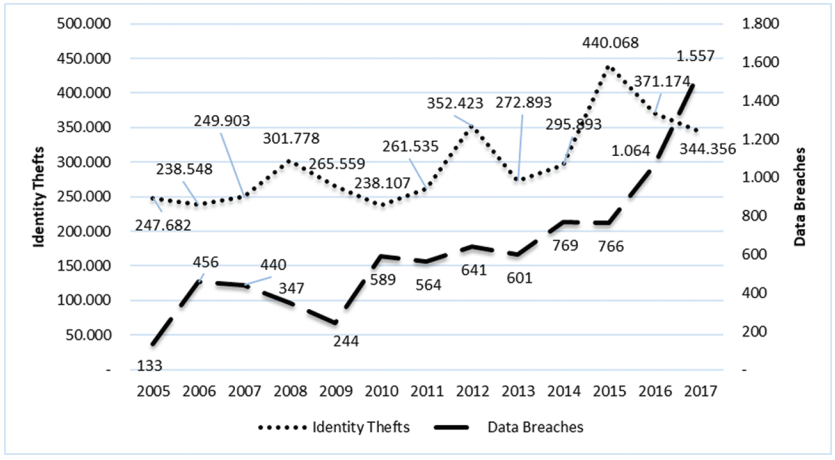

FIGURE 2 US Trends in Data Breaches and Identity Thefts.

The causes of this growth can be traced to many factors, such as growing digitalization, and the increasing ease with which financial transactions are conducted electronically and processes are managed digitally. This development enlarges the opportunity for criminals to act in the digital arena. In addition, the ability to monetize personal information has increased as an incentive to perpetrate data breaches. At the same time, data breaches have been more frequently publicized over time, with higher numbers of Attorneys General publishing the notifications received from breached organizations and therefore increasing the number of breaches fed into relevant databases.[41] As such, the number is not necessarily growing due solely to more frequent data breaches, but could also be growing due simple to the increased reporting of data breaches through public channels.

Figure 2 also illustrates that the number of reported incidences of identity theft follows a more unstable trend: the phenomenon has been generally growing over the 13 years, but with positive and negative peaks. For example, the number of reported identity thefts in 2017 (344,356) was slightly lower than the value reported in 2012 (352,423). The differences in the two time trends indicate the existence of data breaches that do not lead to identity theft, and incidences of identity theft that are not a result of data breaches.

As the scatterplots illustrated in Figure 3—left reveal, the correlation between data breaches and identity theft is significant (Pearson correlation coefficient of 0.77).[42] However, the correlation *considerably weakens if we*

41. see Bisogni, Asghari, and Van Eeten.

42. The correlation is much stronger (0.95) if we take into consideration only the subset of states (and years), where Attorneys General report notifications received from breached

normalize the variable "identity theft" with a state's population (Pearson coefficient 0.42; see Figure 2—right).[43] The correlation *further weakens if we also normalize "data breaches"* by the number of firms in a state (0.23). In other words, the *strong correlation is driven by the size of the state, and once we control for size, the unexplained variance increases*. (This finding is in line with the fact that the causes of identity theft are not limited to data breaches, and that not all breaches are publicly known.)

However, another factor may also be at play: DBNLs may have different effects on the two variables. We investigate this scenario in the following sections.

The Impact of DBNL on Data Breaches

We employed regression analysis to model the impact of DBNL on data breaches.

Difference in Differences

As explained in the Methods section, we first used the *DiD* method. DiD is a well-established method that *under the right assumptions* mimics an experimental design using observational data.[44] The basic requirement is a longitudinal and cross-sectional dataset, with treatments applied at various points in

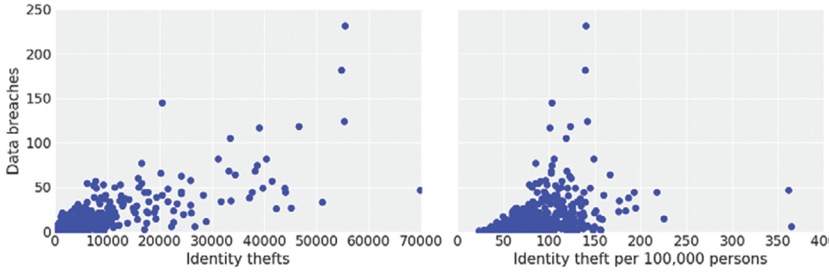

FIGURE 3 Scatter Plots between Data Breaches and Identity Thefts (left); and Data Breaches and Identity Theft per 100,000 Persons (right). The Pearson correlation coefficient is 0.77 for the left and 0.42 for the right scatter plot.

organizations, shrinking the number of data breaches not known to the public and therefore reducing the gap between current data breaches and reported ones (Bisogni, Asghari, and Van Eeten). As of 2019, 22 states require notifications to the Attorney General, but only 10 states publish details of the events and the notification letters. These 10 states include California, Indiana, Maine, Maryland, Montana, New Hampshire, Oregon, Vermont, Washington, and Wisconsin.

43. This correlation is similarly stronger when considering only states with AG reporting (0.58).

44. Angrist and Pischke.

time. The key assumption is that the control and treatment outcomes move *in parallel* in the absence of treatment.

In our case, this means assuming that data breach trends run in parallel across states. As a number of prior studies which examined the impacts of data breaches have used DiD,[45] we *temporarily* accept this assumption. The regression formula is expressed as follows:

$$\text{Breaches}_{s,y} = \delta_{DD}\text{Enacted}_{s,y} + \sum_k \beta_k \ \text{State}_k + \sum_j \gamma_j \ \text{Year}_j$$

The formula includes the enactment effect (δ_{DD}) and add dummies to control for difference by state (β_k) and by year (γ_j).

The left part of Figure 4 depicts the density plot for the dependent variable, data breaches.[46] This variable may be fitted with a negative binomial curve. This choice is conceptually sound because data breaches are rare,[47] discrete events, and counts of such events are best modeled using the negative binomial distribution (see among others: Edwards, Hofmeyr, and Forest).[48]

The regression results are summarized in Table 3. The "enactment effect" (*hasdbnl*) is $e^{0.02 \pm 0.40}$ (coefficients must be interpreted as $e^{\text{coef} \pm 2\text{stderr}}$ due to the GLM specification). This results in 68% to 152% change in the odds of a breach being reported after enactment.

The high standard error means that we cannot reliably estimate the enactment effect using DiD. In fact, the assumption of parallel trends also does not hold if we plot the trends for states before (or only after) enactment. Another problem with the DiD specification is that it assumes that the year and state intercepts (dummies, or fixed-effects) are completely independent of each other. This assumption ignores the fact that certain external effects may impact data breaches across all states.[49]

---

45. For example, Kwon and Johnson; Choi and Johnson.

46. We *normalize* breaches in our models (that is given its correlation with the number of firms in a state) using a regression *offset* and keep the dependent variable as breaches.

47. Rare considering the number of data breaches reported relative to the millions of firms processing data.

48. Edwards, Hofmeyr, Forrest.

49. Some prior work has attempted to resolve the fact that the year and state dummies are not completely independent in this instance using *robust and cluster-corrected* error terms (e.g., Romanosky, Telang, and Acquisti). However, the Bayesian multilevel method that we present next is a more flexible and robust approach.

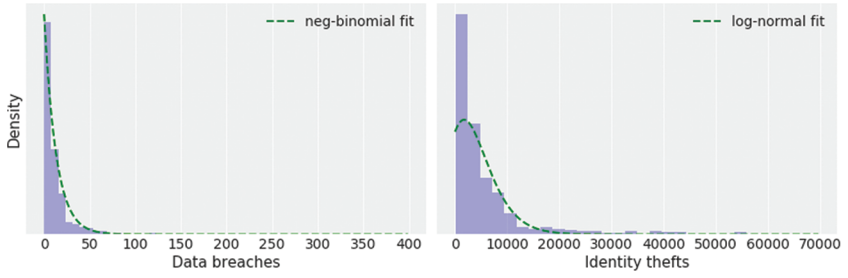

FIGURE 4 Density Plots for the Dependent Variables (Data Breaches, Identity Thefts, and the Normalized Versions). The dashed lines represent fitted distributions: negative-binomial (also known as the gamma-Poisson) for data breaches; log-normal for ID thefts.

TABLE 3 DiD Model Results. Breaches as the Dependent var.; Uses Negative Binomial Regression; See Appendix for Full Results

Generalized Linear Model Regression Results						
Dep. variable:		Breaches	No. observations:		650	
Model:		GLM	Df residuals:		587	
Model family:		Negative binomial	Df model:		62	
Link function:		Log	Scale:		1.0000	
Method:		IRLS	Log-likelihood:		−1929.6	
	coef	Std err	Z	P>|Z|	[0.025	0.975]
Hasdbn1	0.0173	0.203	0.085	0.932	−0.381	0.415
Intercept	−1.0892	0.408	−2.672	0.008	−1.888	−0.290
States dummies		*(See appendix)*				
Years dummies		*(See appendix)*				

Bayesian Multilevel Model

Bayesian multilevel modeling effectively resolves the deficits of DiD. We built a model in which each state (and year) has its own intercept, but the intercepts are *pooled* together by assuming they come from a common underlying distribution (with tight prior variance).[50] We specifically opted

50. McElreath (2016) refers to this as "partial pooling," which is in between "no pooling" (assuming each state acts fully independent of the other) and "complete pooling" (ignoring differences among states and having only a common intercept). Partial pooling strikes a balance by allowing some state differences while assuming there still is a common pattern. These models are also referred to as *random effects models*.

for a *multilevel Poisson model*, because it yields more efficient results (that is better model fits) than the negative binomial distribution *when combined with pooled varying intercepts*.[51]

We modeled the regression using *Stan* platform and programming language. The complete model code can be found in the Appendix. The key lines of the model are the following:

```
// priors
alpha  ~ normal(0, 10);
betas  ~ normal(0, 1);
a_years  ~ normal(0, sigma_y);
a_states ~ normal(0, sigma_s);
sigma_y  ~ cauchy(0, 1);
sigma_s  ~ cauchy(0, 1);
// linear relation
mu = intercept + a_years + a_states + betas*X + offset;
Y  ~ poisson_log(mu);
```

In the model, *Y* is the observed data (breaches per state/year); *X* represents the regression predictors (e.g., whether DBNL has been enacted, DBNL provisions, and control variables) and *mu* is the Poisson rate. *mu* is modeled using a linear relationship between a common intercept, varying year and state intercepts (*a_years*, *a_states*),[52] the predictor coefficients (*betas*) and an *offset*[53] that limits the Poisson rate (here, the number of firms in the state). The year and state intercepts have *weakly informative priors,* in this case, a shared normal distribution and a tight sigma. We plugged in the following predictors:

- *b_enacted* indicates whether a state in a given year enacts a DBNL;

51. A Poisson distribution is (also) a distribution of counts events, but it requires the sample mean and variance to be equal. This isn't the case if we look at all the data breaches together, but it holds if we assume each state to have its *own* rate. The *mixture* of Poisson distributions leads to the negative-binomial (or gamma-Poisson) distribution. It has the advantage of not needing negative binomial's *dispersion* parameter, which makes the model estimations more efficient. This better fit can also be tested with *the widely applicable information criteria (WAIC)*, which indeed holds in this case. Also see McElreath, 350–383.

52. Another common approach is to use a varying intercept *per observation*. This of course risks over-fitting the model; a point that is also reflected in a worse WAIC score here.

53. Using an *offset* is the recommended method for setting limits on Poisson rates (here number of firms). An offset basically fixes the coefficient for the limiting factor to 1. If we use firms as a predictor instead, the model will estimate its coefficient still close to 1, but the estimates will be less efficient (and take much longer to compute).

- *b_agp* indicates whether a state's Attorney General publishes breach notification letters;
- *b_revised* captures whether a state has revised (or amended) its DBNL in a given year;
- *b_ytrend* captures the yearly trend of data breaches;
- *b_gdp_pcap* captures the yearly trend of GDP per capita.

We then run the model:[54]

$$\text{Breaches}_{s,y} \sim \alpha + \beta_{enacted}.\text{Enacted}_{s,y} + \beta_{agp}.\text{AGP}_{s,y} + \beta_{revised}.\text{Revised}_{s,y} + \beta_{y\_trend}.\text{year} + \beta_{gdp\_pcap}.\text{GDP\_pcap}_{s,y} + \alpha_{state\_pooled} + \alpha_{year\_pooled} + \log\left(\text{Firms}_{s,y}\right)$$

The Bayesian models converge well.[55] In Bayesian analysis, the *posterior distribution* of the parameters provides the same results as the coefficient estimates in non-Bayesian regression analysis. The resulting posterior distributions are shown in Figure 5. The mean of each parameter, and the 94% highest posterior density (HPD) interval, also known as the credible interval, are also marked. The HPD visualizes the parameter uncertainty.[56] The variance of the varying intercepts is considerably small which indicates successful pooling (i.e., the states differ, but not too much).[57]

Posterior *b_enacted* reveals an approximate 11% (±12%) increase[58] in reported breaches after a state enacts a DBNL. In other words, the model is not fully certain about the enactment effect; passing a DBNL may have no effect (–1%), or some increase (23%).

The uncertainty around enactment effect (i.e., the coefficient's spread) across states may be explained, foremost, by the fact that the provisions of the DBNL matter, a point we shall return to in the next paragraph. An alternative (or compounding) explanation might be that the effect of

54. *A model specification with only a single enactment effect as the predictor—basically the same as the DiD specification—yields similar results for that parameter as the full model explained here; the only difference is that the common and pooled intercepts have a larger spread since less of the variance is captured by the other predictors.*

55. See the Appendix for more convergence details; a posterior predictive plot is presented later in this section as well.

56. The HPD functions somewhat similar to the standard errors in non-Bayesian regression results. The 94% interval is chosen on purpose by the ArviZ package so as not to be confused with the 95% frequentist significance levels. If one selects a different credible interval (e.g., 80%), then the reported parameter range becomes smaller.

57. The unique year and state intercepts are presented in the Appendix.

58. The coefficients for a Poisson models need to be interpreted as *change in the odds* by $e^{\text{mean}\ (\pm\text{range})}$. Here this is $e^{0.10\ (\pm 0.11)}$, which translates to a breach rate *change* of 99% to 123%, or 11% (±12%) increase.

enacting DBNLs decreases over time, as more states enact them, states that enact a DBNL later will experience less of an impact, given that larger firms active across multiple states will have already adopted breach notification duties (i.e., have procedures and systems in place for it), a phenomenon known as *the "California Effect"*.[59],[60] We will return to this point in the identity theft model.

AGP indicates whether a state's Attorney General publishes breach notification letters,[61] which we know from prior research plays an important role in the public's knowledge of a breach having occurred.[62] The effect of *b_agp* is quite strong: the number of (known) breaches in a state increases on average by 28% (±12%) if it enacts its DBNL with the additional condition that the AG be notified of any breach and the AG subsequently publicizes breaches.

The effect of *b_revised* captures whether a state has revised (or amended) its DBNL in a given year. As previously noted, approximately two-thirds of states followed their DBNL enactment with a revision (or amendment).[63] We hypothesized that in the absence of enforcement,[64] revisions may help maintain a vigilant environment among actors involved in the notification process. On average there are 4% (±7%) more breaches reported in years that DBNLs are revised (excluding revisions that lead to the AG publicizing the notifications, as that is captured by *b_agp*). As the uncertainty around this parameter's estimates are high, much cannot be said about it.[65]

59. Due to its large market share, and preference for strict consumer and environmental regulations, California often leads with regulations which all firms active in California must implement. For larger firms, once they have implemented these changes in their operations, they might prefer to streamline their operations and de facto implement it in other jurisdictions as well.

60. Vogel; Vogel and Kagan.

61. This is set only once the AG starts publishing these letters. NH, MD, and VT were the first states to do so, publishing the letters since 2010; CA followed suit in 2012; In 2017, this number increased to nine states.

62. Bisogni, Asghari, and Van Eeten.

63. In the context of constitution and law, an amendment is a change or addition to an existing law. A revision, on the other hand, is through reexamination of the entire law. This is done to make changes or alterations in the law.

64. Concerning DBNLs the connection between legal sanctions and notification remains indirect and, in practice, weak (Schafer).

65. We also tried an alternative manner of operationalizing DBNL revisions, by creating *b_dbnl_version* variable, which we defined as the square root of the number of times this law has changed (0 = no DBNL, 1 = DBNL enacted, 1.41 = one revision/amendment, and so on. If we use this variable in place of all existing law variables *(b_enacted, b_agp, b_revised),* we find that every unit increase yields a 10% (±8%) increase in breaches. If, however, we add the variable to

From *b_ytrend,* we see that on average the number of (known) data breaches increases every year by 14% (±7%), even after controlling for changes to the regulatory environment (DBNLs). This increase has two potential sources. On the one hand, the growing number of digital threats that are not countered by proper measures produces more breaches. On the other hand, as organizations become better equipped to detect breaches, more reports are produced. Unpacking these two sources requires more data (and maybe of interest for future research).

Finally, GDP per capita is a common state-level control that is a proxy for wealth and infrastructure, among other variables.[66] The number of data breaches increases by approximately 5% (±7%) for every $9,600 increase in GDP per capita.[67] This effect may simply reflect that companies in richer areas are more attractive targets for hackers.

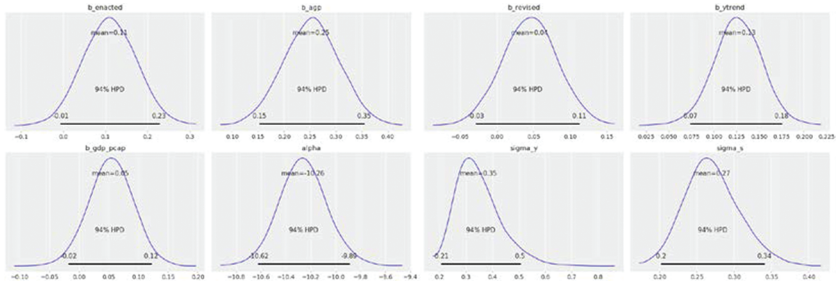

FIGURE 5 Posterior Distributions for the Multilevel Data Breach Model. (The 94% highest posterior densities are marked; The betas are the predictors; other parameters are the common intercept, and the variances of the two varying/random intercepts; All betas need to be interpreted as e^{mean} due to the Poisson log link function). See Appendix for unique intercepts.

the model next to the *b_agp* variable, its effect disappears (while other parameter coefficients stay approximately the same). In other words, a key success factor for a DBNL is that the regulator is placed in the notification loop, and it publicizes the notifications (whether as part of the original law or added in a revision).

66. We exclude some other common controls that are either correlated with GDP_pcap, as they can lead to multicollinearity; or are unrelated to firm behavior, such as crime (which is about the population of a state, while breaches happen over state lines), since they can lead to inefficient estimates. (Multicollinearity and inefficient estimates can mask the actual effects of interest.)

67. The GDP per capita variable has been centered and standardized. Thus, the parameter's value is the increase caused by one standard deviation change in GDP per capita (from the mean GDP per capita), which is approximately $9,600.

The Impact of DBNL on Identity Theft

We followed a similar reasoning process to model the impacts of DBNLs on identity theft–using, once again, a multilevel Bayesian model with pooled varying intercepts.

A key computational difference between identity theft and data breaches is the choice of distribution for the dependent variable (see Figure 3 right). Empirically, a *log-normal* distribution offers the best fit, not a gamma-Poisson distribution. Conceptually, a log-normal distribution points to a *multiplicative* underlying process instead of an additive one.[68] This process can be understood by the fact that a fraudster will (typically) target multiple victims within a single fraud campaign, which explains identity thefts strong correlation with the state population. (Data breaches, on the other hand, are rarer and more independent).

The key lines of the Stan model are as follows; as in the previous model, we used weakly informed priors. (The additional *sigma* parameter captures the overall variance for the log normal distribution):

```
mu = intercept + a_years + a_states + X*betas + offset;
Y ~ lognormal(mu, sigma);
```

This time, the population size is used as the *offset*. The predictors we use are as follows:

- *b_enacted* indicates whether a state in a given year enacts a DBNL;
- *b_agp* indicates whether a state's Attorney General publishes breach notification letters;
- *b_revised* captures whether a state has revised (or amended) its DBNL in a given year;
- *b_records_pcap* is the number of records breached per capita (estimated for state/year);
- *b_pcrime_pcap* captures the yearly trend of property crime per capita;
- *b_gdp_pcap* captures the yearly trend of GDP per capita.

The model includes two new predictors, *breached records per capita,* and *property crime per capita,*[69] both which are both expected to increase identity theft. We estimate the number of breached data records per state by summing

68. Limpert, Stahel, and Abbt.

69. *Property crime results from the sum of burglary, larceny, and motor vehicle theft. Data source: Summary (SRS) Data with Estimates at https://crime-data-explorer.fr.cloud.gov/downloads-and-docs. Another possible control is internet penetration (e.g., Internet users per capita). This variable is highly correlated with GDP per capita; and substituting it in the model does not affect any of the reported predictors.*

up the total data records breached in all reported breaches across the United States in a year (an average of 76 million records) and dividing this total by each state's population. (The rationale is that the sum of records breached per year is strongly driven by the so-called "mega" breaches—breaches that impact millions of customers. These customers are likely spread over all US states). Property crime we include since it can be a cause of identity theft; and also, the socioeconomic factors that lead to a rise of property crime in a region may also lead to increased identity theft. We exclude the yearly trend variable in this model, since identity theft does not show a strong trend in the logarithmic form.[70] Thus we run the following model:

$$\begin{aligned}\text{IdentityTheft}_{s,y} \sim{}& \alpha + \beta_{enacted}.\text{Enacted}_{s,y} + \beta_{agp}.\text{AGP}_{s,y} + \beta_{revised}.\text{Revised}_{s,y} \\ &+ \beta_{records_{pcap}}.\text{Records}_{pcaps,y} + \beta_{crime_{pcap}}.\text{PCrime}_{pcaps,y} \\ &+ \beta_{gdp_{pcap}}.\text{GDP}_{pcaps,y} + \alpha_{state\_pooled} + \alpha_{year\_pooled} + \log\left(\text{Population}_{s,y}\right)\end{aligned}$$

Figure 6 presents the posterior distributions of this model, which again converge well.[71]

Adopting a DBNL results in a 2.5% (±3%) decrease in identity theft. While the direction of this effect is negative as expected, the decrease is quite small, and the credible interval crosses zero, making it also uncertain. *What's interesting is that the effect size is less than half the 6.1% that Romanosky et al.*[72] *reported for the period 2002 to 2009.* This contrast may be evidence of the California effect—larger firms active across multiple states may have already adopted breach notification duties and practices in all states by choice, thus decreasing the effects of DBNL enactment by later states.[73] Another explanation for this small effect size is that data breaches are only a portion of identity theft, for example, Javelin Report[74] estimated that data breaches are the source of 11% of identity theft. In other words, the decrease in breach-related identity theft is several-fold larger: if all incidents of identity theft were driven by data breaches the magnitude of the

70. Additionally, breach records have a strong yearly component to them, and having both predictors would mask the other's effect. If we use yearly trend (in place of records), we find a slight annual growth of 4% (±2%).

71. The posterior predictive plot for this model can be found at the end of the previous section.

72. Romanosky, Telang, and Acquisti.

73. To make this more concrete: in 2005 (the start of our dataset), eight states had passed DBNLs, and these states held approximately a third of all US firms; In other words, a third of US firms were already subject to some DBNL in that year. By 2008, this had increased to 40 states and 84% of all US firms.

74. Javelin Strategy & Research, *2009 Identity Fraud Survey Report.*

identity theft decrease would be about 22.7%, applying the 2.5% decrease to an 11% subset. Nonetheless, it also highlights that *being notified of a breach does not guarantee one can stop the resulting identity theft in time*.[75]

The credible interval for *b_revised* and *b_agp* spreads widely around zero, indicating no clear effect. The fact that the AG publicizing breaches does not further reduce identity theft is, paradoxically, a positive finding: it suggests that firms notify breach-affected customers, as required by law, irrespective of the publicity.[76]

The posterior for *b_records_pcap* suggests that a one percentage increase in the number of breached records equates to a 1% (±1%) increase in identity theft.[77] Finally, *b_pcrime_pcap* is a strong predictor of identity theft: a one percentage point increase in property crime per capita equates to an 11% (±5%) increase in the identity theft rate. (As the mean property crime per capita is 2.8%, a 1% increase is substantial).

Counterfactual Plots for Identity Theft

We can use "counterfactual plots" to visualize and better understand how the three key predictors (*b_dbnl*, *b_records_pcap*, *b_pcrime_pcap*) impact identity theft. This is shown in Figure 7: the model's predicted

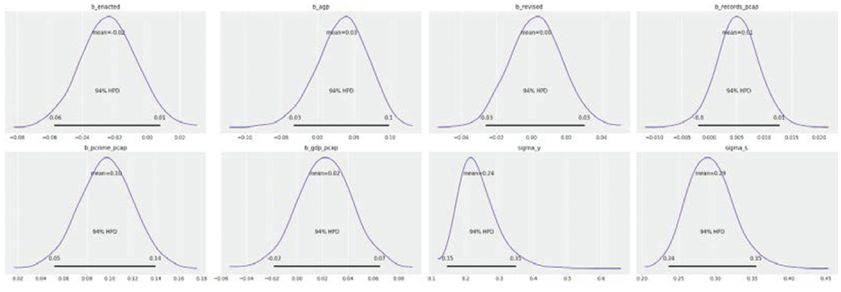

FIGURE 6 Posterior Distributions for the Multilevel Identity Theft Model. (The 94% highest posterior densities are marked. Betas need to be interpreted as e^{mean}).

75. With more precise data on identity theft causes, which currently are not available, this idea can be further explored (future work).

76. Note that using the *b_dbnl_version* variable (explained in a footnote of the data breach model) in place of the three separate law variables yields a similar effect as *b_enacted* alone.

77. The credible interval for this predictor also touches zero, reflecting uncertainty in the effect. This uncertainty is in part because we use a rough estimate for the number of breached records per state (as the actual number of data subjects affected in each state aren't reported). Examining the counterfactual plot for this parameter makes this point evident (e.g., the alignment of the dots).

outcome (identity theft) is shown for imaginary states with varying degrees of breached records and property crime, with and without a DBNL. The plots make it clear that any benefits that come from enacting a DBNL (in terms of a decrease in identity theft) are by far outweighed by a significant increase in the number of breached records (e.g., resulting from a mega breach). In other words, the drop from the black line to the red line is small, compared to the overall upward slope that shows the effect of additional breached records on identity theft.

Posterior Predictive Checks

It is customary in Bayesian statistics to check, as an additional robustness measure, whether the posterior predictions of a model mimic the observed data with reasonable accuracy. The plots of Figure 8 (respectively for data breaches—left and identity theft—right) depicts 100 *simulations*

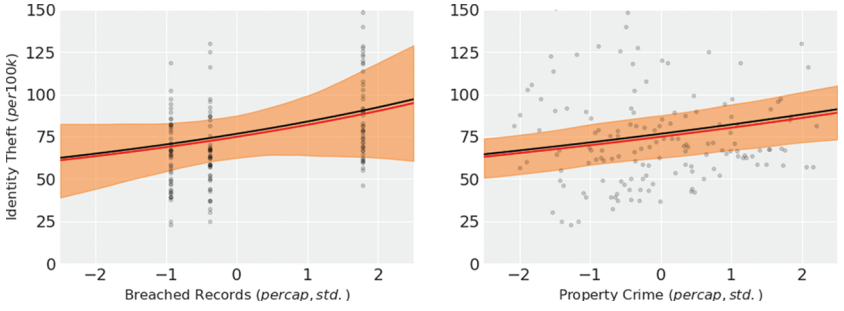

FIGURE 7 Counterfactual Plots for States with Different Breached Records (left) and Property Crime (right). The two solid lines are whether a DBNL is enacted or not (with enactment being the lower line). The shaded area is the 94% HPD; The dots are observed data (plotted for 2005, 2011, and 2017); Counterfactual variables are standardized.

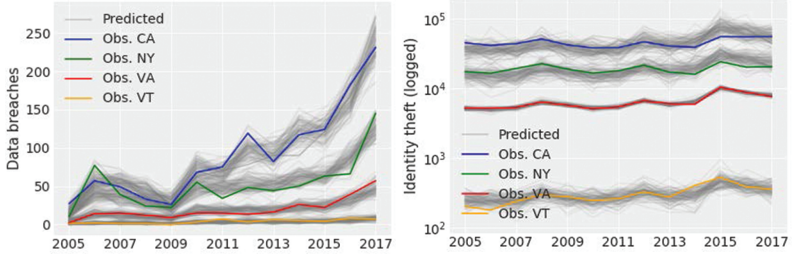

FIGURE 8 Posterior Predictive Plots for the Multilevel Data Breach (left) and Identity Theft (right) Models. They show observed versus predicted over time for select states. The grey shades indicate 100 simulations, and the solid colors indicate observed values.

of breaches over time for California, New York, Virginia, and New Hampshire versus the actual trends for these states. The simulations are in light grey, and the observed trends in solid colors. As visible in the figure, the predictions and observations are reasonably well matched.[78]

Discussion: How the US Situation Can Inform the European Union

We analyzed the correlation between identity theft and data breaches, and found that the correlation between them is primarily driven by the size of a state. That is, the correlation decreases (but does not disappear) when we control for population or the number of firms.

We next used multilevel Bayesian modeling to examine the effects of DBNLs on data breaches and identity theft over 15 years in the United States. We observed an increase in reported (and known) breaches rates after DBNL enactment, and a considerable increase if the Attorney General publicizes the notifications. DBNL enactment slightly reduces identity theft rates as well, and if we consider that data breaches are not the only source of identity theft, the decrease is considerable. These findings are very relevant for the European context, particularly in the present period of implementation of the GDPR.

Starting with identity theft, it is important to underline that currently in Europe there is no common way Member States identify identity thefts internally and there is no procedure to report them centrally (at European level). There are several ways identity thefts are defined, recorded, and subsequently reported, generating important differences in number from one country to the other. The situation for the three most peopled European countries is as follows: in Germany, identity theft falls under Internet crime, and includes phishing, fraud related to services and goods done via Internet, and malicious software. According to the German Institute for Economic Research, in 2015 the identity theft rate ranged between 1,265 and 4,135 per 100,000 inhabitants (based on analyzed regions). In the UK, the number reported by the CIFAS[79] (and recorded in the National Fraud Database is) 169,592 for 2015, and 172,919 for 2016. This statistic

78. These four states were chosen simply because they have very different baseline levels. A more classic posterior predictive plot of *y* to *y_hat,* which includes all the states and years, can be found in the Appendix.

79. Credit Industry Fraud Avoidance System https://www.cifas.org.uk.

only includes the identity theft reported by the 277 CIFAS organizations' members. In France, a survey conducted by Fellowes/ObSoCo[80] in 2015 found that 200,000 identity thefts take place yearly, in line with the figure reported by CREDOC[81] in 2009 (210,000). This overview shows the need for a centralized public repository of such information. A number of European Projects have started initiatives in this direction; For instance, EKSISTENZ[82] promoted the establishment of a European Observatory on Identity Theft[83]. This Observatory brings together those researchers across the EU to create a focal point and repository of knowledge (for anti-identity theft projects). The Observatory and its website inform citizens on methods, procedures, and possibilities to recover his/her identity after theft, serve as a policy adviser to EU Member States, and advance a common view for European identity protection. At present, 22 organizations participate in the Observatory, that is, Universities, research institutes, relevant Member State agencies, police forces, and consultancy companies.

One opportunity to investigate (with some major approximations) identity theft differences within Europe comes from the Commission's Eurobarometer reports, in particular two special reports (2017, 2018).[84] The reports asked EU citizens about whether they had been victims of identity theft,[85] and if they were, if they would contact the police.[86] We multiply these two numbers to have a figure that is comparable to the US statistic that is the number of identity theft actually reported (collected by each state). The results, presented in Table 4, show a large difference between the United States, where 0.11% of the population actually reported identity theft, and the EU where between 0.58% (Greece) and 4.11% (Belgium) of those surveyed said that they suffered an identity theft (and would have reported it to the police). The EU numbers are, in our opinion, should

80. https://www.fellowes.com.

81. Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie https://www.credoc.fr.

82. https://cordis.europa.eu/project/rcn/188570/reporting/en.

83. http://www.idtheftobservatory.eu.

84. These are special reports 464a (2017) and 480 (2018) "Europeans' attitudes towards cyber security." The Standard Eurobarometer was established in 1974. Each survey consists of approximately 1,000 face-to-face interviews per country. Reports are published twice yearly. Special Eurobarometer reports are based on in-depth thematic studies carried out for various services of the European Commission or other EU institutions and integrated in the Standard Eurobarometer's polling waves.

85. We divided that value by three as the question QD10 is structured as follows "In the last three years, how often have you personally experienced or been victim of identity theft" (p. 27).

86. QB13.1 "If you experienced or were a victim of identity thefts, who would you contact?" (p. 92).

be seen as an upper limit of actual identity theft, since they come from a survey rather than actual reported cases.[87]

In the same table we also compare data breaches statistics for Europe (for 2019) based on a DLA Piper report that has collected available aggregate statistics across the EU.[88] To make the numbers comparable with the United States, we divide the total breaches by the number of firms in each country.[89] This difference is stark and revealing: while the number of breaches per 100k firms in the United States is 20.5,[90] in most EU countries this metric is over 100, and in the Denmark, Ireland, and the Netherlands there have been more than 5,000 breaches reported per 100,000 firms. This large difference reflects a difference in the notification regime.

There are significant differences between the breach notification regimes implemented by the GDPR and the US DBNLs. Firstly, the GDPR is regulated at a central European level. In the United States, there is currently a patchwork of DBNLs in place (48 out of 50 states enacted their DBNL before the GDPR). This creates challenges for organization's located in one state following one DBNL, and their "breached" customers residing in a state following a different one. The sanction regime in the United States (the administrative penalties) are two orders of magnitude lower than in the EU,[91] although in the United States there is the possibility to activate privacy class actions. Finally, the US approach focuses more on the "name and shame," or "sunlight as disinfectant"[92] rationale, while the right to protection of personal data (and the right to know) have been the reasons to adopt the GDPR. This means that companies in Europe do not fear the reputational effect related to notification to supervising authorities as dictated by GDPR. As the national authorities do not (yet)

87. Part of the difference might be that survey participants inflate the numbers because of "telescoping" effects (where incidents occurring outside the reference period are inflated when reported to the interviewer). Also, if the cases would be reported, the police might not deem them all to be legally significant to investigate or even count as an identity theft.

88. DLA Piper GDPR Data Breach Survey: January 2020.

89. As per United States we excluded firms with no employee. Source: EUROSTAT business demography by size class (from 2004 onwards, NACE Rev. 2) [bd_9bd_sz_cl_r2].

90. The US statistics is for 2018, the latest ITRC number available at the moment of this publication.

91. see Nieuwesteeg and Faure.

92. The reputation damage resulting from a reported breach would activate "the sunlight as disinfectant" principle, leading companies to invest more in cybersecurity, and disinfect organizations of shoddy security practices (Ranger).

TABLE 4 Breaches per 100k Capita and Firms, Identity Theft Rate

Country	Estimated ID Theft %	Breaches p100k Persons	Breaches p100k Firm	Population	Employer Firms
Netherlands	0.90%	147.20	10,544.49	17,081,507	238,456
Ireland	2.16%	132.52	5,712.05	4,784,383	110,998
Denmark	0.93%	115.43	5,544.44	5,748,769	119,684
Finland	1.23%	71.11	2,881.20	5,503,297	135,825
Germany	1.42%	31.12	1,722.02	82,521,653	1,491,314
Sweden	2.17%	48.14	1,684.10	9,995,153	285,712
Luxembourg	2.64%	56.97	1,671.73	590,667	20,129
Slovenia	1.45%	52.55	1,600.44	2,065,895	67,833
Malta	1.40%	31.00	1,073.03	460,297	13,298
Poland	1.87%	13.74	694.13	37,972,964	751,657
Austria	2.27%	12.10	544.61	8,772,865	194,913
UK	2.61%	17.79	524.16	65,808,573	2,233,560
Belgium	4.11%	7.88	469.14	11,351,727	190,672
Estonia	1.37%	9.74	235.89	1,315,634	54,322
Czech Republic	1.52%	4.03	188.39	10,578,820	226,304
France	2.64%	3.20	188.37	66,989,083	1,138,011
Latvia	1.56%	6.13	169.17	1,950,116	70,662
Lithuania	0.89%	4.18	158.56	2,847,904	75,075
Hungary	3.01%	4.87	129.90	9,797,561	367,328
Cyprus	2.61%	4.80	121.21	854,802	33,852
Romania	2.88%	1.90	100.21	19,644,350	372,471
Italy	1.94%	2.05	90.46	60,589,445	1,373,008
Spain	1.84%	2.08	74.12	46,527,039	1,305,705
Greece	0.58%	1.50	39.25	10,768,193	411,555
USA	**0.11%**	**0.38**	**20.48**	**325,025,206**	**6,073,017**

make the notifications public, companies have an incentive to "over-inform" the authority as a matter of caution, knowing that there will be little reputational damage for doing so. To reinforce this point, the number of breaches reported in the Netherlands alone in 2018 was 20,881 (Dutch DPA, 2019), which is more than 16 times the data breaches recorded in the United States in the same year (1,244 according to the ITRC, 2018).

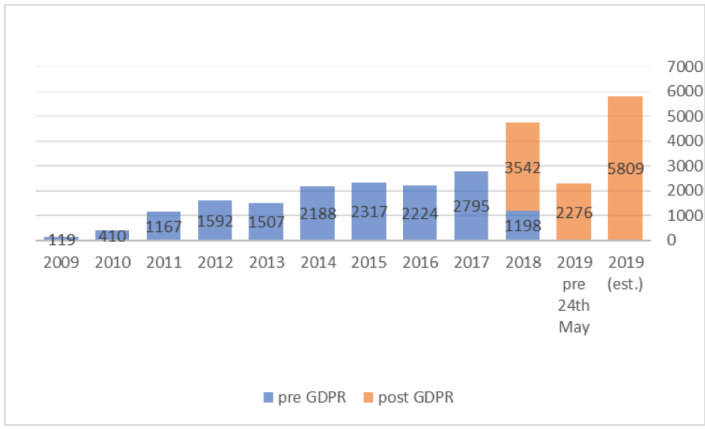

FIGURE 9 Data Breaches by Year for Ireland (2009–2019). Source: Irish Data Protection Officer Annual Reports.[93] The numbers for 2019 are extrapolated from the available period for this year (January to May).

Longitudinal data breach data is even harder to find for Europe, even at aggregated level. The Irish Data Protection Commission is among the few regulators that have released this information, from 2009 (prior to the implementation of the Irish Personal Data Security Breach Code of Practice) to 2019. We plot this data in Figure 9, with a steep growth after the GDPR. The implementation of the GDPR in Ireland lead to an approximate 102% increase in the number of reported breaches (i.e., comparing a year before and after the GDPR). This effect can be compared with the effect of enacting a DBNL (+11%) and also informing the AG (+28%) in the United States. The GDPR effect is much stronger and relates[93] back to the over-reporting point in the previous paragraph.[94]

93. https://www.dataprotection.ie/sites/default/files/uploads/2019-10/Info%20Note_Data%20Breach%20Trends%202018-19_Oct19.pdf

https://www.dataprotection.ie/sites/default/files/uploads/2019-03/DPC%20Annual%20Report%2025%20May%20-%2031%20December%202018.pdf.

https://www.dataprotection.ie/sites/default/files/uploads/2018-11/DPC%20annual%20Report%202018_0.pdf.

https://www.dataprotection.ie/sites/default/files/uploads/2018-11/Annual%20Report%202017.pdf.

https://www.dataprotection.ie/sites/default/files/uploads/2018-11/Annual%20Report%202016.pdf.

https://www.dataprotection.ie/sites/default/files/uploads/2018-11/Annual%20Report%202015.pdf.

https://www.dataprotection.ie/sites/default/files/uploads/2018-11/Annual%20Report%202014.pdf.

If the desired regulatory approach is to enforce the GDPR's provisions at the central European level, two core elements are missing: a common public collection of identity thefts, and a common public collection of data breaches. Currently, the source of identity theft data varies from country to country (e.g., police forces, associations), and it even lacks a common framework to define them. It is not sufficient to rely only on surveys for aggregated data on identity theft. Equally, most data breaches will also not be revealed to the public, since the GDPR Art 59 states:

> *Each supervisory authority shall draw up an annual report on its activities, which may include a list of types of infringement notified and types of measures taken in accordance with Article 58(2). Those reports shall be transmitted to the national parliament, the government and other authorities as designated by Member State law. They shall be made available to the public, to the Commission and to the Board.*

The inclusion of the list of breaches is therefore an option and not a duty.

To conclude, whether the focus is information disclosure or regulation, a central question about data breach notification policy in the 21st century is whether we have appropriately designed institutions and processes to foster and monitor the desired outcomes. There is little question that the current mix of public policies does not always live up to these expectations. Information disclosure policy, such as the GDPR or the US DBNL, have played a fruitful role in the mix of contemporary policy and regulations. However, much can be done to improve program effectiveness, particularly in ensuring that information collected is easily accessible, understandable, and meaningful in terms of real public and private risks faced across the countries.

---

https://www.dataprotection.ie/sites/default/files/uploads/2018-12/Annual%20Report%202013.pdf.

https://www.dataprotection.ie/sites/default/files/uploads/2018-12/Annual_Report_2012.pdf.

94 With the obvious caveat that the effect is only for one country, Ireland. An additional reason for the difference maybe that in the many US states there are minimum thresholds (in terms of affected records or possible harms) before there is a duty to notify of a breach

Bibliography

Angrist, J. D., and J. S. Pischke. *Mastering 'Metrics: The Path from Cause to Effect* (Princeton, NJ: Princeton University Press, 2014).

Bisogni, F. "Proving Limits of State Data Breach Notification Laws: Is a Federal Law the Most Adequate Solution?" *Journal of Information Policy* 6 (2016): 154–205, Penn State University Press.

Bisogni, F., H. Asghari, and M. Van Eeten. "Estimating the Size of the Iceberg from its Tip. An Investigation into Unreported Data Breach Notifications." WEIS 2017—16th Annual Workshop on the Economics of Information Security, La Jolla, San Diego, June 2017.

Bug, M., et al. *WISIND-Datensätze: Kriminalitätsbefragung* (Berechnungen des DIW Berlin, 2015).

Choi, S., and M. E. Johnson. "Do Hospital Data Breaches Reduce Patient Care Quality?" WEIS 2017—16th Annual Workshop on the Economics of Information Security, La Jolla, San Diego, June 2017.

Di Ciccio, F. "Comparison of Identity Theft in Different Countries." 2014. https://courses.cs.ut.ee/MTAT.07.022/2014_fall/uploads/Main/francesco-report-f14.pdf. Accessed November 29, 2019.

DLA Piper. *GDPR Data Breach Survey* (January 2020). https://www.dlapiper.com/it/italy/insights/publications/2020/01/gdpr-data-breach-survey-2020/ Accessed February 28, 2020.

Draper, A. "Identity Theft: Plugging the Massive Data Leaks with a Stricter Nationwide Breach-Notification Law." *Journal Marshall & Law Review* 40 (2006): 681–703.

Dutch DPA (2019): "Overzicht meldingen datalekken eerste kwartaal 2017", Available online at: https://www.autoriteitpersoonsgegevens.nl/nl/nieuws/ap-ontvangt-bijna-21000-datalekken-2018. Accessed November 29, 2019.

Edwards, B., S. Hofmeyr, and F. Forrest. (2015) "Hype and Heavy Tails: A Closer Look at Data Breaches." WEIS 2015: 14th Workshop on the Economics of Information Security, June 2015.

Garrison, C. P., and M. Ncube. "A Longitudinal Analysis of Data Breaches." *Information Management & Computer Security* 19, no. 4 (2011): 216–230. doi:10.1108/09685221111173049

Gavison, R. "Privacy and the Limits of Law." *The Yale Law Journal* 89 (1980): 421–423.

Gordon, G. R., et al. *Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement* (Utica, NY: Center for Identity Management and Information Protection (CIMIP), Utica College, 2007).

Irish Data Protection Commission. *Information Note: Data Breach Trends from the First Year of the GDPR* (October 2019). https://www.dataprotection.ie/sites/default/files/uploads/2019-10/Info%20Note_Data%20Breach%20Trends%202018-19_Oct19.pdf. Accessed February 28, 2020.

Identity Theft Resource Center (ITRC). (2018): "Data Breaches", https://www.idtheftcenter.org/data-breaches/. Accessed November 29, 2019.

Javelin Strategy & Research. *2009 Identity Fraud Survey Report* (February 2009). Pleasanton, CA: Javelin Strategy & Research.

Javelin Strategy & Research. *2018 Identity Fraud Survey Report* (February 2018). Livonia, MI: Escalent.

Kang, J. "Information Privacy in Cyberspace Transactions." *Stanford Law Review* 50 (1998): 1193–1203.

Kruschke, J. K. *Doing Bayesian Data Analysis: A Tutorial with R and BUGS* (New York: Academic Press, 2015).

Kwon, J., and E. Johnson. "The Market Effect of Healthcare Security: Do Patients Care about Data Breaches?" WEIS 2015: 14th Workshop on the Economics of Information Security, Delft, the Netherlands, June 2015.

Limpert, E., W. A. Stahel, and M. Abbt. "Log-normal Distributions across the Sciences: Keys and Clues: On the Charms of Statistics, and How Mechanical Models Resembling Gambling Machines Offer a Link to a Handy Way to Characterize Log-normal Distributions, Which Can Provide Deeper Insight into Variability and Probability—Normal or Log-normal: That is the Question." *BioScience* 51, no. 5 (May 2001): 341–352.

McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* Boca Raton, FL: CRC Press, 2016.

Moor, J. H. "Towards a Theory of Privacy in the Information Age." *Computers and Society* 27 (2010): 27–31. (Outlining the restricted access/limited control approach to privacy).

Nieuwesteeg, B., and M. Faure. "An Analysis of the Effectiveness of the EU Data Breach Notification Obligation." Computer Law & Security Review 34, no. 6 (2018): 1232–1246.

Ranger, Steve. "Data Breach Laws Make Companies Serious about Security." September 3, 2007. Silicon.com. https://www.law.berkeley.edu/article/data-breach-laws-make-companies-serious-about-security/ Accessed November, 29, 2019.

Roberds, William, and Stacey L. Schreft. *Data Breaches and Identity Theft, Working Paper, Federal Reserve Bank of Atlanta, No. 2008-22* (2008). Available at SSRN: https://ssrn.com/abstract=1296131 or http://dx.doi.org/10.2139/ssrn.1296131 Accessed November 29, 2019.

Romanosky, S., R. Telang, and A. Acquisti. "Do Data Breach Disclosure Laws Reduce Identity Theft?" *Journal of Policy Analysis and Management* 30, no. 2 (2011): 256–286.

Schafer, B. "Speaking Truth to/as Victims—A Jurisprudential Analysis of Data Breach Notification Laws." In *The Responsibilities of Online Service Providers. Law, Governance and Technology Series*, vol. 31, edited by M. Taddeo, and L. Floridi. Cham, Switzerland: Springer, 2017.

Simitian, J. "UCB Security Breach Notification Symposium March 6, 2009 'How a bill becomes a law, really.'" *Berkeley Technology Law Journal* 24 (2009): 1009–1018.

Skinner, T. H. "California's Database Breach Notification Security Act: The First State Breach Notification Law is not yet a Suitable Template for National Identity Theft Legislation." *Richmond Journal Law & Technology* 10 (2003): 1–40.

Towle, H. K. "Identity Theft: Myths, Methods, and New Law." *Rutgers Computer & Technology Law Journal* 30 (2003): 237–326.

US Government Accountability Office (GAO). *Report to Congressional Requesters "PERSONAL INFORMATION Data Breaches are Frequent, but Evidence of Resulting Identity Theft Is Limited; However, the Full Extent Is Unknown"* Washington, DC: GAO, July 2007.

Vogel, D. *Trading Up: Consumer and Environmental Regulation in a Global Economy.* Cambridge, MA: Harvard University Press, 1995.

Vogel, D., and R. Kagan. "Introduction." In *Dynamics of Regulatory Change: How Globalization Affects National Regulatory Policies*, edited by D. Vogel, and R. A. Kagan. Oakland, CA: University of California Press, 2004.

Winn, J. K. "Are "Better" Security Breach Notification Laws Possible?" *Berkley Technology Law Journal* 24 (2009): 1133–1165.

Appendix I. Identity Theft versus Breached Records

In the Figure A1 we compare identity thefts with the number of available *records* breached in data breaches.

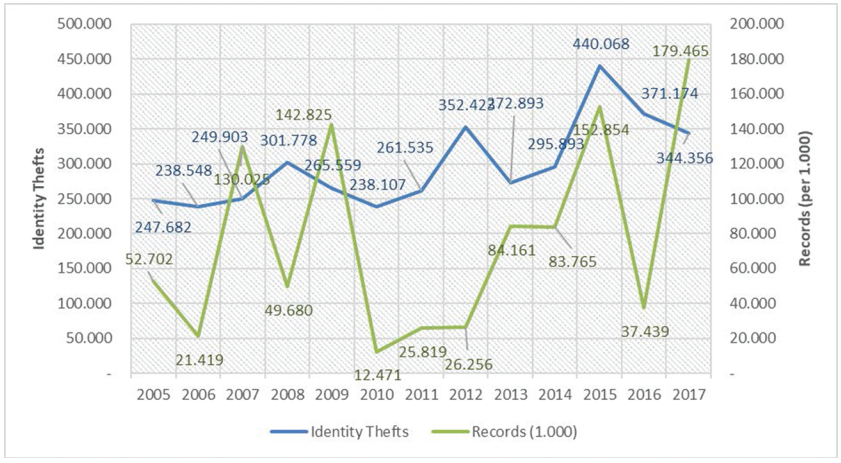

FIGURE A1 Identity Theft and Breached Records.

Appendix II. Difference in Differences Summary

Generalized Linear Model Regression Results						
Dep. Variable:		**Breaches**		No. Observations		650
Model:		GLM		Df Residuals:		587
Model Family:		Negative Binomial		Df Model:		62
Link Function:		log		Scale:		1.0000
Method:		IRLS		Log-Likelihood:		−1929.6
Date:		Tue, 10 Sep 2019		Deviance:		159.72
Time:		13:05:18		Pearson χ^2:		126
No. Iterations:		9				
Covariance Type:		Nonrobust				
	Coef	Std err	*z*	*P*>|*z*|	[0.025	0.975]
Intercept	−1.0892	0.408	−2.672	0.008	−1.888	−0.290
st[T.AL]	1.3936	0.496	2.808	0.005	0.421	2.367
st[T.AR]	0.7319	0.489	1.497	0.134	−0.226	1.690
st[T.AZ]	1.6743	0.470	3.562	0.000	0.753	2.596
st[T.CA]	3.9716	0.462	8.601	0.000	3.067	4.877
st[T.CO]	2.2532	0.466	4.835	0.000	1.340	3.167
st[T.CT]	1.9302	0.469	4.120	0.000	1.012	2.849
st[T.DE]	0.2864	0.503	0.569	0.569	−0.700	1.272
st[T.FL]	2.9928	0.464	6.453	0.000	2.084	3.902
st[T.GA]	2.6037	0.465	5.595	0.000	1.692	3.516

Generalized Linear Model Regression Results						
st[T.HI]	0.3784	0.498	0.760	0.447	−0.598	1.354
st[T.IA]	1.2670	0.475	2.667	0.008	0.336	2.198
st[T.ID]	0.2149	0.505	0.425	0.671	−0.775	1.205
st[T.IL]	2.7511	0.463	5.939	0.000	1.843	3.659
st[T.IN]	2.1382	0.467	4.581	0.000	1.223	3.053
st[T.KS]	0.8172	0.485	1.685	0.092	−0.133	1.768
st[T.KY]	1.6015	0.477	3.354	0.001	0.666	2.537
st[T.LA]	1.0198	0.481	2.118	0.034	0.076	1.963
st[T.MA]	2.5885	0.462	5.600	0.000	1.683	3.495
st[T.MD]	2.1207	0.465	4.559	0.000	1.209	3.032
st[T.ME]	0.6912	0.489	1.413	0.158	−0.268	1.650
st[T.MI]	1.9587	0.467	4.193	0.000	1.043	2.874
st[T.MN]	1.9144	0.469	4.085	0.000	0.996	2.833
st[T.MO]	1.5857	0.470	3.371	0.001	0.664	2.508
st[T.MS]	0.2054	0.505	0.407	0.684	−0.784	1.195
st[T.MT]	0.7114	0.489	1.456	0.145	−0.246	1.669
st[T.NC]	2.4072	0.465	5.177	0.000	1.496	3.319
st[T.ND]	−0.6612	0.558	−1.185	0.236	−1.755	0.433
st[T.NE]	0.8085	0.486	1.663	0.096	−0.144	1.761
st[T.NH]	1.1493	0.478	2.405	0.016	0.213	2.086
st[T.NJ]	2.1932	0.466	4.702	0.000	1.279	3.107
st[T.NM]	0.8111	0.502	1.614	0.106	−0.174	1.796
st[T.NV]	1.1871	0.478	2.482	0.013	0.250	2.124
st[T.NY]	3.4484	0.461	7.479	0.000	2.545	4.352
st[T.OH]	2.6571	0.464	5.730	0.000	1.748	3.566
st[T.OK]	1.1637	0.477	2.440	0.015	0.229	2.098
st[T.OR]	1.8130	0.468	3.874	0.000	0.896	2.730
st[T.PA]	2.5388	0.464	5.468	0.000	1.629	3.449
st[T.RI]	0.9043	0.484	1.869	0.062	−0.044	1.853
st[T.SC]	1.2044	0.476	2.530	0.011	0.271	2.138
st[T.SD]	−0.3078	0.554	−0.556	0.578	−1.393	0.777
st[T.TN]	2.0570	0.469	4.389	0.000	1.138	2.976
st[T.TX]	3.1904	0.460	6.938	0.000	2.289	4.092
st[T.UT]	1.2570	0.476	2.641	0.008	0.324	2.190
st[T.VA]	2.4042	0.463	5.190	0.000	1.496	3.312
st[T.VT]	0.7576	0.486	1.558	0.119	−0.196	1.711

Generalized Linear Model Regression Results						
st[T.WA]	2.1432	0.468	4.579	0.000	1.226	3.061
st[T.WI]	1.5795	0.472	3.344	0.001	0.654	2.505
st[T.WV]	0.0718	0.510	0.141	0.888	−0.928	1.071
st[T.WY]	−0.5431	0.548	−0.992	0.321	−1.616	0.530
ys[T.2006]	1.2275	0.257	4.785	0.000	0.725	1.730
ys[T.2007]	1.3148	0.266	4.938	0.000	0.793	1.837
ys[T.2008]	1.0217	0.280	3.649	0.000	0.473	1.570
ys[T.2009]	0.7473	0.293	2.552	0.011	0.173	1.321
ys[T.2010]	1.5495	0.286	5.416	0.000	0.989	2.110
ys[T.2011]	1.5587	0.288	5.417	0.000	0.995	2.123
ys[T.2012]	1.5507	0.288	5.388	0.000	0.987	2.115
ys[T.2013]	1.4918	0.288	5.177	0.000	0.927	2.057
ys[T.2014]	1.8052	0.289	6.249	0.000	1.239	2.371
ys[T.2015]	1.7391	0.289	6.014	0.000	1.172	2.306
ys[T.2016]	2.1162	0.288	7.358	0.000	1.552	2.680
ys[T.2017]	2.5281	0.289	8.759	0.000	1.962	3.094
hasdbnl	**0.0173**	**0.203**	0.085	0.932	−0.381	0.415

Appendix III. Stan Code and Convergence Details

Multilevel Poisson Model with Varying Intercepts Per State/Year and Offset (for Data Breaches):

```
  data {
   int<lower=1> nY;
   int<lower=1> nS;
   int<lower=1>nP; // number of (individual) predictors
   matrix[nY*nS, nP] X; // predictors (e.g., dbnl enact-
ment, revisions)
   vector[nY*nS] offset;// a rate, has the coef set to 1
   int yy[nY*nS];
   int ss[nY*nS];
   int<lower=0> Y[nY*nS];// outcome/observations
  }
  transformed data {
   int N = nY * nS;
  }
  parameters {
   real alpha; // overall intercept
   vector[nY] a_yy;// unique intercept (poisson level) per
year
   vector[nS] a_ss;// unique intercept per state
   real<lower=0> sigma_y; // pool unique YY intercepts
```

```
   real<lower=0> sigma_s; // pool unique SS intercepts
   vector[nP] beta; // beta for all predictors
  }
  transformed parameters {}
  model {
   vector[N] mu;
   // priors
   target += normal_lpdf(alpha | 0, 10);
   target += normal_lpdf(beta | 0, 1);
   target += normal_lpdf(a_yy | 0, sigma_y);
   target += normal_lpdf(a_ss | 0, sigma_s);
   target += cauchy_lpdf(sigma_y | 0 , 1);
   target += cauchy_lpdf(sigma_s | 0 , 1);
   // linear model
   for ( i in 1:N )
    mu[i] = alpha + a_yy[yy[i]] + a_ss[ss[i]] + X[i] * beta
+ offset[i];
   target += poisson_log_lpmf(Y | mu);
  }
  generated quantities {
   vector[N] yhat;
   vector[N] log_lik;
    for ( i in 1:N ) {
    real mu;
    mu = alpha + a_yy[yy[i]] + a_ss[ss[i]] + X[i] * beta +
offset[i];
    mu = fmin(mu, 20.7944); // max for poisson;
    yhat[i] = poisson_log_rng(mu);
    log_lik[i] = poisson_log_lpmf(Y[i] | mu);
   }
  }
```

Multilevel Log-Normal Model with Varying Intercepts Per State/Year and Offset (for Identity Theft):

```
  data {
   int nY;
   int nS;
   int nP; // number of (individual) predictors
   matrix[nY*nS, nP] X; // predictors, e.g. laws, etc.
   vector[nY*nS] offset; // a rate, has the coef set to 1
(should be logged)
   int yy[nY*nS];
   int ss[nY*nS];
   real<lower=0> Y[nY*nS]; // outcome/observations
  }
  transformed data {
```

```
   int N = nY * nS;
  }
  parameters {
   real alpha; // overall intercept
   vector[nY] a_yy; // unique intercept per year
   vector[nS] a_ss; // unique intercept per state
   real<lower=0> sigma_y; // pool unique YY intercepts
   real<lower=0> sigma_s; // pool unique SS intercepts
   vector[nP] beta; // beta for all predictors
   vector<lower=0>[nS] sigma_l; // log normal sigma (per
state).
  }
  transformed parameters {}
  model {
   vector[N] mu;
   vector[N] sigma;
   // priors
   target += normal_lpdf(alpha | 0, 10);
   target += normal_lpdf(beta | 0, 10);
   target += normal_lpdf(a_yy | 0, sigma_y);
   target += normal_lpdf(a_ss | 0, sigma_s);
   target += cauchy_lpdf(sigma_y | 0, 1);
   target += cauchy_lpdf(sigma_s | 0, 1);
   target += exponential_lpdf(sigma_l | 2); // tighter (re
lognorm)
   // linear model
   for ( i in 1:N ) {
  mu[i] = alpha + a_yy[yy[i]] + a_ss[ss[i]] + X[i] * beta
+ offset[i];
  sigma[i] = sigma_l[ss[i]];
  }
   target += lognormal_lpdf(Y | mu, sigma);
  }
  generated quantities {
   vector[N] yhat;
   vector[N] log_lik;
   for ( i in 1:N ) {
    real mu;
    real sigma;
    mu = alpha + a_yy[yy[i]] + a_ss[ss[i]] + X[i] * beta +
offset[i];
  sigma = sigma_l[ss[i]];
  yhat[i] = lognormal_rng(mu, sigma);
    log_lik[i] = lognormal_lpdf(Y[i] | mu, sigma);
   }
  }
```

The Bayesian chains converge well: the Gelman–Rubin statistic (***rhat***s are equal to 1±0.005) and Stan gives no serious warnings. A complementary posterior predictive plot is shown below (next to the one in the text). The observed *y*'s fall within the light blue posterior predictive band, indicating a reasonable fit.

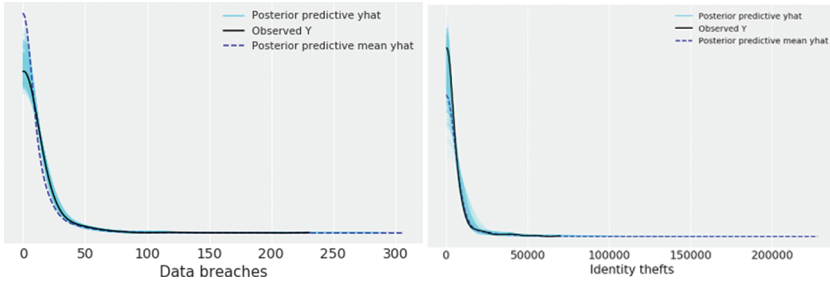

FIGURE A2 Posterior Predictive Plots Based on the *y*/*y*_hat Distribution. Left: data breach model; right: identity theft model.

Appendix IV. Unique Year and State Intercepts

Year Intercepts. The model estimates unique intercepts for each state and year. When a unique intercept's credible interval is around zero, it can be interpreted as random noise. In the data breach model, the year intercepts for 2005 and 2009 are below zero, and for 2006, 2007, and 2010 above zero. These intercepts are what remains after detrending (via y_trend), and they point to unknown influences on breach levels in those years.

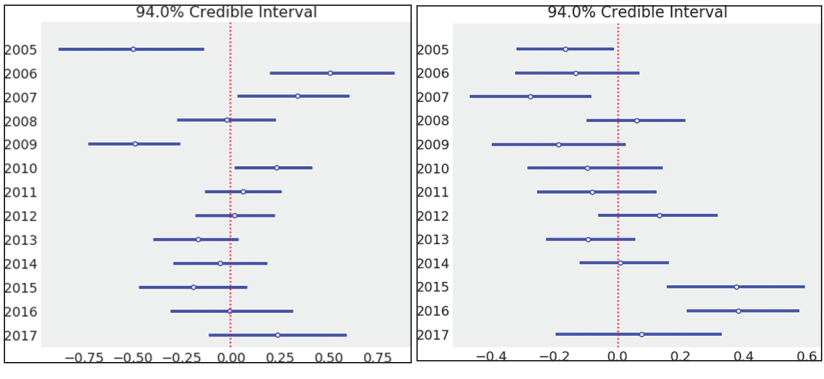

FIGURE A3 Unique Year Intercepts. Left: data breach model. Right: identity theft model. *(As before, the unique effect be estimated using* $e^{\text{mid-point}}$*.)*

State Intercepts. In both multilevel models, the state intercepts for a number of states differ from the baseline. This reflects differences that remain after controlling for the state size and regulatory and control predictors.

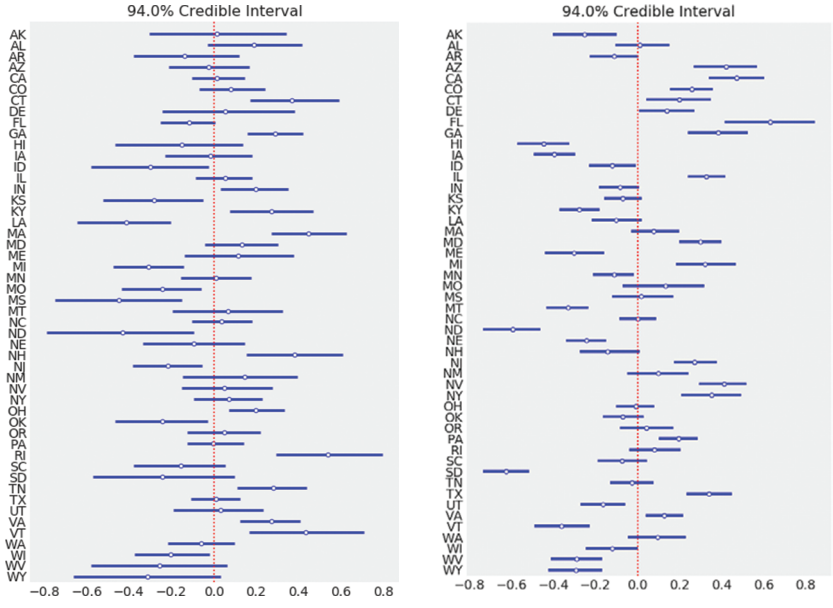

FIGURE A4 Unique State Intercepts. Left: data breach model. Right: identity theft model.