Hype and Heavy Tails: A Closer Look at Data Breaches

Benjamin Edwards
University of New Mexico
bedwards@cs.unm.edu

Steven Hofmeyr
Lawrence Berkeley National Laboratory
shofmeyr@lbl.gov

Stephanie Forrest
University of New Mexico
Santa Fe Institute
forrest@cs.unm.edu

ABSTRACT

Recent widely publicized data breaches have exposed the personal information of hundreds of millions of people. Some reports point to alarming increases in both the size and frequency of data breaches, spurring institutions around the world to address what appears to be a worsening situation. But, is the problem actually growing worse? In this paper, we study a popular public dataset and develop Bayesian Generalized Linear Models to investigate trends in data breaches. Analysis of the model shows that neither size nor frequency of data breaches has increased over the past decade. We find that the increases that have attracted attention can be explained by the heavy-tailed statistical distributions underlying the dataset. Specifically, we find that data breach size is log-normally distributed and that the daily frequency of breaches is described by a negative binomial distribution. These distributions may provide clues to the generative mechanisms that are responsible for the breaches. Additionally, our model predicts the likelihood of breaches of a particular size in the future. For example, we find that in the next year there is only a 31% chance of a breach of 10 million records or more in the US. Regardless of any trend, data breaches are costly, and we combine the model with two different cost models to project that in the next three years breaches could cost up to $55 billion.

1. INTRODUCTION

In February 2015, the second largest health insurer in the United States, Anthem Inc., was attacked, and 80 million records containing personal information were stolen [30]. Just a few months earlier, in September 2014, Home Depot's corporate network was penetrated and over 56 million credit card numbers were acquired [6, 27]. Both incidents made national headlines, the latest in a string of large-scale data breaches ([13, 26, 16]) that have spurred both the United States Congress [12] and the White House [25] to propose new disclosure laws to address what appears to be a worsening situation.

Several studies provide evidence that the problem of electronic data theft is growing. A 2014 Symantec report noted that there was an increase in the number of large data breaches, and a dramatic five-fold increase in the number of identities exposed over a single year [11]. In another study, Redspin reported that the number of breaches in the health care industry increased 29% from 2011 to 2012, and the total number of records compromised increased 138% for 2012 to 2013 [23].

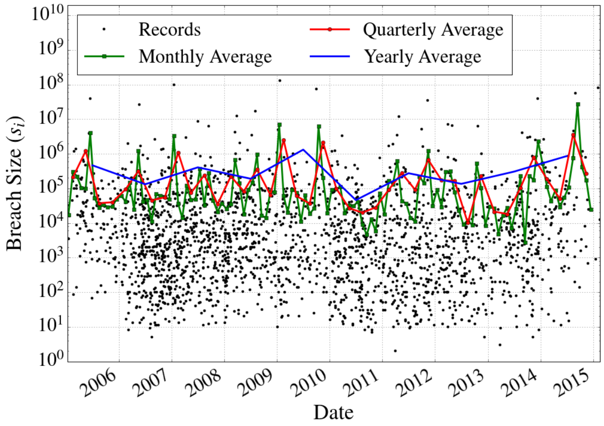


Figure 1: Data breach sizes (records exposed) over a ten-year period. Data taken from [9]

But, is the problem actually growing worse? Or if it is, how much worse is it, and what are the trends? The data used to produce these kinds of reports have very high variance, so simply reporting average values, as in these earlier reports, can be misleading. Figure 1 plots breach sizes over the past ten years using data obtained from a popular dataset published by the Privacy Rights Clearinghouse (PRC) [9]. In the figure, data breach sizes span eight orders of magnitude, which means that the average value can be significantly affected by just a few data points. For example, if we consider the identical data, but plot it on a yearly basis, it appears that breaches have increased in average size since 2012 (blue line on the figure). However, this trend is not at all obvious if we consider the data on a monthly or even quarterly basis, also shown in Figure 1 (green and red lines). Thus, there is a need for statistically sound data analyses to determine what, if any, trends exist, and where possible to make predictions about the future.

To address these issues, we adopt a statistical modeling approach and apply it to the PRC data, showing that in this dataset neither the size nor the frequency of breaches has increased over time. We use a Bayesian approach, which allows us to construct accurate models without overfitting (see subsection 3.1). Our analysis shows different trends for different subsets of the data. We consider two distinct types of breaches: *malicious*, where attackers actively target per-

To appear at the 14th Annual Workshop on the Economics of Information Security June 22-23, 2015. Copyright the Authors. Draft copy, do not distribute.

sonal information, and *negligent*, which occur when private information is exposed accidentally (e.g. misplacing a laptop). In the dataset, the size of malicious breaches has been slowly decreasing over the ten-year period, but the frequency has remained constant. By contrast, negligent breaches have remained constant in size and frequency over the ten-year period (see subsection 3.2 and subsection 3.3).

Beyond assessing trends, this approach enables us to determine the likelihood of certain future events, at least in the United States (see section 4). For example, the model predicts that in the next three years there is 7.8% chance of another Anthem sized (80 million) breach, and only a 0.4% chance of a Anthem and Home depot sized breach occurring within a year of each other. Further, there is a 1.2% chance of another Anthem-sized breach occurring between February 19, 2015 and the date of the Workshop on Information Security (in June 2015), and a 70% probability that there will be a breach of at least one million records in the same time frame. The probabilities are relatively high for breaches of one million records because the distributions that best describe the size of breaches in the dataset are heavy-tailed, meaning that rare events are much more likely to occur than would be expected for normal or exponential distributions.

Another contribution of our paper is identifying the particular forms of the underlying distributions, which may offer insight into the generative processes that lead to data breaches. For breach sizes, we find that the distribution is log-normal (see subsection 2.2); such distributions are known to emerge from multiplicative growth. In fact, the size distribution of companies is best described by a log-normal [40], so we speculate that as a company grows, the number of data records it holds grows proportionally, and breach sizes follow along. By contrast, the breach frequency best fits a negative binomial, which could be generated by a mixture of different types of breaches, with each type occurring at a different but constant rate (see subsection 2.3).

Some of our results seem counter-intuitive given the current level of concern about privacy and the damage that a data breach can cause. However, some simple anecdotal observations about our data lend credence to the results. The largest data breach in our data occurred back in 2009 when cyber-criminals stole 130 million credit card numbers from Heartland payment systems [33]. Additionally, as of March 4, 2015 there had been no breaches of personal information in the past 15 days, less than might be expected given current headlines.

We used the publicly available dataset that we believe is the most complete, but our models could easily be applied to additional datasets, for example, datasets that are not yet in the public domain or those that may arise if new disclosure laws are passed. Moreover, by establishing a baselinel, the models we describe could be extended in the future by incorporating additional data on the nature of the breaches, which could help identify promising areas for technical improvement. Such analysis could also help policy makers make better decisions about which problems are most pressing and how they should be addressed. For example, cybersecurity today is often framed in terms of risk analysis and management [34, 4]. Accurately assessing risk, however, requires quantitative measures of likelihood and cost. In this paper, we use available data and statistically sound models to provide precise estimates of the likelihood of data breaches. Using these estimates, we then incorporate two different cost models (see subsection 4.4 to assess likely future risks. Depending on the cost model, if trends continue we can expect the cumulative cost of data breaches to be between $3 and $55 billion over the next three years.

2. DATA

In this section, we describe the dataset obtained from the *Privacy Rights Clearinghouse* (PRC) and examine the distribution of breach sizes and frequencies. We show that the size of data breaches is log-normally distributed, whereas the daily frequency of breaches follows a *negative binomial*. Finally, we show how those distributions are affected when the data are divided into malicious and negligent breaches.

2.1 Privacy Rights Clearinghouse

The PRC is a California nonprofit corporation focused on issues of privacy [8]. The PRC has compiled a "Chronology of Data Breaches" dataset[1] that, as of February 23, 2015, contains information on 4,486 publicized data breaches that have occurred in the United States since 2005. For each breach, the dataset contains a number of variables including: the date the breach was made public, the name of the entity responsible for the data, the type of entity breached, a classification of the type of breach, the total number of records breached, the location (city and state) where the entity operates, information on the source of the data, and a short description of the breach.

Of the 4,486 breaches in the dataset, only those involving exposure of sensitive information have associated record counts. We restricted our analysis to this subset, which consists of 2,234 breaches. There are two noteworthy limitations to these data. First, the number of records listed in the dataset for each breach is only an estimate of the number of individuals affected, and second, the dataset contains only those breaches that have been publicly acknowledged. However, the PRC dataset is the largest and most extensive public dataset of its type. It is possible that many data breaches are going unreported. Different surveys have indicated that anywhere between 60% [42] to 89% [7] of security incidents go unreported. However, these reports are based on informal surveys of security professionals, their accuracy can't be confirmed (section 6), and there is no obvious reason why their size/frequency distributions should differ from PRC.

2.2 Breach Size

We denote the distribution of breach sizes over the number of records contained in individual breaches as S. For each individual breach i, we denote the number of associated records as s_i. To determine the time-independent distribution that best fits the data, we examined over 20 different distributions, for example, log-normal, log-skewnormal, power-law, log-logistic, and log-gamma.[2] In each case, we estimated the best fit parameters for the distribution using the maximum likelihood, and then performed a Kolomogorov-Smirnov (KS) test to determine if the parameterized distribution and the data were statistically significantly differ-

[1] Available for public download from http://www.privacyrights.org/data-breach.

[2] Specifically, we tested all of the distributions in the scipy stats package that have a domain defined for values > 0. http://docs.scipy.org/doc/scipy/reference/stats.html#continuous-distributions.

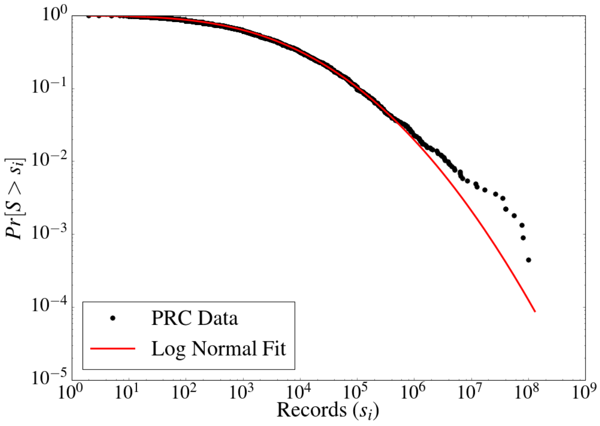


Figure 2: The distribution of breach sizes and the fit to a log-normal distribution.

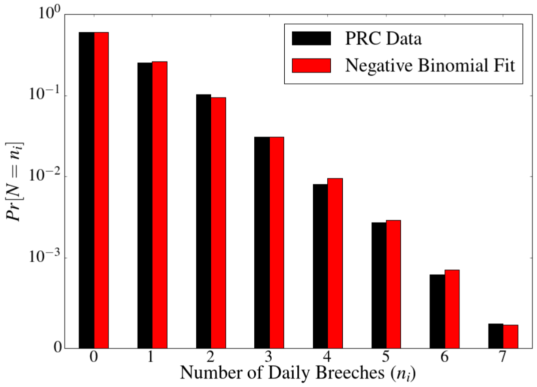


Figure 3: The distribution of the daily number of breaches and the fit to a negative binomial.

ent [29]. Figure 2 shows the fit to log-normal; the KS test gives $p = 0.19$, which means that we cannot reject the null hypothesis that the fit is different than the data.[3] For all other distributions, $p < 0.05$, which tells us that the data were unlikely to have been generated from that distribution. Although the best fit is to the log-normal, we can see in Figure 2 that the data points in the tail (high values of records) deviate from the best-fit line. We return to this issue in section 6.

Log-normal distributions often arise from multiplicative growth processes, where an entity's growth is expressed as a percentage of its current size, independent of its actual size [31]. This process has been used to model the size of companies as measured by annual sales, current employment, or total assets [40], and we speculate that a related process is operating here, because the number of sensitive (customer) records held by a company could reasonably be assumed to be proportional to its size.

2.3 Breach Frequency

We are interested in studying how often breaches occur and whether or not there are interesting trends in breach frequency. The dataset reports the exact date at which each breach became publicly known. For the majority of dates in the dataset, however, there were no publicly reported data breaches, and on days when breaches did occur, there were seldom more than two (Figure 3). Similar to the breach size data, there are no obvious visible trends in the daily frequency (data not shown).

We used a similar approach to the one we employed in subsection 2.2, except that we studied discrete distributions, because the range of daily frequencies is so small. We examined a number of discrete distributions, such as Poisson, binomial, zero-inflated Poisson and negative binomial, and found that the best fit is provided by a negative binomial. Figure 3 shows that the parameterized negative binomial and the data do not differ significantly, according to the KS test for discrete distributions [2], with $p = 0.99$. If we assume that breaches occur independently and at a constant rate, then we would expect the daily frequency to be a Poisson distribution [19]. However, the data are more dispersed than can be explained by a Poisson, which has a very poor fit, with $p = 8 \times 10^{-10}$.

There are a number of random processes that generate a negative binomial distribution [48]. The most likely candidate in this case is a continuous mixture of Poisson distributions, which occurs when events are generated by a Poisson process whose rate is itself a random variable. In our case, breaches at different organizations, perpetrated by different groups could all have different rates, leading to the negative binomial distribution we observe here.

2.4 Negligent and Malicious Breaches

Each breach in the PRC dataset is categorized into one of seven different categories (plus the category *Unknown*). The seven categories naturally divide into two groups. The first are breaches arising from *negligence*, where records were not actively sought by an attacker but were exposed accidentally, for example, through the loss of laptops, or accidental public exposure of sensitive information. The second group includes breaches arising from *malicious* activities that actively targeted private information, for example, attackers hacking into systems, an insider using information for malicious purposes, or payment card fraud. Table 1 contains information on the number of each type of breach in the dataset, and our groupings. It is apparent that negligent breaches occur nearly twice as often as malicious breaches.

We re-applied the data fitting analysis described earlier (subsection 2.2 and subsection 2.3) separately to each of the two groups. We find that even when the data are divided into negligent and malicious categories, each category matches a negative binomial distribution for daily frequency, although with different means. As before, the sizes of negligent breaches are well fit by a log-normal distribution. However, malicious breach sizes have a weaker fit to the log normal. Even though the lumped data (all categories aggregated) are log-normally distributed, it is possible that the weaker fit for malicious breaches arises because this distribution is changing over time. In section 3 we show how such a change in trend could account for this poorer fit.

3. MODELING DATA BREACH TRENDS

Our earlier analysis does not allow for the possibility that

[3] In this case, higher values of p are better, because they indicate that we are *not* rejecting the null hypothesis, i.e. that the data are drawn from a log-normal.

Breach Type	**Description**	**Count**
Negligent Breaches		1408
Portable Device	Lost, discarded or stolen, portable device or media	625
Unintended Disclosure	Information posted in a publicly available place, mishandled, or sent to the wrong party	455
Physical	Lost, discarded, or stolen non-electronic records	195
Stationary Device	Lost, discarded or stolen stationary device or media	135
Malicious Breaches		767
Hacking	Electronic entry by an outside party	458
Insider	Someone with legitimate access intentionally breaches information	279
Payment Card Fraud	Fraud involving debit and credit cards that is not accomplished via hacking	30
Unknown	Other or Unknown	57

Table 1: Types of data breaches as categorized by the PRC, grouped into negligent and malicious breaches.

the distributions are changing over time. In this section, we describe how we use Bayesian Generalized Linear Models (BLGMs) [17] to construct models of trends in the PRC the dataset. We then use Bayesian Information Criteria (BIC) to determine the highest likelihood model, while avoiding overfitting. We use the distributions derived in section 2, as the basis for our time-dependent models.

3.1 Bayesian Approach

We illustrate our approach by focusing on the sizes of negligent data breaches, S_n. The basic strategy assumes an underlying type of distribution for the data (e.g., sizes of negligent breaches), which we found to be log normal in subsection 2.2. Hence $S_n \sim Lognormal(\mu, \tau)$, where μ is the location parameter and τ is the shape parameter (standard deviation).

To incorporate temporal variations, we model the location parameter, μ, as a polynomial function of time, t, i.e. $\mu = \beta_0 + \beta_1 t + \cdots + \beta_d t^d$. Time is expressed as a decimal value in years since January 1, 2005, with a resolution of one day, e.g. $t = 1.2$ would be March 13, 2005. We describe how to determine the degree of the polynomial, d, later. The parameters, β_i, for the polynomial, together with the shape parameter, τ, comprise the prior distributions for the model.

The choice of prior distributions is an important and active area of research in Bayesian statistics. As suggested in the literature [17], we used normally distributed priors for the polynomial parameters, $\beta_0 \sim \mathcal{N}(\overline{\log(S_n)}, 1)$ and $\beta_i \sim \mathcal{N}(0, \frac{1}{Var[t^i]})$, and a gamma-distributed prior for the shape parameter, $\tau \sim Gamma(1, 1)$. These priors are "uninformative," i.e. they assume the least amount of information about the data. Although there are other possible priors, our results did not vary significantly when tested with other reasonable choices. Once the model is defined, we can numerically determine the parameters using maximum-likelihood estimation.

To assess the accuracy of the estimates, we determine confidence intervals for the values of the parameters using a variant of Markov Chain Monte Carlo (MCMC) sampling to ensure robust, fast samples [20]. MCMC is an efficient general method for sampling possible values for the parameters of the model.

The remaining unknown in the model is d, the degree of the polynomial. We determine a model for each $d \in [0, 10]$, and choose the model (and hence the polynomial) with the minimum Bayesian Information Criterion (BIC) [39]. The BIC balances the likelihood of the model, which is increased by adding parameters, with the number of parameters and size of data, and hence prevents overfitting. This enables us to chose a model that best fits changes in the data, rather than modeling statistical noise. This is an important feature when the distributions are heavy-tailed.

To summarize, our modeling approach involves the following steps:

1. Define a BGLM similar to Equation 1, as shown in subsection 3.2.
2. Find the maximum likelihood estimates for the parameters of the model (e.g. β_i, τ) for polynomial trends d up to degree 10.
3. Select the model that has the minimum BIC for the maximum likelihood estimates of the parameters.
4. Sample from the distribution of β_i using MCMC to determine the confidence intervals for the parameters.
5. Randomly sample the model to generate a distribution, and compare that to the actual distribution, using the KS test.

3.2 Modeling Breach Size

As derived in subsection 3.1, the model for breach sizes is

$$
\begin{aligned}
S_n &\sim Lognormal(\mu, \tau) \\
\mu &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_d t^d \\
\beta_0 &\sim \mathcal{N}(\overline{\log(S_n)}, 1) \\
\beta_i &\sim \mathcal{N}(0, \frac{1}{Var[t^i]}) \\
\tau &\sim Gamma(1, 1)
\end{aligned}
\tag{1}
$$

The best fit model for malicious breaches, as determined by the minimum BIC, gives $d = 1$, which indicates a linear trend in the data. Surprisingly, the trend is negative. By contrast, for negligent breaches, the best fit is at $d = 0$, which indicates that the distribution of sizes is constant. Figure 4 shows the median values for models, plotted against the PRC data[4]. Maximum likelihood estimates for the parameters are given in Table 2.

To summarize, we find that the distribution of negligent breach sizes has remained constant with a median size of 2731, while malicious breaches have declined in size, with the median declining at a rate of 15.6% a year over the ten-year period represented by the dataset. Random samples generated using Equation 1 and the estimates found in Table 2, indicate that the predicted distribution of sizes by

[4] We show median rather than the mean because it better represents the typical values in heavy tailed distributions.

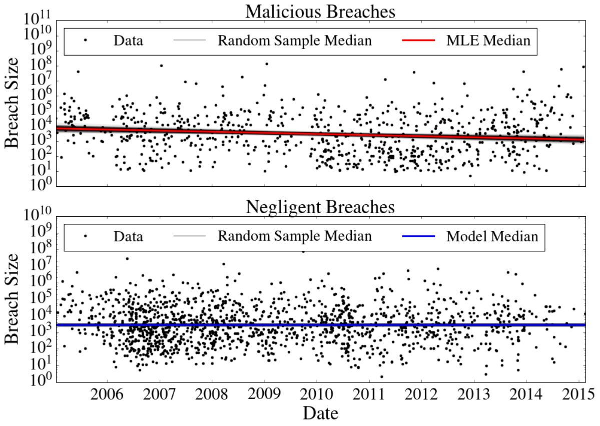


Figure 4: The size of data breaches from the PRC dataset, versus the maximum likelihood estimate of the median size.

Variable	Estimate	95% Confidence Interval
	Negligent	
β_0	7.894	[7.753, 8.030]
τ	0.138	[0.128, 0.148]
	Malicious	
β_0	8.56	[8.077, 9.032]
β_1	-0.130	[-0.208, -0.046]
τ	0.0991	[0.089, 0.109]

Table 2: Maximum likelihood estimates and 95% confidence intervals for models of breach size.

the model does not significantly differ from the data, i.e. our model generates data that are indistinguishable from the actual data. The KS test gives $p = 0.33$ for the fit to the negligent breach sizes, and $p = 0.52$ for the fit to the malicious breach sizes.

3.3 Modeling Breach Frequency

We use the same methodology to model the frequency of data breaches, with a negative binomial as the basic distribution, as determined in subsection 2.3.[5] The daily frequency, B_n of negligent breaches is given by

$$\begin{aligned} B_n &\sim NegativeBinomial(\mu, \alpha) \\ \log(\mu) &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_k t^k \\ \beta_0 &\sim \mathcal{N}(\overline{\log(D_n)}, 1) \\ \beta_i &\sim \mathcal{N}(0, Var[t^i]) \\ \alpha &\sim Gamma(1, 1) \end{aligned} \tag{2}$$

The same model is used for malicious breaches, replacing B_n with B_m, the daily number of malicious breaches.

For the daily frequencies of both negligent and malicious breaches, the models with the lowest BIC are polynomials of degree $d = 0$, indicating that the daily frequency of breaches has remained constant over the past ten years. The maximum likelihood estimates and 95% confidence intervals are shown in Table 3. Random samples generated using the Equation 2 are not statistically significantly different from the data for both negligent and malicious breaches; which have $p = 1.0$ and $p = 0.96$, respectively, for the KS test.

Variable	Estimate	95% Confidence Interval
	Negligent	
e^{β_0}	0.383	[0.360, 0.407]
α	1.028	[0.841, 1.292]
	Malicious	
e^{β_0}	0.208	[0.193, 0.223]
α	1.738	[1.113, 3.225]

Table 3: Maximum likelihood estimates and 95% confidence intervals for models of daily breach counts. We report e^{β_0} as this is the mean number of breaches of each type per day.

3.4 Modeling Large Breaches

It is possible that the models developed above are dominated by smaller breaches, which have experienced little change over the last ten years, while larger breaches are increasing in size or frequency. We define *large* breaches as those involving 500,000 or more records. This threshold was chosen because it includes a large enough sample size for us to fit reasonable models (93 malicious and 121 negligent breaches), but the threshold is high enough that the breach would likely be reported widely in the press.

Using this definition, we find that large breach sizes still fit a log-normal distribution, and that neither malicious nor negligent large breaches show a significant trend over the past ten years. Given that there is a slight downward trend for malicious breaches of all sizes, this result implies that small breaches must actually be declining even more on average.

The frequency of large breaches, both malicious and negligent, fits a Poisson distribution, rather than the negative binomial observed for breaches of all sizes. This could indicate that different processes are responsible for generating large versus small breaches. Alternatively, it could simply be that the very low probability of a large breach results in a distribution that is difficult to distinguish from the negative binomial. In this case, we would expect the BIC of the Poisson model to be lower because it has one less parameter than the negative binomial. Regardless of whether the best model mathematically is a negative binomial or Poisson, the trends for large breaches are the same as the overall trends, with the frequency of malicious and negligent large breaches remaining constant over the ten years covered by the dataset.

4. PREDICTION

The power of a good statistical model is that it can be used to make predictions about the likelihood of future events. In this section we discuss what types of predictions models like ours can legitimately make, and point out some of the ways in which naive interpretations of the data can lead to erroneous conclusions. We then demonstrate how the model can be used to quantify the likelihood of some of the large breaches that were experienced in 2014, and we make some predictions about the likelihood of large breaches in the future. Finally, we project the possible cost of data breaches over the next three years.

4.1 Variance and Prediction

[5] We also test a Poisson model, but found it had a higher BIC than a negative binomial model.

Because the distributions of both the breach sizes and frequencies in the PRC dataset are heavy-tailed, it is difficult for any model to make precise predictions about the exact number of breaches or their average size. This is different from a dataset that is, for example, normally distributed, where, with sufficiently large sample size, one can say with high probability that samples in the future will cluster around the mean, and estimate the chances of samples falling outside one standard deviation from the mean. However, in the PRC dataset, common statistics like the mean or the total number of records exposed are much less predictable. The data often vary wildly from year to year, even if the process generating the breaches has not changed at all. This phenomenon is common in many complex systems, including many security-relevant datasets, e.g., [15].

We illustrate the effect of the high variability in Figure 5a and Figure 5b. These figures show the result of measuring the total number of malicious breaches and average breach size annually for the historical data (black line) and a single simulation using the models presented in section 3 (red line). Although our model indicates a simple trend in the mean, the distribution can generate large year-to-year variations. These changes are often reported as though they are significant, but our results suggest that they are likely artifacts of the heavy-tailed nature of the data.

For example, a number of industry reports, some using the PRC dataset, have pointed to large changes in the size or number of data breaches from year to year [45, 11]. One of the most alarmist is the Symantec Threat Report which noted a 493% increase in the total number of records exposed from 2012 to 2013, and a 62% increase in the number of breaches in the same time frame.[6] The 493% number includes the large Court Ventures data breach, which was initially reported as revealing 200 million records, but later reports reduced that that number to 3.1 million records [16]. Even with this correction, the report implies a 282% increase in the total number of breached records. These increases sound startling, and a naive interpretation might suggest that both the number and size of data breaches are skyrocketing.

We can test for the likelihood of such extreme changes using our model. To do so, we used the model to generate 10,000 samples of possible annual totals, both for the number of breaches and the number of records, from 2005-2014. We find that a 62% year-to-year increase in the total number of breaches is relatively common in simulation, occurring 15.3% of the time. Similarly, an increase of 282% in total records occurs in 14.7% of year-to-year transitions. These results suggest that the large changes identified in these reports are not necessarily significant and could be natural variations arising from the underlying observed distributions of data breaches.

Although our model cannot accurately predict the total number or typical size of data breaches in any given year, it can assess the likelihood of different sizes of breaches. That is, we can predict the probability of a breach of a specific size within a given time-frame, as we show in the next subsection.

4.2 2014 Breaches

To assess the likelihood of the breaches that occurred in 2014, we fit the model using data from 2005 to the end of 2013, and used it to "predict" the events of 2014. The MLEs of this smaller dataset are virtually identical to those found for the whole range, suggesting that the 2014 data are not significantly different from those of the previous nine years.

We used the models derived from the 2005 to 2013 data to generate 50,000 simulations of breaches from Jan. 1, 2014 through February 18, 2015. For each day in this simulated timespan we generated a random number of breaches using Equation 2, and then for each simulated breach we generated a random breach size using Equation 1. We plot the cumulative number of records breached in Figure 6.

The mean cumulative number of breached records roughly matches the actual cumulative number of records up to September 2014, when the Home Depot breach exposed 56 million credit card numbers. Less than six months later an 80-million record breach of the healthcare provider Anthem led to a large increase in the cumulative number of breaches, well outside the model's 95% confidence interval.[7]

As discussed in subsection 4.1, large data breaches are expected to occur occasionally due to the heavy-tailed nature of the distribution from which they are drawn. However, in our experiments with the model, two breaches of this size or larger occurred in the same year in only 0.07% of simulations, suggesting that the co-occurrence of these two breach sizes was indeed rare. Although this event was unlikely, it is unclear whether or not it represents a statistically significant change in the overall pattern exhibited by the rest of the data.

Indeed, this result appears to be in line with the heavy-tailed distribution of the data, as shown in Figure 2. Inspecting the data in Figure 2, there is 0.0022 probability of a breach of 56 million or more records, and 0.0013 probability of a breach of 80 million or more records. Hence, the probability of two breaches of this magnitude happening within the 413 days of our prediction is $413 \times 0.62 \times 0.0022 \times 0.0013 = 0.0007$, where 0.62 is the average number of breaches per day over the course of the 10 years. This resulting value of 0.07% agrees with the simulations of the model.

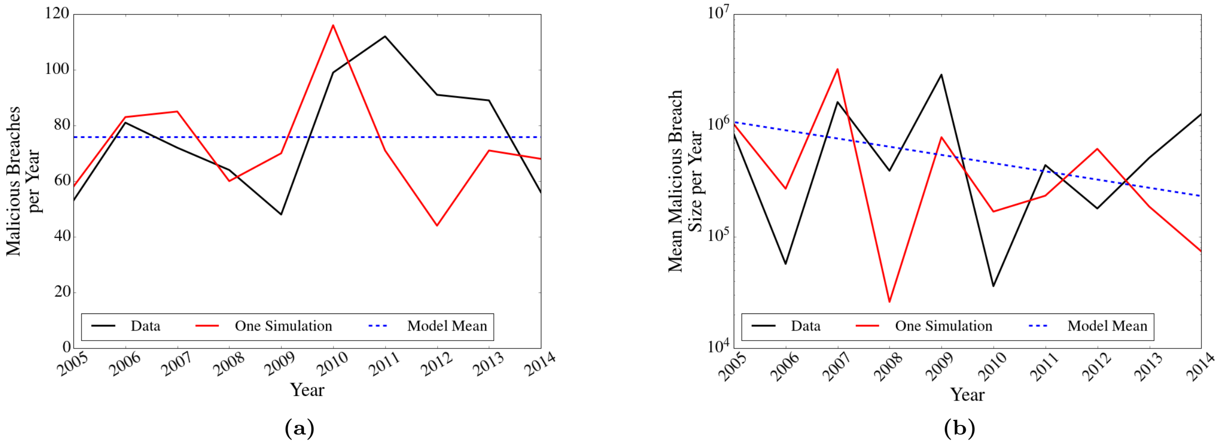
4.3 Future Breaches

We now use our model built on the past decade of data breaches to simulate what breaches we might expect in the next three years in the United States. With the current climate and concern over data breaches, there will likely be changes in practices and policy that will change data breach trends. However, this gives us an opportunity to examine what might occur if the status quo is maintained. Once again we use the same methodology, predicting from February 19, 2015, through Feb 19, 2018. We predict the probability of several different sizes of breaches. The results can be seen in Figure 7a and Figure 7b.

Breaches of 750,000 records or more are almost certain (97.6%) within the next year, but larger breach frequency does not increase as quickly as intuition might suggest. For example, there is a 7.7× increase in the probability of the largest breach (130 million) occurring in the next year, from 0.23% to 1.78%, whereas for the year after that, the probability only increases by a factor of 1.7, to 3.1%. This drop-off is a consequence of the decreasing trend in malicious breach sizes that we identified earlier. This is especially clear in

[6] These reports use a combination of public and private data, so comparison of exact numbers is not feasible.

[7] In the absence of the Home Depot and Anthem breaches, the median value of our simulations provides an excellent estimate of the cumulative number of records breached.



(a) (b)

Figure 5: (a) The number of breaches reported each year throughout the dataset, together with a single simulation sampled from our model and the average number of breaches. (b) The average breach size reported for each year of data along with simulated sizes of breaches, and the model's average breach size.

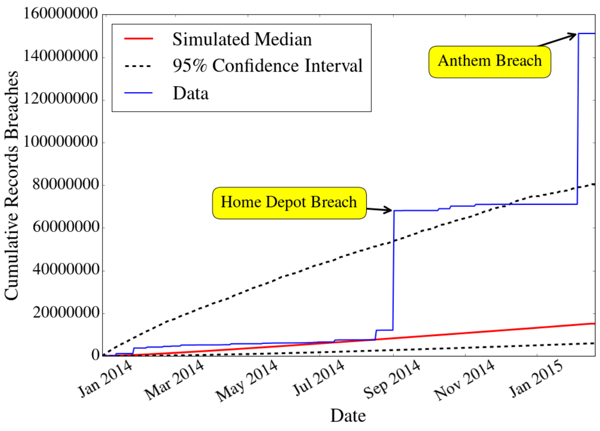


Figure 6: The cumulative number of breached records, both historically (shown in blue) and as predicted by our model. The simulated median (shown in red) is computed over 50,000 independent simulations. The dashed lines represent the 95% confidence interval.

Breach size	**% Chance**	
(millions)	**Before WEIS**	**In 2016**
1	70	97
1.5	57	91
2	48	84
5	23	52
10	12	31
25	4.6	12
56	1.8	4.8
80	1.2	3.2
130	0.6	1.7

Table 4: Chance of the occurrence of various size malicious breaches by the start of WEIS. The breach size is in millions of records.

Figure 7b, which shows that we are almost certain to see a breach of five million records or more in the next three years (86.2%), but above that size the probability drops off rapidly, e.g. a breach of size greater than 60 million has less than a 10% chance of occurring in the next three years.

Predictions like this could be relevant for policy makers interested in the problem of reducing data breaches. For example, the results suggest that it might be more sensible to address the problem of smaller breaches that are almost certain to happen, than to focus on the very large and infrequent headline-grabbing events. Disclosure laws at the Federal level, that force small, local organizations to consistently report breaches, could be one way of doing this.

As with most efforts to model dynamic, real-world phenomena, we expect the predictions to lose accuracy over time. So although our predictions for the next three years could be off, we expect the model to work better for the short term. As a demonstration, beginning February 19, 2015 we predict the probability of various breach sizes in the next year and before June 22, 2015, which is the start of the Workshop on the Economics of Information Security (WEIS). The vertical line in Figure 7a is the date of WEIS. The exact probabilities are given in Table 4. Thus, we can say with high probability (70%) that a breach of at least one million records will occur before WEIS, and we do not expect to see a breach equivalent to Anthem (1.2% chance). In the next year we expect only a 31% chance of a breach of 10 million records or more.

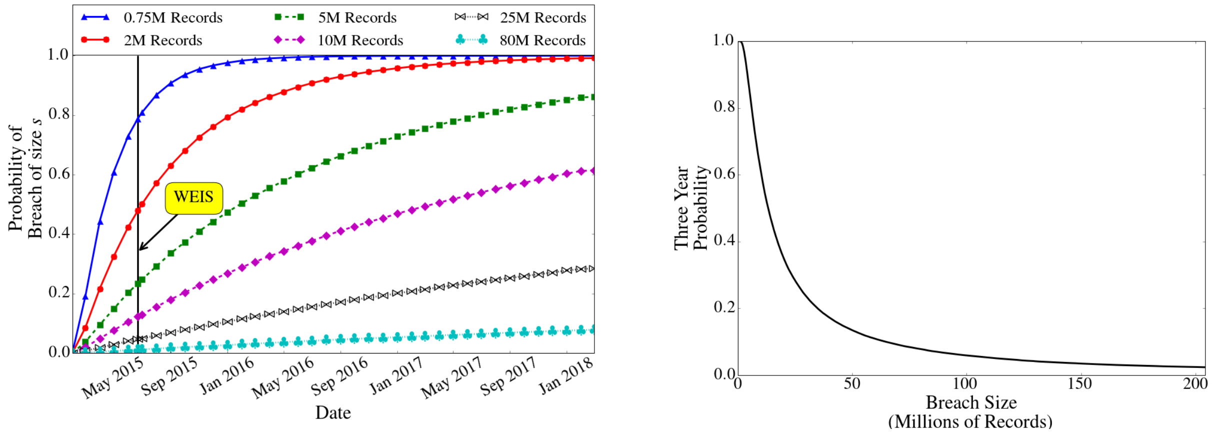
4.4 Predicting Future Costs

We can estimate the total expected cost of breaches in the future by incorporating data and other models related to cost. The Ponemon Institute publishes annual costs of data breaches, and found an average \$201 cost per record breached in 2014 [28]. Further analysis by others argues that such a flat rate is not the most accurate model for costs. Using non-public data, for example, Jacobs showed that the cost of a breach can be better estimated with a log-log model of the form [24]

$$\log(c) = 7.68 + 0.7584 * \log(s) \qquad (3)$$

where c is the cost of the breach in data, and s is the size of the breach.

In Equation 3 the cost of a breach grows less than linearly, resulting in overall lower costs than those predicted



(a) **(b)**

Figure 7: (a) The predicted probability of breaches of various sizes over the next three years. Each line represents the probability of at least one breach of the size denoted in the legend occurring before the date on the horizontal axis. We do not include smaller breach sizes, as they will almost certainly occur within the next few months. (b) The predicted probabilities of breach size after three years.

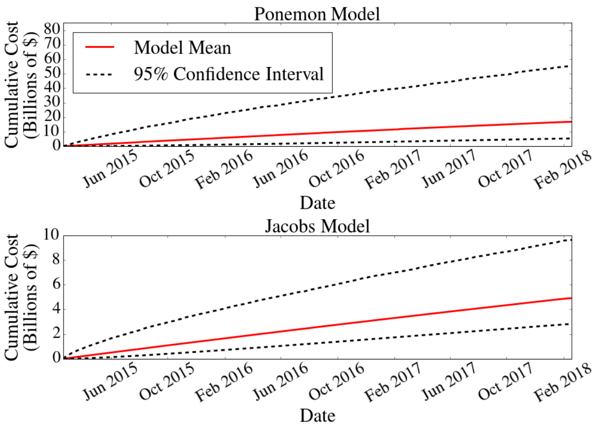


Figure 8: Predicted cumulative cost of data breaches in the next three years using two different cost models.

by the Ponemon model. Because the data used to create this model are not public, it is hard to assess its validity, but if it is valid, then it can help us estimate future costs of data breaches. Combining this model with Equation 1 and Equation 2 produces the predicted cumulative cost of data breaches over the next three years, as shown in Figure 8.

The flat rate cost model (Ponemon) suggests that in the next three years we can expect anywhere between $5.36 billion and $55 billion in losses associated with public data breaches. Jacob's model gives a more modest estimate of somewhere between $2.8 and $9.64 billion. b

5. RELATED WORK

According to the PRC, over 90 reports and articles reference the data used in our study [9]. However, only a few of those reports perform quantitative analysis, and most do not investigate trends in the size or frequency of data breaches. There are a few exceptions, for example, the Symantec Threat Report mentioned earlier. Another example is a Verizon report released in 2014 [45], which examines trends in the relative frequency over time of various types of attacks and motivations. However, the methodology for determining the trends is not described, and the report makes no predictions about the future. Many reports from security companies, such as those from Trustwave [43], focus on classifying the various attack vectors, without attempting to model trends.

There has been little focus on trends in breaches in the academic literature. Some older research investigated trends in the relative frequency of various categories of breaches from 2005-2007, but they found that the limited sample size prevented them from making statements about the significance of their results [14]. More recently, in 2010, Widup examined yearly trends in different types of data breaches [47]. However, no statistical analysis was conducted to estimate the underlying distribution or to separate out normal variations from distinct trends. Some papers investigate predictions about future events. For example, Bagchi and Udo developed a general statistical model for predicting the cumulative number security incidents of a specific type [3], and Condon et. al used a time series model to predict security incidents [10]. However neither of these two studies focused specifically on data breaches.

Numerous reports focus on the health care industry. The U.S. Department of Health and Human Services released a 2014 report examining breaches of protected health information [32]. The report includes basic counts of different types of breaches but does not identify any clear trends. Redspin has published three annual reports on data breaches in the healthcare industry [21, 22, 23]. In 2011, they reported a 97% increase in the number of breaches from the previous year, and a dramatic 525% increase in the number of total records breached [21]. The following year, they report an increase in the number of large breaches (22%) and a decrease in the number of total records breached. These variations fit well with our observations of the heavy-tailed nature of the underlying data.

Some reports focusing on the cost of data breaches were described in subsection 4.4. Similar studies focused on hospitals claim that breaches can cost organizations an average of $2.4 million over the course of two years.

Other work has focused on the overall cost of security breaches. Acquisti et al. found a negative impact on the stock value of companies experiencing privacy breaches [1]. Thomas et al. built a branching activity model which measures the impact of information security breaches beyond a breached organization [41]. Studies such as these could be combined with our methodology to infer future overall costs of breaches.

A number of other studies have examined the possible policy implications of data breach notification laws. Picanso suggested a framework for legislation of uniform data breach notifications [36]. Romanosky et al. analyzed the economic and legal ramifications of lawsuits when consumer data is compromised [37]. Later, Romanosky et al. created an abstract economic model to investigate the effect of mandatory data breach disclosure laws [38]. Using older parameter estimates, their model shows that if disclosure were made mandatory, then costs would be higher for companies experiencing breaches and that companies would likely increase their investment in security infrastructure. Graves et al. use PRC data to conclude that credit card companies should wait until fraud occurs before reissuing credit cards in the wake of a breach [18].

6. DISCUSSION

Our results suggest that publicly reported data breaches in the U.S. have not increased significantly over the past ten years, either in frequency or in size. Because the distribution of breach sizes is heavy-tailed, large (rare) events occur more frequently than intuition would suggest. This helps to explain why many reports show massive year-to-year increases in both the aggregate number of records exposed and the number of breaches [23, 45, 43, 11]. All of these reports lump data into yearly bins, and this amount of aggregation can often influence the apparent trends (Figure 1).

The idea that breaches are not necessarily worsening may seem counter-intuitive. The Red Queen hypothesis in biology [44] provides a possible explanation. It states that organisms not only compete within their own species to gain reproductive advantage, but they must also compete with other species, leading to an evolutionary arms race. In our case, as security practices have improved, attacks have become more sophisticated, possibly resulting in stasis for both attackers or defenders. This hypothesis is consistent with observed patterns in the dataset. Indeed, for breaches over 500,000 records there was no increase in size or frequency of malicious data breaches, suggesting that for large breaches such an arms race could be occurring. Many large breaches have occurred over the past decade, but the largest was disclosed as far back as 2009 [26], and the second largest was even earlier, in 2007 [5]. Future work could analyze these breaches in depth to determine whether more recent breaches have required more sophisticated attacks.

Even if breaches are stable in size and frequency, their impact is likely growing. The ability to monetize personal information, and the increasing ease with which financial transactions are conducted electronically could mean that the cost of data breaches will rise in the future. To address this issue, we considered two different models taken from the literature, which give wildly different projections. Reconciling these two models is an important area of future work. With improved cost models, however, integration with our models to produce more accurate projections would be straightforward.

Our results are based on publicly available data. It may be that the data are incomplete, and therefore our model is biased downwards, as some breaches will go unreported, but few reported breaches will prove not to have occurred. As more data become available, it will be straightforward to incorporate and update trend analyses and predictions. Given new data, from private sources or other countries other than the United States, it would be important not only to reanalyze trends, but also to revisit the underlying distributions. Despite this caveat, we expect that the PRC data is reasonably complete for the U.S., because most U.S. states already have disclosure laws (48 out of 50 as of January 2015 [35]) that require organizations to report the compromise of sensitive customer information. These laws vary in their requirements so it is possible that many breaches still go unreported. Future work could use interrupted regression to test whether reporting laws change the rate of reporting [46].

As we described earlier, the data are well-modeled by certain distributions, and these distributions could arise from underlying processes related to the breaches (section 2). However, Figure 2 illustrates that there is some deviation in the tail, suggesting that the log-normal fit is not exact for breaches that exceed 1,000,000 records. There are several possible explanations. It could simply be statistical noise, which is a known consequence of the rarity of large breaches. Alternatively, it could be that large breaches are generated by a different process from smaller breaches, a hypothesis that we rejected in subsection 3.4. Another possibility is that large breaches are more likely to be reported than smaller ones, either because there is a higher likelihood that the breach is noticed or because it is more likely that some of the records are covered by a disclosure law.

This paper focuses on identifying trends in the size and frequency of data breaches over time, and predicting the likelihood of future breaches. However, it may be possible to identify other factors that influence breaches, for example, the size of an organization. It is reasonable to expect that the number of records that an organization holds is related to its size, and that this factor alone would affect expected breach size. We conducted a preliminary investigation of U.S. universities with breaches in the PRC dataset but found no significant correlation between university enrollments (proxy for size of institution) at the time of the breach and the size of the breach itself. This unanticipated result bears additional study. In the future we plan to identify features of organizations that are predictive of the size and frequency of breaches they will experience, with the goal of helping policy makers focus their attention where it can have the most impact.

Our model provides estimates of the probability of breaches of specific sizes occurring in the past and the future through simulation. Given its relative simplicity, it may be possible to construct analytic solutions for these probabilities, and not have to rely on simulation. However, in general we cannot expect all such models to be tractable analytically.

7. CONCLUSION

It is popular today to frame the cybersecurity problem in terms of risk analysis and management. For example, the U.S. National Institute of Standards (NIST) has developed and promulgated its cybersecurity framework, which is based almost entirely on the concept of risk assessment [34]. To evaluate these risks, however, requires an accurate assessment of both cost and likelihood. In this paper, we focused on the likelihood component, showing how widely available datasets can be used to develop more nuanced estimates and predictions about data breaches than the typically alarmist reports and headlines produced by security companies and the popular press. As we have shown here, simply comparing last year's data with this year's is unlikely to provide an accurate picture.

Our analysis of the PRC dataset shows that neither the size nor the frequency of two broad classes of data breaches has increased over the past decade. It is, of course, possible that the PRC dataset is not representative of all breaches or that there has been a significant transition in the underlying probabilities in the recent past which is not yet reflected in our data. A third possible explanation for this surprising result is that data privacy practices have improved at roughly the same rate as attacker prowess—Red Queen effect [44]. Under this scenario, we are in an arms race, and can expect continual pressure to increase defenses just to stay even. It will take extraordinary efforts if we are ever to get ahead.

In conclusion, data breaches pose an ongoing threat to personal and financial security, and they are costly for the organizations that hold large collections of personal data. In addition, because so much of our daily lives is now conducted online, it is becoming easier for criminals to monetize stolen information. This problem is especially acute for individual citizens, who generally have no direct control over the fate of their private information. Finding effective solutions will require understanding the scope of the problem, how it is changing over time, and identifying the underlying processes and incentives.

8. ACKNOWLEDGEMENTS

The authors would like to thank Robert Axelrod for his many useful comments on early versions of this paper. The authors gratefully acknowledge the partial support of NSF CNS 1444500, DARPA (P-1070-113237), DOE (DE-AC02-05CH11231) and the Santa Fe Institue.

9. REFERENCES

[1] A. Acquisti, A. Friedman, and R. Telang. Is there a cost to privacy breaches? an event study. *ICIS 2006 Proceedings*, page 94, 2006.

[2] T. B. Arnold and J. W. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39, 2011.

[3] K. Bagchi and G. Udo. An analysis of the growth of computer and internet security breaches. *Communications of the Association for Information Systems*, 12(1):46, 2003.

[4] B. Blakley, E. McDermott, and D. Geer. Information security is information risk management. In *Proceedings of the 2001 workshop on New security paradigms*, pages 97–104. ACM, 2001.

[5] M. H. Bosworth. Tjx data breach victims reach 94 million. *Consumer Affairs*, Oct. 2007.

[6] B. X. Chen. Home depot investigates a possible credit card breach. *The New York Times*, Sept. 2014.

[7] T. Claburn. Most security breaches go unreported. *Information Week*, July 2008.

[8] P. R. Clearinghouse. Mission statement. https://www.privacyrights.org/content/about-privacy-rights-clearinghouse#goals, May 2014.

[9] P. R. Clearinghouse. Chronology of data breaches: Faq. https://www.privacyrights.org/content/chronology-data-breaches-faq, 2015.

[10] E. Condon, A. He, and M. Cukier. Analysis of computer security incident data using time series models. In *Software Reliability Engineering, 2008. ISSRE 2008. 19th International Symposium on*, pages 77–86. IEEE, 2008.

[11] S. Corporation. Internet security threat report 2014. http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf, Apr. 2014.

[12] Covington and B. LLP. Data breach notification bills introduced in house and senate. *The National Law Review*, Feb. 2015.

[13] J. Creswell and E. Dash. Banks unsure which cards were exposed in breach. *The New York Times*, June 2005.

[14] M. Curtin and L. T. Ayres. Using science to combat data loss: Analyzing breaches by type and industry. *ISJLP*, 4:569, 2008.

[15] B. Edwards, T. Moore, G. Stelle, S. Hofmeyr, and S. Forrest. Beyond the blacklist: modeling malware spread and the effect of interventions. In *Proceedings of the 2012 workshop on New security paradigms*, pages 53–66. ACM, 2012.

[16] J. Finkle. Experian enmeshed in litigation over business that was breached. *Reuters*, Apr. 2014.

[17] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[18] J. T. Graves, A. Acquisti, and N. Christin. Should payment card issuers reissue cards in response to a data breach? *2014 Workshop on the Economics of Information Security*, 2014.

[19] F. A. Haight. *Handbook of the Poisson distribution*. Wiley New York, 1967.

[20] M. D. Homan and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[21] R. Inc. Redspin breach report 2011: Protected health information. http://www.redspin.com/docs/Redspin_PHI_2011_Breach_Report.pdf, Feb. 2012.

[22] R. Inc. Redspin breach report 2012: Protected health information. http://www.redspin.com/docs/Redspin_Breach_Report_2012.pdf, Feb. 2013.

[23] R. Inc. Redspin breach report 2013: Protected health information. https://www.redspin.com/docs/Redspin-2013-Breach-Report-Protected-Health-Information-PHI.pdf, Feb. 2014.

[24] J. Jacobs. Analyzing ponemon cost of data breach. http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/, Dec. 2014.

[25] J. Kosseff. Analysis of white house data breach notification bill. *The National Law Review*, Jan. 2015.

[26] B. Krebs. Payment processor breach may be largest ever. *The Washington Post*, Jan. 2009.

[27] B. Krebs. Home depot: Hackers stole 53m email addresses. *Krebs on Security*, Nov. 2014.

[28] P. I. LLC. 2014 cost of data breach study: Global analysis. `http://www.ponemon.org/blog/ponemon-institute-releases-2014-cost-of-data-breach-global-analysis`, May 2014.

[29] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[30] A. W. Mathews and D. Yadron. Health insurer anthem hit by hackers. *The Wall Street Journal*, Feb. 2015.

[31] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[32] U. D. of Health and H. Services. Annual report to congress on breaches of unsecured protected health information. `http://www.hhs.gov/ocr/privacy/hipaa/administrative/breachnotificationrule/breachreport2011-2012.pdf`, 2014.

[33] U. D. of Justice. Alleged international hacker indicted for massive attack on u.s. retail and banking networks. `http://www.justice.gov/opa/pr/alleged-international-hacker-indicted-massive-attack-us-retail-and-banking-networks`, Aug. 2009.

[34] N. I. of Standards and Technology. Framework for improving critical infrastructure cybersecurity. `http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf`, Feb. 2014.

[35] N. C. of State Legislatures. Security breach notification laws. `http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx`, Jan. 2015.

[36] K. E. Picanso. Protecting information security under a uniform data breach notification law. *Fordham L. Rev.*, 75:355, 2006.

[37] S. Romanosky and A. Acquisti. Privacy costs and personal data protection: Economic and legal perspectives. *Berkeley Tech. LJ*, 24:1061, 2009.

[38] S. Romanosky, A. Acquisti, and R. Sharp. Data breaches and identity theft: When is mandatory disclosure optimal? TPRC, 2010.

[39] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[40] J. Sutton. Gibrat's legacy. *Journal of economic literature*, pages 40–59, 1997.

[41] R. C. Thomas, M. Antkiewicz, P. Florer, S. Widup, and M. Woodyard. How bad is it?–a branching activity model to estimate the impact of information security breaches. *A Branching Activity Model to Estimate the Impact of Information Security Breaches (March 11, 2013)*, 2013.

[42] T. Track. Majority of malware analysts aware of data breaches not disclosed by their employers. `http://www.threattracksecurity.com/press-release/majority-of-malware-analysts-aware-of-data-breaches-not-disclosed-by-their-employers.aspx`, Nov. 2013.

[43] Trustwave. Trustwave 2013 global security report. `https://www2.trustwave.com/2013gsr.html`, 2013.

[44] L. Van Valen. A new evolutionary law. *Evolutionary theory*, 1:1–30, 1973.

[45] Verizon. 2014 data breach investigations report. `http://www.verizonenterprise.com/DBIR/2014/`, 2014.

[46] A. K. Wagner, S. B. Soumerai, F. Zhang, and D. Ross-Degnan. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of clinical pharmacy and therapeutics*, 27(4):299–309, 2002.

[47] S. Widup. The leaking vault: Five years of data breaches. *Digital Forensics Association*, 2010.

[48] M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Learning*, 2013.