
Research paper

Hype and heavy tails: A closer look at data breaches

Benjamin Edwards^{1*}, Steven Hofmeyr² and Stephanie Forrest^{1,3}

¹Department of Computer Science, University of New Mexico, 1 University of New Mexico Albuquerque, NM 87131, United States; ²Computer Languages & Systems Software, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, United States; ³Santa Fe Institute 1399 Hyde Park Road Santa Fe, NM 87501, United States

*Corresponding author. E-mail: bedwards@cs.unm.edu

Received 2 October 2015; revised 10 June 2016; accepted 21 July 2016

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 article, 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 the size of data breaches is well modeled by the log-normal family of distributions 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 between 15 September 2015 and 16 September 2016 there is only a 53.6% chance of a breach of 10 million records or more in the USA. Regardless of any trend, data breaches are costly, and we combine the model with two different cost models to project that in the next 3 years breaches could cost up to \$179 billion.

Key words: data breaches; heavy tails; log-normal; negative binomial; Bayesian linear model.

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 [1]. A few months later, the US Office of Personal Management announced that personal information, including the background checks of 21.5 million federal employees was compromised [2]. Ten months earlier, in September 2014, Home Depot's corporate network was penetrated and over 56 million credit card numbers were acquired [3, 4]. These incidents made national headlines, the latest in a string of large-scale data breaches [5, 6, 7] that have spurred both the US Congress [8] and the White House [9] 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 recent report by TrendMicro concludes

that the frequency of data breaches has increased since 2009 [10]. A report published the same month by Gemalto, indicates that the total number of breaches increased by 10% while the number of records breached in the first half of 2015 declined compared to 2014 [11]. A 2014 Symantec report noted that there was an increase in the number of large data breaches, and a dramatic 5-fold increase in the number of identities exposed over a single year [12]. 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 2012–2013 [13].

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 10 years using data obtained from a popular dataset published by the Privacy Rights Clearinghouse (PRC) [14]. 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 2013 (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 Fig. 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 “Bayesian Approach” subsection). Our analysis shows different trends for different subsets of the data. We consider two distinct types of breaches: “malicious,” where attackers actively target personal information, and “negligent,” which occur when private information is exposed accidentally (e.g. misplacing a laptop). In the dataset, both the size of malicious and negligent breaches have remained constant over the last 10 years. Similarly, the frequency has also remained constant (see “Modeling Breach Size” and “Modeling Breach Frequency” subsections). While our approach is simple, our univariate model does a remarkably good job of reproducing the distribution of data breaches over time.

Beyond assessing trends, this approach enables us to determine the likelihood of certain future events, at least in the USA (see “Prediction” section). For example, the model predicts that in the next 3 years there is 25.7% chance of another Anthem sized (80 million) breach, and only a 4.0% chance of a Anthem and Home Depot sized breach occurring within a year of each other. However, there is an 75.6% chance of a breach of at least five million records in the next year. The probabilities are relatively high for breaches of five 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 article is identifying the particular forms of the underlying distributions, which may offer insight into the generative processes that lead to data breaches. For malicious breach sizes, we find that the distribution is log-normal (see “Breach Size” subsection); such distributions are known to emerge from multiplicative growth. In fact, the size distribution of companies is best described by a log-normal [15], so we speculate that as a company grows, the number of data records it holds grows proportionally, and breach sizes follow along. We find that negligent breaches are better described by a log-skewnormal distribution [16]. The log-skewnormal distribution is similar to log-normal distribution except it allows for a further skew of the data toward larger breaches. This skew may represent different underlying features of breaches at different organizations. In contrast, the breach frequency for both negligent and malicious breaches 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 “Breach Frequency” subsection). Future investigations could validate the specific nature of the process which generates these distributions.

Some of our results seem counterintuitive 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 [17].

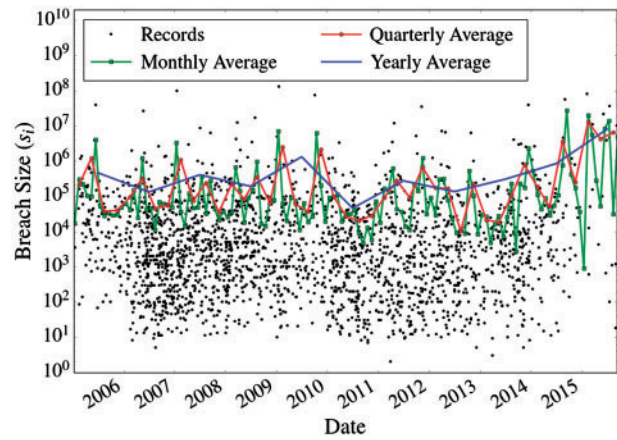


Figure 1. Data breach sizes (records exposed) over a 10-year period. Data taken from [14].

We used the publicly available dataset that we believe is the most complete, but our models could easily be applied to additional datasets, e.g., datasets that are not yet in the public domain or those that may arise if new disclosure laws are passed. Moreover, by establishing a baseline, 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 [18, 19]. Accurately assessing risk, however, requires quantitative measures of likelihood and cost. In this article, 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 “Predicting Future Costs” subsection) 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 \$4 and \$179 billion over the next 3 years.

Data

In this section, we describe the dataset obtained from the PRC and examine the distribution of breach sizes and frequencies. We show that the size distribution is well-fit by a log-normal or log-skewnormal distributions, whereas the daily frequency of breaches is well-fit by a negative binomial. Finally, we show how these distributions are affected when the data are divided into malicious and negligent breaches.

PRC

The PRC is a California nonprofit corporation focused on issues of privacy [20]. The PRC has compiled a “Chronology of Data Breaches” dataset (Available for public download from <http://www.privacyrights.org/data-breach>, 12 August 2016, date last accessed) that, as of 15 September 2015, contains information on 4571 publicized data breaches that have occurred in the USA 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 4571 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 2253 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% [21] and 89% [22] of security incidents go unreported. However, these reports are based on informal surveys of security professionals, their accuracy cannot be confirmed (“Discussion” section), and there is no obvious reason why their size/frequency distributions should differ from PRC.

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, e.g., log-normal, log-skewnormal, power-law, generalized pareto, log-logistic, and log-gamma (Specifically, we tested all of the distributions in the `scipy stats` package that have a domain defined for values greater than 0 (<http://docs.scipy.org/doc/scipy/reference/stats.html#continuous-distributions>, 12 August 2016, date last accessed). 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 different [23]. Figure 2 shows the fit to log-normal; the KS test gives $P = 0.21$, which means that we cannot reject the null hypothesis that the data were generated by this distribution (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.). For all other tested 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 Fig. 2 that the data points in the tail (high values of records) deviate from the best-fit line. We return to this issue in “Discussion” section.

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 [24]. Under this assumption and at steady state, the distribution of entity sizes is known to be log-normally distributed. For example, this process has been used to model the size distribution of companies as measured by annual sales, current employment, or total assets [15]. We speculate that a related process is operating here, if the number of sensitive (customer) records held by a company is proportional to its size, or the number of stored records is increasingly multiplicatively over time.

Breach frequency

We are interested in studying how often breaches occur and whether or not there are any 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 (Fig. 3).

We used a similar approach to the one we employed in “Breach Size” subsection, 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

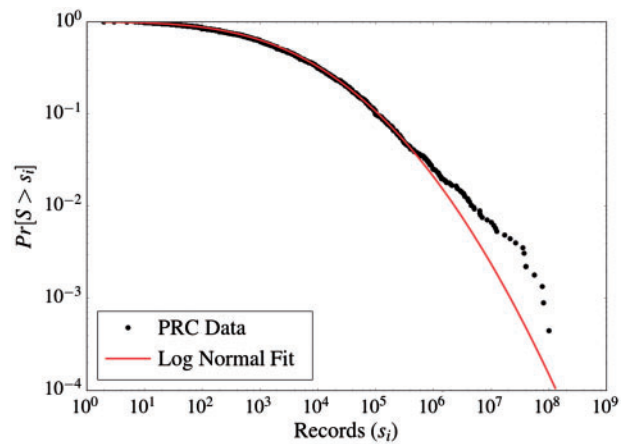


Figure 2. The distribution of breach sizes and the fit to a log-normal distribution.

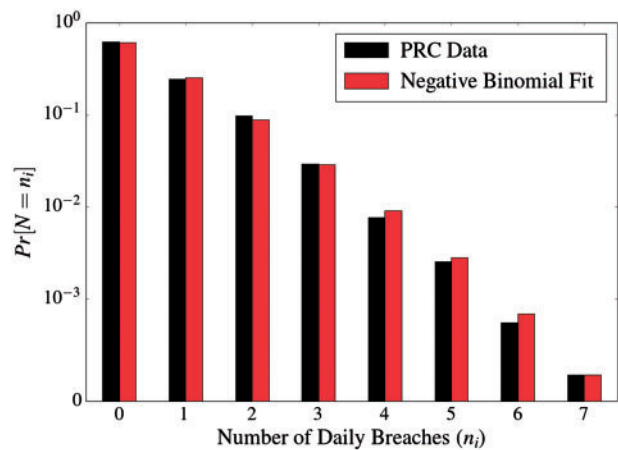


Figure 3. The distribution of the daily number of breaches and the fit to a negative binomial.

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 [25], 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 [26]. 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 [27]. 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. It is also possible that breaches are announced on specific dates to reduce their impact in the media. This could lead to a clustering of breach reports on Fridays or before holidays.

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

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

Breach type	Description	Count
<i>Negligent breaches</i>		
Portable device	Lost, discarded or stolen, portable device or media.	1412
Unintended disclosure	Information posted in a publicly available place, mishandled, or sent to the wrong party.	456
Physical	Lost, discarded, or stolen non-electronic records.	196
Stationary device	Lost, discarded or stolen stationary device or media.	135
<i>Malicious breaches</i>		
Hacking	Electronic entry by an outside party.	469
Insider	Someone with legitimate access intentionally breaches information.	282
Payment card fraud	Fraud involving debit and credit cards that is not accomplished via hacking.	30
Unknown	Other or unknown.	58

but were exposed accidentally, e.g., 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, e.g., 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 reapplied the data fitting analysis described earlier (“Breach Size” and “Breach Frequency” subsections) 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. However, malicious and negligent breaches fit different distributions. Specifically, the sizes of malicious breaches are well fit by a log-normal distribution, while negligent breaches are well fit by a log-skewnormal distribution. Even though the lumped data (all categories aggregated) are log-normally distributed, it is possible that the different distributions arise because this distribution is changing over time, or that different processes are producing different breach sizes. We provide evidence against the former hypothesis in the next section.

Modeling data breach trends

Our earlier analysis does not allow for the possibility that the distributions are changing over time. In this section, we describe how we use Bayesian Generalized Linear Models (BGLMs) [28] to construct models of trends in the PRC dataset. We then use Bayesian Information Criteria (BIC) to determine the highest likelihood model, while avoiding overfitting. We use the distributions derived in “Data” section, as the basis for our time-dependent models.

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-skewnormal in “Breach Size” subsection. Hence $S_n \sim \text{LogSkewNormal}(\mu, \tau, \alpha)$, where μ is the location parameter, τ is the shape parameter (the inverse of the variance), and α is the skew parameter.

To incorporate temporal variations, we model the location parameter, μ , as a polynomial function of time, t , i.e. $\mu = \beta_0 + \beta_1 t + \dots + \beta_d t^d$. Time is expressed as a decimal value in years since 1 January 2005, with a resolution of 1 day, e.g. $t = 1.2$ would be 13 March 2006. We describe how to determine the degree

of the polynomial, d , later. The parameters, β_i , for the polynomial, together with the shape parameter and skew parameter (τ and α respectively), comprise the free variables of the model. For each free parameter we need to define a prior distribution.

The choice of prior distributions is an important and active area of research in Bayesian statistics. As suggested in the literature [28], we used normally distributed priors for the polynomial parameters, $\beta_0 \sim \mathcal{N}(\log(S_n), 1)$ and $\beta_i \sim \mathcal{N}(0, \frac{1}{\text{Var}[t^i]})$, a gamma-distributed prior for the shape parameter, $\tau \sim \text{Gamma}(1, 1)$, and a generalized student’s T distribution for the skew parameter, $\alpha \sim T(2.5, 0, \frac{\beta_0^2}{4})$ [29]. 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 [30]. 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 $d \in [0, 6]$, and choose the model (and hence the polynomial) with the minimum BIC [31]. We compute the BIC as $\text{BIC} = -2L + k * \log(n)$, where L is the likelihood of the model when the parameters are set to their MLE, k is the number of parameters (the degree of the polynomial plus any shape parameters), and n is the number of data points. 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 choose 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. Another common model selection tool is Akaike Information Criteria (AIC), but we obtained the same results using AIC.

To summarize, our modeling approach involves the following steps:

1. Define a BGLM similar to Equation (1), as shown in “Modeling Breach Size” subsection.
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 free parameters (i.e. β_i, τ, α) using MCMC to determine the confidence intervals for the parameters.

- Randomly sample the model to generate a distribution, and compare that to the actual distribution, using the KS test.

We assume that size and frequency are conditionally independent in our model. That is, we assume that the distribution of breach sizes does not change when multiple breaches are disclosed on the same day. To verify this, we conducted a two sample KS test between the distribution of breach sizes on days when n breaches were announced and when m breaches were announced, for all $\{n, m\} \in [1, 7]$ with $n \neq m$. We found no statistically significant difference in each combination of distributions ($P > 0.1$) (recall as before a larger P value indicates that there is not enough evidence to reject the null hypothesis, i.e. the two sets of data are drawn from the same distribution), providing evidence that this assumption is reasonable.

Modeling breach size

As derived in “Bayesian Approach” subsection, the model for negligent breach sizes is

$$\begin{aligned}
 S_n &\sim \text{Log Skew Normal}(\mu, \tau, \alpha) \\
 \mu &= \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_d t^d \\
 \beta_0 &\sim \mathcal{N}(\log(S_n), 1) \\
 \beta_i &\sim \mathcal{N}\left(0, \frac{1}{\text{Var}[t^i]}\right) \\
 \tau &\sim \text{Gamma}(1, 1) \\
 \alpha &\sim T\left(2.5, 0, \frac{\pi^2}{4}\right)
 \end{aligned} \tag{1}$$

For malicious breaches we fit a similar model, except using a log-normal distribution

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

The best fit model for both malicious and negligent breaches, as determined by the minimum BIC, gives $d=0$, which indicates that the distribution of sizes is constant. Figure 4 shows the median values for models, plotted against the PRC data (We show median rather than the mean because it better represents the typical values in heavy tailed distributions.). Maximum likelihood estimates for the parameters are given in Table 2.

To summarize, we find that the distribution of negligent and malicious breach sizes has remained constant with a median size of 383 and 3141, respectively, over the 10-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 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 $P=0.55$ for the fit to the negligent breach sizes, and $P=0.11$ for the fit to the malicious breach sizes.

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 “Breach Frequency” subsection (We also test a

Poisson model, but found it had a higher BIC than a negative binomial model.). The daily frequency, B_n of negligent breaches is given by

$$\begin{aligned}
 B_n &\sim \text{Negative Binomial}(\mu, \alpha) \\
 \log(\mu) &= \beta_0 + \beta_1 t + \beta_2 t^2 + \dots + \beta_k t^k \\
 \beta_0 &\sim \mathcal{N}(\log(D_n), 1) \\
 \beta_i &\sim \mathcal{N}(0, \text{Var}[t^i]) \\
 \alpha &\sim \text{Gamma}(1, 1)
 \end{aligned} \tag{3}$$

The same model is used for malicious breaches, replacing B_n with B_m , the daily number of malicious breaches. We use a log link function for the mean value of the negative binomial distribution, which ensures that the mean value is always positive regardless of the value of the polynomial [28].

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 10 years. The maximum likelihood estimates and 95% confidence intervals are shown in Table 3. Random samples generated using the Equation (3) are not statistically significantly different from the data for both negligent and malicious breaches; which have $P=1.0$ and $P=0.99$, respectively, for the KS test.

Modeling large breaches

It is possible that the models developed above are dominated by smaller breaches, which have experienced little change over the last 10 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 fit a log-normal distribution for both negligent and malicious breaches, and that large breaches in both categories do not show a significant trend over the past 10 years.

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 10 years covered by the dataset.

Alternative modeling approach

Alternative modeling approaches could be taken with the data. ARIMA [32] and GARCH [33] models are frequently used to model financial time series data and could potentially be applied to breach size data or, with modification [34, 35], frequency data. We tested these types of models on log-transformed data breach size. We used the Box-Jenkins methodology [36] to identify the correct order for the ARIMA and GARCH models and found that, using this particular type of model, the parameters that best describe

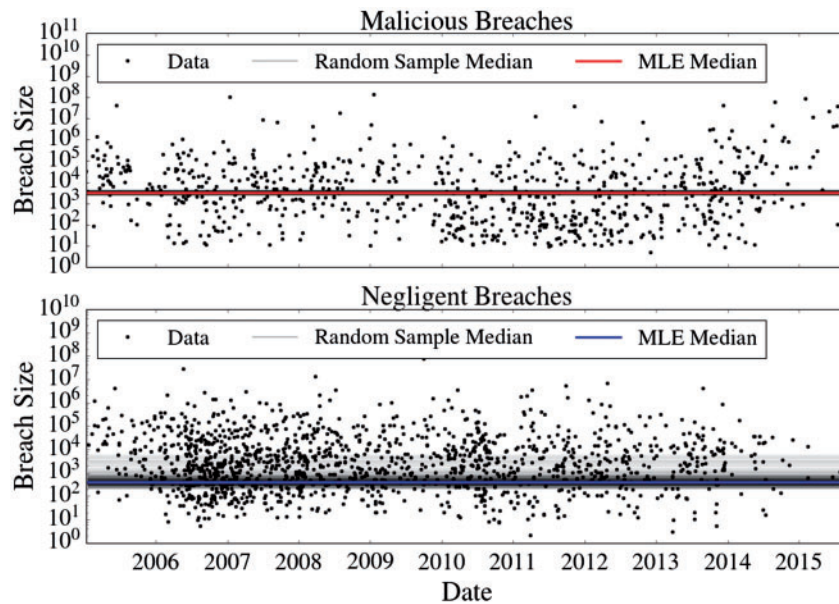


Figure 4. The size of data breaches from the PRC dataset, versus the maximum likelihood estimate of the median size.

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

Variable	Estimate	95% Confidence Interval
Negligent		
β_0	6.186	[5.453, 8.111]
τ	0.098	[0.075, 0.139]
α	0.959	[-0.11, 1.521]
Malicious		
β_0	8.052	[7.827, 8.282]
τ	0.093	[0.084, 0.103]

Table 3. Maximum likelihood estimates and 95% confidence intervals for models of daily breach counts

Variable	Estimate	95% Confidence Interval
Negligent		
e^{β_0}	0.364	[0.343, 0.388]
α	0.944	[0.762, 1.170]
Malicious		
e^{β_0}	0.200	[0.191, 0.216]
α	1.897	[1.157, 3.107]

We report e^{β_0} as this is the mean number of breaches of each type per day.

the data are $ARIMA(0, 0, 0)$ and $GARCH(0, 0)$. That is, a model that does not include autoregressive components, differencing, or moving average components in the conditional mean and no autoregressive components in the residuals or their variance. Under a log transformation, this model is very similar in form to Equation (2) when $k=0$, providing further evidence that the distribution of breach sizes has not changed.

Prediction

The power of a good statistical model is that it can be used to estimate 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 3 years.

Variance and prediction

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, e.g., 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., [37].

We illustrate the effect of the high variability in Figs 5 and 6. These figures show the result of measuring the total number of malicious and negligent breaches and the total number of records contained in those breaches annually for the historical data (black line) and a single simulation using the models presented in “Modeling Data Breach Trends” section (red line) (We use data through 2014 as it was the last complete year we have data. Our 2015 data only runs to September.). Although our model indicates no trend in the size or frequency of breaches, 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 [38, 12]. 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

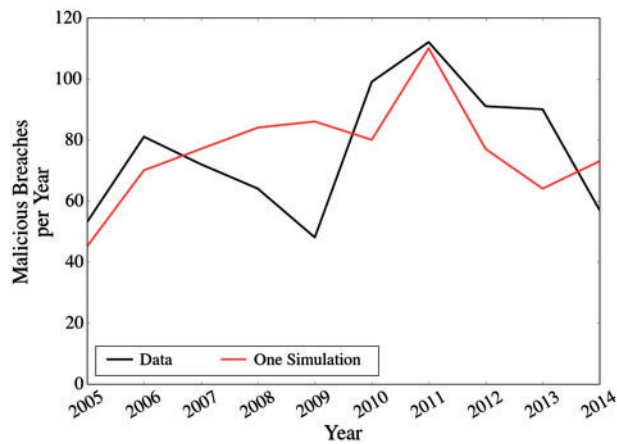


Figure 5. The number of malicious breaches reported each year throughout the dataset, together with a single simulation sampled from our model.

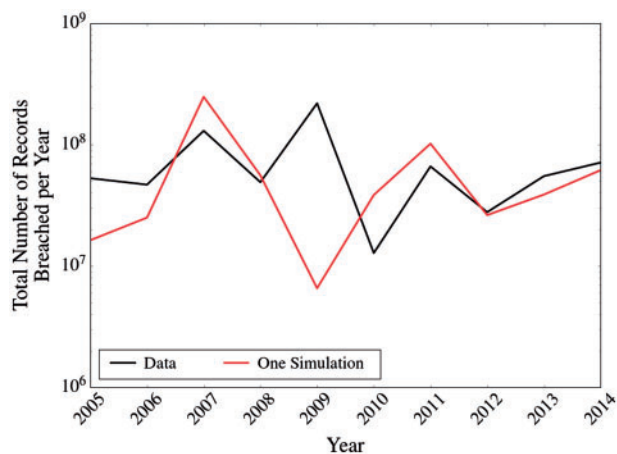


Figure 6. The total number of records breached for each year of data along with simulated total sizes of breaches.

in the number of breaches in the same time frame (These reports use a combination of public and private data, so comparison of exact numbers is not feasible.). 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 [7]. 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 to 2014. We find that a 62% year-to-year increase in the total “number” of breaches is relatively common in simulation, occurring 14.0% of the time. Similarly, an increase of 282% in total “records” exposed occurs in 17.6% 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.

“Predicting” the last year of breaches

To assess the likelihood of the breaches that occurred in 2014, we fit the model using data from 2005 to the September of 2014, and used it to “predict” the events of the last year. 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 and a half years.

We used the models derived from the 2005 to September 2014 data to generate 50 000 simulations of breaches from 15 September 2014 through 15 September 2015. For each day in this simulated timespan we generated a random number of breaches using Equation (3), and then for each simulated breach we generated a random breach size using Equation (1). We plot the cumulative number of records breached in Fig. 7.

The mean cumulative number of breached records roughly matches the actual cumulative number of records up to February of 2015, when the Anthem Breach exposed 80 million medical records. In the next 6 months, Premera/Blue Cross experienced a breach of 11 million health care records, the US office of Personal Management experienced a breach containing 21.5 million records, and Ashley Madison experienced the exposure of 37 million user accounts resulting in a significant increase in the total number of records lost. However, all of these breaches are well within the 95% confidence interval of our model.

As discussed in “Modeling Data Breach Trends” subsection, 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, breaches of the size of the Anthem and Ashley Madison breach occurred in the same year in only 1.08% 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.

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 3 years in the USA. 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 15 September 2015, through 15 September 2018. We predict the probability of several different sizes of breaches. The results can be seen in Figs 8 and 9.

Breaches of 1 000 000 records or more are almost certain (99.32%) within the next year. However, in the next year the probability of exceptionally large breaches decreases quickly, with only a 9.77% chance of an Anthem sized breach in the next year. However, in the next 3 years we can expect to have more large breaches. This is especially clear in Fig. 9, which shows that we are almost certain to see a breach of 10 million records or more in the next 3 years (86.2%), but above that size the probability drops off rapidly, e.g. a breach of size greater than 80 million has less than a 25.7% chance of occurring in the next 3 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.

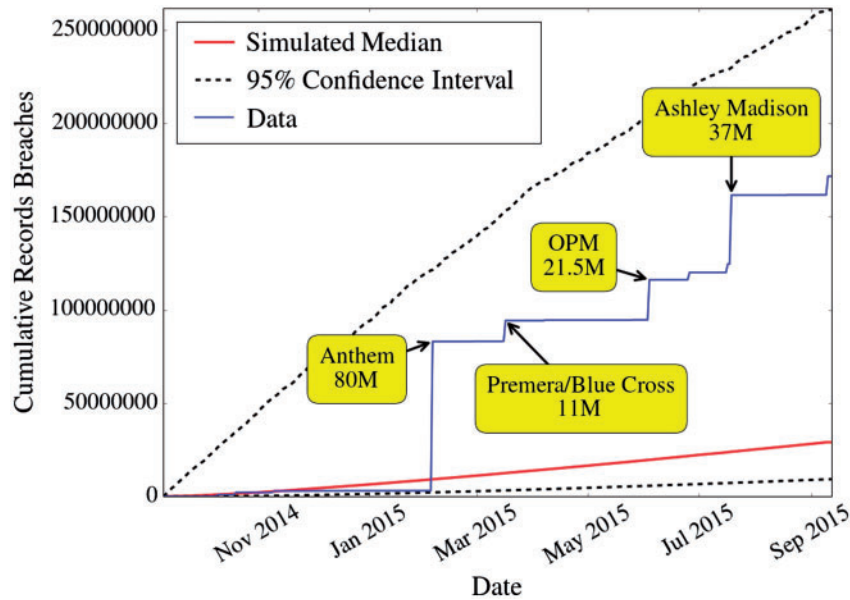


Figure 7. The cumulative number of breached records, both historically (shown in blue) and as predicted by our model. The simulated median (shown in red) is computed over 50 000 independent simulations. The dashed lines represent the 95% confidence interval.

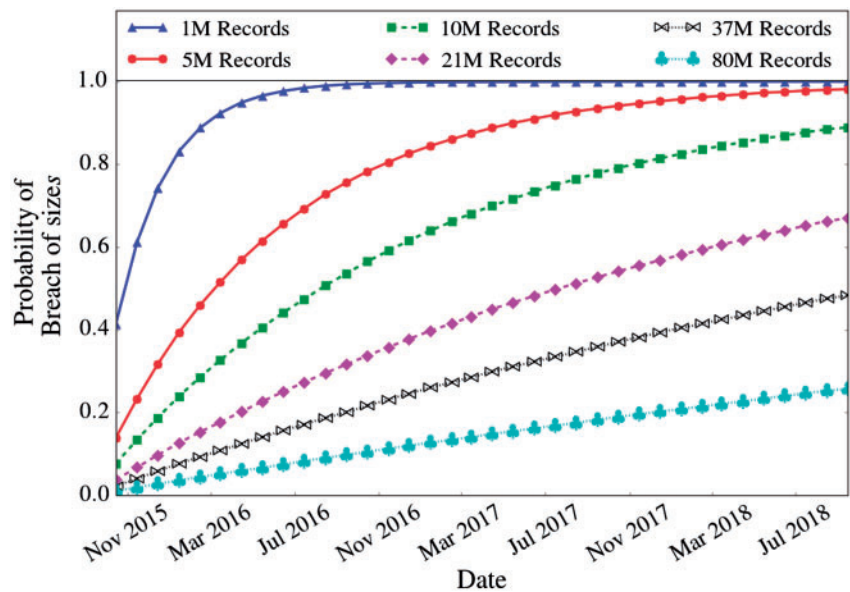


Figure 8. The predicted probability of breaches of various sizes over the next 3 years. Each line represents the probability of at least one breach of the size denoted in the legend occurring before the date on the horizontal axis. We do not include smaller breach sizes, as they will almost certainly occur within the next few months.

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 3 years could be off, we expect the model to work better for the short term. As a demonstration, beginning 15 September 2015 we predict the probability of various breach sizes in the next year and the next 3 years. The exact probabilities are given in Table 4. Thus, we can say with high probability (99.3%) that a breach of at least one million records will occur in the next year, and we do not expect to see a breach equivalent to Anthem (9.77% chance). In the next year, we expect only a 53.6% chance of a breach of 10 million records or more.

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 [39]. Further analysis by others argues that such a flat rate is not the most accurate model for costs. Using nonpublic data, e.g., Jacobs showed that the cost of a breach can be better estimated with a log-log model of the form [40]

$$\log(c) = 7.68 + 0.7584 * \log(s), \tag{4}$$

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

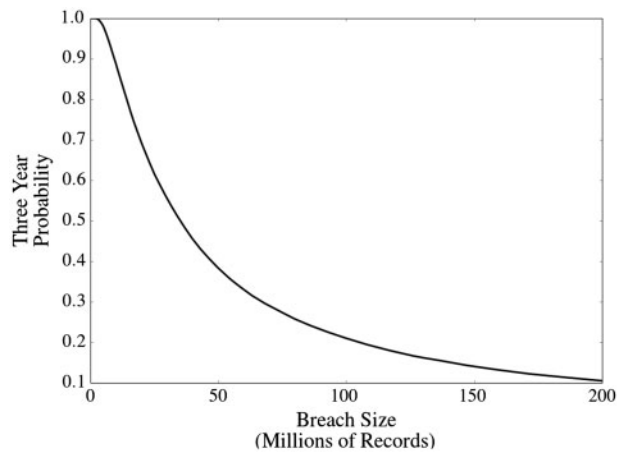


Figure 9. The predicted probabilities of breach size after 3 years.

Table 4. Chance of the occurrence of various size malicious breaches by in the next year and 3 years

Breach size (millions)	% Chance One year	Three years
1	99.3	100
5	75.6	98.2
10	53.6	88.9
21.5	31.6	67.0
37	20.1	48.3
80	9.77	25.7
130	5.82	16.2

The breach size is in millions of records.

In Equation (4) the cost of a breach grows less than linearly, resulting in overall lower costs than those predicted by the Ponemon model. Because the data used to create these models are not public, it is hard to assess their validity, but they illustrate how any cost model can be combined with our results to estimate the future costs of data breaches. Combining these models with Equation (1) and Equation (3) produces the predicted cumulative cost of data breaches over the next 3 years, as shown in Fig. 10.

The flat rate cost model (Ponemon) suggests that in the next 3 years we can expect anywhere between \$8.90 billion and \$179 billion in losses associated with public data breaches. Jacob's model gives a more modest estimate of somewhere between \$3.87 and \$19.9 billion.

Related work

According to the PRC, over 90 reports and articles reference the data used in our study [14]. 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, e.g., the Symantec Threat Report [12] and the TrendMicro report [10] mentioned earlier. Gemalto reports data breach trends, but does not use the PRC data [11]. Another example is a Verizon report released in 2014 [38], 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 [41], focus on classifying the various attack vectors, without attempting to model trends.

Trends in data breaches have received little attention in the academic literature; one exception is Maillart et al.'s analysis of a related dataset [42]. By focusing on the tail of the data their analysis reveals a power-law, which is indistinguishable from the tail of the log-normal distribution we found by considering the entire dataset. Heavy-tailed datasets have also been studied in other domains using similar methods, e.g., [43]. Earlier research investigated trends in the relative frequency of various categories of breaches from 2005 to 2007 but found that the limited sample size prevented them from making statements about the significance of their results [44]. More recently, in 2010, Widup examined yearly trends in different types of data breaches [45]. 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 of security incidents of a specific type [46], and Condon et al. used a time series model to predict security incidents [47]. However, neither of these two studies focused specifically on data breaches.

Numerous reports focus on the health care industry. The US Department of Health and Human Services released a 2014 report examining breaches of protected health information [48]. 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 [49, 50, 13]. 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 [49]. 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 "Predicting Future Costs" subsection. Similar studies focused on hospitals claim that breaches can cost organizations an average of \$2.4 million over the course of 2 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 [51]. Thomas et al. built a branching activity model which measures the impact of information security breaches beyond a breached organization [52]. 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 [53]. Romanosky et al. analyzed the economic and legal ramifications of lawsuits when consumer data is compromised [54]. Later, Romanosky et al. created an abstract economic model to investigate the effect of mandatory data breach disclosure laws [55]. 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 [56].

Discussion

Our results suggest that publicly reported data breaches in the USA have not increased significantly over the past 10 years, either in

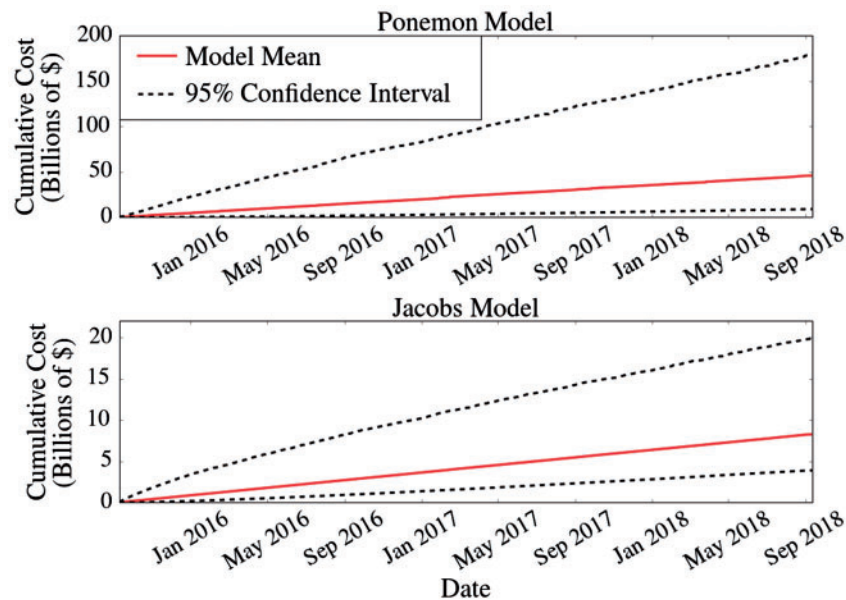


Figure 10. Predicted cumulative cost of data breaches in the next 3 years using two different cost models.

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 [13, 38, 41, 12, 10, 11]. All of these reports lump data into yearly bins, and this amount of aggregation can often influence the apparent trends (Fig. 1).

The idea that breaches are not necessarily worsening may seem counterintuitive. The Red Queen hypothesis in biology [57] 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 [6], and the second largest was even earlier, in 2007 [58]. 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

countries other than the USA, 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 USA, because most US states already have disclosure laws (48 out of 50 as of January 2015 [59]) 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. Moreover, different sectors have different reporting laws. For example, the US Department of Health and Human Services requires hospitals to report breaches of medical information containing more than 500 records [60]. This may lead to an over representation of medical breaches in the data. Future work could use interrupted regression to test whether reporting laws change the rate of reporting [61].

As we described earlier, the data are well-modeled by certain distributions, and these distributions could arise from underlying processes related to the breaches (“Data” section). However, Fig. 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 “Modeling Large Breaches” subsection. 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. The negative binomial distribution we observe in breach frequency could be the result of a mixture of different random Poisson processes. For example, breaches from different organization types on different days of the week may be Poisson distributed with different rates, resulting in the appearance of a negative binomial.

More complex behavioral models unique to security may also provide insight. Models which include processes such as data collection, organizational growth, the deployment of defenses, the capabilities of attackers, and the notification process may produce the distributions we see here. This is a rich area for future work.

Different modeling paradigms such those which model large and small breaches differently may result in better predictions. Models which allow a continuous variation in the frequency of breaches of different sizes may provide further insight [62]. It is also possible that large breaches have become more common very recently, representing a discrete jump in the data, rather than the continuous change used in our models here. Models which account for different days of the week for the frequency of reporting and discrete changes may provide a better explanation for the data.

This article 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, e.g., the size of an organization. Our univariate approach here can serve as a basis for future investigations which incorporate more information on data breaches. For example, 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 US 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. For example, the presence of out of date software or poor security training within an organization may contribute to the likelihood of major data breaches.

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.

Conclusion

It is popular today to frame the cybersecurity problem in terms of risk analysis and management. For example, the US National Institute of Standards and Technology (NIST) has developed and promulgated its cybersecurity framework, which is based almost entirely on the concept of risk assessment [18]. To evaluate these risks, however, requires an accurate assessment of both cost and likelihood. In this article, 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 [57]. 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.

Acknowledgements

The authors would thank Robert Axelrod, Marshall Kuypers, and Cris Moore for their many useful comments on early versions of this article. The authors gratefully acknowledge the partial support of NSF (1518878,1444871), DARPA (FA8750-15-C-0118), AFRL (FA8750-15-2-0075), DOE (DE-AC02-05CH11231), and the Santa Fe Institute. A previous version of this work was presented at the 2015 Workshop on the Economics of Information Security. This paper reports data updated through September 2015.

References

1. Mathews AW, Yadron D. Health insurer anthem hit by hackers. *The Wall Street Journal*, February 2015.
2. Zengerle P, Cassella M. Millions more Americans hit by government personnel data hack. *Reuters*, July 2015.
3. Chen BX. Home depot investigates a possible credit card breach. *The New York Times*, September 2014.
4. Krebs B. Home depot: Hackers stole 53m email addresses. *Krebs on Security*, November 2014.
5. Creswell J, Dash E. Banks unsure which cards were exposed in breach. *The New York Times*, June 2005.
6. Krebs B. Payment processor breach may be largest ever. *The Washington Post*, January 2009.
7. Finkle J. Experian enmeshed in litigation over business that was breached. *Reuters*, April 2014.
8. Covington and LLP B. Data breach notification bills introduced in house and senate. *The National Law Review*, February 2015.
9. Kosseff J. Analysis of White House data breach notification bill. *The National Law Review*, January 2015.
10. Huq N. Follow the data: dissecting data breaches and debunking myths. *TrendMicro Research Paper*, September 2015.
11. Gemalto. 2015 first half review. *Findings from the Breach Level Index*, September 2015.
12. 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 (April 2014, date last accessed).
13. R. Inc. Redspin breach report 2013: protected health information. <https://www.redspin.com/docs/Redspin-2013-Breach-Report-Protected-Health-Information-PHI.pdf> (February 2014, date last accessed).
14. P. R. Clearinghouse. Chronology of data breaches: Faq. <https://www.privacyrights.org/content/chronology-data-breaches-faq> (13 September 2015, date last accessed), 2015.
15. Sutton J. Gibrat's legacy. *J Econ Lit* 1997;35:40–59.
16. Hcine MB, Bouallegue R. Fitting the log skew normal to the sum of independent lognormals distribution. *arXiv preprint arXiv:1501.02344*, 2015.
17. 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> (August 2009, date last accessed).
18. N. I. of Standards and Technology. Framework for improving critical infrastructure cybersecurity. <http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf> (February 2014, date last accessed).

19. Blakley B, McDermott E, Geer D. Information security is information risk management. In: *Proceedings of the 2001 Workshop on New Security Paradigms*, pp. 97–104. ACM, 2001.
20. P. R. Clearinghouse. Mission statement. <https://www.privacyrights.org/content/about-privacy-rights-clearinghouse#goals> (May 2014, date last accessed).
21. Track T. 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> (November 2013, date last accessed).
22. Claburn T. Most security breaches go unreported. *Information Week*, July 2008.
23. Massey FJ, Jr. The Kolmogorov-Smirnov test for goodness of fit. *J Am Stat Assoc* 1951;46:68–78.
24. Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. *Internet Math* 2004;1:226–51.
25. Arnold TB, Emerson JW. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal* 2011;3:34–39.
26. Haight FA. *Handbook of the Poisson Distribution*. New York: Wiley, 1967.
27. Zhou M, Carin L. Negative binomial process count and mixture modeling. *IEEE Trans Pattern Anal Mach Learn* 2013;37:307–320.
28. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press, 2006.
29. Bayes CL, Branco M. Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian J Probab Stat* 2007;21: 141–63.
30. Homan MD, Gelman A. The No-U-Turn Sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J Mach Learn Res* 2014;15:1593–623.
31. Schwarz G *et al.* Estimating the dimension of a model. *Annals of Stat* 1978;6:461–64.
32. Hibon M, Makridakis S. ARMA models and the Box-Jenkins methodology. 1997.
33. Bollerslev T. Generalized autoregressive conditional heteroskedasticity. *J Econometrics* 1986;31:307–27.
34. Brandt PT, Williams JT. A linear Poisson autoregressive model: The Poisson ar (p) model. *Polit Anal* 2001;9:164–84.
35. McKenzie E. Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv Appl Probab* 1986;18:679–705.
36. Brockwell PJ, Davis RA. *Time Series: Theory and Methods*. New York, NY: Springer Science & Business Media, 2013.
37. Edwards B *et al.* Beyond the blacklist: modeling malware spread and the effect of interventions. In *Proceedings of the 2012 Workshop on New Security Paradigms*, pp. 53–66. ACM, 2012.
38. Verizon. 2014 data breach investigations report. http://www.verizonenterprisec.com/resources/reports/rp_Verizon-DBIR-2014_en_xg.pdf (12 August 2016, date last accessed), 2014.
39. 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, date last accessed).
40. Jacobs J. Analyzing ponemon cost of data breach. <http://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/> (December 2014, date last accessed).
41. Trustwave. Trustwave 2013 global security report. <https://www.trustwave.com/Resources/Library/Documents/2013-Trustwave-Global-Security-Report/?dl=1> (12 August 2016, date last accessed), 2013.
42. Maillart T, Sornette D. Heavy-tailed distribution of cyber-risks. *Eur Physical J B-Condensed Matter Complex Syst* 2010;75:357–64.
43. Clauset A *et al.* Estimating the historical and future probabilities of large terrorist events. *Annals Appl Stat* 2013;7:1838–65.
44. Curtin M, Ayres LT. Using science to combat data loss: analyzing breaches by type and industry. *ISJLP* 2008;4:569.
45. Widup S. The leaking vault: five years of data breaches. *Digital Forensics Assoc* 2010. http://static1.1.sqspcdn.com/static/fi/305562/7846948/1279954456270/The_Leaking_Vault-Five_Years_of_Data_Breaches.pdf (12 August 2016, date last accessed)
46. Bagchi K, Udo G. An analysis of the growth of computer and internet security breaches. *Commun Assoc Informat Syst* 2003;12:46.
47. Condon E, He A, Cukier M. Analysis of computer security incident data using time series models. In: *19th International Symposium on Software Reliability Engineering, 2008. ISSRE 2008*, pp. 77–86. IEEE, 2008.
48. U.S. Department of Health and Human Services. Annual report to congress on breaches of unsecured protected health information. <http://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/breachnotificationrule/breachreport2011-2012.pdf> (12 August 2016, date last accessed), 2014.
49. R. Inc. Redspin breach report 2011: Protected health information. http://www.redspin.com/docs/Redspin_PHI_2011_Breach_Report.pdf (February 2012, date last accessed).
50. R. Inc. Redspin breach report 2012: Protected health information. http://www.redspin.com/docs/Redspin_Breach_Report_2012.pdf (February 2013, date last accessed).
51. Acquisti A, Friedman A, Telang R. Is there a cost to privacy breaches? an event study. *ICIS 2006 Proceedings* 2006;94. Page
52. Thomas RC *et al.* 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 (11 March 2013)*, 2013.
53. Picanso KE. Protecting information security under a uniform data breach notification law. *Fordham L Rev* 2006;75:355.
54. Romanosky S, Acquisti A. Privacy costs and personal data protection: economic and legal perspectives. *Berkeley Tech L J* 2009;24:1061.
55. Romanosky S, Acquisti A, Sharp R. Data breaches and identity theft: when is mandatory disclosure optimal? *TPRC* 2010.
56. Graves JT, Acquisti A, Christin N. Should payment card issuers reissue cards in response to a data breach? *2014 Workshop on the Economics of Information Security*, 2014.
57. Van Valen L. A new evolutionary law. *Evolutionary Theory* 1973;1:1–30.
58. Bosworth MH. TJX data breach victims reach 94 million. *Consumer Affairs*, October 2007.
59. N. C. of State Legislatures. Security breach notification laws. <http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx> (January 2015, date last accessed).
60. C. S. 164.400-414. Health insurance portability and accountability act, August 1996.
61. Wagner AK *et al.* Segmented regression analysis of interrupted time series studies in medication use research. *J Clin Pharmacy Therap* 2002;27:299–309.
62. Park E, Lord D. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transport Res Record J Transport Res Board* 2007;2019:1–6.