The True Cost of a Data Breach

Although data breaches are among the most common damaging cyber events, ascertaining their costs remains challenging. These incidents expose sensitive data—including personally identifiable information (PII), medical records and financial records—physically and electronically. Costs can result from lost business, customer turnover, new business acquisition, breach detection, notifications, legal fees, civil or criminal fees and post-breach responses. For chief information security officers (CISOs) who understand how serious these events could be to their organizations, conveying the economic impact of negative cyber events is critical. Many CISOs attempt to leverage cyberrisk quantification (CRQ) methods, typically with the Factor Analysis of Information Risk (FAIR) standard.[1] CRQ can assist in economic impact assessment. However, a major drawback is the availability of representative data.[2] Where there are missing data in publicly available record sets, an approach is needed to appropriately impute loss amounts from the number of lost records.

To fully understand the impact of data breaches and create models to predict future expenses, one must understand the associated costs, both direct and indirect. However, data breaches are underreported for a multitude of reasons. For example, organizations often do not want to disclose breaches that may negatively impact their reputations. Although some industries are mandated to disclose breaches, reporting requirements vary by sector, regulatory environment and event type. In addition, organizations may not know the full extent of a data breach's impact, or they may not detect the breach at all. As a result, information about a data breach may be incomplete or entirely absent. A methodology to extrapolate any missing information is needed.

Researchers reviewed existing breach cost studies and built new models using Advisen's comprehensive data set of historical cyber events. Their analysis resulted in two regression models reflecting changes in the landscape of data breach events since 2019.

Examining the Literature

The Ponemon Institute publishes annual reports on the preceeding year's data breaches and estimates data breach costs through enterprise interviews. To derive the average per-record cost of a data breach, Ponemon researchers divide the total monetary losses by total records breached in the year. For example, the average

NATALIE JORION | PH.D.

Is a principal data scientist at BitSight, a cybersecurity ratings enterprise. She has been involved in the validation and refinement of algorithms used for the financial quantification and risk vectors models.

JACK FREUND | PH.D., CISA, CRISC, CISM, CGEIT, CDPSE, NACD.DC

Is vice president and head of cyberrisk methodology for BitSight. He is a coauthor of *Measuring and Managing Information Risk*, a 2016 inductee into the Cybersecurity Canon, an Information Systems Security Association (ISSA) Distinguished Fellow, a FAIR Institute Fellow, an International Association of Privacy Professionals (IAPP) Fellow of Information Privacy, an (ISC)[2] 2020 Global Achievement Awardee and the recipient of the ISACA 2018 John W. Lainhart IV Common Body of Knowledge Award.

© 2023 ISACA. All rights reserved. *www.isaca.org*

The model estimating cost as a function of annual revenue better predicted direct costs, but it did not better predict indirect costs.

cost from May 2020 to March 2021 was US$4.24 million, or US$161 per record.[3]

This simple model fits the data poorly. Researchers analyzed the Ponemon Institute's data (which are not available publicly) and found that their estimated cost per record explained only a small percentage of the variance in the observed values. For 2013, the price per record yielded an r-squared value of 0.13, and for 2014, the r-squared value was 0.02. In other words, the model explained 13 percent and 2 percent of the variance in the data set.[4]

The Ponemon analyses have several other limitations regarding the small sample size, sampling methods, nonresponse bias and extrapolated cost results by respondents. According to the Cyentia Institute, "A single cost-per-record metric simply doesn't work and shouldn't be used. It underestimates the cost of smaller events and (vastly) overestimates large events."[5] This analysis indicates that other variables are necessary to build a defensible model besides simply calculating data breach cost as a function of the number of records breached.

Researchers have attempted to create additional explanatory models to predict data breach cost, such as a simple linear regression using the Ponemon Institute's data:[6]

> For 2013: (US dollar amount of losses) = 2,330,000 + $107*(Record count)
>
> For 2014: (US dollar amount of losses) = 2,862,000 + $103*(Record count)

The equation for 2013 explains 29 percent of the variance, and the equation for 2014 explains 24 percent of the variance.

A log-log linear regression can also be used to explain 50 percent of the variance: [7]

> log(US dollar amount of losses) = 7.68 + 0.76*log(Record count)

Building on this work, researchers applied a regression model using the Advisen data set (n = 265). They used log(record count), log(enterprise revenue), whether the organization faced previous data breaches, whether the breach was malicious, whether there was a resulting lawsuit, and whether the industry was part of a government, private or public to predict log(US dollar amount of losses).[8] Of these variables, only the record count was significant. This equation explains 46 percent of the variance.

An alternative method of predicting costs using a percentage of annual enterprise revenue was also suggested. The researchers found that most cyber events cost enterprises less than 0.4 percent of their yearly revenue (although they did not assess the fit of this model).[9] The model estimating cost as a function of annual revenue better predicted direct costs, but it did not better predict indirect costs. Although data breaches had an impact on indirect costs such as stock prices, "the trend was isolated and, in general, had minimal impact on annual revenue trends over time."[10]

Cost Factors

Devising a model with more explanatory power requires integrating variables that significantly impact overall breach costs. The Ponemon Institute *Cost of a Data Breach Report 2022* lists multiple factors related to breach expenses, including:[11]

- Number of records breached
- Type of record lost and type of data breach
- Time to contain the breach
- Enterprise size (measured by employee count or revenue)
- Compliance features
- Industry
- Enterprise location
- Organizational maturity posture and system complexity

Some of these variables are easier to measure than others, and many are related. The researchers decided to systematically test a number of variables, which were also a function of the available data, including:

1. Number of records breached
2. Type of record lost (i.e., personally identifiable information [PII], personal financial information [PFI], protected health information [PHI])

© 2023 ISACA. All rights reserved. www.isaca.org

3. Time to contain the breach
4. Enterprise size: revenue, Fortune 500 status, employee count
5. Whether there were legal fees associated with the breach
6. Enterprise industry (e.g., finance or healthcare)

The later the year, the flatter the line, indicating that the cost of breaches, over time, becomes less dependent on the number of records lost.

Advisen Analysis

The Advisen data were filtered to include cases with affected counts (i.e., number of records lost) that occurred after 2012. The research looked exclusively at privacy and data loss cases (n = 62,306). Of those with an affected count, only 1.8 percent had an associated cost. Cost was adjusted for inflation based on data from the World Bank.

A preliminary analysis explored whether the Advisen data were missing at random or not at random. For data missing completely at random (MCAR), the data handling technique has fewer limitations. For data not missing at random (NMAR), any imputations can yield biased results. However, there are also no techniques available to handle data NMAR.

The researchers investigated whether there was a relationship between dependent variables (e.g., year, Fortune 500 status, finance, healthcare sector) and the total amount missing. They determined that there was a statistically significant relationship between total amount missing and the year in which the cyber event took place. There were more associated costs in earlier years, likely because enterprises had more lingering costs associated with older breaches. There was a minor association between missing cases and sector: Healthcare had more missing cases. There was no association with Fortune 500 status. Given these findings, there is a probable difference between the missing and nonmissing samples. The imputation method described herein should be used with this caveat in mind.

The main analysis filtered out cases that had a missing financial loss amount or a financial loss of zero, leading to a total count of 1,101 cases. A linear regression was run to investigate the relationship between records lost and data breach cost in dollars. Generally, the more records lost, the greater the cost to the enterprise. The confidence interval widens as the record count exceeds 100,000, which is likely why the Ponemon Institute analysis excluded those cases from its analysis and classified them as "mega breaches." **Figure 1** is a scatter plot of the two variables log-transformed.

Figure 1 shows that the greater the number of records lost, the wider the confidence interval. The minimum number of records lost had a wide range of associated costs.

The researchers tested different models to see which would have the most explanatory power and adjusted the data breach cost for inflation. The Ponemon Institute's US$180 cost per record model accounted for 8 percent of the variance. A simple linear regression of the two variables produced a similar r-squared value, likely because the relationship between the two variables is not linear. Using the formula[12] resulted in a model explaining only 13 percent of the variance:

Exp(7.68 + 0.76*log(records))

A similar log-log approach modeled to this data resulted in a model accounting for 29 percent of the variance:

log(US dollar amount of losses) =
-3.82 + 0.32*log(Record count)

In addition to the log of record count, the researchers investigated 13 variables in the analysis:

1. Whether there were third-party mitigation fees
2. Whether there were legal proceedings
3. Fortune 500 status
4. Employee count
5. Whether the enterprise was in the healthcare industry
6. Whether the enterprise was in the finance industry
7. Repeater status (whether the enterprise previously had an event)

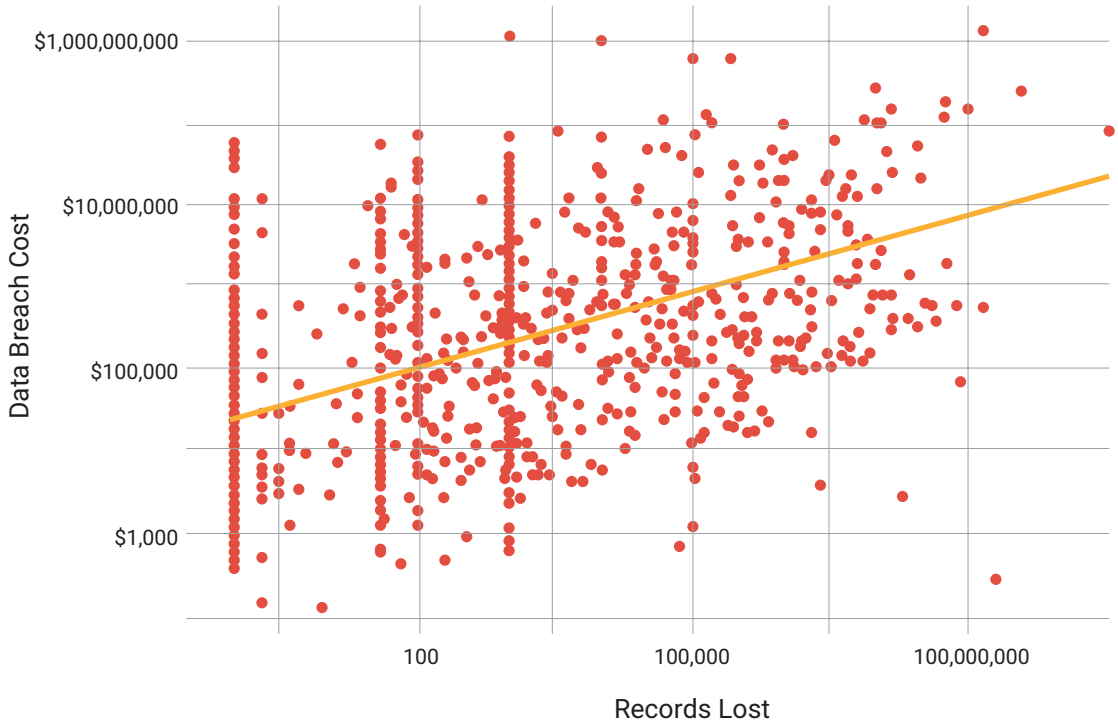
LOOKING FOR MORE?

- Read *Privacy in Practice 2021. www.isaca.org/privacy-in-practice-2021*
- Learn more about, discuss and collaborate on information and cybersecurity in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*

© 2023 ISACA. All rights reserved. *www.isaca.org*

FIGURE 1

Records Lost and Data Breach Costs in US Dollars



8. Whether credit cards were mentioned
9. The log of the enterprise's revenue
10. The difference between accident and discovery date
11. Whether the case involved PII
12. Whether the case involved PHI
13. Whether the case involved PFI

Stepwise regression was used and the model with the lowest Akaike information criterion (AIC) and Bayesian information criterion (BIC) values, which are goodness-of-fit measures that penalize for the number of model parameters, was chosen.[13] The variance inflation factor (VIF) was also checked to ensure that there was no significant collinearity among the variables. The resulting equation explained 36 percent of the variance:

log(US dollar amount of losses) =
-4.7 + 0.3*log(Record count) + 1.0*Legal –
0.2*log(Employee count) + 1.5*Fortune500
Status + 0.6*Finance Industry –
0.4*Repeater Status

Algorithm Refinement

To further refine the model, the researchers looked for major trends in the loss of data to determine what other variables could be used to predict the total cost. It was apparent that the rise in ransomware would likely impact the model. Cases with ransom in the case descriptions for the overall data set (without filters) were flagged. The percentage of cases with ransom in the description spiked in 2020 to 10.82 percent of all cases (**figure 2**).

It was also helpful to determine the relationship between records lost and breach costs by year (**figure 3**).

The later the year, the flatter the line, indicating that the cost of breaches, over time, becomes less dependent on the number of records lost. After 2017, the number of mega-breaches (those with more than 100,000 cases) also declined. Changes in yearly patterns may also be due to regulatory factors. For example, EU General Data Protection Regulation (GDPR) fines rose by 40 percent in 2020 and there has been an increase in the number of lawsuits for negligence since 2020.[14]

© 2023 ISACA. All rights reserved. www.isaca.org

FIGURE 2

Percentage of Ransomware-Related Cases by Accident Year

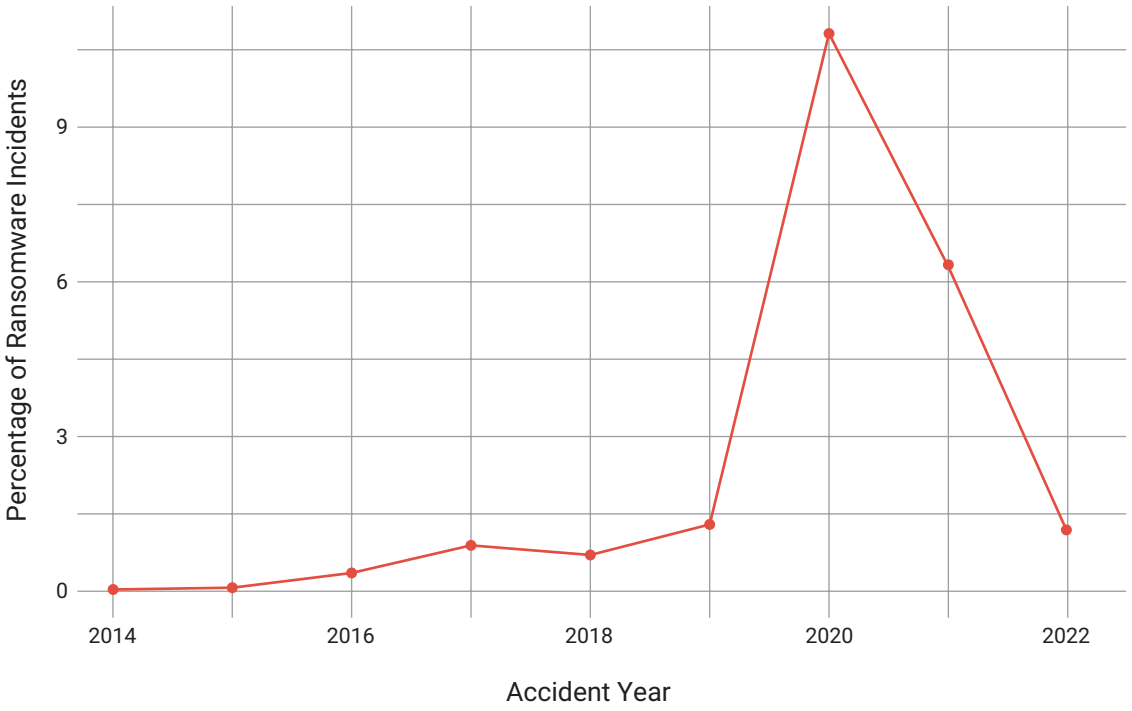
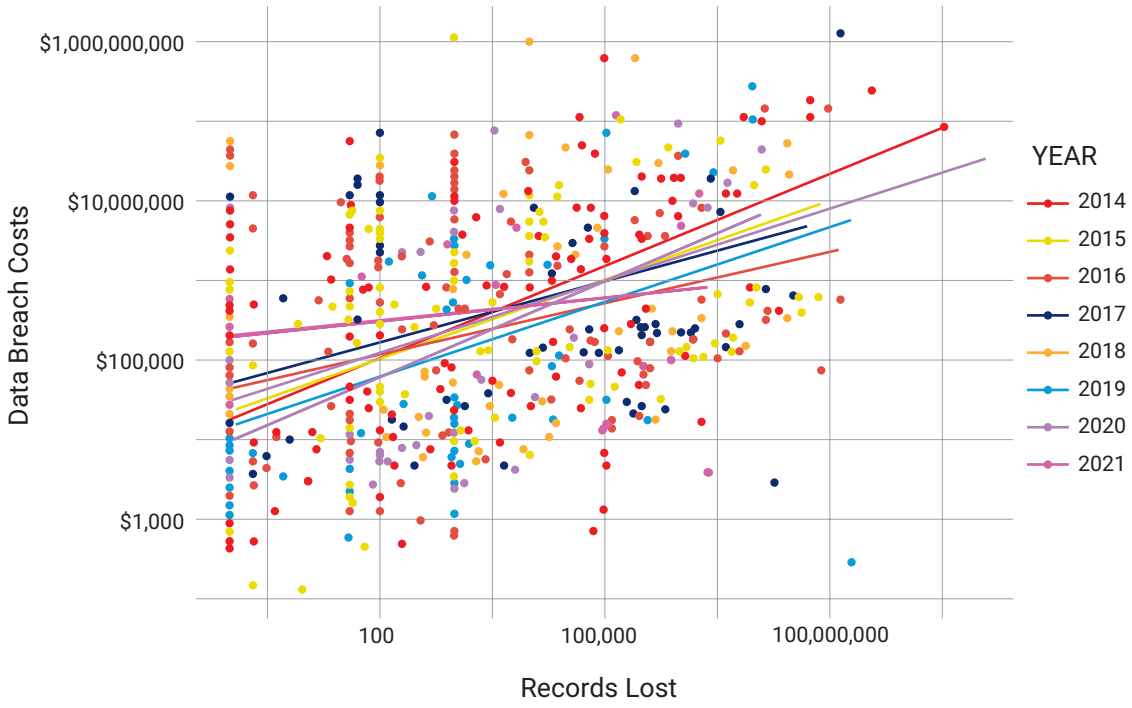


FIGURE 3

Records Lost and Data Breach Costs in US Dollars by Year



© 2023 ISACA. All rights reserved. *www.isaca.org*

As more data become available, these models must be tested, and the fit of additional variables investigated.

The researchers then filtered the data to 2019 and later (n = 164) to account for the increase in ransomware cases and ran a stepwise regression using the same 13 variables used previously. In addition, a ransomware flag variable was added.

The final model for this set explained 42 percent of the variance:

log(US dollar amount of losses) = 9.005 + 0.307*log(Record count) + 0.894*Legal - 0.163*PII + 0.172*log(Employee count)

For cases in the year prior to 2019 (n = 937), the best-fitting model explained 37 percent of the variance:

log(US dollar amount of losses) = 9.6953 + 0.279*log(Record count) + 1.614*Legal - 0.494*PII + 0.877*Fortune500 + 0.321 *Finance + 0.135*log(Employee count)

Discussion, Limitations and Next Steps

The two proposed models are relatively simple, yet they explain a fair amount of variance in the data. These models perform better than the ones proposed by previous researchers, and they encompass a greater span of years. For the most recent years, this model explains more of the variance. Previous research included more predictors, most of which were insignificant.

The fact that the two ranges of dates had slightly different significant variables suggests that there has been a change in the factors that influence the cost of cyber events. The recent rise in ransomware attacks might explain this effect. Fortune 500 and finance enterprises were targeted more frequently in the past, but this seems to have changed in recent years. One possible reason is that these enterprises might have more resources and better safeguards in place against potential data breaches, causing bad actors to target lower-tier enterprises.

There are several pitfalls in using these models to impute values. The more time that has passed since a breach, the more likely it is that the case will have additional costs, such as litigation fees. As the missing data analysis demonstrates, there are significant missing data and differences between some nominal variables in the data set (year and industry). The industry matters in terms of the availability of data and the loss amounts, which are likely a function of the regulatory requirements. Given that the data might not be missing at random, imputing values uniformly can be problematic.

Moreover, the results were not cross-validated because of the limited amount of data; therefore, the model may overfit the data. The missing explanatory variables could increase the variance explained in the model. As more data become available, these models must be tested, and the fit of additional variables investigated. It is also unclear if this model can predict future events.

Conclusion

Estimating data breach costs is not as simple as calculating the cost per record lost. This heuristic has become even less accurate in recent years, especially given the rise of ransomware. The Red Queen Effect in cybersecurity is based on the idea that cybersecurity defense evolves in response to innovation in hacker strategies.[15] Changing hacker strategies will lead to different effects and, ultimately, outdated models. To maximize model accuracy, many factors should be considered when extrapolating potential losses. Further, these factors should be checked for relevance on an ongoing basis.

This research proposes a new way of extrapolating data for loss modeling. Normally, when modeling loss using historic data, cases with missing costs are excluded from the dataset. Sometimes only a small fraction of cases are retained to create these models, resulting in overreliance on a small data set. This research proposes a way to calculate missing costs, resulting in more extracted information and a much larger data set for further model building. Accurately forecasting future losses from historical data sets requires continued discipline around testing models and making adjustments as needed. These proposed models should be subjected to future evaluation. It is only through continued model validation that the industry can advance the maturity of cybersecurity risk management practices.

© 2023 ISACA. All rights reserved. www.isaca.org

Endnotes

1 Blum, D.; L. Voicu; "How FAIR Risk Quantification Enables Information Security Decisions at Swisscom," *ISACA® Journal*, vol. 5, 2020, *https://www.isaca.org/archives*

2 Ebersbach, J.; M. Powers; "Quantifying the Qualitative Technology Risk Assessment," *ISACA Journal*, vol. 5, 2022, *https://www.isaca.org/archives*

3 Ponemon Institute, *Cost of a Data Breach Report 2022*, USA 2021, *https://www.ibm.com/reports/data-breach*

4 Jacobs, J.; "Analyzing Ponemon Cost of Data Breach," Data-Driven Security, 11 December 2014, *https://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/*

5 Cyentia Institute, *Information Risk Insights Study: A Clearer Vision for Assessing the Risk of Cyber Incidents*, USA, 2020, *https://www.cyentia.com/wp-content/uploads/IRIS2020_cyentia.pdf*

6 *Op cit* Jacobs

7 *Ibid*.

8 Romanosky, S.; "Examining the Costs and Causes of Cyber Incidents," *Journal of Cybersecurity*, vol. 2, iss. 2, December 2016, p. 121–135

9 Hershberger, P.; Data Breach Impact Estimation, SANS Institute, USA, 2021, *https://sansorg.egnyte.com/dl/hAT7YW07Sh*

10 *Ibid*.

11 *Op cit* Ponemon Institute

12 *Op cit* Jacobs

13 Akaike, H.; "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, vol. 19, iss. 6, 1974, *https://ieeexplore.ieee.org/document/1100705*

14 ForgeRock, *2021 ForgeRock Consumer Identity Breach Report*, 2021, USA, *https://www.forgerock.com/resources/2021-consumer-identity-breach-report-information-security-breach*

15 Zimmerman, C.; *Ten Strategies of a World-Class Cybersecurity Operations Center*, The MITRE Corporation, USA, 2014

© 2023 ISACA. All rights reserved. *www.isaca.org*