2021

A Monte-Carlo Analysis of Monetary Impact of Mega Data Breaches

Mustafa Canan

Omer Ilker Poyraz
Old Dominion University

Anthony Akil
Naval Postgraduate School

Original Publication Citation

Mustafa, C., Omer Ilker, P., & Anthony, A. (2021). A Monte-Carlo analysis of monetary impact of mega data breaches. *International Journal of Cyber Warfare and Terrorism (IJCWT)*, 11(3), 58-81, Article 5. https://doi.org/10.4018/IJCWT.2021070105

International Journal of Cyber Warfare and Terrorism
Volume 11 • Issue 3 • July-September 2021

A Monte-Carlo Analysis of Monetary Impact of Mega Data Breaches

Mustafa Canan, Naval Postgraduate School, USA
Omer Ilker Poyraz, Old Dominion University, USA
Anthony Akil, Naval Postgraduate School, USA

ABSTRACT

The monetary impact of mega data breaches has been a significant concern for enterprises. The study of data breach risk assessment is a necessity for organizations to have effective cybersecurity risk management. Due to the lack of available data, it is not easy to obtain a comprehensive understanding of the interactions among factors that affect the cost of mega data breaches. The Monte Carlo analysis results were used to explicate the interactions among independent variables and emerging patterns in the variation of the total data breach cost. The findings of this study are as follows: The total data breach cost varies significantly with personally identifiable information (PII) and sensitive personally identifiable information (SPII) with unique patterns. Second, SPII must be a separate independent variable. Third, the multilevel factorial interactions between SPII and the other independent variables elucidate subtle patterns in the total data breach cost variation. Fourth, class action lawsuit (CAL) categorical variables regulate the variation in the total data breach cost.

KEYWORDS

Cyber Attacks, Data Breach, Economics of Cybersecurity, Information, Monte Carlo, Personal Information, PII

INTRODUCTION

Data breach incidents have become a critical risk item in cybersecurity risk assessment. Data security plays an essential role in keeping companies' reputations and avoiding financial fees or litigations. A primary concern of data breaches for companies is severe financial consequences. Recent data privacy laws have enabled government organizations such as the Securities and Exchange Commission and Federal Trade Commission to issue financial fees on companies in case of a data breach. Class-action lawsuits and settlements with the government can exceed a hundred million dollars, evidenced by the Equifax case. The increasing dependency on cyber systems and interdependency among assets makes cyber-attacks a legitimate concern. This dependency put the cyber-attacks one of the top 10 global economic risks (WEF, 2019). As a result of this, to reduce the financial impact of data breaches, cyber insurance has become a way to minimize data breaches' monetary impact.

Quantifying data breaches into a monetary value is a point of interest for insurers and risk managers that they still try to decipher the impact due to the lack of data and latent costs. The monetary impact

DOI: 10.4018/IJCWT.2021070105

of data breaches may exceed hundreds of millions of dollars that can harshly reduce an organization's profit, if not bankrupt them. Therefore, decision-makers and cyber insurance companies need to understand better that loss of information has financial consequences and impacts on business. This increased situational awareness can ameliorate companies' investment strategies in cybersecurity tools and techniques and consider transferring the data breach risk by purchasing cyber insurance. The insurance industry also needs to figure out the probability and impact of data breaches to define premiums and sell cyber insurance.

This study adopts the bifurcated categorization of personally identifiable information (PII) as PII and sensitive PII (SPII) based on Department of Homeland Security definitions (2017) (Poyraz et al., 2020). Hence, the scope of data breaches is limited, with the ones that include PII and SPII. Although there are myriad data breach incidents and a few data breach datasets, there are not enough comprehensive public datasets that shed light on the details of the incidents, such as stolen information, causes, type, and costs. This obscurity precludes decision-makers and insurers from fathoming the multiple implications of data breaches. Thus, they have been struggling to determine companies' cyber risk exposure, and assessing PII and SPII data breaches' monetary impact is crucial for organizations to forecast and manage the risk.

Data breach risk is an integral part of the cyber risk due to the enforcement of governments. Because of multiple cyber risk implications such as monetary loss, business interruption, loss of a customer, and loss of confidential information, organizations have been integrating cyber risk into overall enterprise risk management. Cyber risk must be well understood, and this can be achieved by data categorization that can capture the quirks of the cyber risk.

This study aims to explicate the effects of separate categorization of PII and SPII on the cost of mega data breaches. In this paper, we expand the previous research (Poyraz et al., 2020), which introduced a model to demonstrate the significance of the SPII category, in three aspects. First, a new mega data breach data point has been added to the previously used dataset (Poyraz et al., 2020). Second, using the new dataset, a robust stepwise regression analysis was conducted. Third, using the new dataset and the developed model, a Monte Carlo analysis was conducted to investigate the interaction among independent variables and emerging patterns.

The structure of the paper is as follows. The literature review summarizes the background of this work. The methodology section includes the dataset we utilized, robust stepwise regression, and a predicted R-squared study. The methodology section also includes a Monte Carlo analysis to explain the interaction among the four independent variables. The conclusion part reviews the results and further research directions.

Literature review

Foundations

PII data breaches have been a disturbing concern for individuals due to fraud cases. The finance and healthcare industries are the primary targets of malicious actors because organizations in these industries use, store or transmit individuals' credit card numbers, social security numbers, or passport numbers more often than other industries (Ablon, 2018). PII and SPII are shown in Tables 1 and 2 (Poyraz, et al., 2020).

Recent mega data breach cases such as Marriot, Target, Anthem, Equifax, and Capital One urged the governments, individuals, and media to scrutinize these events to understand the diverse set of impacts. Table 3 shows the five most notorious mega data breaches and their cost to the victim companies.

Data breach attacks may be conducted by individual hackers, insiders, hacktivists, or state-affiliated groups. A study asserts that hackers go after money, fun, challenge, a*nd "to do good in the world."* (Tatar and Celik, 2015). Malicious actors hunt for sensitive personal information to earn

Table 1. Example of PII

PII
Name
Account name/ user ID
Password
Email
Address
Telephone number
Education credentials/certificates
Date/place of birth
Vehicle title number

Table 2. Examples of SPII

SPII
Social security numbers
Medical history
Credit/ debit card numbers
Driver's license numbers
Bank account numbers
Passport numbers
Alien registration numbers
Biometric identifiers
Taxpayer identification number

Table 3. Mega Data Breaches

Company	Number of affected people (in millions)	Cost of Data Breach ($ in millions)
Anthem	78	406.50
Equifax	147	1,445.00
Target	110	310.00
Capital One	106	150.00 – 300.00
Marriot	300 +	220.00

money by selling credit cards and passports or I.D.s, espionage, or profiling purposes for a foreign government (Tatar et al., 2016). The value of the information in the dark web changes depending on the information's sensitivity; for example, passports or diplomas are more expensive.

RELATED WORK

Researchers have been studying data breaches for a while especially focusing on the impact of data breaches on stock prices (Mcshane and Nguyen, 2020). Edwards, Hofmeyr, and Forrest (2016) examine data breach trends focusing on data breaches' size and frequency. Another study claims that the frequency of data breaches is steady, whereas the size of the breach increases over time (Wheatley, Maillart, and Sornette 2016). Eling and Loperfido (2017) analyze the distribution of data breaches and state that data breach modeling should cover different risk types. Carfora et al. (2019) extend Elling and Loperfido's (2017) study by pointing out cyber insurance pricing issues. Another study develops stochastic process models to understand the inter-arrival times and size of the breaches (Xu et al. 2018)

The study of data breach cost is a recent research field and requires a comprehensive understanding of the cost factors. Few practical studies shed light on the cost of data breaches; for example, Jacobs (2014) used the Ponemon Cost of Data Breach information to develop a linear regression model to identify the association between cost and the number of records. He observed heteroskedasticity; therefore, he moved to log-log regression and stated that a 10% increase in the compromised number of records causes a 7.6% increase in the data breach cost. In another study, the Advisen dataset was used to develop a multiple regression model to determine the factors correlated with cost (Romanosky, 2016). Romanosky (2016) claims that a 10% increase in revenue increases the cost by 1.3%; a 10% increase in the number of affected counts increases the cost by 2.9%. Layton and Watters (2014) estimated the tangible costs by using case studies based on a salary guide to the expense of labor hours of employees who took part in dealing with a data breach. The authors calculated labor costs as tangible, whereas loss of reputation as taken into account under intangible cost. A recent study introduced a new categorization of the stolen personal information as PII and SPII (Poyraz et al., 2020). The study claims that the categorization of personal information as PII and SPII can explain the cost variance compared to the affected number of people-based models. By capitalizing on PII and SPII distinction, a more recent study (Poyraz et al., 2020) proposed four new independent variables to calculate the cost of a mega data breach. These independent variables are revenue, PII, SPII, and class-action lawsuits and claims. The analysis conducted in (Poyraz et al., 2020) demonstrates the strong multilevel factorial interactions among the four independent variables.

METHODOLOGY

This study implemented a three-step analysis approach. First, predictive screening was conducted to identify the significant independent variables. Second, a stepwise regression analysis is used to develop the model. Because of the small size of the dataset, in addition to the regression analysis and the adjusted R-squared study, a predicted R-squared study was conducted to monitor overfitting. In the final step, a Monte-Carlo simulation was conducted using steps one and two to examine interactions between the independent variables of the developed model.

DATA COLLECTION

The dataset created by Poyraz et al. (2020) is used by adding one more data point. The dataset is created by integrating the data points from annual financial reports of companies, Privacy Rights Clearinghouse, websites, SEC filings, news media, and case studies. The dataset comprises the following features:

company name, industry, incident year, revenue, total PII, total SPII, a class-action lawsuit, and total cost of the breach. Due to the limited availability of the data, this data set includes 31 incidents, and the number of affected people is more than 1 million. Please see appendix for the details of the dataset.

INDEPENDENT VARIABLE SCREENING

The initial predictive screening is conducted to check the individual contributions of each independent variable. As shown in Table *4*, the cost is the dependent variable, and the four independent variables are PII, SPII, Revenue, and Class-Action Lawsuit (CAL). Two significant independent variables are PII and SPII; the total explanatory contribution of the two independent variables, PII and SPII, is 78%.

The correlation analysis, Figure 1, includes distribution histograms, correlation values (with heat map), and simple regression analysis. Histograms show the distribution of each variable, and the patterned part of each histogram demonstrates the cases, which do not involve a class-action lawsuit. The correlation values and heat maps are included in the upper triangle of Figure 1. The red lines in Figure 1 demonstrate the fit lines for the simple regression analyses. The shaded area around the fit line represents the 95% confidence interval for the linear fit. The R-Squared values for the simple regression analyses of Figure 1 are shown in Table 5. The dependent variable, cost, has the highest correlation values with SPII and PII. Although the correlation values in Figure 1 are not significant among the independent variables, a pairwise correlation probability matrix was created, Table 6. All of the pairwise correlation probability values are significantly higher than the threshold value of 0.05. Therefore, collinearity is not an issue among the independent variables.

STEPWISE REGRESSION ANALYSIS AND PREDICTOR SELECTION

In this analysis, the stepwise regression technique was used. Stepwise regression technique provides an interactive approach to select a different number of predictors; it provides fits for crossed, interaction, or polynomial terms. In this stepwise regression analysis, the JMP statistical analysis software was used with the following step stopping rules: minimum Bayesian Information Criteria (BIC), the Forward Direction Rule, and the Combined Rule. BIC values of models with different predictors were compared using the built-in JMP tools, and the model with the smallest BIC value (Figure 2) was selected. The Forward Direction Rule enters the terms (simple, polynomial, factorial to a degree (=3)) with the lowest p-value. The Combined Rule calculates p-values for two separate tests considering an entry for a term that has precedents.

In addition, Stepwise regression analysis includes a parameter estimate analysis. Depending on the analyst's preferences, polynomial terms can be included in the model. In this study, polynomial degree terms were included. One potential issue with this aspect of stepwise regression is the possibility of overfitting. Although higher-order significant predictors can be the reason for overfitting, these terms are included in the model because of possible interaction effects. Thus, to monitor overfitting in this study, two additional analyses were conducted to monitor overfitting. These are adjusted R-squared and predictive R-squared analyses.

Table 4. Predictive screening analysis results

Predictor	Contribution	Portion	Rank
PII	2.624e+17	0.3918	1
SPII	2.574e+17	0.3845	2
Revenue	1.19e+17	0.1778	3
Class-Action	3.071e+16	0.0459	4

Figure 1. Multivariate Scatter Plot of the dataset. The dashed line represents the simple regression line fit. The shaded area around the fit line represents the %95 confidence interval. The numbers in the upper triangle represent the correlation coefficients.

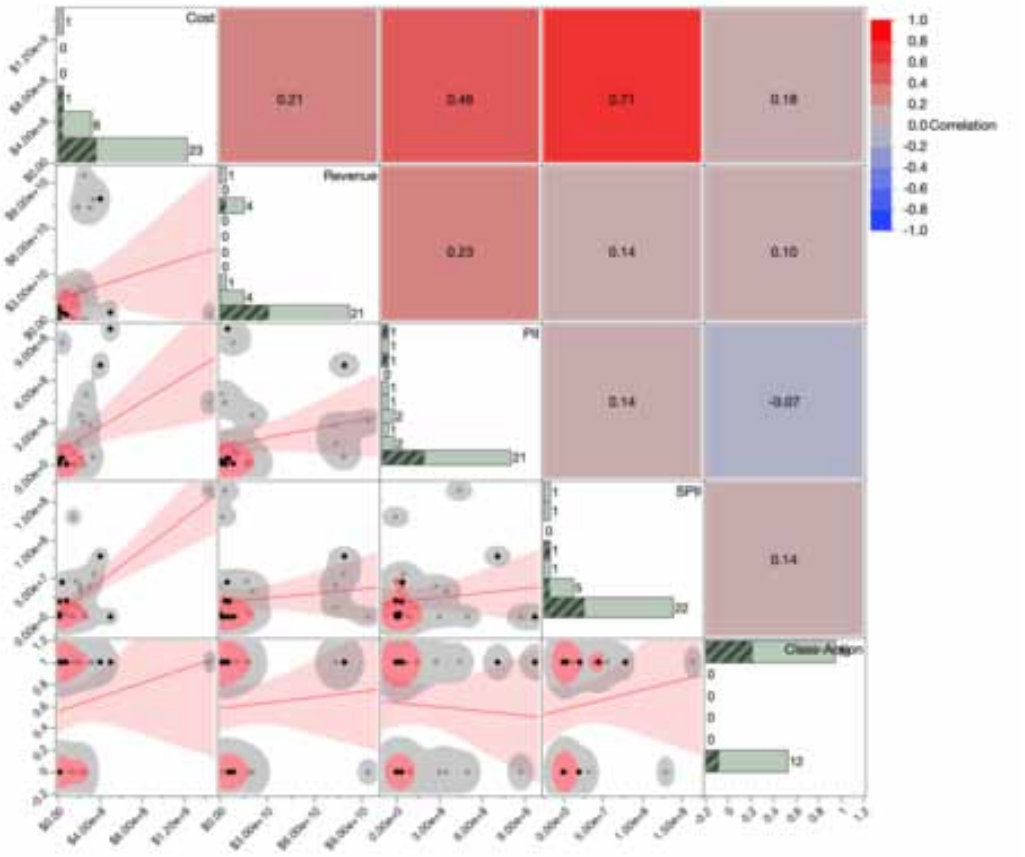


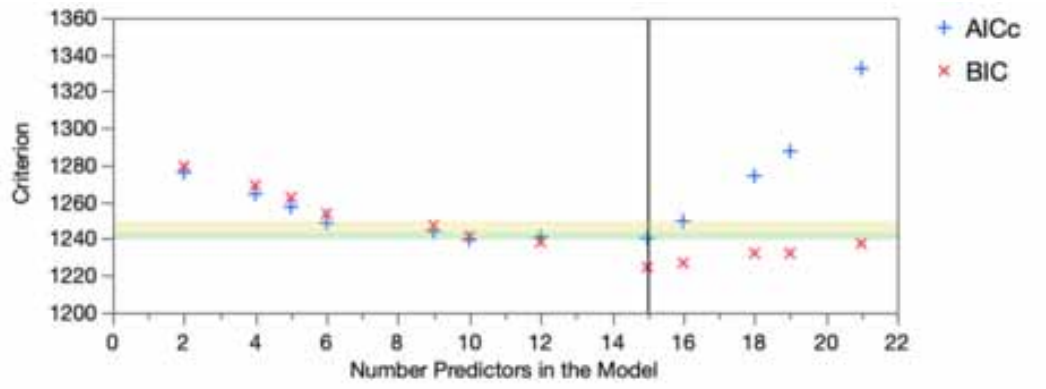
Table 5. Correlation and R-Squared values. The first row shows the independent variables of the study. The variable column includes the dependent variable and three of the independent variables.

Variable	Class-Action		PII		SPII		Revenue	
	Correlation	R-Squared	Correlation	R-Squared	Correlation	R-Squared	Correlation	R-Squared
Cost	0.1775	0.032	0.4629	0.214	0.7148	0.511	0.2111	0.045
Revenue	0.0992	0.010	0.2332	0.054	0.1398	0.020		
SPII	0.1436	0.021	0.1367	0.019				
PII	-0.0694	0.005						

Table 6. Pairwise Correlation Probability

	Cost	Revenue	PII	SPII	Class-Action
Cost	<.0001				
Revenue	0.2542	<.0001			
PII	0.0087	0.2067	<.0001		
SPII	<.0001	0.4533	0.4634	<.0001	
Class-Action	0.3394	0.5955	0.7106	0.4410	<.0001

Figure 2. Criterion History for BIC and Akaike Information Criterion (AIC) values for a different number of predictors. The lowest BIC value was obtained with 15 predictors.



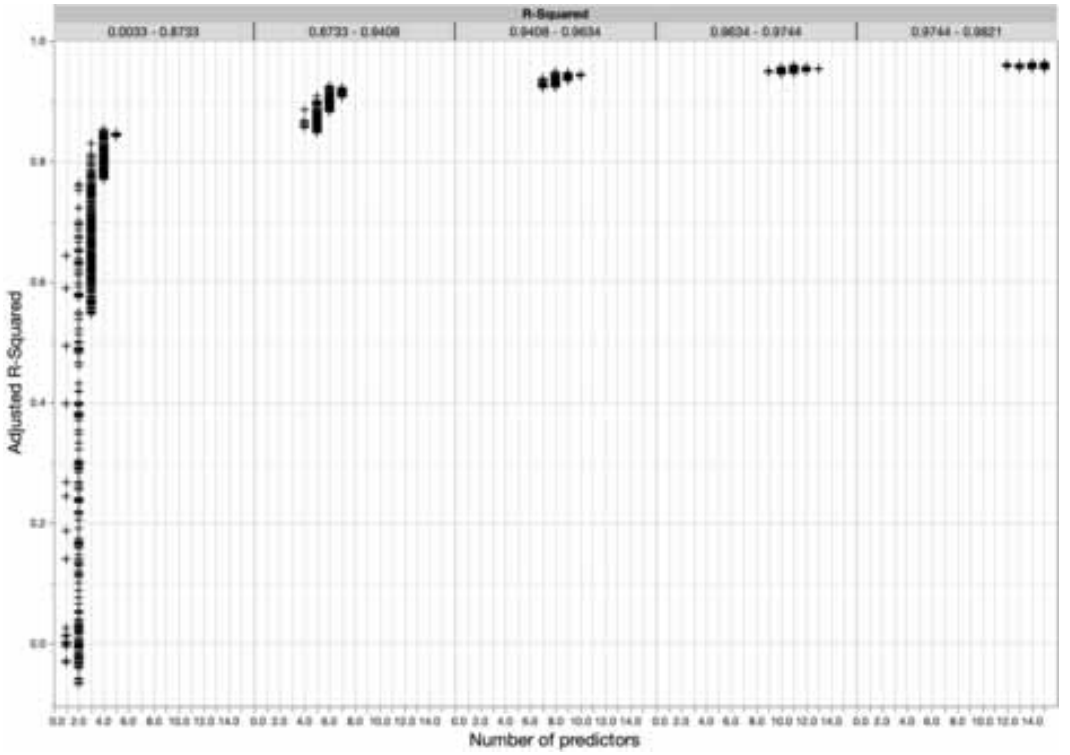
ADJUSTED R-SQUARED

As part of the stepwise regression approach, 5,410 models were created. The maximum number of predictors included in these models is 22 (the developed model has 15 predictors). For each model, the adjusted R-squared was calculated, and the results of these calculations are shown in Figure 3. The adjusted R-Squared was grouped into five R-Squared intervals. In the interval, which includes the R-Squared values of the best model, adjusted R-squared increases. The R-squared value of the fit of the developed model is 0.980324 (Table 7). As shown in Table 7, the new data point increased the values of the fit; all of the metrics, R-squared, adjusted R-squared and predicted R-squared increased. Since the adjusted R-squared increases in the R-squared interval, which includes the R-squared value of the fit (Figure 3), the adjusted R-squared analysis does not demonstrate a possibility of overfitting.

Table 7. Summary of the current analysis and the previous analysis (Poyraz et al., 2020)

	Current Analysis	Previous Analysis
R-Squared	0.980324	0.96
Adjusted R-Squared	0.963108	0.94
Predicted R-Squared	0.8271	0.76
Observations	31	30

Figure 3. Variation of Adjusted R-Squared with the number of predictors in a model



PREDICTED R-SQUARED

Because of this data set's small sample size, machine learning techniques cannot be used to test the model. To maintain the robustness of the fit, in addition to the adjusted R-squared analysis, a predicted R-squared analysis was conducted. For a small sample size, the predicted R-squared technique is used to monitor the possibility of overfitting. Predicted R-squared algorithms systematically remove one observation at a time, and for each step, the algorithm estimates the regression equation and determines how well the reduced model predicts the removed observation. The predicted R-squared value of this analysis (equation shown in Figure 4) is 0.82; this indicates that for the equation in Figure 4, overfitting is not a concern in this analysis. Therefore, based on adjusted R-squared and predicted R-squared results, the dataset, and developed model can be used in a Monte-Carlo simulation study.

REGRESSION ANALYSIS RESULTS

The predictive screening results, multivariate and pairwise correlation analyses indicate that the PII and SPII independent variables have high significance, and the two independent variables, Revenue, and CAL have relatively small significance. These two variables were included in the analysis because factorial predictors with higher significance (Table 8) include these two independent variables.

The dataset of this analysis includes one more observation than the one used in a previous study (*Poyraz et al., 2020)*. Including a new observation improved the significance of the developed model. For example, as shown in Table 7, including this new data point, the R-squared increased to 0.980 (was 0.96), adjusted R-Squared increased to 0.963 (was 0.94), and the predicted R-squared increased to 0.827 (was 0.76).

Figure 4. Prediction expression of the developed model. As can be seen, three of the independent variables are continuous variables, and CAL independent variable appears as Class-Action, which is a categorical variable.

$$
\begin{aligned}
&-420712995.4 \\
&+ 0.0373935603 \bullet \text{Revenue} \\
&+ -1.078805848 \bullet \text{PII} \\
&+ 10.216539152 \bullet \text{SPII} \\
&+ \text{Match}\left(\text{Class-Action}\right)\begin{cases} 0 & \Rightarrow 169723037.23 \\ 1 & \Rightarrow -169723037.2 \\ \text{else} & \Rightarrow . \end{cases} \\
&+ \left(\text{Revenue} - 17832983871\right) \bullet \left(\left(\text{SPII} - 22838877.419\right) \bullet 1.5717526e\text{-}9\right) \\
&+ \left(\text{Revenue} - 17832983871\right) \bullet \text{Match}\left(\text{Class-Action}\right)\begin{cases} 0 & \Rightarrow 0.0358734668 \\ 1 & \Rightarrow -0.035873467 \\ \text{else} & \Rightarrow . \end{cases} \\
&+ \left(\text{PII} - 165991258.06\right) \bullet \left(\left(\text{SPII} - 22838877.419\right) \bullet -8.27996e\text{-}8\right) \\
&+ \left(\text{PII} - 165991258.06\right) \bullet \text{Match}\left(\text{Class-Action}\right)\begin{cases} 0 & \Rightarrow -1.891841987 \\ 1 & \Rightarrow 1.8918419869 \\ \text{else} & \Rightarrow . \end{cases} \\
&+ \left(\text{SPII} - 22838877.419\right) \bullet \text{Match}\left(\text{Class-Action}\right)\begin{cases} 0 & \Rightarrow 9.972561545 \\ 1 & \Rightarrow -9.972561545 \\ \text{else} & \Rightarrow . \end{cases} \\
&+ \left(\text{Revenue} - 17832983871\right) \bullet \left(\left(\text{SPII} - 22838877.419\right) \bullet \text{Match}\left(\text{Class-Action}\right)\begin{cases} 0 & \Rightarrow 1.5767784e\text{-}9 \\ 1 & \Rightarrow -1.576778e\text{-}9 \\ \text{else} & \Rightarrow . \end{cases}\right) \\
&+ \left(\text{PII} - 165991258.06\right) \bullet \left(\left(\text{SPII} - 22838877.419\right) \bullet \text{Match}\left(\text{Class-Action}\right)\begin{cases} 0 & \Rightarrow -6.492021e\text{-}8 \\ 1 & \Rightarrow 6.492021e\text{-}8 \\ \text{else} & \Rightarrow . \end{cases}\right) \\
&+ \left(\text{PII} - 165991258.06\right) \bullet \left(\left(\text{PII} - 165991258.06\right) \bullet 2.2964654e\text{-}9\right) \\
&+ \left(\text{PII} - 165991258.06\right) \bullet \left(\left(\text{PII} - 165991258.06\right) \bullet \left(\left(\text{PII} - 165991258.06\right) \bullet -4.06213e\text{-}18\right)\right) \\
&+ \left(\text{SPII} - 22838877.419\right) \bullet \left(\left(\text{SPII} - 22838877.419\right) \bullet 8.3636476e\text{-}8\right)
\end{aligned}
$$

Table 8 shows the model's predictors' fit parameters and p-values; all of the predictors have significant p-values –less than the significance threshold value (0.05). The notable improvement in this study (comparing the findings of (*Poyraz et al., .2020)* is that the p-values of the intercept and CAL have significantly improved, shown in Table 8, both of them are less than 0.05. This

Table 8. Stepwise regression Parameter estimates for all of the predictors of the model

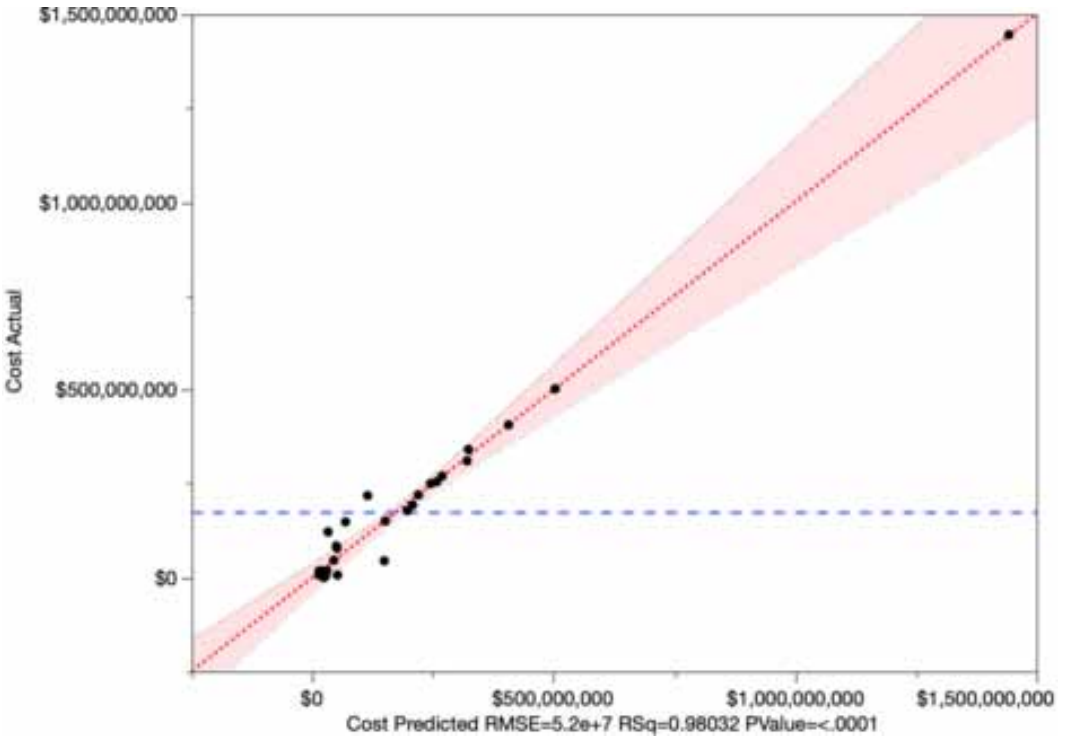
Term	Estimate	P-value
Intercept	-4.207e+8	0.0002
Revenue	0.0373936	0.0005
PII	-1.078806	0.0286
SPII	10.216539	0.0016
Class-Action	169723037	0.0107
Revenue*SPII	1.5718e-9	0.0008
Revenue*Class-Action	0.0358735	0.0008
PII*SPII	-8.28e-8	0.0012
PII*Class-Action	-1.891842	0.0010
SPII*Class-Action	9.9725615	0.0023
Revenue*SPII*Class-Action	1.5768e-9	0.0007
PII*SPII*Class-Action	-6.492e-8	0.0065
PII*PII	2.2965e-9	0.0416
PII*PII*PII	-4.06e-18	0.0077
SPII*SPII	8.3636e-8	<.0001

analysis supports the proposed additional PII category that is SPII; the predictors that include SPII independent variable contribute to the explanation of the variation of the total data breach cost with higher significance. For example, the p-value of SPII*SPII is less than 0.0001 whereas, the p-value of SPII is 0.0016 (Table 8).

The details of the prediction equation are shown in Figure 4; three of the independent variables are continuous variables, and CAL independent variable is shown as Class-Action, which is a categorical variable. Total cost vs. predicted cost values are shown in Figure 5, in which 27 of the 31 observations are in the 95% confidence interval. The blue horizontal line demonstrates the mean total actual cost, which is $172.6 million. The diagonal line represents the values where the Actual Cost = Predicted cost. The shaded band in Figure 5 represents the significance test's confidence level at the 0.05 level.

An additional analysis of the model's fit in Figure 4 was conducted by examining the total cost residual (= actual total cost – predicted total cost). Figure 6 shows the distribution of the residual total cost and normal quantile plot. The skewness of the residual total cost distribution is 0.64. This indicates that some of the residuals are not in the 95% confidence interval. This can be seen in the residual quantile plot in Figure 6; four data points are outside the confidence interval. These four data points are Equifax, eBay, Excellus Blue Cross Blue Shield, and JP Morgan Chase. The studentized residuals, Figure 7, of this analysis indicates that these four data points are within the acceptable limits and are not outliers. Therefore, although the residual total cost distribution cannot be claimed to be normal because of the skewness value of 0.64, the studentized residuals indicate that the model's accuracy is reliable. In the preceding study (Poyraz et al., 2020), the residual distribution skewness was 0.51. Since the newly added data point is not among the four data points outside of the confidence interval in Figure 5, with this dataset, it is not possible to scrutinize the possible causes; more data points are required for further examination.

Figure 5. Actual Cost vs. Predicted Cost. The red dotted diagonal line represents where the values of actual and predicted costs are equal. The dashed blue line represents the mean of the actual cost. The shaded area represents the 0.95 confidence interval. The predicted R-Squared of this fit is 0.8271.



INTERACTION PROFILER

The result of the stepwise regression analysis consists of predictors, which may include independent variables that interact. Interaction means that two or more independent variables can describe the dependent variable's behavior with higher significance. An interaction study was conducted to better understand interacting independent variables' effect on the variation of the total cost. Figure *8* shows the matrix of interaction plots. Each cell in Figure *8* shows the interaction of the row effect with the column effect. The horizontal axis is scaled for effect displayed in each column. In Figure *8*, a plot (cell) illustrates the interaction of the row effect with the column effect. Line segments in each cell show the interaction of that effect with the corresponding row's effect. Non-parallel line segments in the cells of Figure *8* are indicators of interaction.

A red line in an interaction plot corresponds to the lower limit of that column's independent variable; a blue line in an interaction plot corresponds to that row's independent variable's upper limit. These upper and lower values are shown in Table 9. For example, the bottom row first interaction plot from the left represents the response of the cost for two values of the CAL independent variable. The red line demonstrates how cost varies while revenue increasing and CAL is 0; the blue line represents how cost varies while revenue increasing, and CAL is 1. In the case of CAL=1, the total cost does not increase with revenue as it increases in the case of CAL=0.

As shown in the second-row first interaction plot from the left, the line segments are parallel. This is an indicator that these two independent variables do not interact. This supports the findings of the stepwise regression, which is shown in Table *8*. Within the p-value limit of this study, there is no PII*Revenue predictor in the regression equation (Figure *4*). On the other hand, the third-row second

Figure 6. Normal quantile plot and residual distribution. The dashed line is the Normal 2 Mixture fit curve; the dotted line is the Normal fit curve. The skewness of the normal distribution is 0.65.

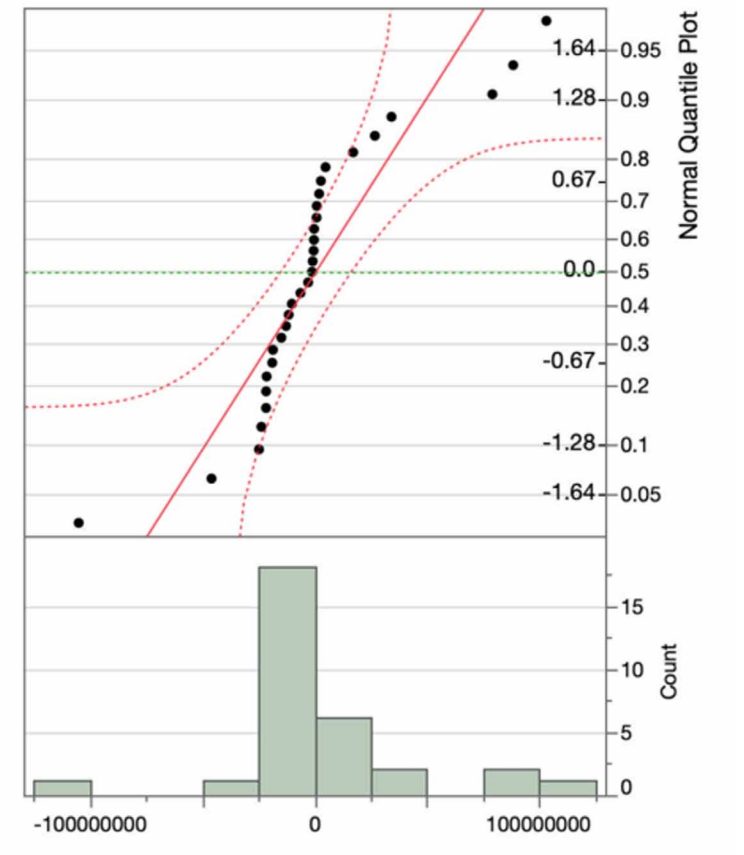


Figure 7. Studentized residuals with 95% simultaneous limits (Bonferroni) and individual limits are represented with red dotted and green dashed lines, respectively. Instead of company names, the row numbers in the dataset were used.

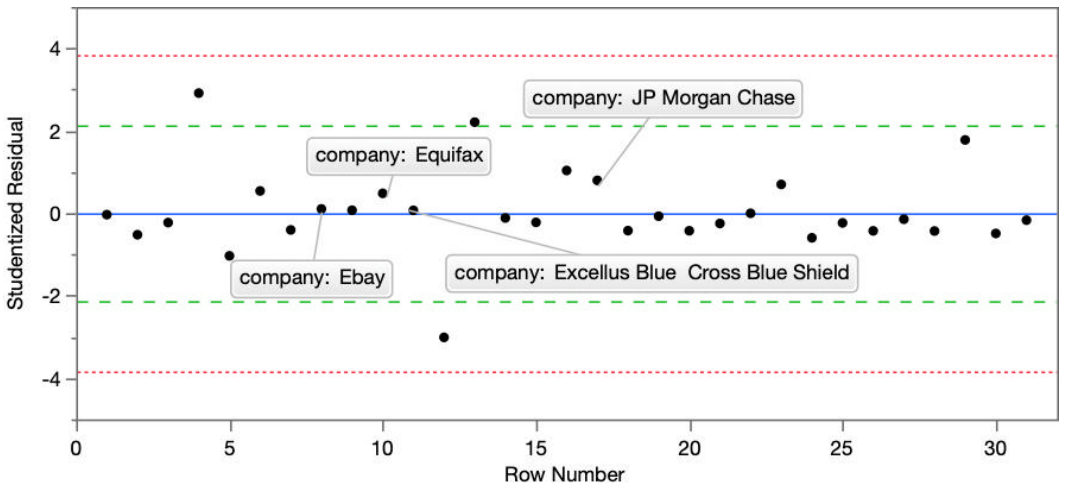
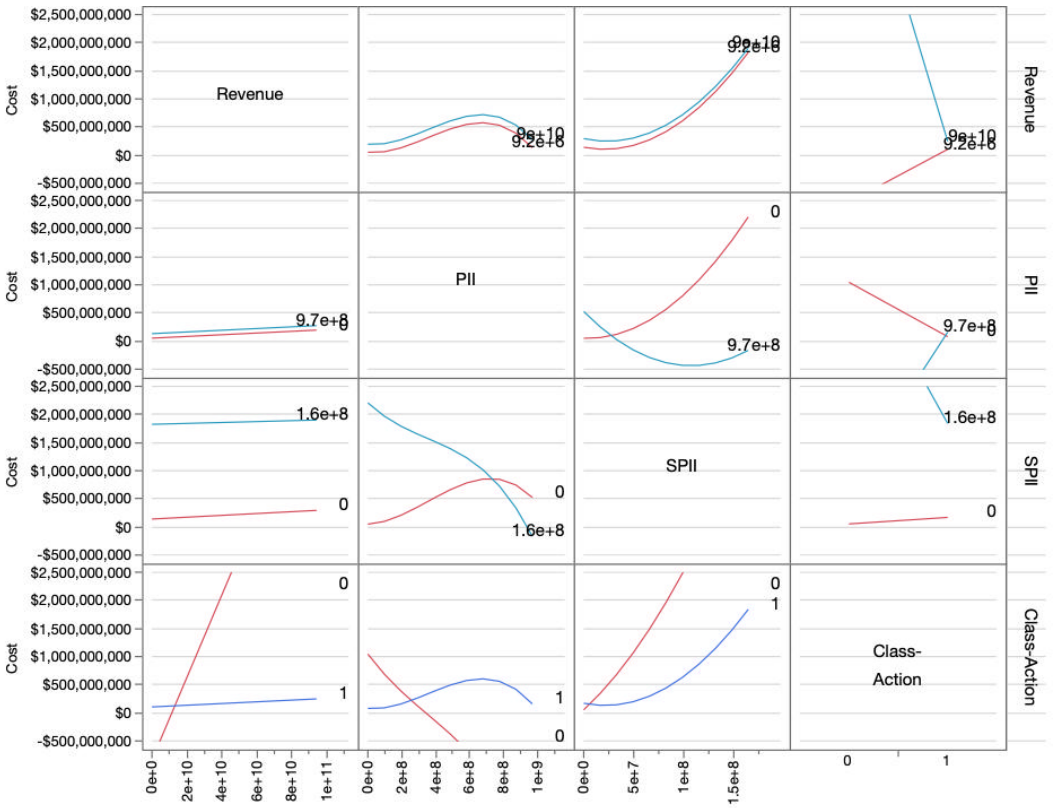


Table 9. Factor Settings for interaction profiler (Figure 8).

Factor	Revenue	PII	SPII	CAL
Minimum	$9.2x10^6$	0	0	0
Maximum	$9.4x10^{10}$	$9.7x10^8$	$1.64x10^8$	1

Figure 8. Interaction Profiler for all of the independent variables in this analysis.



interaction plot from the left demonstrates the interaction effect between SPII and PII independent variables. This cell includes non-parallel line segments, which is an indicator of interaction effects between these two independent variables. As shown in Table *8*, the PII*SPII predictor is significant, with a p-value = 0.0012.

SIMULATION

Monte Carlo simulation is a powerful tool to estimate expected values, quantiles, and any other properties of interest in the model. The Monte Carlo approach capitalizes on the independent random sampling method. There are four independent variables in this analysis. Three of these independent variables (Revenue, PII, and SPII) are continuous, and Class Action Lawsuit is a categorical variable. This Monte Carlo simulation was conducted to examine the distinction between SPII and PII information categories.

DISTRIBUTION ANALYSIS

As part of the Monte Carlo simulation study, a distribution fit analysis was conducted. For each distribution, the minimum AIC criterion is applied for the best-fitted distribution selection. The distribution of the Revenue independent variable can be seen in Figure *9*, the selected best fit is Weibull, and its AIC measure is 1500.7882. Figure *10* demonstrates the distribution of the PII, and Figure *11* demonstrates the SPII distribution. The selected best fit for these two independent variables is Johnson Sb. The parameter estimates for each fit can be seen in Table *10*. The CAL independent variable is included as a categorical variable in the simulation analysis. Therefore, instead of having fit parameter estimates, a probabilistic distribution is included in the simulation analysis. Using the built-in JMP tools, error terms were added to the simulated response based on the developed model (Figure *4*).

SIMULATION RESULTS

By using the random sampling approach, 2 million events were generated. These events include negatives responses and negative values for the three continuous independent variables. Negative values were removed after running the full simulation; these negative valued events were filtered by

Figure 9. Revenue Distribution is best described by Weibull distribution; the fitted distribution measures are AIC =1500.7882, BIC =1503.2276.

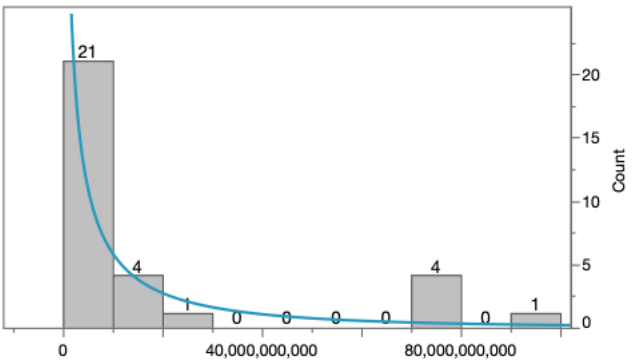


Figure 10. PII Distribution is best described by Johnson Sb distribution; the measures of fitted distribution are AIC =1199.991, BIC = 1204.1885.

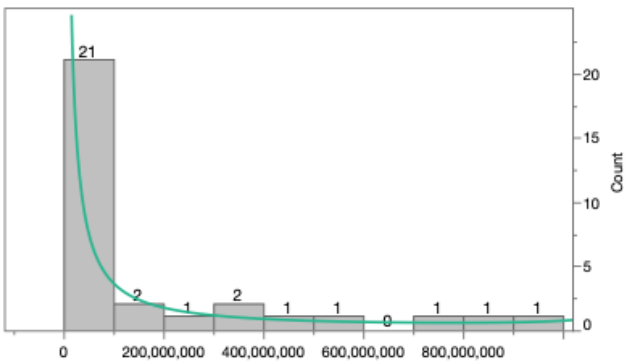


Figure 11. SPII distribution is best described by Johnson Sb distribution; the measures of this fitted distribution are AIC 601.55232, BIC = 605.704981

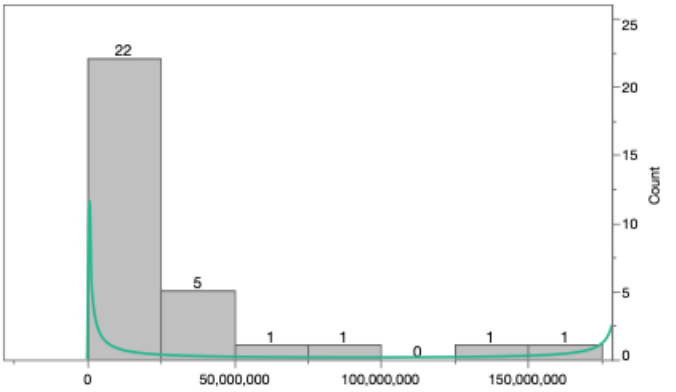


Table 10. Fit Parameter Estimates for three of the independent variables: Revenue, PII, and SPII. These parameter estimates were used in the Monte-Carlo simulation

Revenue (Weibull)		PII (Johnson Sb)		SPII (Johnson Sb)	
Parameter	**Estimate**	**Parameter**	**Estimate**	**Parameter**	**Estimate**
Scale α	9.9472e+9	Shape γ	1.1482802	Shape γ	0.5666907
Shapeβ	0.5170113	Shape δ	0.2991946	Shape δ	0.029366
		Location θ	-123717.9	Location θ	-1.49e-8
		Scale σ	1.0992e+9	Scale σ	181459750

creating a subset of initially generated events. After filtering, the total number of events that were included in the analysis is 1,079,930.

Figure 12 shows the total cost variation for all three continuous independent variables for the categorical variable's two values. The dependent variable, total cost, responds to the changes of independent variables differently with and without CAL (Class-Action). For a uniform scaling, In Figure 12, the CAL=0 case total cost axis is set to the same maximum value. The response of the total cost varies significantly with the presence of the CAL. The summary statistics and quantiles are shown in Table *11* and Table *12*. Two salient interpretations are as follows. First, although SPII and PII categorically close independent variables, their effects on the response variable are unique. Second, the CAL independent variable regulates the response; thus, prediction models can be significantly improved by including predictors that include CAL and interacting independent variables.

Figure 13 shows the total cost variation with PII and SPII independent variables for different revenue intervals. The first two columns from the left show the CAL=0 case; the last two columns show the CAL=1 case. Figure 13 shows that CAL's presence regulates the dependent variable's response to independent variables' changes. In the CAL=1 column, each column has the same pattern for all of the revenue intervals. For example, the third column from the left, CAL=1, independent variable = SPII, starts with a decrease; after the SPII value 40M, the total cost increases as the number of SPII increases; whereas, the total cost decreases as the number of PII increases. Although, in the case of CAL=0, the total cost increases with an increase in SPII and decreases with an increase in PII. The distinct patterns observed in the CAL=1 column are not observed in the CAL =0 case. The response of the total cost is significantly different with PII and SPII independent variables.

Figure 12. Monte-Carlo simulation results based on the prediction equation in Figure 4, and simulation is smeared by random noise based on the model

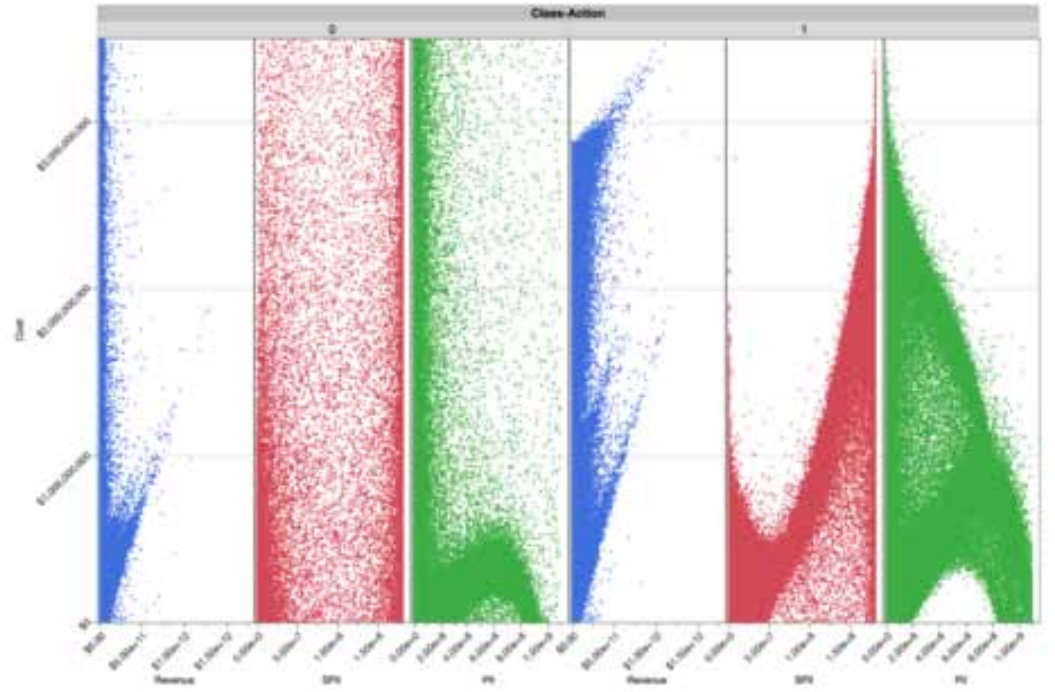


Table 11. Quantile and distribution statistics of the total cost when the Class Action Lawsuit (CAL) =1

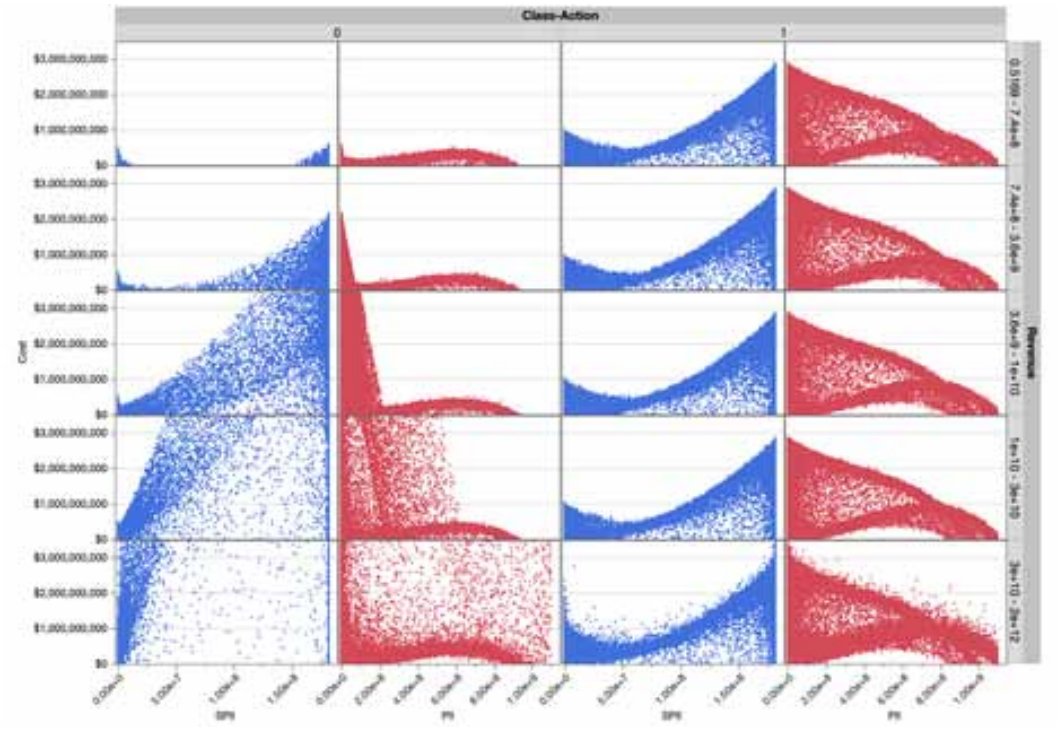
Quantile			Statistics	
100%	MAX	$3,562,163,088.3	Mean	$1.1154e+9
90%		$2,666,775,918.6	Std Dev	$1.1036e+9
80%		$2,576,465,720.8	Std Err Mean	$1,298,024.1
70%		$2,255,115,106.7	Minimum	$372.26916
60%		$1,175,974,758.3	Maximum	$3.5622e+9
50%	Median	$576,730,505.77	Median	$576,730,506
40%		$242,691,709.1	Geometric Mean	$404,858,122
30%		$122,002,755.07		
20%		$73,360,528.033		
10%		$37724874.72		
0%	MIN	$372.2691553		

The PII column in the CAL=1 section of Figure 13 demonstrates three overlapped patterns. As discussed earlier, SPII and PII are two interacting independent variables (Figure 8). The existence of the CAL independent variable accentuates the interaction pattern between SPII and PII. As shown in Figure 14, in the case of CAL=1, the variation of the total cost demonstrates different patterns. As the number of SPII increases, the initial increase of the total cost with the PII independent variable's

Table 12. Quantile and distribution statistics of the total cost when the Class Action Lawsuit (CAL) =0.

Quantile			Statistics	
100%	MAX	$606,892,201,832.0	Mean	$5.2427e+9
90%		$13,776,318,589	Std Dev	$1.7e+10
80%		$4,148,342,502.9	Std Err Mean	$28,452,466
70%		$989,319,275.85	Minimum	$2,771.4592
60%		$284,177,129.89	Maximum	$6.069e+11
50%	Median	$162,451,047.64	Median	$162,451,048
40%		$108,697,727.63	Geometric Mean	$333,840,296
30%		$77,603,209.877		
20%		$52,565,266.412		
10%		$28,125,254.813		
0%	MIN	$2,771.4592247		

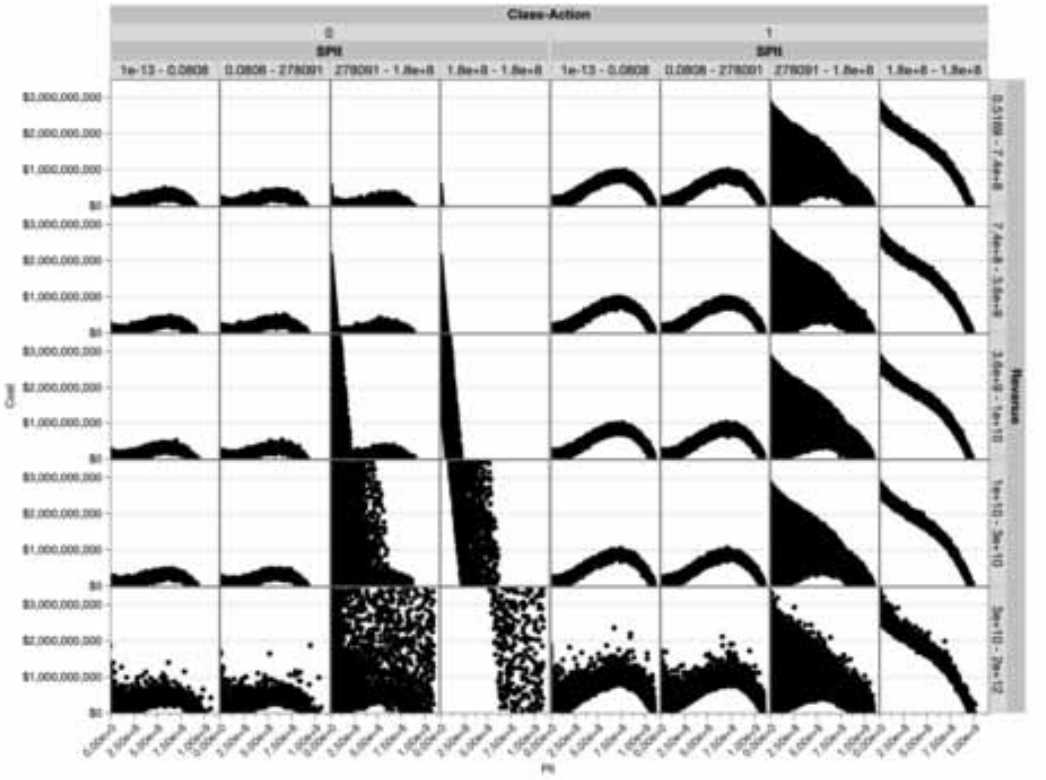
Figure 13. Variation of the total cost with PII and SPII independent variables for different revenue intervals



increase changes to a decrease in the total cost. The SPII independent variable accentuates the variation of the total cost for mega-breaches; therefore, it is of paramount importance to include SPII as a unique information category.

Figure 15 shows the variation of the total cost with the Revenue and PII independent variables. The columns of the matrix in Figure 15 represent PII and Revenue for two values of the CAL

Figure 14. Cost vs. PII for three categories. The columns represent the SPII intervals and CAL values, and there are five rows based on Revenue intervals.



independent variable; the matrix's rows represent five different SPII intervals. Figure 15 provides two important insights. First, it corroborates the distinct PII patterns in Figure 14 for different SPII intervals. Column three from the left demonstrates the variation of the total cost as the PII increases. For all SPII intervals, the observed patterns in the variation of the total cost are different. Second, there is no significant pattern change in column four from the left, representing the Revenue independent variable. Figure 15 also shows that the observed patterns in the variation of the total cost with the PII independent variable do not originate from any possible interaction with revenue; instead, PII explanatory power is different from SPII, and they interact.

Figure 16 shows the variation of the total cost with the SPII and Revenue independent variables. The columns in the matrix represent SPII and Revenue for two values of the CAL independent variable; the matrix's rows represent five different PII intervals. As shown in Figure 16, the third column's bottom cell from the left demonstrates a different pattern than the other three cells in the same column. Although this confirms the interaction between SPII and PII, a detailed analysis has not been conducted to scrutinize this pattern in the variation of total cost in the context of this study.

CONCLUSION

This study studied the significance of categorizing personally identifiable information as PII and SPII with a newer data set by advancing the earlier model (Poyraz et al., 2020) with an interaction profiler and Monte Carlo simulation study. The types of PII stolen during mega data breaches are; name, address, email, login information, non-sensitive medical information, insurance membership

Figure 15. Variation of the total cost with Revenue and PII independent variable for five SPII intervals, CAL =1. For a uniform scaling, the total cost axis maximum value is set to $3,500 Million.

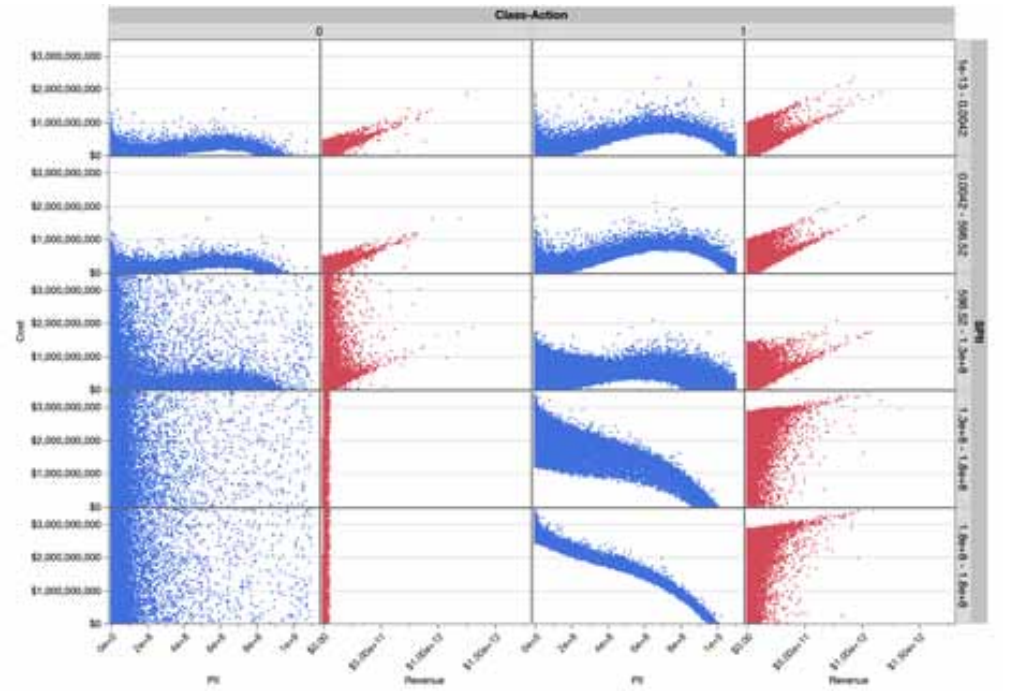
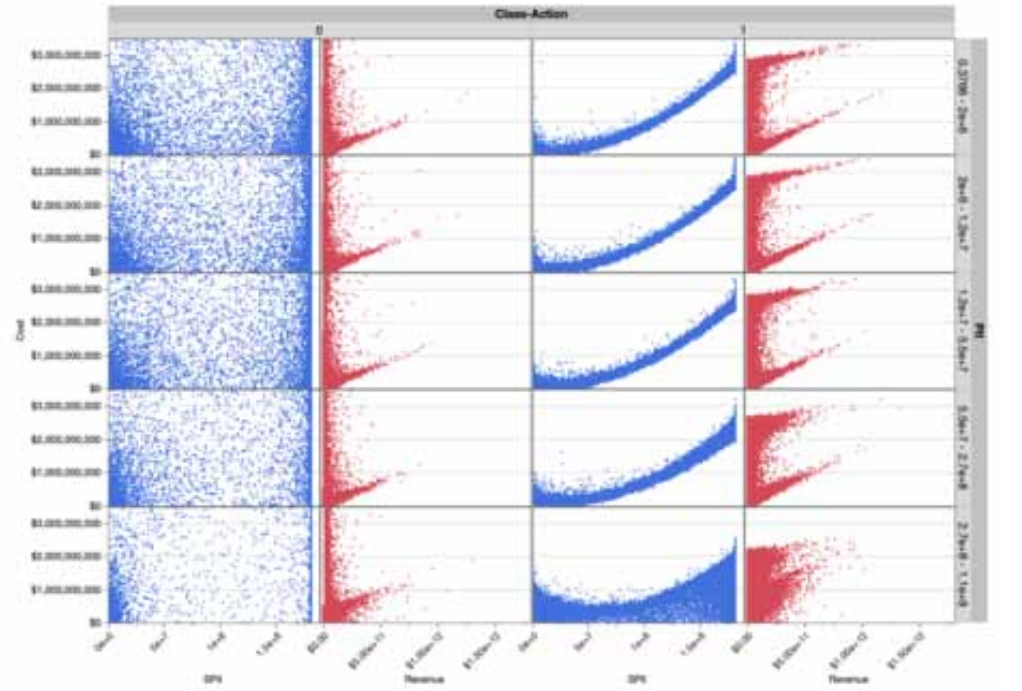


Figure 16. Variation of the total cost with Revenue and SPII for different PII intervals, CAL =1. For a uniform scaling, the total cost axis maximum value is set to $3,500 Million.



number, employment information, income, date of birth, driver's license plate. Stolen type of SPII is social security numbers, debit/credit card numbers, driver's license numbers, tax I.D., passport number, and bank account. With the increased sample size, PII and SPII independent variables become the main contributing independent variables. The combined contribution of these two independent variables to understanding the dependent variable is 78% (Table 4). As discussed, these mega data breaches incur higher costs due to the loss of SPII (Poyraz et al.,2020), and the results of this study corroborate these findings. The results of multivariate analysis (Figure 1) accentuate two points. First, collinearity is not a concern to build the model by using the new independent variables. Second, based on the simple linear regression analysis results, SPII has the highest correlation with the total cost (Figure 1 and Table 4).

In the analysis, the stepwise regression analysis technique was used. A common concern for this technique is the possibility of overfitting. Two additional studies were conducted to monitor the possibility of overfitting. First, 5410 models with different predictor numbers were developed, and the trend of the adjusted R-Squared values between R-Squared 0.95-0.99 Figure 3 is increasing. Second, a predicted R-squared analysis was conducted. The predicted R-squared value and the adjusted R-Squared values indicate that overfitting is not an issue for the model developed in this analysis. As demonstrated in Table 7, the R-squared (=0.98), adjusted R-squared (=0.96), and predicted R-squared (=0.82) values increased.

This study's parameter estimates demonstrate the importance of the SPII independent variable and the predictors that include SPII. For example, as shown in Table 8, SPII*SPII predictors have the highest significance with a p-value of less than 0.0001.

The skewness of the residual total cost distribution is 0.64, which means that some of the predictions are not in the 95% confidence interval. This can be seen in the residual quantile plot in Figure 6; four data points are outside the confidence interval. The studentized residuals, Figure 7, of this analysis indicates that these four data points are within the acceptable limits and are not outliers. Although the residual total cost distribution cannot be claimed to be normal because of the skewness value of 0.64, the studentized residuals indicate that the model's accuracy is reliable.

The result of the stepwise regression analysis consists of predictors, which may include interacting independent variables. Interaction means that two or more independent variables can describe the dependent variable's behavior with higher significance. The findings of the interaction analysis support the findings of the stepwise regression (Table 8). Within the p-value limit of this study, there is no PII*Revenue predictor in the regression equation (Figure 4). On the other hand, the third-row second interaction plot from the left demonstrates the interaction effect between SPII and PII independent variables. This cell includes non-parallel line segments, an indicator of interaction effects between these two independent variables. As shown in Table 8, the PII*SPII predictor is significant, with a p-value = 0.0012.

The findings of the regression and interaction profiler analyses were used in the Monte-Carlo simulation analysis. For this analysis, a robust distribution analysis was conducted for each of the continuous independent variables. For each distribution, the minimum AIC criterion was applied for the best-fitted distribution selection Table 10. Simulation analysis was conducted for 2 Million events, and besides the distribution parameters limit, no restrictions were applied. In doing so, possible biases were eliminated. After removing the negative values in the simulation results, the total cost variation was studied for different conditions.

According to the Monte-Carlo simulation results:

- Although SPII and PII are categorically similar independent variables, PII and SPII are two different independent variables.
- SPII and PII independent variables interact and complement each other.
- As the number of SPII increases, the total cost of data breaches increases.
- As the number of PII increases, the total cost of a data breach demonstrates three distinct behavior.

- Five SPII intervals have accentuated the difference between the three patterns in the total cost variation due to PII number changes.
- CAL, SPII, and PII independent variables interact.
- The CAL independent variable regulates the variation of the total cost significantly for all of the independent variables. As a result, prediction models can be significantly improved by including predictors that include CAL and interacting independent variables.
- CAL =1 situation accentuates the patterns in the variation of the total cost with PII.
- CAL =1 situation highlights the interaction between PII and SPII independent variables.

The simulation study demonstrates that adding SPII as a new independent variable improves the understanding of the total cost variation. To further enhance the result of this study, data collection continues.

REFERENCES

Ablon, L. (2018). Data Thieves: The Motivations of Cyber Threat Actors and Their Use and Monetization of Stolen Data. In *Data Thieves: The Motivations of Cyber Threat Actors and Their Use and Monetization of Stolen Data*. RAND Corporation. doi:10.7249/CT490

Carfora, M. F., Martinelli, F., Mercaldo, F., & Orlando, A. (2019). Cyber Risk Management: An Actuarial Point of View. *Journal of Operational Risk*, *14*(4), 77–103. doi:10.21314/JOP.2019.231

DHS. (2017). *DHS Handbook Safeguarding Sensitive PII*. www.dhs.gov/privacy

Edwards, B., Hofmeyr, S., & Forrest, S. (2016). Hype and Heavy Tails: A Closer Look at Data Breaches. *Journal of Cybersecurity*, *2*(1), 3–14. doi:10.1093/cybsec/tyw003

Eling, M., & Loperfido, N. (2017). Data Breaches: Goodness of Fit, Pricing, and Risk Measurement. *Insurance, Mathematics & Economics*, *75*, 126–136. doi:10.1016/j.insmatheco.2017.05.008

Jacobs, J. (2014). *Analyzing Ponemon Cost of Data Breach*. Data Driven Security. https://datadrivensecurity.info/blog/posts/2014/Dec/ponemon/

Layton, R., & Watters, P. A. (2014). A Methodology for Estimating the Tangible Cost of Data Breaches. *Journal of Information Security and Applications*, *19*(6), 321–330. doi:10.1016/j.jisa.2014.10.012

Mcshane, M., & Nguyen, T. (2020). Time Varying Effects of Cyberattacks on Firm Value. *Geneva Papers on Risk and Insurance: Issues and Practice*.

Poyraz, O. I., Bouazzaoui, S., Keskin, O., McShane, M., & Pinto, C. A. (2020). Cyber-assets at Risk (CAR): The Cost of Personally Identifiable Information Data Breaches. *ICCWS 2020 15th International Conference on Cyber Warfare and Security*, 402.

Poyraz, O. I., Canan, M., McShane, M., Pinto, C. A., & Cotter, T. S. (2020). Cyber assets at risk: Monetary impact of U.S. personally identifiable information mega data breaches. *The Geneva Papers on Risk and Insurance. Issues and Practice*, *45*(4), 616–638. Advance online publication. doi:10.1057/s41288-020-00185-4

Romanosky, S. (2016). Examining the costs and causes of cyber incidents. *Journal of Cybersecurity*, *2*(2), 121–135. doi:10.1093/cybsec/tyw001

Tatar, U., & Çelik, M. M. (2015). Hacktivism as an emerging cyberthreat: case study of a Turkish hacktivist group. In *Terrorism Online* (pp. 66–83). Routledge.

Tatar, U., Karabacak, B., & Gheorghe, A. (2016) An Assessment Model to Improve National Cyber Security Governance, In *11th International Conference on Cyber Warfare and Security: ICCWS2016* (p. 312). Academic Press.

WEF. (2019). *The Global Risks Report 2019 14th Edition Insight Report*. http://wef.ch/risks2019

Wheatley, S., Maillart, T., & Sornette, D. (2016). The Extreme Risk of Personal Data Breaches and the Erosion of Privacy. *The European Physical Journal B*, *89*(1), 1–12. doi:10.1140/epjb/e2015-60754-4

Xu, M., Schweitzer, K. M., Bateman, R. M., & Xu, S. (2018). Modeling and Predicting Cyber Hacking Breaches. *IEEE Transactions on Information Forensics and Security*, *13*(11), 2856–2871. doi:10.1109/TIFS.2018.2834227

APPENDIX: MEGA DATA BREACH INCIDENTS

Table 13. The dataset of this study

Company	Industry	Year	Revenue	PII	SPII	Class action	
Anthem	MED	2015	7.92E+10	7.09E+08	7.88E+07	1	4.07E+08
AOL	BSO	2004	8.70E+09	9.20E+07	0.00E+00	0	1.00E+06
Blue Cross Blue Shield of Tennessee	MED	2009	6.72E+09	3.31E+06	2.40E+05	1	1.85E+07
CardSystems Solutions	BSF	2005	1.00E+07	4.00E+07	4.00E+07	1	2.18E+08
CareFirst Blue Cross Blue Shield	MED	2015	8.80E+09	4.40E+06	0.00E+00	0	6.50E+06
Community Health Systems	MED	2014	1.86E+10	2.25E+07	4.50E+06	1	7.81E+07
DSW Shoe Warehouse	BSR	2005	9.61E+08	1.40E+06	1.50E+06	0	6.80E+06
eBay	BSO	2014	8.80E+09	8.70E+08	0.00E+00	0	4.60E+07
Epsilon	BSO	2011	8.47E+08	5.00E+08	0.00E+00	0	2.70E+08
Equifax	BSF	2017	3.40E+09	4.42E+08	1.64E+08	1	1.45E+09
Excellus Blue Cross Blue Shield	MED	2015	5.94E+09	4.00E+07	2.00E+07	0	1.73E+07
Experian	BSF	2015	4.80E+09	4.50E+07	4.50E+07	1	4.42E+07
Global Payments	BSF	2012	2.20E+09	1.50E+06	1.50E+06	0	1.21E+08
Hannaford Bros Supermarket (Delhaize Group)	BSR	2008	1.90E+10	0.00E+00	4.20E+06	0	2.55E+08
Heartland Payment Systems	BSF	2009	1.65E+09	0.00E+00	1.30E+08	0	1.51E+08
Home Depot	BSR	2014	7.88E+10	5.30E+07	5.60E+07	1	3.41E+08
JP Morgan Chase	BSF	2014	9.42E+10	3.04E+08	0.00E+00	0	2.50E+08
LinkedIn	BSO	2012	9.72E+08	1.30E+07	0.00E+00	1	1.30E+06
Marriott	BSO	2018	2.08E+10	3.50E+08	3.29E+07	0	2.20E+08
Medical Informatics Engineering	MED	2015	9.00E+06	2.73E+07	3.90E+06	1	1.00E+06
Nationwide Mutual Insurance & Allied Insurance	MED	2012	1.90E+07	2.54E+06	2.54E+06	1	5.50E+06
PNI (Staples)	BSO	2015	2.30E+07	2.80E+06	2.80E+06	1	1.80E+07
Premare Blue Cross	MED	2015	3.68E+09	0.00E+00	2.08E+07	1	8.40E+07
Sony PSN	BSR	2011	7.38E+10	1.48E+08	1.23E+07	1	1.93E+08
Target Corp	BSR	2013	7.33E+10	2.80E+08	4.00E+07	1	3.11E+08
TD Ameritrade Holding Corp	BSF	2007	2.18E+09	2.52E+07	0.00E+00	1	6.50E+06
Ticketfly (Eventbrite)	BSR	2018	2.23E+08	1.04E+08	0.00E+00	0	7.30E+06
TJ Stores	BSR	2007	1.71E+10	0.00E+00	4.62E+07	1	1.78E+08
Uber	BSO	2017	1.11E+10	7.68E+07	6.07E+05	1	1.48E+08
Vtech	BSO	2015	1.88E+09	1.80E+07	0.00E+00	1	7.00E+05
Yahoo	BSO	2016	5.17E+09	9.70E+08	0.00E+00	1	5.03E+08

Mustafa Canan (Ph.D) is an Assistant Professor in the Department of Information Sciences at the Naval Postgraduate School. His research focuses on decision making, information operations, situational awareness, and complex adaptive system behavior. Before joining the NPS, he was a National Research Council (NRC) Postdoctoral Fellow at the Air Force Research Laboratory at Wright-Patterson AFB. He holds two Ph.D. degrees in Particle Physics and Engineering Management. Dr. Canan teaches decision-making in complex situations and has published articles in Nature, Nature Communication, Nature Physics, IEEE, and other leading journals and conferences.

Omer Ilker Poyraz earned his Ph.D. from the Engineering Management and Systems Engineering Department at Old Dominion University in 2020. He graduated from Yeditepe University with a BA degree in Business and got his MBA degree from Old Dominion University. He worked as a research officer for the Turkish Army for five years. His main fields of research include machine learning, cyber risk and risk analysis.

Anthony Akil is from Windsor Locks Connecticut, a small town outside of Hartford. He enlisted in the United States Navy in 2002 after graduating from Windsor Locks High School. He spent five years enlisted as a Personnel Specialist reaching the rank of Petty Officer Second Class. While enlisted, he served with VFA-192 "The World Famous Golden Dragons" homeported at Naval Air Station Atsugi. While assigned to VFA-192 he completed three deployments embarked onboard USS Kitty Hawk. In 2007 Anthony Akil was selected for the "Seamen to Admiral" commissioning program and attended the University of Washington. In 2010 he graduated with a Bachelor of Arts in Economics and reported to the Navy Information Operations Command (NIOC) San Antonio, Texas. At NIOC Texas he functioned as Battle Watch Division Officer responsible for providing indications and warnings in the support of units entering 4th Fleet's area of operations. In 2012 he volunteered to deploy in support of Operation Enduring Freedom as a Cryptologic Support Team Officer in Charge. Assigned to Camp Nathan Smith located in Kandahar City Anthony Akil supported the 2nd Infantry Brigade 4th Division and the 2nd Stryker Brigade 2nd Division. Following his first tour at NIOC TX, Anthony Akil served aboard USS Oscar Austin (DDG-79) as the Signals Warfare Officer. While attached to Oscar Austin, Anthony Akil supported BALTOPS 2015 and a counter-piracy deployment to the Indian Ocean. Anthony Akil graduated from the Naval Postgraduate School earning a master's degree in Cyber Systems & Operations. He is currently stationed at PMW-120, supporting NAVWAR the SYSCOM for Information Warfare, while attending law school at California Western School of Law.