Code Challenge – Tenchi Security

ABIN Report - Cyber Breach Analysis

Pedro H P M Lopes

06 February 2022

Introduction

This reports analyses cyber breaches data collected between 01/01/2011 and 12/31/2021. Provided data consists of a unique breach identifier, the number of data records affected and the dollar cost of said breach, as well as the NAICS code of the industry sector and NAICS code of the affected company. A text description of the sector is also given, alongside with the breach date and its cause. A total of twenty sectors were analyzed here

All this information was collected for each individual breach. A total of 10.000 samples were collected, 1374 of them (13.74%) did not affect data records nor resulted in any direct financial cost for the companies.

Economic Impact Analysis

During this decade long period, these breaches costed a total of 46,61 B$, with the cost of each breach ranging between 0 and 7,92 B$, with a total average dollar cost per breach of 4,66 M$. The distribution of dollars by sector can be seen in the tables below, alongside with the number of breaches and average cost per breach.

Sector	Economical Impact (US$)	Number of Security Breaches	Average Cost per Breach (US$)
Other	9,84 B	415	23,71 M
Administrative	7,14 B	1089	6,56 M
Financial	5,63 B	1258	4,47 M
Professional	3,97 B	1232	3,22 M
Education	3,87 B	719	5,38 M
Information	3,32 B	579	5,73 M
Healthcare	3,09 B	1158	26,68 M
Retail	2,02 B	586	3,44 M
Public	1,85 B	655	2,82 M
Manufacturing	1,41 B	556	2,54 M

Sector	Economical Impact (US$)	Number of Security Breaches	Average Cost per Breach (US$)
Trade	884,15 M	343	2,58 M
Hospitality	755,90 M	368	2,05 M
Real Estate	612,17 M	191	3,21 M
Management	455,24 M	195	2,33 M
Utilities	416,17 M	53	7,85 M
Construction	305,70 M	151	2,02 M
Entertainment	204,60 M	145	1,41 M
Transportation	160,19 M	128	1,25 M
Mining	76,70 M	28	2,47 M
Agriculture	10,32 M	16	645 k

Table 1 – Dollar cost distribution and analysis per sector

Sectors are ordered by the economic impact suffered. The companies comprised in the "Others" sector (NAICS Sector Code 81) were the most affected economically, even though they had a rather low occurrence of breaches. Because of this, this sector also had the highest average cost per breach. However, it's interesting to note that the most expensive breach, of 7,92 B$ was observed in this sector, meaning that one single breach represented 80,04% of the sector's economic impact. This breach can be considered an outlier, an abnormal observation when compared to the others in this sector. By disregarding it, the total economic impact would become 1,92 B$, with an average cost per breach of 4,63 M$. This would place the 'Others' sector right below the Public, but still in the ten most affected sectors.

Five other breaches resulting in costs of over 1 B$ were registered. Two of them belonging to the Financial sector, with a total of 3.14 B$ (55,77% of the sector's total), other two from the Administrative, of 3,26 B$ (45,66% of the sector's total), and the last one from the Education sector, of 2,32 B$ (43,12% of the sector's total). The Financial, Professional, Healthcare and Administrative, in this order, were the most targeted, having more than 1000 breaches each in this ten years period. They also had high losses when compared to the rest of the sectors, all being above 3 B$. All high impacting breaches (breaches resulting in costs of over 500 M$) were recorded in the top 10 most economically affected sectors, the ones that are also the most targeted by breaches.

These breaches, a total of 11 (0,11% of the total samples), were all considered outliers, and thus removed. Keeping this data could mislead the analysis, specially when considering year-by-year growth of economical impact average cost per data record, since most of these happened after 2015.

Cyber Breaches during the decade

The figure below shows the economic analysis and how it evolved over the years. The blue lines represent the actual content of the graph, detailed in the vertical axis of each one of them, while the red indicates the linear trendline, calculated via simple linear regression, a method that creates a straight line that better adapts to and represents the samples in question. This is done to identify either an increasing, decreasing or stable trend in the data.

It's possible to identify a slight growing trend on the occurrence of breaches, affected data records and annual cost, indicating that it can be expected that they keep growing over the next years. The average cost per data record, however, shows a stable trend, so one should not expect neither a decrease or increase of the cost of data records.

At first glance, the amount of data records affected per breach does not seem to have a connection with the annual cost of data breaches. Apart from the two peaks of both data records and costs in the years of 2017 and 2019, no other correlation between the two can be made by this graph. This will be further investigated in the next section.

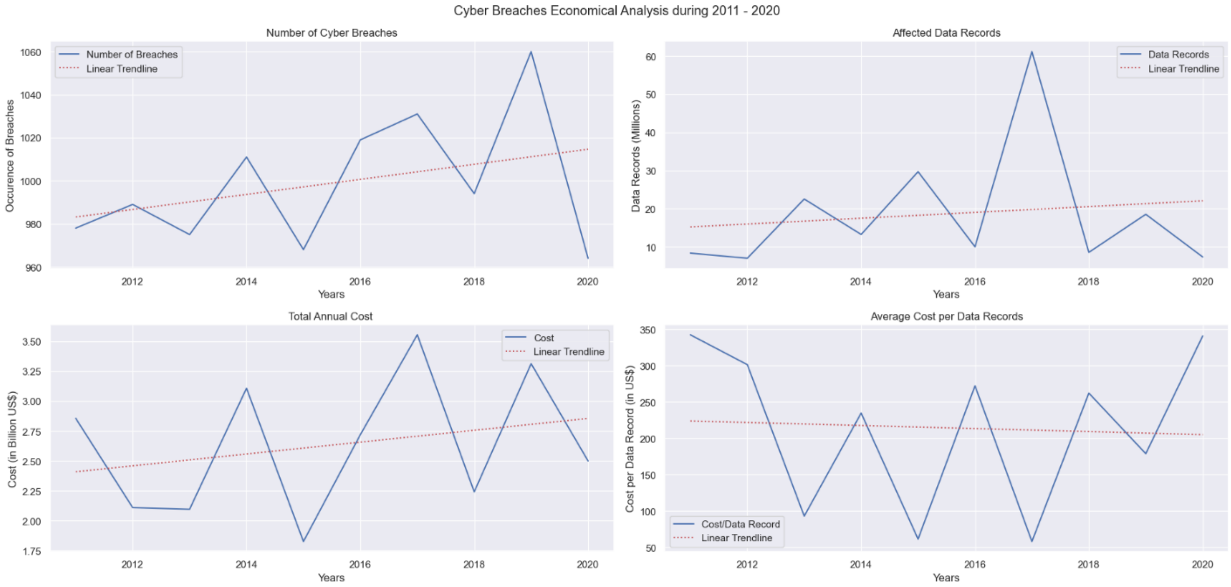

Figure 1 – Cyber Breaches analysis over the period of 2011 to 2020

Cost of Data Records

In order to search for a relationship between the cost of a breach and how many records were affected by it, a scatter plot was made one against another. In the figure below, each of the dots represents a sample of a breach. It's placement in the graph is determined by its cost, displayed on the vertical axis, and the affected record, on the horizontal axis. Since the data is highly skewed, a logarithmic scale was proven necessary for the correct visualization. The distance between each tick is ten times bigger than the previous one.

In this graphic however, it is not possible to extract a direct linear correlation between the two, so a direct increase in affected data records is not transferred to an increase in the dollar cost. One other technique may be more suited to this situation, the spearman correlation coefficient, assessing the relationship between two variables using a monotonic function. This means that, this relationship should follow the overall same direction, either increasing or decreasing. In this case, the Spearman's correlation coefficient was of 0.58, with a high degree of statistic significance. This indicates that, the data records and the dollar cost are somewhat positive correlated, meaning that the more data records are affected by a breach, its dollar cost will tend to increase as well.

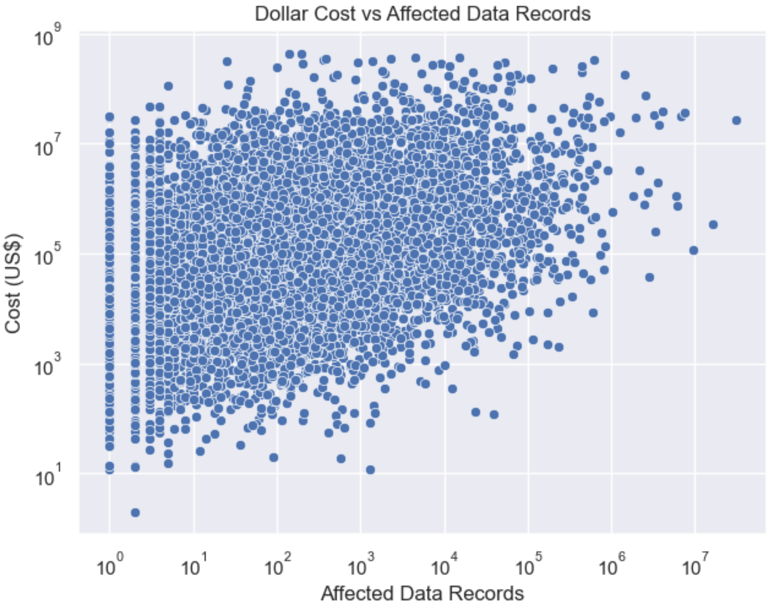

Figure 2 – Cyber Breaches analysis over the period of 2011 to 2020

Conclusions

The top 10 most economically affected sectors as displayed in table 1 should be the target of incentives of cybersecurity solutions. Not only they had dollar costs of over 1 B$ each, they also seem to be the target of large scale breaches that may result in individual breaches of over 500 M$. A growing trend on the occurrence of breaches, their dollar cost, and data leakage was also noted. As of yet, this is a rather weak trend, and does not seem to be too concerning. Nevertheless, actions should be made in order to halt the situation before it escalates.

The dollar cost of and the data leakages are somewhat positively correlated, as the first tend to increase alongside the second. Projections of the economical impact of breaches based on data records for the following years may be possible. However, the use of more advanced techniques such as Artificial Intelligence (Ai) are recommended. These techniques allow the use of more variables to make a projection of the dollar cost. The Random Forest algorithm, in particular, can be suited for this, since it can handle categorical (like the Sectors) and non-categorical data (like the affected data records). A study of more information to be collected from these companies would be interesting, as it would provide the database with more information that could possibly be fed to these AI models. The use of feature engineering in the time series of these data (as shown in figure 1) may also provide valuable information.