An Analysis of Cybersecurity Data Breach in The State of California

Zakaria Tayeb Bey and Michael Opoku Agyeman[0000-0002-3734-4451]

Centre for Advanced and Smart Systems, University of Northampton, UK
Michael.OpokuAgyeman@northampotn.ac.uk

Abstract. As the wave of data breaches continues to crash down on organisations, we will analyse the largest publicly available database of data breaches in the state of California using the public data breach database functioning under the state notification law of data breach. The dataset contains records since January 2012. These records were analysed in order to classify and identify California data breaches by multiple company types, attack vectors and stolen personal information. The main findings were that Software vulnerability is the most common attack vector due to third-party software, the financial industry is the most targeted industry while both large and small organisations are equally targeted by attackers. The analysis also found that credit/debit card information and social security numbers represent the most stolen personal information.

Keywords: Cybersecurity, Data Breach, Cyber-attack, Ethical Hacking, Internet of Things, IoT, Security, California Data Breach Database.

1 Introduction

The extent of data breaches is expanding, impacting more organisations and people which put a lot on the line as data breaches may cost individuals and businesses money, reputational damage, and lost opportunities. These breaches put key infrastructure at risk and national security at jeopardy [1] [2].

In 2013, A group of hackers gained access to Target's security and payments system and they were able to install a malicious software (malware) to steal credit card information used at the company's 1,797 U.S. stores. The cybercriminals were able to steal 40 million credit and debit records and 70 million customer records which caused the company to pay $18.5 million to settle claims by 47 states and their earnings dropped 46% [3]. Target is only one of many companies that got affected by malicious software.

There was also another huge breach in 2012, cybercriminals compromised the professional networking site LinkedIn using a SQL injection attack and leaked over 100 million records of personal information and hashed passwords, Attackers were able to crack the passwords and sold them online even though the passwords were hashed in the database but the encryption keys were found by attackers on the same compromised server [4].

Nowadays, cloud computing and the use of cloud storage are widely used due to the popularity of social media, e-commerce, and mobile services meaning that more personal consumer information and data are exposed to the public network which explain the increase of data breach costs from $3.86 million to $4.24 million in 2021, the highest in the past 17 years [4]. COVID-19 played an important role too by increasing the remote work which opens the door to insecure home networks and devices which lead to an increase in the average cost of a data breach by $1.07 million and a 300% increase in reported cybercrimes by the US FBI [6]. Another attractive target to threat actors is Internet of Things (IoT) devices. 75 billion IoT devices will be connected by 2025 [8], IoT is the future but unfortunately with the increase of IoT devices in the network there'll be an increase of cyber-attacks because IoT devices are the most vulnerable devices in the network. Also, more than 93% of healthcare organisations experienced a data breach in the past few years [9] because healthcare enterprises often use legacy software due to budget constraints even though it is a highly critical industry to secure. As a result of this increasing concern, California has obliged organisations to report residents of the state when their personal information is breached, beginning in 2003. Businesses and government institutions have been required to notify the state governments of data breaches impacting more than 500 California residents since 2012 [10]. These data breach reporting rules made it possible to identify the numbers of data breaches in the state of California where 2017 was the worst year in the record of the state with over 1,200 breaches and 3.4 billion leaked records according to Risk Based Security [11].
Hotels, schools, dentists, spas, universities, restaurants, hospitals, doctors, retailers, government agencies, banks were all affected by data breaches and cyber security attacks. Mostly, third party service providers or employees were responsible for most of the reported breaches. Insiders' unintentional and intentional activities, as well as stolen and lost devices storing unencrypted data, all contributed to data breaches [11].

2 Aims and Objectives

2.1 Aim

An analysis of the California data breach database will be conducted in order to extract different statistics related to data breaches and cyber security attacks in 98 different industries. Three types of data analysis will be conducted in this research, Companies Analysis (By : Industries, Sectors, Sizes, Financial loss, Breach detection), Attacks Analysis (Unauthorised Access, Software Vulnerability, Stolen Computer or Data, Data Found Publicly, Wrong Data Sent, Exposed Data, Compromised Machine, Phishing Email, Insider Theft, Stolen Credentials, Compromised Email, Lost Computer or Data, Ransomware, Social Engineering) and Personal Information Analysis (Social Security Number, Payment Card, Medical Record, Password, Bank Account, Health Insurance Driver's Licence).

2.2 Specific Objectives

- To classify California data breaches by multiple company types.
- To classify California data breaches by common attack vectors.
- To classify California data breaches by common stolen personal information.

2.3 Research Question

- What are the patterns in company's types in California data breaches?
- What are the common attack vectors in California data breaches?
- What are the common stolen personal information in California data breaches?

3 California's Data Security Breach Reporting Law

In 2003, California established a data breach notification law first. 46 other states and international authorities across the world have passed similar legislation in a twelve-year period [10].

Over the last five years, hackers and attackers with malicious intent have penetrated the security of a remarkable number of organisations in nearly every state in America. In 2016, Yahoo, situated in California, suffered a huge data security breach in which online hackers stole the private information of around 500 million customers. The hackers were so skilled that Yahoo was unaware of the breach for two years until it was detected. Fortunately, most consumers who were affected did not face any long-term consequences. Therefore, the state legislature of California chose to introduce legislation requiring corporations to assume responsibility for the security of their customers' personal information [12].

As the frequency of data breaches keeps rising worldwide [1], organizations doing business in California must show that they are taking reasonable precautions to protect their customers' personally identifiable information. This data set contains a variety of sources of credentials for the authentication of individual's and is frequently used to gain access to online accounts or perform financial transactions.

If a data security breach occurs, firms must notify all affected individuals and the California Attorney General if the incident affects more than 500 people. Businesses that fail to report a data breach or who delay notification without reasonable cause may face severe financial penalties under the law.

Today, California's regulations and reporting obligations for data breaches are based on the California Consumer Privacy Act (CCPA). On June 28, 2018, Governor Jerry Brown signed the Act into law [10].

Finally, these rules have generated entire new enterprises dedicated to assisting businesses in preventing data breaches and responding effectively when they do occur. Cyber insurance, for example, is a relatively new profession that protects organisations

against data breaches and cyber-attacks. Cyber insurance premiums totaled $5 billion in 2018, with the market likely to double in the past five years [13].

4 Methodology

4.1 Data Source

We looked at all publicly available California data breach notices since January 2012 on the California Department of Justice website for this analysis. As the California state government maintains a complete database of all enterprises and citizens cybersecurity breaches. Therefore, all data breaches recorded since January 2012 are listed on the website, along with complete breach notification reports.

4.2 Data Scraping

The California state government website does not provide any public application programming interface (API) [11]. Therefore, web scraping was used to extract all the publicly available data security breaches on the website since the data shown on the website does not represent any underlying organised structure like JSON or XML.
PHP programming language is used to code the web scraper. This script will fetch the HTML page content of the data security breaches website. All breaches since 2012 are displayed in one page without any pagination. Therefore, only one request is required to fetch the HTML string and extract all the organisation name, date of breach (if known) and reported date. However, all records in the PDF report located in the details page 2 requests are required for each breached organisation (2N) in order to download the pdf report. First request to the details page and the second request to download the report.
Text content will be extracted from the pdf, the content will be processed using regular expression in order to extract all the records. Finally, the parsed records will be inserted into a MySql database.

4.3 Data Structure

Using the company name from the extracted records and the LinkedIn api, we managed to fetch the company type (nonprofit, partnership, government, privately-held, public-company, sole-proprietorship, educational-institution, self-employed) with 75.5% match, the company size (1-10, 11-50, 51-200, 201-500, 501-1,000, 1,001-5,000, 5,001-10,000, or 10,000+) with 83.4% match, and finally the industry with 100% match. The "What Happened?" section is one of the required sections in the data breach reports according to the California law, this section describes the breach in general. The attack vector will be assigned to the breach according to the content of this section as follow:

- **Compromised Email:** Once attackers gain access to the email address, They can compromise all services assigned to this address and possibly escalate

from personal accounts to company accounts which can eventually lead to a cybersecurity breach.

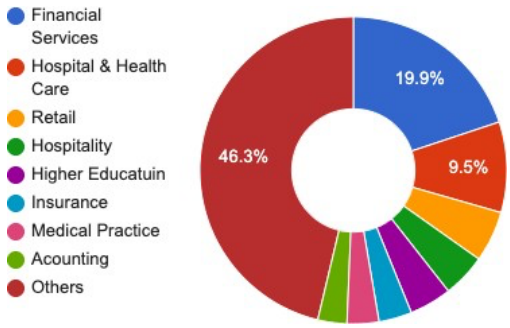
- **Compromised Machine:** Hacked physical machines like servers, ATM's or Printers within the compromised network.
- **Data Found Publicly:** publicly available personal information online or in physical documents.
- **Exposed Data:** expose confidential information to an unauthorised person due to misconfiguration or software bug.
- **Insider Theft:** intentionally or unintentionally leaking confidential credentials by a current or former employee via unsecured email address or cloud storage.
- **Lost Computer or Data:** employees lose unencrypted physical devices or physical records containing confidential information.
- **Phishing Email:** stolen employee credentials via an email that appears to be from a trusted source.
- **Ransomware:** company critical data is encrypted over the network; decryption key is required to decrypt the company data and can be obtained only by paying a ransom or using reverse engineering techniques which can cause a risk of losing data.
- **Social Engineering:** Manipulating an employee to send confidential credentials or give access to customers account by impersonating a high-level company manager.
- **Software Vulnerability:** known or 0day vulnerabilities on the company network and servers like remote code execution, SQL injection, Windows SMB privilege escalation. Third-party library vulnerabilities are included.
- **Stolen Computer or Data:** stolen employee unencrypted physical devices or physical records containing confidential information.
- **Stolen Credentials:** Employees use the same password in a previous data breach or a weak password which can be easily brute forced.
- **Unauthorised Access:** Any unauthorised access to the company network that may cause a risk of exposing confidential information.
- **Wrong Data Sent:** Employees may send confidential credentials to an unauthorised party by accident.

5 Presentations of Findings

5.1 Analysis: Companies

Industries. Top 8 industries represent 53.7% of the total breached companies where the financial industry represents 19.9% followed by hospital and healthcare, Retail, Hospitality, Higher Education, Insurance, Medical Practice, Accounting. The other 90 industries represent 46.3% of the total breached companies with a low frequency under 3% (Figure 1).

American Express (5.9 %) and Discover Financial Services (1.8 %) were the most frequently compromised companies. These companies are required to track all the publicly exposed personal information including debit/credit cards credentials in order to notify



their customers and block the hacked cards.

Fig. 1. Frequency of breached companies for the top 8 industries.

Sectors. Most of the companies that were breached were either privately owned (37.0%) or public companies (34.8 %). The rest of the companies were nonprofit (11.5 %), educational institutions (6.9 %), government agencies (6.0 percent), and 3.8% for other businesses (Figure 2).

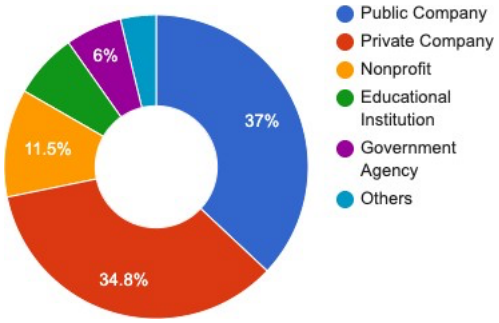


Fig. 2. Frequency of data breaches by company sector in 98 industries.

Sizes. Sizeable companies with +10.000 employees are by large the most breached companies representing 30.3% of total breaches. Over half of cybersecurity breaches affect large companies with +1,000 employees. The frequency of data breach in small to medium sized companies range from 6 to around 12% which can be identical to large companies (Figure 3).

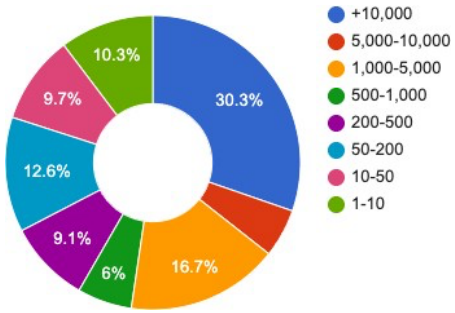
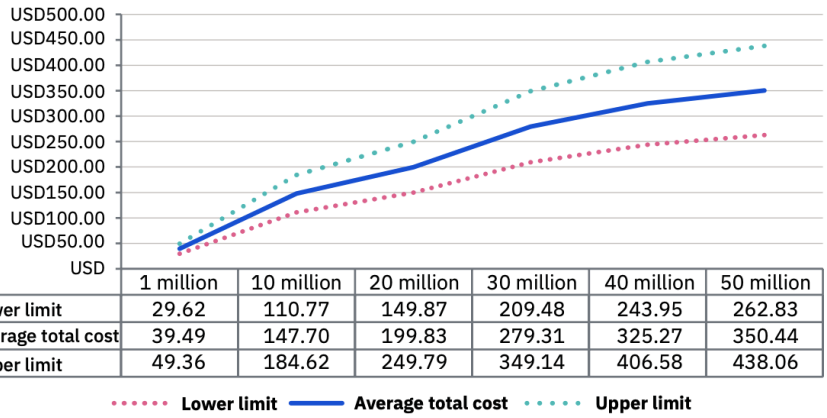


Fig. 3. Frequency of data breaches by company sizes in 98 industries.

Financial loss. The healthcare sector suffers the most from data breaches. The average cost grew by 29.3 percent between 2020 and 2021, from $7.13 million to $9.23 million. Data breaches that were discovered and resolved within 200 days cost an average of $3.61 million. However, breaches that took more than 200 days to uncover and contain cost an average of $4.87 million, a $1.26 million difference.

Breaches with at least 50 million records cost 100 times as much as the ordinary data breach. Moreover, Data breaches with 50 million to 65 million records cost an average of $401 million in 2021, up from $392 million in 2020 (Figure 4).

Only 21.5% of cybersecurity incidents are reported within 30 days. However, prior literature review found that data breach incidents reported within 30 days of the incident may save over 1 million dollars.



	1 million	10 million	20 million	30 million	40 million	50 million
Lower limit	29.62	110.77	149.87	209.48	243.95	262.83
Average total cost	39.49	147.70	199.83	279.31	325.27	350.44
Upper limit	49.36	184.62	249.79	349.14	406.58	438.06

Fig. 3. Cost of cybersecurity data breach measured in USD million by breached records. (ibm.com)

Breach detection. There's an average of 108 data breaches per semester. However, almost 19% of businesses that reported data breaches were unable to determine the specific date(s) of the incident.
The amount of data breaches has been gradually increasing, with 1.08% more data breaches occurring each semester than the previous one.

5.2 Attack Analysis

Due to missing explanation in some reports, Unauthorised Access is the most common attack with 387 frequencies. Software Vulnerability is the second most common attack representing 13.1% of total attack vectors as many softwares or third party softwares are unprotected and open to remote access. The 2017 Ukraine ransomware attacks is a great example of third-party vendors where the attacker hacked a vendor that serviced a large number of companies to use the auto-update feature to upload a malicious code and gain remote access to all clients associated with the vendor. Stolen Computer or Data is the third most common attack vector representing 11.4% with 163 frequencies, followed by Data Found Publicly (159), Wrong Data Sent (105), Exposed Data (103), Compromised Machine (78), Phishing Email (63), Insider Theft (43), Stolen Credentials (42), Compromised Email (35), Lost Computer or Data (31), Ransomware (20), Social Engineering (16) (Table 1).

Table 1. Frequency of data breaches per attack vector

Attack Vector	Frequency	Rate
Unauthorised Access	387	27%
Software Vulnerability	188	13.1%
Stolen Computer or Data	163	11.4%
Data Found Publicly	159	11.1%
Wrong Data Sent	105	7.3%
Exposed Data	103	7.4%
Compromised Machine	78	5.4%
Phishing Email	63	4.4%
Insider Theft	43	3%
Stolen Credentials	42	2.9%
Compromised Email	35	2.4%
Lost Computer or Data	31	2.2%
Ransomware	20	1.4%
Social Engineering	16	1.1%

5.3 Personal Information Analysis

Throughout the top 25 industries, the two most common types of personal information stolen were social security numbers (31.1%) and payment cards (30.1%) followed by Medical Record (10.8%), Password (8.5%), Bank Account (7.5%), Health Insurance (6.4%), Driver's Licence (5.8%) (Table 2).

Table 2. Frequency of stolen personal information stolen across the top 25 industries.

Personal Information Type	**Frequency**	**Rate**
Social Security Number	609	31.1%
Payment Card	590	30.1%
Medical Record	212	10.8%
Password	167	8.5%
Bank Account	143	7.3%
Health Insurance	125	6.4%
Driver's Licence	114	5.8%

6 Discussion of Findings

Financial services industry received the highest number of data breaches following the healthcare sector because hackers target organisations that have what they want, which is usually money. Therefore, industries with data that can be sold for money are the most targeted industries. The fact that financial services organisations are often breached doesn't mean that they aren't as careful about security as their peers. It means that they are being targeted more than other businesses because of their rich financial and data assets, and a small percentage of attacks are successful because of their cybersecurity issues, which leads to an overall higher number of attacks. conducting biannual comprehensive security awareness training that goes beyond the basics to teach employees how to recognise sophisticated threat strategies. Advanced phishing techniques, a broad spectrum of social engineering strategies, indicators of insider threat activity (along with anonymous reporting systems), and physical security should all be addressed throughout training. The training programme should include custom modules addressing the unique characteristics of each group inside the organisation and how they may be targeted.

Every industry has vulnerabilities, and they are all being attacked. Financial services, on the other hand, face the brunt of attacks due to the financial and data assets they

oversee. By focusing security efforts on the right areas and properly training employees, the financial services industry can reduce cybersecurity threats.
Most companies are either public or private companies. Therefore, most affected companies by data breaches are public and private companies. However, public companies tend to be an easier target for attackers and attract cybercriminal organisations due to the poor funding, outdated technology and insufficient staff training as it's essential to implement integrated and comprehensive protection that enables a public sector organisation to identify and effectively respond to multiple threat vectors.
The data demonstrates that no business is too large or too small to be a victim of a cybersecurity attack or data breach. In fact, small organisations can affect large organisations because they're always working together as suppliers, business partners or providing services. Attackers can use privilege escalation to escalate the breach from small companies to large companies. Due to the fact that larger firms invest considerably in network security equipment, hackers are constantly looking for any entry points into large organisations networks. Therefore, several cybercriminals target smaller organisations that collaborate with larger corporations in order to escalate the privilege from small to large companies network using information exchange protocols between the two organisations like auto software update or advanced social engineering and fishing using corporate email credentials.
Data breaches can take a wide range of shapes and sizes. Each of these may result in financial loss. All industries are moving to automation with technology. Therefore, new opportunities are created for hackers and fraudsters to benefit from users' sensitive financial information. Due to data breaches the financial industry lost an average of $5 million in 2021. Even though the financial industry is the most valuable and targeted industry. In 2021, the financial industry cybersecurity cost is slightly lower than in 2020.
Since 2019, There's a high demand for cloud solutions, especially in the banking industry and cloud security has become more frequent and resulted in an increase in compromised credentials.
A growing number of softwares are developed and compiled using commercially available or open-source code. Therefore, hackers can freely analyse the open-source code and find 0day vulnerabilities which make software vulnerability the most frequent attack vector.
Software vulnerability and related attacks were the most common attack vectors, aside from unlawful access. Ransomware and phishing emails are two relatively new attack vectors that have become increasingly frequent since 2016.

7 Limitations

This analysis is limited to the publicly available dataset provided by the office of the attorney general under the notification law, Other publicly available dataset in the state of California are available and they can be used in order to have more accurate and precise analysis which can be used to identify other data breach factors and patterns, However, this research focused only on analysing companies types, attack vectors and

stolen personal information. The result of the attack vector analysis can be further discussed in order to create an effective cybersecurity threat detection process based on the most common attack patterns. Moreover, an effective incident response plan can be extracted from the analysis to help enterprises prepare for, detect, and respond to data breaches incidents. Previous years records can also be used to train a machine learning model in order to predict future data breaches frequencies by companies' profile (Industry, Sector, Size, Financial loss), attack types and stolen personal information.

8 Conclusion

Eight industries found to be the most commonly affected by cyber-attacks represent more than 50% of breaches while 25 industries represent 80% of all California data breaches when looking at the company's analysis. Over half of all data breaches were caused by large firms (those with 1,000 or more employees). The healthcare sector suffers the most from data breaches and software vulnerability is the most common attack vector due to third-party software, the financial industry is the most targeted industry while both large and small, public and private organisations are equally targeted by attackers and credit/debit card information and social security numbers represent the most stolen personal information. Breaches with at least 50 million records cost 100 times as much as the ordinary data breach.

References

1. Data breaches reach record levels worldwide. (2019). Network Security, 2019(3), pp.1–2.
2. Dean, Andrew, and Michael Opoku Agyeman. "A study of the advances in iot security." Proceedings of the 2nd International Symposium on Computer Science and Intelligent Control. 2018.
3. Why you should care about the Target data breach, 2016 https://www.sciencedirect.com/science/article/pii/S0007681316000033
4. The Cryptographic Implications of the LinkedIn Data Breach, 2017 https://arxiv.org/abs/1703.06586
5. Cost of a Data Breach Report 2021 explores https://www.ibm.com/security/data-breach
6. FBI 2020 Internet Crime Report, Including COVID-19 Scam Statistics
7. https://www.fbi.gov/news/pressrel/press-releases/fbi-releases-the-internet-crime-com plaint-center-2020-internet-crime-report-including-covid-19-scam-statistics
8. Statistics of IoT connected devices installed base worldwide from 2015 to 2025 https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwid e/
9. September 2020 Healthcare Data Breach Report: 9.7 Million Records Compromised https://www.hipaajournal.com/september-2020-healthcare-data-breach-report-9-7-mil lion-records-compromised/
10. California Civil Code.
11. "Data Breach List" State of California Department of Justice https://oag.ca.gov/privacy/databreach/list

12. Trautman, L.J. (2016). Corporate Directors and Officers Cybersecurity Standard of Care: The Yahoo Data Breach. SSRN Electronic Journal.
13. Xie, X., Lee, C. and Eling, M. (2019). Cyber Insurance Supply and Performance: An Analysis of the U.S. Cyber Insurance Market. SSRN Electronic Journal.