# VISUALIZING AND ANALYSING THE NATURE OF SECURITY DATA BREACHES IN THE WORLD

Lixuan LIM, Qi Xun YEO, Xinyi TAN

**Abstract**—Cyber-attacks, amongst natural disasters, extreme weather, data fraud, and failure to address climate change have made its way to being one of the top five risks to global stability. In addition, there has been a growing trend of the use of cyber attacks to target critical infrastructure and strategic industrial sectors, possibly leading to a breakdown in the systems that keep societies functioning.[1] In view of this, governmental organizations and multinational corporations have stepped up their investment efforts in cybersecurity to minimize the damage caused by cyber-attacks, specifically data breaches. With the use of an exploratory visualization tool, this project aims to help security personnel discover visually interesting patterns that could assist them in understanding the intention behind every cyber attack and hence better channel resources to effectively tackle future data breaches. To achieve this, a visualization has been crafted after extensive data exploration and preparation. This visualization includes a scatter plot, boxplot, calendar heat map and a comparison treemap. With this exploratory visualization tool, users are able to derive insights to the correlation between the culprit behind the breach(source) and types of data compromised, even within industries, when the attacks are likely to occur, and whether other countries are also experiencing similar attacks.

**Index Terms** — Data Breaches, Cyber-attacks, Cyber Security.

## 1 INTRODUCTION

Along with the growing importance of data in this age of information technology comes an increasing focus on data protection from the point of its initial collection. Many multinational corporations, with their operations spanning across different continents, hold an abundance of global customer data. As such, they must adhere to the different data protection rules specific to the country in which they are operating in, such as the Personal Data Protection Act in Singapore and the General Data Protection Regulation in the European Union. Data breaches have been on the rise and are costly to recover from, when it occurs. [2] Through this project, we are interested to find out the correlations between the type and/or sources of attacks along with the industry or country suffered a data breach We aim to help the users of our web application to discover visually interesting patterns that could assist them in understanding the intention behind every attack and hence aid better channelling of resources to effectively tackle future data breaches.

Lim Lixuan is an undergraduate student at the School of Information Systems, Singapore Management University. (e-mail: lixuan.lim.2016@sis.smu.edu.sg)
Tan Xinyi is an undergraduate student at the School of Information Systems, Singapore Management University. (e-mail: xinyi.tan.2016@sis.smu.edu.sg)
Yeo Qi Xun is an undergraduate student at the School of Information Systems, Singapore Management University. (e-mail: qixun.yeo.2016@sis.smu.edu.sg)

By doing so, respective parties can take steps to resolve urgent issues regarding data breaches and related cybersecurity threats. All in all, this paper attempts to report on the research and development efforts taken to design and implement a web visualization application to help respective users to gain insights into cyber-attacks. More specifically, how the source and type of attacks are correlated within industries when the attacks occur, and if other countries are also experiencing similar attacks. Section 1 provides a general introduction to the project scope and issue at hand. An overview of the motivation and objectives of this research is elaborated in Section 2. The next section will expound on past works related to cyber data breaches. In Section 4, the approach taken to complete this project is discussed, along with a brief overview of the web application's architecture design. Sections 5 and 6 will include explanations on the process of selecting, acquiring, exploring and preparing the data required of the visualization, followed by the design considerations of developing the web visualization application. Next, the key findings derived from the visualization are discussed. This paper will conclude by highlighting the directions in which future research can take for possible extensions of this project.

## 2 MOTIVATION AND OBJECTIVE

Our research is motivated by a general lack of visualization tools and efforts pertaining to cyber breaches across different multinational corporations or countries. Through our research and development, we hope to develop a visualization web application that could help security personnel at the managerial level put past breaches to perspective and derive the motivations and intentions behind attacks on their own systems in the future. This will allow them to implement and prioritize necessary mitigation plans to counter these attacks. Specifically, the visualization focuses on analysing the correlation between the sources and/or the type of attack with respect to the industry or country that suffered a data breach.

Through the use of our visualization tool, we attempt to equip users with the ability to:
1) Identify when are attackers likely to execute an attack
2) Explore the relationship between the source and type of attack
3) Compare attacks between two countries
4) Establish possible correlations of a country or an organization's data breach with the time or intent via the source and/or type of breach.

## 3 RELATED WORKS

Through our research, we have noticed that there is an obvious lack of visualisations for cyber breaches in general. However, some useful examples which we have come across and gained inspiration and ideas from are: Breach Level Index [3] and World's Biggest Data Breach Hacks [4].

### 3.1 Breach Level Index

The Breach Level Index serves as a global database that tracks and analyses data breaches, the type of data compromised and how it was accessed, lost or stolen. Not only did we attain our main source of data from this site, it also contains various infographics pertaining to data breaches. Their focus is on displaying time and geographical charts and information about the data breaches. In the Fig 1, they highlighted the frequency of data breaches accompanied by icons.

They also provided an map chloropeth view of data breaches from 2013, which are segmented by continents and differentiated by the colours as shown in Fig 2.


Fig.1. Infographic from breach level index on time


Fig.2. Infographic from breach level index on location

### 3.2 World's Biggest Data Breach Hacks

This site presented data breaches in an interactive application, with focus as to which industry the organisation is from and the type of data breach in a bubble plot. The number of records affected is shown with respect to the size of the bubble and details when mouse over, the colour only indicates an interesting story. They are plotted with respect to time as shown in Fig 3. However, we felt that it is not very efficient in terms of how user-friendly it is, given the vertical presentation of the bubble plot with time.

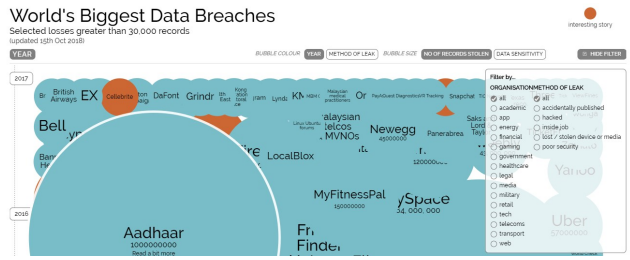


Fig.3. Screenshot from World's Biggest Data Breaches

Both examples allowed users to have an overview of data breaches from 2013 to date. They gave users insights on the size and impact of the breach, within the user's focus of exploration. These infographics and interactive visualizations primarily allow users to analyze potential trends of data breaches over the years. We felt that, although they are visually appealing, they are not able to provide information as to the motive or intention of the attacks, which will uncover how and where companies or governments should focus on to further step up cybersecurity efforts to prevent future data breaches. This research project attempts to augment current efforts by focusing on the motive and intentions of these attacks along with its patterns in a less aggregated manner.

## 4 Visualization Approach

To develop a visualization web application that assists users in understanding the data and to derive impactful insights, 3 main activities were conducted with respect to our dataset. These 3 activities include the process of (1) selecting, acquiring, exploring and preparing the data, (2) planning the visualization tool and (3) creating the visualization tool. The process of selecting, exploring and preparing the data forms the foundation of the research done for the project. The current section will elaborate on the stage of selecting, exploring and preparing the data.

### 4.1 Data Selection

After selecting our area of focus for the project, we sourced various websites for possible datasets that we can explore on and visualise. In the end, we selected the data provided by Gemalto's Breach Level index, as mentioned previously in Section 3a. It contained reported global data breaches from 2013 to date. We appreciated how they allocated a risk score to each attack and found that it better represented the impact of each reported data breach. The risk score is tabulated by the formula: *Log 10 (Type of Data * Source of Breach * Whether Action was taken).*

**Log10 (N x t x s x A)**

***Where:***

**N= the total number of records breached, or, in the case of intellectual property loss the equivalent dollar loss.**

**t= the type of data in the records**

*values*

1 Nuisance (email addresses, affiliation, etc.)
2 Account access (username/passwords to social media, websites, etc.)
3 Financial access (bank account credentials, credit card data)
4 Identity theft (information that can be used to masquerade as someone)
5 Existential data (information of national security value or threatens business survival)

**s= source of the breach**

*values*

1 Lost device such as a laptop, DVD, or USB thumb drive
2 Stolen device
3 Malicious insider
4 Malicious outsider
5 State espionage

**Action= whether or not the stolen data has been used to cause harm be it identity theft, credit application, or bank account withdrawals**

*values*

1 No action
5 Publication of embarrassing or harmful information (Wikileaks, hacker logs, etc.)
10 Use of financial identity to obtain funds or apply for loans

Fig.4. Table of Risk Score Formula

| Category | Breach Level Index Score | Characterization |
|---|---|---|
| 5 | 9-10 | Breach with immense long term impact on breached organization, customers and/or partners. Very large amount of highly sensitive information lost (usually 10-100+ million records). Massive notification process. Potentially existential financial loss for breached organization in remediation and related costs. Use of lost sensitive information seen. |
| 4 | 7-8.9 | A breach with significant exposure to business, Legal and/or regulatory impact. Large amount of sensitive information lost (usually hundreds of thousands to millions of records). Significant notification process costs involved and public image impact. |
| 3 | 5-6.9 | A breach with likely short to midterm exposure to business. Legal and/or regulatory impact. Usually tens of thousands of records of moderate sensitive information involved . Some breach notification and financial loss. |
| 2 | 3-4.9 | A breach with low long-term business impact. Usually involves the loss several thousands of records of semi sensitive information. Limited breach notification and financial exposure. |
| 1 | 1-2.9 | A breach with no material effect. Less than one thousand records. breach notification required, but little damage done. |

Fig.5. Table of Risk Score Interpretation

### 4.2 Data Acquisition

Python was used to help us crawl for data from the webpage, clean the raw data source to be read as a CSV. format the required data and even create the necessary new columns for our consumption and visualizations.

### 4.3 Data Exploration

Tableau, a software that produces interactive data visualizations focused on business intelligence, was used to help us analyse the nature of our data columns and developed suitable types of data visualizations given the nature of our data. We used it to remove duplicates and normalize dates. We were also able to validate our web application's end results and ensure that our visualisation was pulling the correct information given the user's inputs, with the resultant visualizations produced by Tableau.

### 4.4 Data Preparation

R was also used to clean the raw data sources and more importantly allowed us to include logic to extract relevant data as per user's inputs.

## 5 Our Design

Our team researched various technologies and designs related to creating effective visualisations. Initially, we agreed upon using D3.js as the main framework for the visualisations. However, after consulting with Professor Kam Tin Seong who recommended using R Shiny for the revised visualisations, our team decided to work with Rstudio R Shiny framework, due to its user-friendly interface for both developers and users and its wide array of resources published by both the company and its community for easier incorporation of various interactive elements. We have deployed our web application onto shinyapps.io which is a deployment server for Shiny applications.

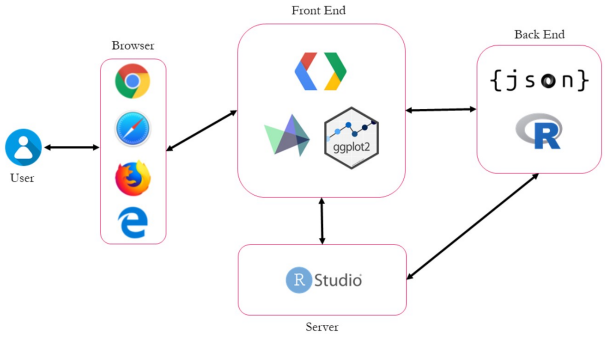


Fig.6. The Defender's Interactive Web Application Architecture Diagram

To explore the different charts that would be considered for the visualization, an iterative process of researching and brainstorming was done. A storyboard was then developed to create the final web visualization application before we commenced on the development phase. The following illustrates the design considerations, during our planning process, for each tab in our web visualization application

### 5.1 Iteration One

Our initial plan was to develop an application to analyse the data breach index, along with selected human development index (HDI) indicators provided by the United Nations Development Programme. This is shown via:

*Parallel coordinates*, showing the distribution of each feature amongst the different countries

*Radar chart*, showing the breakdown of different Human Development Index indicators of both countries shown side-by-side

*Line chart*, comparing the breaches between the two countries determined via the dropdown list

*Tree map* comparison chart, to display a high-level view of our breaches and displaying the breach details at the same time.

We hypothesize that there would be correlations between a country's HDI and the frequency of data breach attacks within the nation. However, we were unable to derive any notable results to back our hypothesis. Moreover, through reviews with our Professor, it was established that the additional HDI data is too insignificant to garner much useful insight from. Hence, this iteration was concluded with a plan to revise this hypothesis, changing the types of charts visualised.

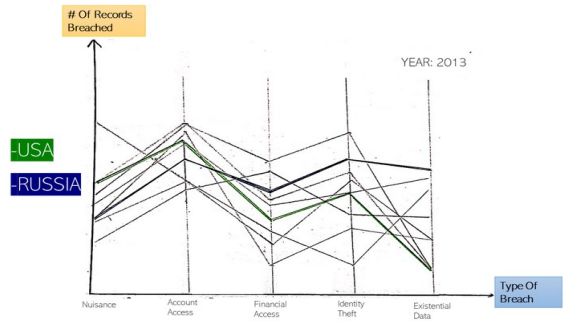


Fig.7. Parallel Coordinate Sketch

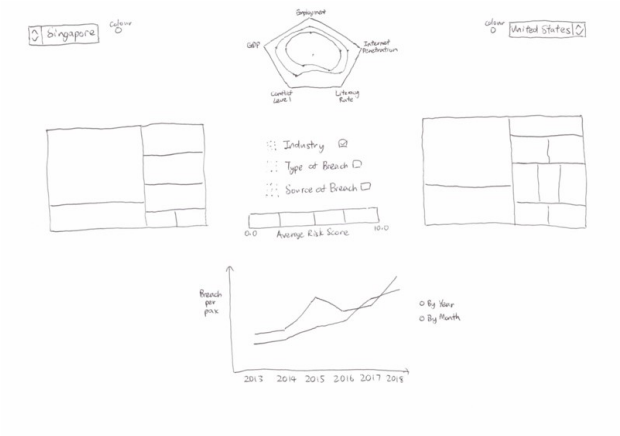


Fig.8. Sketch of initial risk comparison

### 5.2 Iteration Two

After further exploration and ideation on the feedback received from Iteration 1, we have decided to focus only on the data breach dataset and make sense of its relationship between the different data columns. This is shown through:

#### 5.2.1 Scatterplot

The distribution of the sources of attacks for each type of attacks and vice versa are shown, in a scatterplot. This tab gives an overview of the relationship between the source and type of attack and complemented details via the tooltip. It allows the user to identify the company, risk score and industry with respect to the data breach point in the plot. Finally, the risk score is denoted by the colour intensity of the plotted point.

#### 5.2.2 Boxplot

To understand across industries, what are the distribution of the types and sources of breaches, along with the measured risk score or the number of records breached. Additionally, users can also narrow down to the year and industry of the breach they wish to explore. The risk score shows the user to determine the extent of damage in relation to the attack, this presentation enables users to better prioritize on the type of attacks they wish to defend themselves against. Users can toggle between source or type of breach to observe the risk scores or number of records breached accordingly as shown in Fig 9.

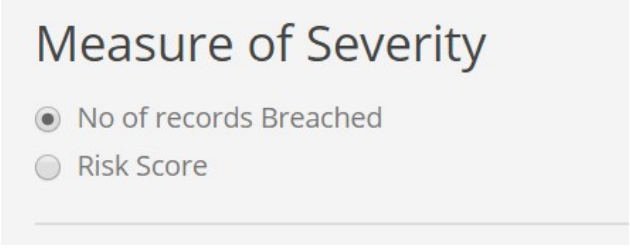

Fig.9. Screenshot from The Defender's Risk of Attacks Tab to show radio toggle

### 5.2.3 Calendar Heatmap

To display time series data of the security breaches over the years, we used a calendar heat map which can be filtered by the different industries as shown in Fig 10. Users can then observe any spikes or anomalies within the same time frame, across the different industries and take possible precautions for future data breaches. Mouse over to observe the number of records breached in that day, the colour intensity of each day also depicts the number of records breached in that day.

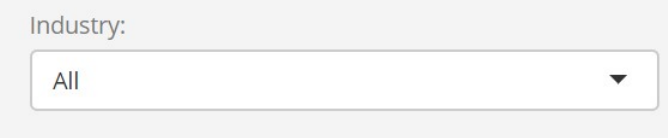


Fig.10. Screenshot from The Defender's Timeline of Attacks Tab drop list toggle

### 5.2.4 Tree Map

To understand in greater detail the characteristics of breaches affecting each country, we used two tree maps compared side-by-side. The tree map provides a macro view of the breaches before drilling down into the micro details at an industry level by clicking on the desired segment in the tree map. Users can compare alone or between countries to gain more insights about the selected countries. Multi-toggles functionality is as shown in Fig 11: radio button to filter Tree map by source or type of breach as the first level and view the other as a second level of the analysis. Toggle between which countries and/or continents to select for comparison, slider to select the period to be observed. Tooltip shows more details when mouse over to show: Number of records breaches, respective risk score and the computed risk score per record.

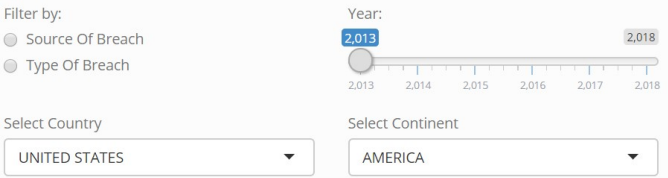


Fig.11. Screenshot from The Defender's Treemap Comparison Tab, multi-toggle

## 6 Key Findings and Insights

### 6.1 Scatterplot

Ever wondered behind every breach used for financial gains, is it more likely to be done by a malicious insider or outsider?

Gemalto uncovered that identity theft represented the leading type of data breach, accounting for 69% of all data breaches. Hence, we decided to study how the types of breach correlate with the sources of breach.

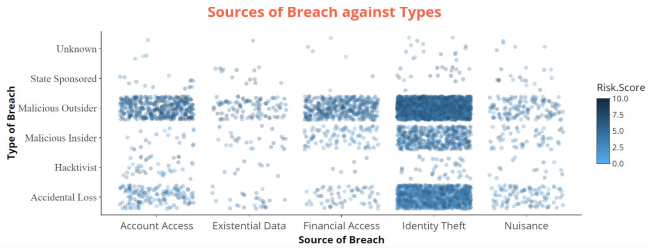


Fig.12. Screenshot from The Defender's Overview Tab

From our results, we were able to determine that the highest occurrence is between identity theft by malicious outsiders followed by accidental lost. Secondly, it is observed that the main source of breach is Identity theft, while the main type of breach is identified as the malicious outsider. From this, we can conclude that identify theft seems to be easy to commit and this can possibly be attributed to the fact that we are giving away our information to the web too easily nowadays, every purchase or creation of an account for the use of an application demands for our personal information.[5] Hence this makes everyone vulnerable to the threat of identity theft to malicious outsiders. To tackle the situation, on top of the current measures to tackle threats from malicious outsiders, organizations may consider educating the consumers or to enhance internal security to better prevent an ID Theft Breach.

### 6.2 Boxplot

Ever wondered for the healthcare industry, how susceptible are you to breaches used for financial access?

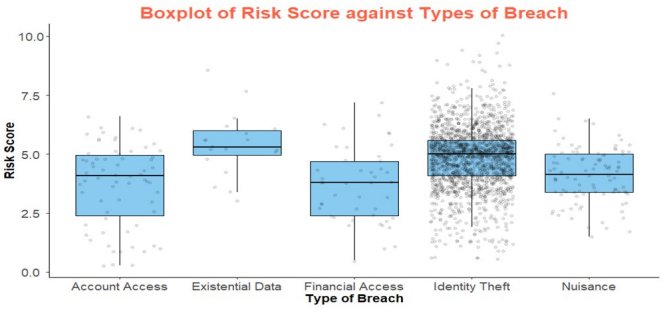


Fig.13. Screenshot from The Defender's Risk of Attacks Tab

From our results, we were able to gather some interesting perspectives that are contrary to conventional wisdom. For example, one would expect that the risk scores for financial access to be one of the highest considering that items considered under financial access such as account number, pin number or bank records are extremely important items that could result in considerable damage once breached. Surprisingly, financial access was the type of breach which resulted in one of the lower risk scores across all industries. This could be a result of more stringent standards imposed in the said documents or possibly that the definition of financial access breaches extends to other documents which may not be as dangerous such as audit records. These observations provide a good platform for the users of the system to reason about these unexpected findings.

### 6.3 Calendar Heatmap

Do you think we will be able to spot correlations between the intended industry and the time of the attack?

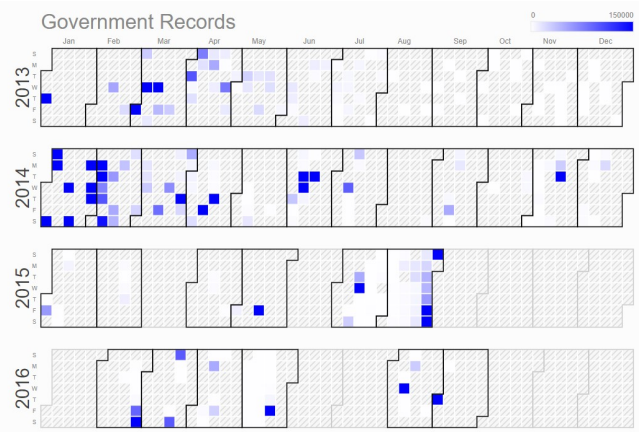


Fig.14. Screenshot from The Defender's Timeline of Attacks Tab

Some notable observations from the Calendar Heatmap are:

- 2014 is the year that most data breaches occur and most aggressive in January, across all industries
- 2017 was a lull period with reported data breaches in only 3 industries
- Within the healthcare industry, data breaches usually happen in the beginning and end of the year
- Data breaches are most common in the Healthcare industry, next by governmental records and lastly technology records which are closely followed by financial records

Upon further research on the trend observed, we have hypothesized that the healthcare industry could have lower standards of cybersecurity mitigation efforts given the use of legacy systems. Coupled with the fact that it is more lucrative and easier to commit a healthcare fraud as compared to credit card frauds. [7] Even though there is no significant period of the year which attacks execute data breaches, from this visualization we can conclude which industries are the main focus of the attackers and hence determine where to prioritize strengthening cybersecurity efforts

### 6.4 Tree Map

Which industry is most threatened by breaches in a country and what are the causes of these breaches?

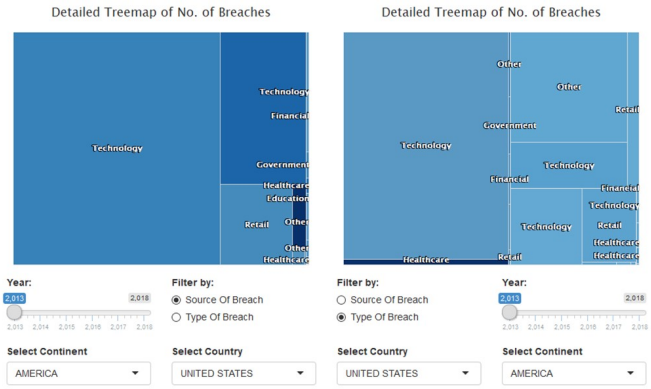


Fig.15. Screenshot from The Defender's Treemap Comparison Tab

The side-by-side comparison is useful in showing hidden characteristics in the data either for a single country or across 2 different countries. For example, countries generally thought of to be similar in terms of characteristics, such as the United States (US) and the United Kingdom (UK), might differ in terms of the distribution of breaches. US has technology as the industry with the largest number of breaches while UK has financial as the industry with the largest number of breaches. Moreover, it is easy to compare between the characteristics of the breaches between 2 different years to possibly come out with hypotheses for the differences or similarities for these years. Some relationships can be observed about the individual relationships between the type and source of breaches in each country broken down in specific industries. For instance, we can deduce that most breaches affecting the US in 2013 are due to malicious outsiders for identity theft in the Technology industry. We can also deduce that malicious outsiders seem to use records stolen from Others industry for account access in the US. There are many more observations that can be made just that one needs to patiently sieve through the treasure trove of data and focus on the relevant areas for the information that one requires.

## 7 Conclusion

Data breaches have become increasingly frequent with growing aggressiveness [8], this research project aims to complement the efforts by security personnel at managerial level to identify possible motivations of these attacks. Through this research, different countries seem to have very different characteristics in terms of breaches encountered. Different industries also seem to have very different characteristics for the breaches encountered. The strength of our visualization application lies in its exploratory nature and it represents a means to guide users in discovering more information which may otherwise, not be presented in our visualization. Rather than concluding about the nature of the visualisations, it implores us to think of the reason behind the observations found.

In future, this research can be further extended to analyse other possible data breaches and investigate any relationships to any other crimes, cyber or not. Additionally, our application can be used to examine other macroeconomic factors and indicators of a country's cybersecurity performance. These determinants include percentage of employment in cybersecurity sectors, private and public cybersecurity expenditure. This application seeks to yield more targeted and effective approaches (by understanding the nature of the breaches) to help our users, personnel from the security departments of private or governmental organizations to defend against cyber-attacks. In conclusion, we would like to emphasize the importance of good practices when releasing personal information on an individual level, to help protect defend against the threat of data breaches more effectively.

## Acknowledgments

The team would like to thank Professor Kam Tin Seong for his guidance throughout this project.

## References

[1] Graham, A. (2018, January 24). Cyber-attacks are now one of the biggest threats to global society. Retrieved from https://www.itgovernance.co.uk/blog/cyber-attacks-are-now-one-of-the-biggest-threats-to-global-society.

[2] NortonOnline. (n.d.). What is a data breach? Retrieved from https://us.norton.com/internetsecurity-privacy-data-breaches-what-you-need-to-know.htm1979.

[3] Gemalto. (n.d.). Data Breach Statistics by Year, Industry, More. Retrieved from https://breachlevelindex.com/.

[4] McCandless, D. (2017, March 21). World's Biggest Data Breaches & Hacks. Retrieved from http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

[5] Why is Identity Theft so Common? (2011, December 27). Retrieved from https://www.identitytheftmanifesto.com/why-is-identity-theft-so-common/

[6] Insider vs. Outsider Data Security Threats: What's the Greater Risk? (2018, April 06). Retrieved from https://digitalguardian.com/blog/insider-outsider-data-security-threats

[7] Humer, C. (2014, September 24). Your medical record is worth more to hackers than your credit card. Retrieved from https://www.reuters.com/article/us-cybersecurity-hospitals/your-medical-record-is-worth-more-to-hackers-than-your-credit-card-idUSKCN0HJ21I20140924

[8] Armerding, T. (2018, January 26). The 17 biggest data breaches of the 21st century. Retrieved from https://www.csoonline.com/article/2130877/data-breach/the-biggest-data-breaches-of-the-21st-century.html