Research papers

Statistical analysis of relationships of US organisations' size, popularity, age and location to frequency of data breaches

Received: 24th March, 2021

Ohud Saud Alqahtani
Department of computer science, King Khalid University, KSA

Ohud Saud Alqahtani is PhD student in Information Systems department with minor in data science at University of Maryland, Baltimore County, and Lecturer in Computer Science Department at King Khalid University. Research interests include machine learning, privacy preserving and statistical analysis. She is especially interested in detecting and preventing data breaches in organisations by analysing risk factors of reported data breaches from unintentional disclosure or hacking.

Guraiger, Abha 62529, KSA
Tel: +966 55-375-3620; E-mail: oh8@umbc.edu

Zhiyuan Chen
Department of Information Systems, University of Maryland Baltimore County, USA

Zhiyuan Chen is a Professor, Information Systems Department, University of Maryland, Baltimore County. He has PhD from Cornell University in Computer Science, an MS in Computer Science from Fudan University, China, and a BS in Computer Science from Fudan University, China.

1000 Hilltop Circle, Baltimore, USA
Tel: +1 410-455-8833; E-mail: zhchen@umbc.edu

Abstract Given the widespread occurrence of data breaches, it is useful for consumers to learn which factors of an organisation, for example, size, popularity or location, will contribute to increased data breach risks. Existing work on risk assessment requires detailed internal information of an information system, which is not available to the public. Furthermore, organisations typically do not want results of such analysis of their IT systems to be made public. This paper conducts comprehensive statistical analyses of the relationships between publicly available information to frequency of data breaches. The publicly available information includes size-related characteristics such as revenue, number of employees, population served and enrolment, popularity-related characteristics such as number of Google Search results, age of the organisation and location of the organisation. We used Pearson, Spearman and Kendall correlation analysis methods to test whether these characteristics are indicators for frequent data breaches for different types of US organisations. We also used linear regression to predict the frequency of data breaches. The results verified that many of these indicators have significant correlation to organisations' frequency of data breaches. The result of this paper can help consumers make more informed decisions with respect to risks of data breaches.

KEYWORDS: data breaches; statistical analysis, correlation and multiple regression models, security and privacy

INTRODUCTION

A data breach is an intentional or unintentional disclosure of sensitive and confidential data, for example, personally identifiable information, payment card information or personal health information to be viewed and misused by unauthorised parties. Data breach can be theft of information, a malicious act on systems and unauthorised access to the network or in any form it affects individual's privacy and businesses' popularity. In the United States alone, the number of recorded and publicly announced data breaches is over 1,000 in 2016[1] and over 8,840 since 2005, with over one billion personal records being revealed.[2] Data breaches often lead to identity thefts, and over 15 million people in the United States were hit by identity theft in 2016 alone.[3]

As consumers are exposed to constant risk of data breaches, it is useful to find out the factors that contribute to higher number of data breaches such that consumers can make more conscious decisions to reduce risks and control their personal information. For example, suppose a consumer wants to choose a bank to open account, shall she use a bigger national or international bank or a smaller community bank considering risks of data breaches? Should she go to bank that is located in city X or city Y?

There exists work on vulnerability and risk analysis of a given information system[4–7]). Such work, however, requires detailed internal information of information systems being analysed, which is typically not available for ordinary consumers. Furthermore, companies or organisations typically do not want the results of risk or vulnerability analysis being revealed to consumers because they may lose business.

This paper conducted a comprehensive analysis of the relationship between an organisation's publicly available information such as revenue and location with frequency of data breaches. The results can be used to help consumers make more informed decisions based on data breach risks.

This paper makes the following contributions:

- We analysed the relationship between various publicly available characteristics of an organisation with the frequency of data breaches. The analysis is based on both Privacy Rights Clearinghouse dataset[8] and data collected from various sources about other characteristics (eg revenue, number of employees) of an organisation.
- We used linear regression to predict an organisation's risks of data breaches using the abovementioned characteristics.
- We analysed relationship between location of an organisation and its data breach risks.
- We discussed our findings which can help consumers make more informed choices based on data breach risks.

The rest of the paper is organised as follows. We first present related work following by a description of our dataset. A methodology is explained before an analysis of relationships between certain characteristics specified in the methodology found in results section. The paper ends with a conclusion and a discussion of research future directions and current limitations.

RELATED WORK

The extensive literature review can be roughly divided into six categories: (1) privacy protection techniques and privacy policy; (2) data breach analysis; (3) the impact of data breaches; (4) behavioural analysis of human error; (5) data breach prediction and (6) vulnerability and risk analysis.

There has been a lot of work on protecting data privacy, including methods for anonymising data before being shared,[9,10] methods for privacy preserving data mining[11–13] and work on privacy policy issues related to data privacy.[14] Such work, however, only focuses on protecting privacy during data collection and sharing. Although some data breaches happen when data is

collected or shared, most data breaches are results of hacking or insider attacks where existing privacy protection techniques have limited use.

Privacy policies and standards are important and critical part to protect organisation's internal system. Many researchers studied privacy policies and standards and how to utilise it to reduce avoidable breaches. Kobsa[15] suggested that privacy policy needs to be dynamically customised to organisation's needs and then enforced within the organisation for better privacy protection. Culnan and Williams[16] discussed the morality of organisations' responsibility to protect data privacy. On one hand, Soomro, Shah and Ahmed[17] suggested that management role should be considered in information security. While Bauer, Bernroider and Chudzikowski proposed that user's compliance with security policies is crucial to reduce breaches.[18]

Several studies focus on analysing data breaches.[19–21] The Verizon Data Breach Report[22] reported some statistics on the types of attackers, victims, means of attack, etc.[23] and analysed statistics of lawsuits filed by victims of data breaches. Ponemon conducted annual global survey to explain breach trends and costs. These studies focus on cause of breaches and trends rather than examining relationships between data breaches and characteristics of an organisation.

Firms experiencing data breaches can be a subject to fine, lawsuits or other recovery costs. A lot of work investigated cost of breaches in business organisations[24–26] Similarly, Acquisti et al.[27] and Gatzlaff et al.[28] examined the impact of breaches on business market value. Acquisti et al.[29] found that, in the case of privacy breaches, the source of the announcement has an influence on the magnitude of the effect of a breach on a company. Martin et al.[30] focused on breaches' impact on customers and firms' performance. Ponemon Institute collaborated with IBM to conduct an annual cost of data breaches' study and created a calculator to estimate breach cost.[31] They reported that an average cost of breach is US$3.86m and average time is 196 days to identify a breach with additional 69 days to contain it globally. In the United States, it takes on average 201 days to identify a breach but only 52 days to contain it. Their research also showed that the cost of failing to protect customers' private information is on the rise and that a single breached record cost US$197. The impact of breach announcement on stock prices was analysed specifically to show its short-term effect.[32]

Behavioural analysis studies data breaches caused by human error. To learn how to reduce human error[33,34] applied different behavioural studies and analysed errors stages to suggest how to avoid future errors. Researchers studied breaches[35,36] by insider attacks and employees. These studies focus on one type of breach which is caused by insider or employees whereas, in our study, we study all breaches caused by internal or external entities.

There is also work predicting future data breaches. Gao, Cheng, He, Susilo and Li built Naïve Bayes classifier to predict specifically substitution-then-comparison (STC) attack.[37] Bai, Jiang and Flasher applied regression analysis on dataset of 141 hospitals within the United States to link size of hospital to higher breach risk[38,39] and analysed data breaches in higher education institutions. Liu et al.[40] built a classifier to predict security incidents using an organisation's network information.[41] Such information, however, is often difficult to obtain and is usually not available to general public. Unlike existing work, this paper focuses on correlations between publicly available characteristics of organisations with data breach frequency.

There is a rich literature on vulnerability and risk analysis of a given information system.[42–44] while an overview can be found can be found at.[45] Existing risk

analysis methods, however, require detailed information of internal of information systems, which is typically not available for ordinary consumers. Furthermore, companies or organisations typically do not want the results of risk or vulnerability analysis to be revealed to consumers because they may lose business.

Alqahtani et al.[46] divided organisations into small and big ones and showed that big ones have higher data breach risks. They also used Pearson correlation analysis to study the relationship between a few size-related factors (such as revenue and enrolment) to the frequency of data breaches. The data used in their study, however, has very small sample sizes. More importantly, their data does not follow normal distribution, so Pearson correlation is not the most appropriate method. In this paper, we use data with larger sample sizes and consider more factors such as those related to popularity, age and location. In terms of analysis, we use Spearman rank correlation methods in addition to Pearson because they do not require data following normal distribution. We also used linear regression to predict frequency of data breaches.

DATASET

The full Privacy Rights Clearinghouse dataset

There are many data breach datasets available collected based on organization type sector or breach method. We reviewed five datasets: Information is beautiful,[47] Data. World,[48] VERIS community database,[49] Breach Portal[50] and Privacy Rights Clearinghouse[51] to select dataset for this study. As we focus on data breaches in United States and want to study significant data breaches with different causes, we excluded datasets that contain global data, contain only hacking incidents or have small number of records. Thus, we selected Privacy Rights Clearinghouse dataset because it covers public US data breaches that have been collected from different sources and represent all sectors chronologically since 2005 and updated regularly.

Privacy Rights Clearinghouse dataset contains public data breaches since 2005 within United States. It has over 8,000 data breaches, making it one of the largest US data breaches dataset available online.[52] It has columns showing date, name of organisation, organisation type, location (city, state), cause of the data breach, number of records breached and a short text description including information such as what is leaked and source of information. Next, we show some summary statistics about this dataset.

Figure 1 shows the number of data breaches by organisation type from 2005 to 2017. The organisation types include government (GOV), medical (MED), education (EDU), business – other (BSO), businesses-financial and insurance services (BSF), businesses-retail/merchant — including online retail (BSR), and nonprofit (NGO). It is obvious that breaches are increasing for the past four years with medical breaches leading the trend.

Table 1 shows the breakdown of data breaches with respect to types of organisations. It reports number of breaches, percentage of breaches, number of records breached and percentage of breached records for educational, governmental, medical, business (we combined BSO, BSF and BSR into business type) and nonprofit organisations in the United States collected in the dataset. Medial sector accounts for almost 50 per cent of breaches in the dataset, whereas over 95 per cent of breached records belong to business and financial organisations. Nonprofit organisations received the least number of breaches and least number of breached records.

Table 2 shows the trends of data breaches for different types of organisations. The results show that medial organisations have more breaches in recent years. On the other hand, governmental organisations have fewer breaches recently, possibly due to

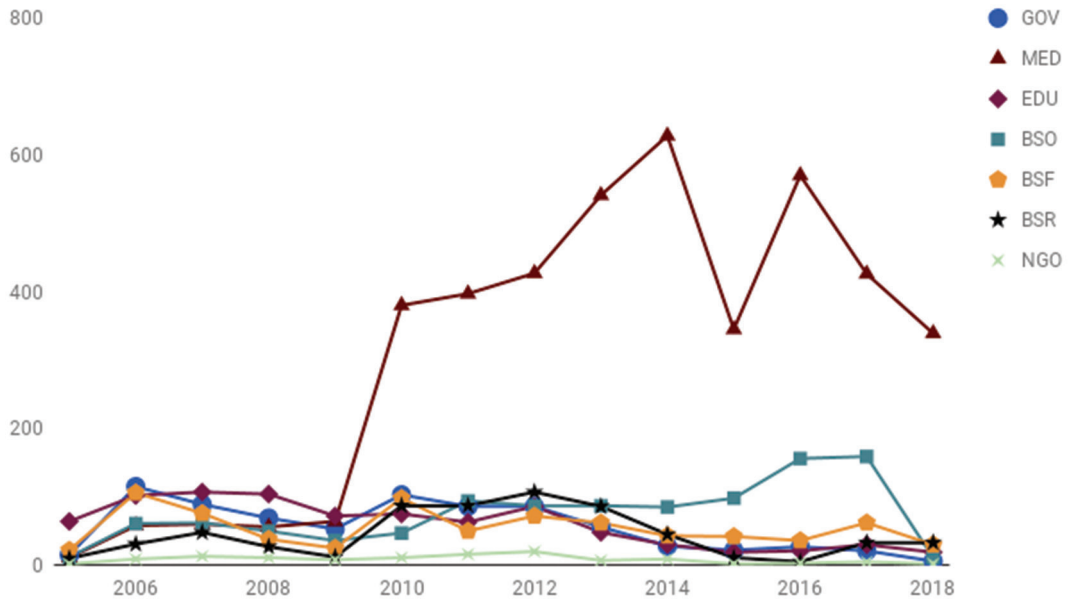

Figure 1: Number of data breaches by organisation type between 2005 and 2017

TABLE 1: Breaches and records breached in Privacy Rights Clearinghouse dataset by organisation types

Organisation type	Number of breaches	Breaches%	Number of records breached	Records %
Education	818	9.98	25,162,790	0.23
Government	776	9.46	227,458,042	2.07
Medical	4,068	49.6	230,737,451	2.09
Business	2,417	29.5	10,514,112,568	95.5
Nonprofit	118	1.43	8,434,545	0.07

TABLE 2: Number of breaches from 2005 to 2017 by organisation type

Year	GOV	MED	EDU	BSO	BSF	BSR	NGO	Unknown
2005	15	11	64	12	22	10	2	0
2006	115	58	102	61	106	31	9	0
2007	89	60	107	62	76	48	13	0
2008	69	56	104	50	38	27	11	0
2009	53	64	72	36	25	12	8	0
2010	103	380	75	47	98	87	11	0
2011	86	397	63	94	50	87	16	0
2012	86	427	86	87	72	107	20	0
2013	56	541	48	87	62	86	7	2
2014	28	628	30	85	43	45	9	0
2015	22	345	19	98	42	11	2	1
2016	27	570	21	156	36	5	3	4
2017	21	426	30	159	62	33	5	29

Notes: GOV, government; MED, medical; EDU, education; BSO, business – other; BSR, businesses-retail/merchant; NGO, nonprofit organisation.

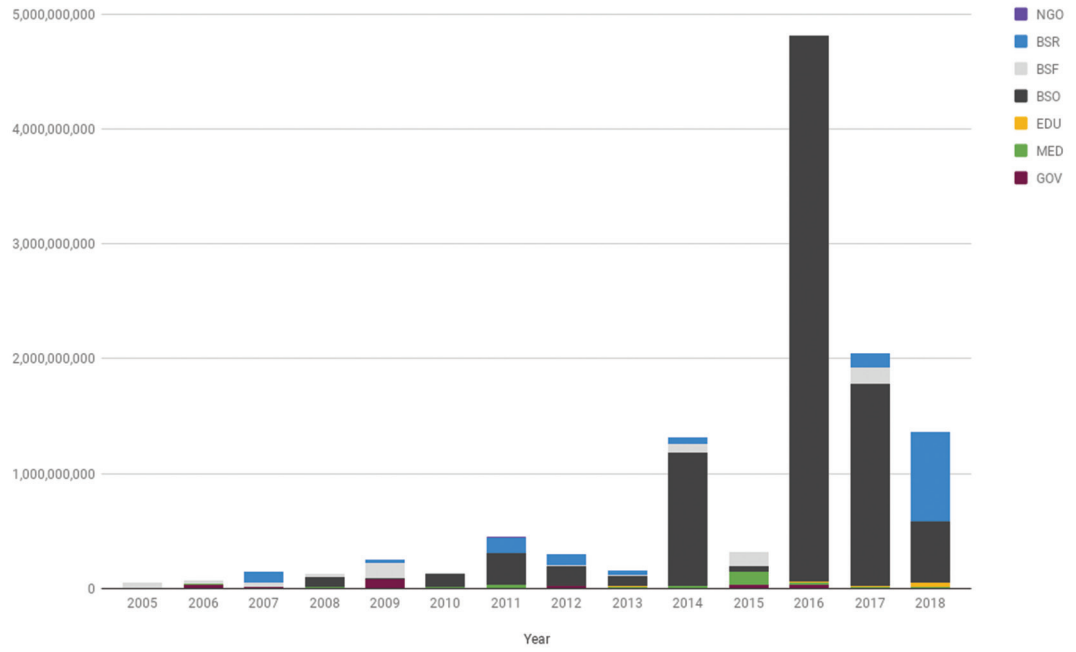

Figure 2: Records of breaches by organisation type from 2005 to 2017

stepped-up effort on cybersecurity. On the other hand, nonprofit organisations show the least number of breaches.

Figure 2 reports the number of records breached per year by organisation types. Despite the fact that medical organisations have the most breaches, business organisations have the highest number of breached records as shown in Figure 2. A peak of BSO breaches occurred in 2016 due to six major breaches in social networking/browsing websites: (1) Myspace breach that resulted in 360 million records breached, (2) Facebook breach with minimum of 32 million records breached, (3) Yahoo breach with one billion records hacked, (4) LinkedIn breach with 117 million records breached, (5) 65 million records breached as a consequence of Tumblr hack and (6) Friend Finder hack that caused 412 million breached records.

The Privacy Rights Clearinghouse dataset contains a breach type column which explains the cause of the data breach. Table 3 shows breaches categorised by breach type from 2005 to 2017. Breach types include card fraud (CARD), unintended disclosure (DISC), hacking or malware (HACK), insider (INSD), physical theft (PHYS), portable device (PORT) and stationary device (STAT).

On one hand, the number of data breaches attributed to hacking has been increasing and is the most dominant cause since 2014. On the other hand, physical theft and unintended disclosure are also among the top causes recently.

Sample data and additional characteristics

The focus of this paper is to study correlations between publicly available characteristics of organisations and data breaches. Privacy Rights Clearinghouse dataset, however, does not contain many characteristics of organisations such as those related to size and popularity. So, we need to collect additional characteristics from other sources. There are thousands of organisations in the dataset, so it is infeasible to collect additional characteristics for all of them. So, we created randomised samples

TABLE 3: Number of breaches from 2005 to 2017 by breach type

Year	CARD	DISC	HACK	INSD	PHYS	PORT	STAT	Unknown
2005	0	20	48	10	8	38	10	2
2006	3	83	75	32	39	186	48	16
2007	2	98	71	25	43	163	36	16
2008	5	78	57	31	53	99	22	10
2009	4	52	53	30	53	62	10	6
2010	13	109	106	103	258	142	37	33
2011	14	112	162	95	226	119	29	36
2012	11	137	246	88	230	112	21	40
2013	11	176	210	101	214	111	24	42
2014	2	195	342	52	204	39	9	34
2015	0	156	189	13	130	45	1	6
2016	0	232	402	12	119	46	2	6
2017	2	160	368	11	69	10	0	42

Notes: CARD, card fraud; DISC, unintended disclosure; HACK, hacking or malware; INSD, insider; PHYS, physical theft; PORT, portable device; STAT, stationary device.

from each type of organisations including governmental, educational, medical, business and nonprofit organisations. Each sample has around 30–100 organisations. Each record contains organisation history of breaches since 2005, which might have single or multiple breaches.

We then manually collect publicly available characteristics of each organisation in the sample. The first category of characteristics is related to the size of the organisation, including revenue, the number of employees,[53] size of budget and size of population served (applicable to government agencies). For educational organisations, we also collected enrolment in 2018.

The second category of characteristics is popularity. This feature represented by Google counts, which is the number of search results returned by google when we use it to search the name of the organisation. For educational organisations, we also collect latest US News Ranking.

The third category of characteristics is location. We aggregate each organisation's location to state level. It is also not meaningful to compare states directly as some states have much bigger population than others and is more likely to have more data breaches. So, we compute a location rate feature which is the number of breaches per million population in that state.

The last category is an organisation's age (number of years since the organisation was first established). We also ensure each organisation in our samples was founded before 2005 and still operating in mid-2018 for fair representation.

We collect most of these characteristics by visiting the organisation's website and extract information from there. We assume that these organisations publish accurate numbers on their website.

METHODOLOGY

We want to answer the following research questions:

1) Are characteristics related to an organisation's size positively correlated with the frequency of data breaches?

2) Are characteristics related to an organisation's popularity positively correlated with the frequency of data breaches?
3) Is the age of organisation correlated with the frequency of data breaches?
4) For a given organisation, is the frequency of data breaches correlated with the number of data breaches per million people in the state the organisation is located (location rate)?
5) Can frequency of data breaches be accurately predicted by the previous characteristics?
6) Are there states having significantly higher number of data breaches per million population than average?

To answer these questions, we first conducted Shapiro–Wilk normality test to check whether data follows normal distribution. We then describe analysis methods used in the paper.

The Shapiro–Wilk normality test

We performed Shaprio–Wilk test of normality to check whether data follows normal distribution. We tested three samples: medical, government and nonprofit organisations , and the p-value is extremely low (≤ 0.001) for all characteristics. So, the null hypothesis (data is normally distributed) is rejected and the data does not follow normal distribution. So, it is important to use analysis methods that do not assume normality of data.

Analysis methods

We used three different correlation analysis methods in this paper:

1) Pearson correlation: It measures the linear relationship between two variables and is the most popular way to measure correlation. As the data does not follow normal distribution, we also consider two other correlation methods.
2) Spearman rank correlation: This is a popular rank correlation method, that is, it analyses monotonic relationship between rankings of two different variables. It is also nonparametric, meaning it makes no assumption of data distribution. So, it works well for data not following normal distribution. It is also quite robust to outliers.
3) Kendall rank correlation (also called Kendall's τ): This is another rank correlation coefficient. It checks how similar orders of two variables are. It is also nonparametric and appropriate for data not following normal distribution. Nevertheless, as Spearman is sufficient, we have not listed Kendall correlation results.

We also conducted linear regression to predict the frequency of data breaches based on characteristics of an organisation. All analyses were done using R.

RESULTS

We report results for each type of organisation, including governmental (Section 5.1), educational (Section 5.2), medical (Section 5.3), business (Section 5.4) and nonprofit organisations (Section 5.5). We also analyse relationship between location (the state an organisation is located) and frequency of data breaches in relationship between location (state) and frequency of data breaches section.

Results for governmental organisations

Since 2005, a total of 773 data breaches occurred in governmental organisations, which accounts for 9 per cent of the total breaches reported in the dataset. These data breaches resulted in 227,407,542 breached records. We created a random sample of 42 governmental organisations and used the following features for governmental dataset to test their relation to breaches frequency:

1) Popularity measured by Google counts.
2) Size measured by population served, number of employees and size of budget. For state government, we use the population of the state. For federal government, we use the entire US population. For city or other local government, we use the population in that city or local jurisdiction.
3) Age of the organisation (number of years since the organisation was first established).
4) Location rate, that is, number of breaches per million people at the state the organisation is located.

A screenshot of government dataset is shown in Figure 3. We test correlation of the features to the dependent variable (breaches frequency) using Pearson and Spearman methods; results are listed in Table 4.

All features except age show positive correlation with frequency of data breaches. This means organisations with larger sizes (larger budget, more employees, serving more people) are more likely to have more data breaches. These correlations are also statistically significant in most analysis methods.

In addition, organisations with higher Google counts will have higher number of breaches, but this correlation is not statistically significant.

Age of an organisation is negatively correlated with number of data breaches, but the correlation is not statistically significant.

Interestingly, location rate is positively correlated with number of data breaches, and the correlation is statistically significant for the two rank correlation methods. This means that organisations located in states that have higher rate of data breaches are more likely to have more data breaches themselves.

Table 5 shows the result of linear regression model to predict the frequency of data breaches using the earlier characteristics:

The model has an adjusted R square of 0.383. This means that there is clear correlation between the independent

Name	founded	age	location	LocRate	Population	Google coun	RecordSum	Budget	EmpNum	breachesNum
Accomack County Virginia residents	1634	384	Virginia	23.02	0.03	327	35000	60148743	14137	1
Alabama Department of Public Health	1875	143	Alabama	16.05	4.88	83200	9754	481300000	4000	3
U.S. Department of Agriculture (USDA)	1862	156	District Of Co	217.58	323.1	724000	376350	1.37E+11	100000	4
Internal Revenue Service (IRS)	1862	156	Distrit of Col	217.58	239	81500000	29163650	1.12E+10	79890	11
U.S. Department of Defense	1947	71	District Of Co	217.58	323.1	356000	52253	5.98E+11	1300000	2

Figure 3: Screenshot of government dataset

TABLE 4: Pearson and Spearman rank correlation for government data breaches

	Correlation			
	Spearman		Pearson	
Predictors	**Coefficient**	**p-value**	**Coefficient**	**p-value**
Budget	0.58	***	0.22	0.14
Google count	0.26	*	0.21	0.15
#Employees	0.62	***	0.30	*
Population served	0.59	***	0.42	**
Age	–0.23	0.12	–0.12	0.42
Location rate	0.53	***	0.40	*

*p<10%; **p<5%; ***p<1%.

TABLE 5: Linear regression over government data. Independent variables are number of employees, Google counts, age, budget, population served, Location rate

Multiple regression models			
R	**R square**	**Adjusted R square**	**Std. error**
0.664	0.440	0.383	1.975

variables and the dependent variable (frequency of data breaches), but there are probably other factors affecting the frequency of data breaches, for example, factors that are not publicly available.

Results for educational organisations

Privacy Rights Clearinghouse dataset[54] contains data breaches reported from mainly US universities and colleges. A total of 818 breaches were reported in educational organisations, which resulted in over 25 million records breached since 2005. The breaches make up close to 10 per cent of total number of data breaches. They, however, account for only 0.23 per cent of records breached. We randomly selected a sample of 132 educational institutions (Figure 4), which have 373 individual breaches. We collected the following characteristics:

1) Popularity measured by Google counts, US News Ranking.
2) Size measured by enrolment in 2018 and number of employees, and size of endowment.
3) Age of the organisation.
4) Location rate, that is, number of breaches per million people at the state that the organisation is located.

Table 6 shows the results of the correlation methods. All these characteristics except ranking have a positive correlation with frequency of data breaches. Ranking is negatively correlated with frequency of data breaches, meaning higher ranked schools (with smaller value of ranking) have more data breaches. The correlations are significant in most cases except for location rate

Name	EmpNum	studentsNum	EnrollNum	LocRate	rank	GoogleCount	RecordsSum	age	EndowmentB	NumBreache
West Virginia University	1870	29175	28776	12	500	6330000	53	151	566420000	1
Washington State University	2261	30614	29686	26	143	56500000	1000300	128	974000000	2
University of Virginia	16000	24360	21985	23	24	45300000	18799	199	1390000000	10
Virginia Commonwealth University	3279	31242	31242	23	164	4160000	196442	50	1843000000	5
Virginia Polytechnic Institute and State	1395	32304	30598	23	74	90400	145333	146	1700000000	2

Figure 4: Screenshot of the education dataset

TABLE 6: Pearson and Spearman rank correlation for education data breaches

	Correlation			
	Spearman		Pearson	
Predictors	**Coefficient**	**p-value**	**Coefficient**	**p-value**
Endowment size	0.62	***	0.37	***
Google count	0.63	***	0.38	***
Number of employees	0.65	***	0.35	***
US News Ranking	−0.49	***	−0.38	***
Enrolment	0.54	***	0.18	**
Age	0.39	***	0.34	***
Location rate	0.06	***	0.013	0.88

**p<5%; ***p<1%.

using Pearson. Overall, we find that most characteristics are correlated with frequency of data breaches except location rate.

Table 7 shows the result using linear regression to predict frequency of data breaches. We also varied the subset of characteristics used in prediction. The best result was given when google counts, age, location rate, number of employees and endowment size were used. The adjusted R square is very high (0.833), suggesting that these characteristics do explain most of the variance of data breaches.

TABLE 7: Linear regression result for educational organisations

Multiple regression models			
R	**R square**	**Adjusted R square**	**Std. error**
0.926	0.857	0.833	0.469

Predictors: Google counts, age, location rate, employees' number, endowment size.

Results for medical organisations

There are 3,967 breaches from medical organisations in Privacy Rights Clearinghouse dataset, which account for 49 per cent of the total breaches. We randomly selected a sample of 61 medical organisations, shown in Figure 5, with 242 breaches in total. The selected organisations include hospitals, insurance companies and clinics. We collected the following characteristics of each organisation:

1) Popularity measured by Google counts.
2) Size measured by number of employees and revenue.
3) Age of organisation.
4) Location rate, that is, number of breaches per million people at the state the organisation is located.

Figure 6 shows causes of data breaches for medical organisations. A large portion

A	B	C	D	E	F	G	H	I	J	K
Name	RevenueB	Google	EmpNum	RecordsSum	Location	LocRate	locRank	age	year founded	Noofbreaches
American Fam	0.45	168000	299	9788	Alabama	16.05	27	36	1982	3
Children's Nati	1	446000	6000	22107	District Of C	217.58	17	148	1870	2
Humana Group	54.38	55900	51600	435291	Kentucky	24.85	25	57	1961	10
Advocate Heal	1.52	577000	70000	812	Illinois	25.58	6	23	1995	5
Suburban Lung	0	52700	49	2984	Illinois	25.58	6	33	1985	1
Advantage Hea	0.5	61600	299	437678	Indiana	30.96	11	19	1999	3

Figure 5: Screenshot of the medical dataset

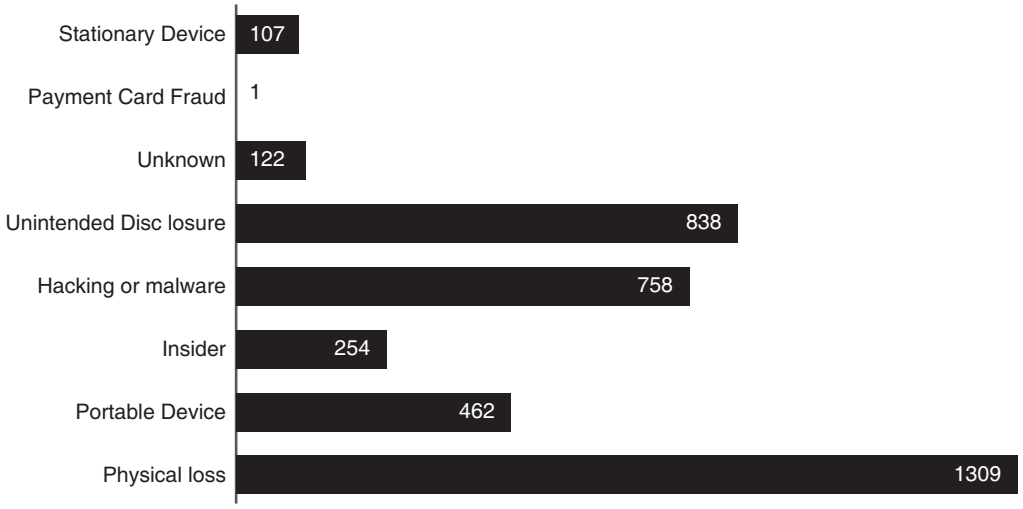

Figure 6: Causes of data breaches for medical organisations

TABLE 8: Pearson and Spearman rank correlation for medical data breaches

	Correlation			
	Spearman		Pearson	
Predictors	Coefficient	p-value	Coefficient	p-value
Revenue	0.73	***	0.656	***
Google count	0.478	***	0.49	***
Number of employees	0.72	***	0.75	***
Age	0.21	0.1	0.15	0.2
Location rate	0.24	*	0.05	0.7

*p<10%; ***p<1%.

TABLE 9: Linear regression using age, Google counts, location rate, number of employees and revenue

Multiple regression models			
R	R square	Adjusted R square	Std. error
0.775	0.601	0.562	2.950

of breaches are caused by insiders (including unintended disclosure and insider).

Table 8 summaries correlation results using Pearson and Spearman. As the table shows, revenue, Google count and number of employees are all positively correlated with frequency of data breaches and the correlation is statistically significant. The correlation between frequency of data breaches and age and location rate is not significant. Number of employees has very strong correlation with breach frequency for medical organisations, probably because insider breaches account for a large portion for medical organisations. Revenue is also highly linked to frequency breaches. One possible explanation is that majority of medical breaches are financially motivated.

Table 9 shows results for linear regression. Adjusted R square is 0.56, which means these characteristics account for more than half of the variance.

Results for business organisations

Business breaches recorded in Privacy Rights Clearinghouse dataset are divided into three categories: BSF, BSR and BSO. There are a total of 2,417 business breaches accounting for 29 per cent of total breaches. They, however, account for 95 per cent of total records breached with over 10 billion records breaches since 2005. We created a randomised sample of 58 organisations with 165 data breaches. The sample contains organisations from BSF, BSR and BSO organisations.

We collected the following characteristics:

1) Popularity measured by Google counts.
2) Size measured by number of employees and revenue.
3) Age of organisation.
4) Location rate, that is., number of breaches per million people at the state that the organisation is located.

Figure 7 shows a screenshot of business dataset and Table 10 shows the results of the correlation methods. Revenue is positively correlated to breach frequency, and the correlation is significant. One possible reason is that attackers hack business for financial gains. The correlation between number of employees and frequency of data breaches is significant as well. The correlations for other characteristics are not significant.

Table 11 shows result of linear regression. Again, we varied subset of predictors, and the best subset contains revenue, location rate, age and number of employees. The adjusted R square is only 0.184, suggesting there are possibly other factors deciding frequency of data breaches.

Name	EmpNum	founded	age	Location	LocRate	RevenueB	Google	RecordsSum	BreachesNum
7-Eleven	45000	1927	91	Texas	19.8	5.67	33000000	9708	4
Adobe	17000	1982	36	California	31.98	5.85	563000000	3060230	6
Rite Aid	87000	1962	56	Pennsylvania	19.91	30.74	14300000	16127	13
CVS	246000	1996	22	Rhode Island	40.7	117.5	87200000	27527	17
Forever 21	30000	1984	34	California	31.98	4.4	26700000	98930	2
Direct TV	10000	1994	24	California	31.98	12	474000	87	2

Figure 7: Screenshot of business breaches dataset

TABLE 10: Pearson and Spearman rank correlation for business data breaches

	Correlation			
	Spearman		Pearson	
Predictors	Coefficient	p-value	Coefficient	p-value
Revenue	0.42	***	0.45	**
Google count	0.28	*	0.12	0.4
Number of employees	0.37	**	0.36	**
Age	–0.04	0.7	–0.13	.3
Location rate	–0.09	0.5	0.05	0.7

*p<10%; **p<5%; ***p<1%.

TABLE 11: Linear regression result for business organisations

Multiple regression models			
R	R square	Adjusted R square	Std. error
0.535	0.286	0.184	3.079

Predictors: revenue, location rate, age, number of employees.

Results for nonprofit organisations

Nonprofit organisations account for 1.43 per cent of the breaches reported since 2005. They have 118 breaches resulted in 8,434,545 records breaches. We randomly selected 36 organisations and collected the following characteristics:

1) Popularity measured by Google counts.
2) Size measured by number of employees and revenue.
3) Age of the organisation.
4) Location rate, that is, number of breaches per million people at the state that the organisation is located.

Figure 8 shows a screenshot of the data. Table 12 summarises the result of three analysis methods. Revenue is positively correlated with frequency of data breaches in all three methods and the correlation is statistically significant. Google count and number of employees are positively correlated using the two rank correlation methods (Spearman and Kendall's correlation) but the correlation is not significant for Pearson. The correlation with age and location rate is not significant.

For multiple regression models, we use Google count, revenue and number of employees as predictors. Table 13 shows the results. The adjusted R square is 0.83, meaning the prediction is quite accurate.

Relationship between location (state) and frequency of data breaches

We applied rank and percentile analysis to study the relationship between the state an organisation is located and frequency of data breaches. Figure 9 shows the number of data breaches and number of records breached per state. There are a few states such as California and New York that have more data breaches than others. In terms of

Company	State	LocRate	Year	Google	Revenue	employee	total_record	Size(emp)	founded	age	BreachNum
211 LA Coun	California	31.98	2018	118000000	66414049	278	30	small - mid	1981	37	2
American Ca	Kentucky	24.85	2006	46700000	5920000000	20000	80649	Large	1913	105	6
Alabama Crir	Alabama	16.05	2013	190000000	0	14	300	small	1975	43	1
Alaskan AIDS	Alaska	31.14	2010	21500	10000000	11	2000	small	1985	33	1
Amateur Ath	Florida	21.21	2008	1070000000	413000000	1375	828	large	1888	130	2

Figure 8: Screenshot of nonprofit breaches dataset

TABLE 12: Pearson and Spearman rank correlation for nonprofit data breaches

	Correlation			
	Spearman		Pearson	
Predictors	**Coeffcient**	**p-value**	**Coefficient**	**p-value**
Revenue	0.63	***	0.87	***
Google count	0.47	**	0.09	0.6
Number of employees	0.60	***	−0.08	0.6
Age	0.23	0.1	0.25	0.1
Location rate	−0.06	0.7	−0.06	0.7

**p<5%; ***p<1%.

TABLE 13: Linear regression result for NGO using Google counts, revenue, employee number, revenue, location rate, age as predictor

Multiple regression models			
R	**R square**	**Adjusted R square**	**Std. error**
0.929	0.862	0.832	0.476

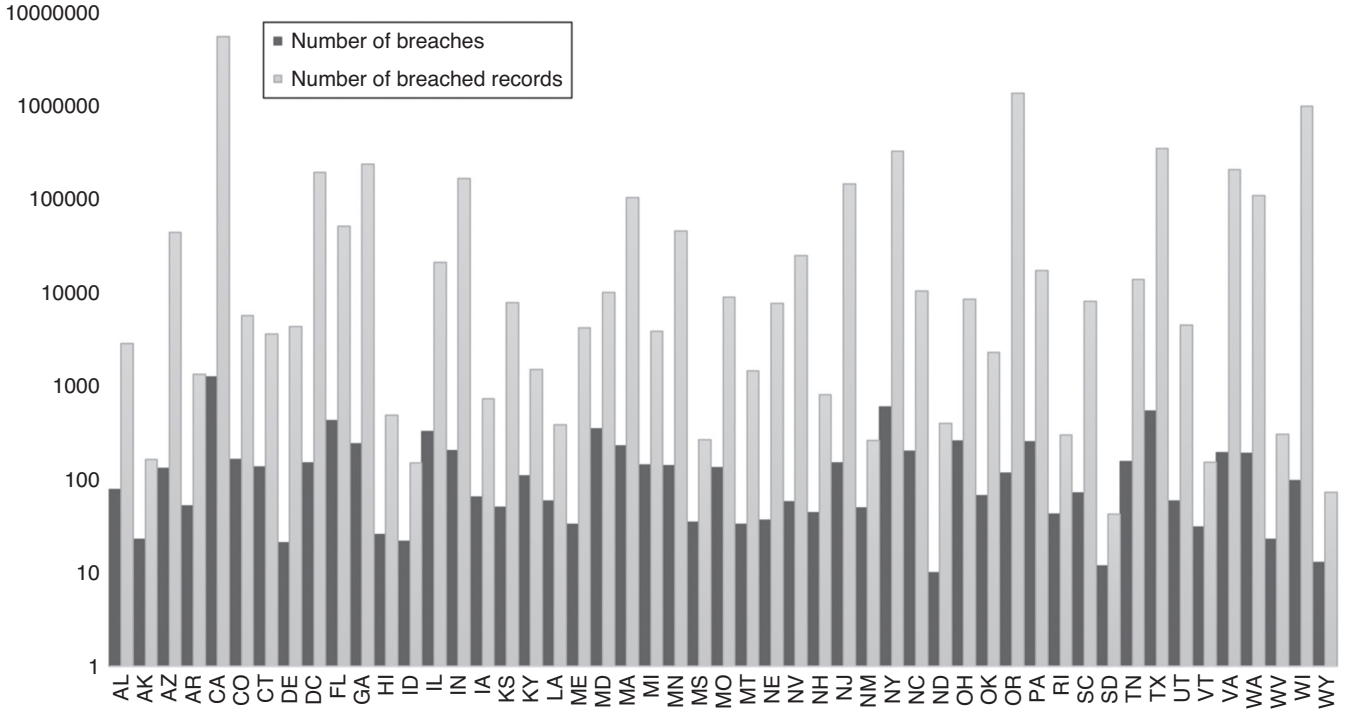

Figure 9: Number of breaches versus number of breached records by state on logarithmic scale with base 2

number of records breaches, California also has far more than the other states.

Table 14 shows rank and percentile results of each state based on number of records breached and number of breaches. A percentile of 50 per cent means that state is ranked higher (either on number of records breached or number of breaches) than 50 per cent of the other states. It also shows number of breaches per million

TABLE 14: Rank and percentile and breaches rate per million population

State	# Breached records	Rank	Percentile	# Breaches	Rank	Percentile	Breaches rate*
Alabama	2,863,834	34	36.54%	78	27	50.00%	16.05
Alaska	162,796	49	7.69%	23	47	9.62%	31.14
Arizona	44,474,925	16	71.15%	132	23	57.69%	19.33
Arkansas	1,341,022	38	28.85%	52	35	34.62%	17.46
California	5,497,071,574	1	100.00%	1252	1	100.00%	31.98
Colorado	5,745,433	28	48.08%	164	15	73.08%	30.05
Connecticut	3,618,086	33	38.46%	137	21	61.54%	38.15
Delaware	4,326,431	30	44.23%	21	50	5.77%	22.2
District of Columbia	194,503,875	9	84.62%	151	17	67.31%	217.58
Florida	51,412,187	14	75.00%	430	4	94.23%	21.21
Georgia	239,724,179	6	90.38%	240	9	84.62%	23.49
Hawaii	489,743	42	21.15%	26	46	13.46%	18.16
Idaho	151,068	51	3.85%	22	49	7.69%	13.29
Illinois	21,107,738	18	67.31%	329	6	90.38%	25.58
Indiana	167,556,191	10	82.69%	205	11	80.77%	30.96
Iowa	735,880	41	23.08%	65	30	44.23%	20.8
Kansas	7,835,189	26	51.92%	51	36	32.69%	17.51
Kentucky	1,512,532	36	32.69%	110	25	53.85%	24.85
Louisiana	389,492	44	17.31%	59	32	38.46%	12.63
Maine	4,246,207	31	42.31%	33	42	19.23%	24.82
Maryland	10,151,593	22	59.62%	352	5	92.31%	58.6
Massachusetts	104,872,042	13	76.92%	228	10	82.69%	33.55
Michigan	3,867,930	32	40.38%	144	19	65.38%	14.51
Minnesota	45,927,264	15	73.08%	142	20	63.46%	25.86
Mississippi	267,347	47	11.54%	35	41	23.08%	11.69
Missouri	8,965,706	23	57.69%	135	22	59.62%	22.19
Montana	1,469,687	37	30.77%	33	42	19.23%	31.94
Nebraska	7,768,565	27	50.00%	37	40	25.00%	19.51
Nevada	24,818,545	17	69.23%	58	34	36.54%	20.06
New Hampshire	811,669	40	25.00%	44	38	28.85%	33.06
New Jersey	147,050,841	11	80.77%	151	17	67.31%	16.85
New Mexico	264,385	48	9.62%	50	37	30.77%	23.97

(*Continued*)

TABLE 14: Rank and percentile and breaches rate per million population (*continued*)

State	# Breached records	Rank	Percentile	# Breaches	Rank	Percentile	Breaches rate*
New York	325,990,547	5	92.31%	597	2	98.08%	30.15
North Carolina	10,382,153	21	61.54%	202	12	78.85%	20.11
North Dakota	401,686	43	19.23%	10	53	0.00%	13.21
Ohio	8,479,903	24	55.77%	258	7	88.46%	22.21
Oklahoma	2,307,081	35	34.62%	67	29	46.15%	17.12
Oregon	1,371,801,409	2	98.08%	117	24	55.77%	29.03
Pennsylvania	17,350,624	19	65.38%	255	8	86.54%	19.91
Rhode Island	301,432	46	13.46%	43	39	26.92%	40.7
South Carolina	8,136,977	25	53.85%	72	28	48.08%	14.7
South Dakota	43,486	53	0.00%	12	52	1.92%	13.97
Tennessee	13,800,505	20	63.46%	156	16	71.15%	23.63
Texas	348,282,678	4	94.23%	544	3	96.15%	19.8
Utah	4,519,425	29	46.15%	59	32	38.46%	19.69
Vermont	153,134	50	5.77%	31	44	15.38%	49.51
Virginia	208,352,253	7	88.46%	193	13	76.92%	23.02
Washington	110,223,982	12	78.85%	190	14	75.00%	26.49
West Virginia	307,996	45	15.38%	23	47	9.62%	12.47
Wisconsin	1,001,339,478	3	96.15%	98	26	51.92%	16.98
Wyoming	73,018	52	1.92%	13	51	3.85%	22.18

*Breach rate is number of breaches per million population.

TABLE 15: Outlier states based on number of breaches per million population

State	Number of breaches per million population	Type of outlier
District of Columbia	217.58	Extreme outlier
Maryland	58.6	Mild outlier
Vermont	49.51	Mild outlier

population. We use the rate of breaches to population per million as a feature (location rate) and the regression test listed in Results section.

The result shows that California has the highest population, highest number of breaches; DC, however, topped the list as the highest number of breaches per million population. We found three outliers in terms of number of breaches per million population (Table 15).

District of Columbia is an extreme outlier with p-value less than 0.05. Majority of government agencies' headquarters are located in District of Columbia and that may explain why it becomes a target of cyberattacks. Maryland also has a lot of government agencies, which may attract hackers as well. Vermont has higher number of employees hired by government compared to the national average based on US Census Bureau records for number of employment; mostly in education sector. This may explain why it also has more data breaches.

DISCUSSION

This study has some limitations. First, we only analysed a sample of data due to the difficulty of collecting all listed characteristics manually. Secondly, we

TABLE 16: Characteristics and whether they have significant correlation with frequency of data breaches using each method

Characteristics	Government	Educational	Medical	Business	Nonprofit
Revenue/budget/ endowment	Two rank based (are significant)	All three	All three	All three	All three
Number of employees	Two rank based	All three	All three	All three	Two rank based
Population served/ enrolment	All three	All three	NA	NA	NA
Google count	None	All three	All tree	Only Kendall	Two rank based
US News Ranking	NA	All three	NA	NA	NA
Age of organisation	None	All three	None	None	None
Location rate	Two rank based	Spearman	None	None	None
Adjusted R square	0.383	0.833	0.562	0.184	0.832

NA means not applicable (the feature is available for that type of organisation)

were not able to collect information about an organisation's internal properties such as system vulnerabilities, which could explain why the linear regression models built on public information are not always accurate.

Table 16 summarises the characteristics and whether their correlations with frequency of data breaches are significant (if p-value<0.05) using each method. The results show that:

1. Characteristics such as revenue, budget, endowment, number of employees, population served and enrolment have positive correlation with frequency of data breaches. So, consumers cannot assume better security measures when it comes to large organisations. The answer to the first research question is yes.
2. Money-related measures such as revenue, budget and endowment size have very strong positive correlation, and most analysis methods find the correlation significant. One possible explanation is that cyberattackers are financially motivated in majority of cases. Therefore, money-related measures have positive correlation to frequent of breaches regardless of organisation type.
3. Most analysis methods also find number of employees has strong positive correlation with frequency of data breaches. This could be explained by a large number of data breaches which are due to insiders. For instance, for medical organisations, 79.6 per cent of breaches were caused by employees in that medical organisation. With alarming number of insider breaches, training employees becomes the first line of defence to an organisation's information security.
4. Population served and enrolment are also positively correlated with number of data breaches by most analysis methods for the two types of organisations that they are applicable. This can be explained similarly due to the larger possible gains by the attackers as they can get more information from such organisations.
5. For educational, medical and NGO organisations, characteristics related to popularity of an organisation such as google count and US News Ranking are also positively correlated with frequency of data breaches by most methods. The correlation, however, is not significant for business and government organisations. One possible explanation is, for educational, medical and NGOs, reputation is very important for consumers and more reputable organisations may attract more attackers.

For business and government agencies, however, there are other factors that are more important. For example, for business, price of service/product may be more important. So, for the second research question, the answer is yes if the organisation belongs to educational, medical and NGO.

6. Organisation's age only has positive correlation with frequency of data breaches for educational organisations. For all other types of organisations, the correlation is insignificant. So, the answer for the third research question is no except for educational organisations. One possible explanation is that old organisations may attract more attackers, but they also have better security measures. As older educational organisation also tends to be more reputable, this may explain that age is correlated with data breaches for educational organisations.
7. Whether an organisation is located at a state with higher possibility of data breaches (measured as location rate, which is number of breaches per million population) does not have significant correlation with frequency of data breaches except for government organisations. So, answer to the fourth research question is no except for government organisations. One possible reason is that as shown in relationship between location(state) and frequency of data breaches section, the few states with exceptionally high location rate are those states having a lot of government agencies.
8. The adjusted R square is quite low for most types of organisations except NGO. So linear regression based on the previous characteristics cannot accurately predict the number of data breaches except for educational and NGO. There could be a couple of reasons for this. One is that the relationship between these factors and number of data breaches may not be linear. The other possible reason is that there are other factors that we are not considering, for example, nonpublic information about the organisation such as vulnerabilities of their IT system. So, the answer to the fifth research question is yes only for NGO and educational organisations and no for other types of organisations. There are, however, still strong correlations between many of the publicly available characteristics and frequency of data breaches, and consumers can still use these characteristics to help reduce data breach risks. For example, they can choose smaller banks (eg credit unions) rather than big banks to reduce their data breach risks.
9. Finally, some states with many government agencies are outliers and have high number of data breaches. Hacking especially those from other nations may be the reason.

CONCLUSION AND FUTURE WORK

This paper analysed correlations between a US organisation's publicly available characteristics and chances of data breaches. The results show that many characteristics related to size and popularity are correlated with chances of data breaches. The results can be used to help consumers make better informed decision to reduce data breach risks.

There are several possible future research directions. First is to study other factors and automate the process of extracting these factors for a given organisation. Furthermore, appropriate natural language processing (NLP) techniques can be used for processing unstructured and semi-structured data found in data breaches reports more information such as what information is leaked. It is also possible to try nonlinear prediction models to estimate frequency of data breaches.

References

1. Federal Trade Commission (2017) 'Identity theft resource center: facts and statistics: find out more about the nation's fastest growing crime.'

2. Clearinghouse, P. R. (2019) 'A chronology of data breaches [Online]', available at: https://www.privacyrights.org/ (accessed 5th April, 2021).
3. Mello, S. (2018) 'Data breaches in higher education institutions', available at: https://www.semanticscholar.org/paper/Data-Breaches-in-Higher-Education-Institutions-Mello/aff02f3868ed559b430b4a802c3ec9339f86758d (accessed 5th April, 2021).
4. Ammann, P., Wijesekera, D. and Kaushik, S. (2002) 'Scalable, graph-based network vulnerability analysis', *ACM*, Vol. 2002, pp. 217–224.
5. Aven, T. (2007) 'A unified framework for risk and vulnerability analysis covering both safety and security', *Reliability Engineering & System Safety*, Vol. 92, pp. 745–754.
6. Pascual, A., Miller, S. and Marchini, K. (2016) '2016 identity fraud: fraud hits an inflection point', Javelin Strategy, available at: https://www.lexisnexis.com/risk/downloads/assets/id-fraud-prevention-playbook.pdf (accessed 5th April, 2021).
7. Soomro, Z. A., Shah, M. H. and Ahmed, J. (2016) 'Information security management needs more holistic approach: a literature review', *International Journal of Information Management*, Vol. 36, pp. 215–225.
8. Clearinghouse, see ref. 2 above.
9. Gkoulalas-Divanis, A. and Loukides, G. (2015) 'A Survey of Anonymization Algorithms for Electronic Health Records', Medical Data Privacy Handbook. Springer, Cham.
10. Widup, S. (2013) 'The Veris Community Database [online],' available at: http://vcdb.org (accessed 5th April, 2021).
11. Aggarwal, C. C. and Yu, P. S. (2008) 'Privacy-Preserving Data Mining: Models and Algorithms', Springer Publishing Company, Incorporated.
12. Dwork, C. (2008) 'Differential privacy: a survey of results', in International conference on theory and applications of models of computation. Springer, Berlin Heidelberg, pp. 1–19.
13. U. S. Department of Health and Human Services (2017) 'Breach portal [online]', available at: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf (accessed 5th April, 2021).
14. Bennett, C. J. and Raab, C. D. (2006) 'The Governance of Privacy: Policy Instruments in Global Perspective', 2nd and updated ed. MIT Press, Cambridge, MA.
15. Kobsa, A. (2001) 'Tailoring privacy to users' needs', Proceedings of the 8th International Conference on User Modeling, Springer, Cham, pp. 301–313.
16. Culnan, M. J. and Williams, C. C. (2009) 'How ethics can enhance organizational privacy: lessons from the choicepoint and TJX data breaches', *Mis Quarterly*, Vol. 2009, pp. 673–687.
17. Romanosky, S., Hoffman, D. and Acquisti, A. (2014) 'Empirical analysis of data breach litigation', *Journal of Empirical Legal Studies*, Col. 11, pp. 74–104.
18. Zhou, B., Pei, J. and Luk, W. (2008) 'A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data', ACM Sigkdd, Beijing, China.
19. Peltier, T. R. (2005) Information Security Risk Analysis', CRC Press, Boca Raton, FL.
20. Ponemon Institute (2019) '2018 Cost of Data Breach Study: Impact of Business Continuity Management', Ponemon Institute.
21. Vaidya, J., Zhu, Y. M. and Clifton, C. W. (2005) 'Privacy Preserving Data Mining (Advances in Information Security),' Springer-Verlag, New York.
22. Verison Enterprise Solution (2019) 'Data Breach Investigations Reports (DBIR)', [online], available at: http:// www.verizonenterprise. com/DBIR (accessed 5th April, 2021).
23. Ibid ref 17
24. Garg, A., Curtis, J. and Halper, H. (2003) 'Quantifying the financial impact of IT security breaches', *Information Management & Computer Security*, Vol. 11, pp. 74–83.
25. Gatzlaff, K. M. and Mccullough, K. A. (2010) 'The effect of data breaches on shareholder wealth', *Risk Management and Insurance Review*, Vol. 13, pp. 61–83.
26. Goel, S. and Shawky, H. A. (2009) 'Estimating the market impact of security breach announcements on firm values', *Information & Management*, Vol. 46, pp. 404–410.
27. Acquisti, A., Friedman, A. and Telang, R. (2006) 'Is there a cost to privacy breaches? An event study', ICIS 2006 Proceedings, p. 94.
28. Gatzlaff and Mccullough, see ref. 25 above.
29. Acquisti et al., see ref. 27 above.
30. Liu, Y., Sarabi, A., Zhang, J., Naghizadeh, P., Karir, M., Bailey, M. and Liu, M. (2015) 'Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents', USENIX Security Symposium, pp. 1009–1024.
31. Peltier, see ref. 19 above.
32. Cavusoglu, H., Mishra, B. and Raghunathan, S. (2004) 'The effect of internet security breach announcements on market value: capital market reactions for breached firms and internet security developers', *International Journal of Electronic Commerce*, Vol. 9, pp. 70–104.
33. Gatzlaff and Mccullough, ref. 25 above.
34. Liginlal, D., Sim, I. and Khansa, L. (2009) 'How significant is human error as a cause of privacy breaches? An empirical study and a framework for error management', *Computers & Security*, Vol. 28, pp. 215–228.
35. Garkoti, G., Peddoju, S. K. and Balasubramanian, R. (2014)' Detection of insider attacks in cloud based e-healthcare environment', IEEE, pp. 195–200.
36. Kurt, A. (2015) 'Effectiveness of Cyber Security Regulations in the US Financial Sector: A Case Study', Carnegie Mellon University.
37. Gao, C.-Z., Cheng, Q., He, P., Susilo, W. and Li, J. (2018) 'Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack', *Information Sciences*, Vol. 444, pp. 72–88.
38. Bai, G., Jiang, J. X. and Flasher, R. (2017) 'Hospital risk of data breaches', *JAMA Internal Medicine*, Vol. 177, pp. 878–880.
39. McCandless, D. (2012) 'Information Is Beautiful', Collins, London, pp. 978-0007294664.

40. Liu, see ref. 31 above.
41. *Ibid*.
42. Ammann, see ref. 4 above.
43. Aven, see ref. 5 above.
44. Soomro, ref. 7 above.
45. Pascual, see ref. 6 above.
46. Alqahtani, O., Chen, Z., Huang, Q. and Gottipati, K. (2018) 'Is Bigger Safer? Analyzing Factors Related to Data Breaches using Publicly Available Information'. International Conference on Information Systems Security and Privacy, pp. 373–378.
47. Martin, K. D., Borah, A. and Palmatier, R. W. (2017) 'Data privacy: effects on customer and firm performance', *Journal of Marketing*, Vol. 81, pp. 36–58.
48. Alice Corona (2017) Biggest-Data-Breaches (1ad877bd), available at: https://data.world/alice-c/biggest-data-breaches (accessed 5th April, 2021).
49. Verison Enterprise Solution, see ref. 22 above.
50. Swiler, L. P., Phillips, C. and Gaylor, T. (1998) 'A Graph-Based Network-Vulnerability Analysis System', Sandia National Labs, Albuquerque, NM.
51. Acquisti, see ref. 27 above.
52. Clearinghouse, see ref. 2 above.
53. Vaidya, see ref. 21 above.
54. Clearinghouse, see ref. 2 above.