Spring 2018

Data Breaches in Higher Education Institutions

Samantha Mello
University of New Hampshire, Durham

Data Breaches in Higher Education Institutions

Prepared by: Samantha Mello

Advised by: Professor Kholekile Gwebu

Spring 2018

Table of Contents

Introduction.

In today's rapidly evolving technological world, data security is among the top priorities for all types of businesses and institutions. Holding an immense amount of personal data can pose a large threat to any type of institution in the form of a data breach. Data breaches come in many forms such as payment card fraud, hacking or malware, insider breach, physical document loss, portable device breach, stationary device breach, or unintended disclosures (Data Breaches n.d.). This study explores data breaches in higher education institutions. From a data security perspective, such institutions are important because they hold vast amounts of data belonging to a large portion of the population. In fact, the National Center of Education Statistics reports that higher education institutions enroll approximately 20.4 million students (NCES, 2017a) and 1.6 million faculty (NCES, 2017b). In many cases, while in college, students begin to prepare themselves, financially, for the rest of their lives. They apply for jobs, rent apartments, and purchase vehicles. Such endeavors require financial stability, therefore, having personal data stolen could be detrimental.

Educational institution data breaches have not been fully explored and consequently, pose many unanswered questions. Research on higher education data breaches is important as it has the potential to identify factors that make such institutions more prone to data breaches. Additionally, given significant losses associated with breaches and educational institutions' inevitable vulnerability to such incidents, understanding how to effectively manage and recover from a breach is likely to be of importance to educational institutions.

To study data breaches in higher education, analysis was conducted on 604 breach announcements between 2005 and 2007, having been made public by Privacy Rights

Clearinghouse. These breached announcements were then merged with College Scorecard data to help identify factors that lead to breaches. Once merged, statistical analysis was performed to gain a deeper understanding of the relationships between the factors.

The remainder of this thesis is organized as follows. The next section reviews the extant literature and identifies some of the gaps that this study intends to fill. Thereafter, a set of hypotheses are developed followed by a description of the methodology used to collect, clean, and analyze the data. Next, the findings and their implications are discussed. Finally, avenues for future research are presented.

Literature Review

The literature considers work that has been conducted on data breaches from a variety of industries including healthcare, corporate (often referred to as business), government and the education sector. This approach was adopted to permit the identification of gaps in the data breach literature. Each research paper was categorized by sector and then summarized, presented in Table 1.

The corporate/business sector has benefited from the most research on data breach management. More than half of the relevant papers found, focused on the cost of data breaches to a company. Corporate data breaches are particularly important to understand because they tend to be extremely public and have a direct relationship with a company's reputation. Within a business, there are many parties that can be affected by a data breach such as consumers of the product or service, the business entity itself and the internet security developers. The cost of a data breach is easily quantifiable due to publicly available stock prices per company. Most of the

papers found that data breaches had a negative, statistically significant impact on the market values of a company (Acquisti, Friedman & Telang, 2006; Gatzkaff & McCullough, 2010; Cavusoglu, Mishra & Raghunathan, 2004; Goel & Shawky, 2009; Garg, Curtis & Harper, 2003). Most papers find that the market value of a company is only impacted directly after a breach. More specifically in research conducted by Kevin Gatzlaff and Kathleen McCullough, they found that 40 days directly after a breach, the market values tend to return to pre-breach levels (Gatzkaff & McCullough, 2010). Another paper found that two days after a security breach, a firm, on average, loses 2.1% of their market value (Cavusoglu, Mishra & Raghunathan, 2004). And similarly, research was conducted to find that companies average about a 1% loss in market value after a data breach (Goel & Shawky, 2009). While the business entities themselves were found to have a negative loss to market value, the research found that security developers' market value was positively impacted in the timeframe directly after a data breach. One study found that, on average, the developers receive a 1.36% increase in market value in the two days directly after a breach (Gatzkaff & McCullough, 2010). Research conducted by Ashish Garg, Jeffrey Curtis, and Hilary Harper found security companies' market value was positively impacted by a data breach (Garg, Curtis & Harper, 2003). Data breaches to firms with higher market-to-book ratios tend to have larger negative returns while firm size and subsidiary status also play a role in mitigating the negative effects of a breach (Gatzkaff & McCullough, 2010). The study by Hovav & D'Arcy, (2003) contradicts the above findings and reports that in general, the market does not penalize companies for an attack. However, they did find that the market tends to react more toward interest specific companies (Hovav & D'Arcy, 2003). Overall, consensus shows that data breaches negatively impact businesses.

Data breaches in the healthcare sector were second most prevalent behind corporate breaches. However, most articles tend to look at technical methodologies for improving data security in the healthcare field. Understanding data breaches in this field is important because medical information is highly sensitive to the individual and can do a lot of damage, financially. For example, insurance information combined with medical information can be used to file claims and thus contribute to insurance fraud (Appari & Johnson, 2010). One study proposed a technical solution to malicious insiders modifying patient data. They suggested using a three-tiered method of a watermarking module, logging module and security module hoping to limit insider attacks in the healthcare industry (Garkoti, Peddoju & Balasubramanian, 2014). Regulation plays a vital role in all aspects of the healthcare industry, however, after the passage of reporting legislation the number of incidents, reported increased (Collins, Sainato & Khey, 2011). This is beneficial for the future because once incidents are reported, they can be researched to identify preventative measures to data breaches in the future. The current healthcare industry research focuses on technical preventative measures more than identifying the root cause of data breaches.

Both the government and education sectors, severely lack research. The public sector tends not to be researched as much, compared to the private sector, perhaps because it is harder to quantitatively measure a breach and the associated damage. However, it is important to look at government data breaches because governments, at the state and federal levels, hold an immense amount of varying types of data. A current study on government data breaches found that human and software incompetence were the most common breach type. However, it is difficult to understand how often these occur because there is no unified reporting system (Froomkin, 2009). In the education sector, universities and other educational institutions hold personal data on

students, faculty, and alumni. However, in a recently completed study, a slight decrease in the number of overall incidents was found (Collins, Sainato & Khey, 2011). Similar to the healthcare industry, the education sector has no definitive way of reporting breaches, making it difficult to fully understand data breaches in this sector. Because there is a lack of research in both sectors, it identifies a gap on this topic. Consequently, this honors thesis will focus on education sector data breaches.

Table 1: A Summary of Related Studies

Sector	Title	Issues	Findings/Solution	Author
Healthcare	"Detection of Insider Attacks in CloudBased e-Healthcare Environment"	Malicious insiders modify the patient data which creates false data. The overarching issues include privacy, reliability, and security.	The proposed method includes: 1. Watermarking module 2. Logging module 3. Security module	Garkoti, Peddoju, and Balasubramanian (2014).
Business (Financial)	"Effectiveness of Cyber Security Regulations in the US Financial Sector: A Case Study"	Data breaches are more prevalent in the financial sector despite having cybersecurity regulations. To address this gap, regulation and actual practices need to be researched and addressed along with defining where the breaches come from.	The major cause of the data breaches were miscellaneous errors and insider misuse. They found different sub-sectors had the same threat patterns with different frequencies. There was a large gap between federal banking regulations and recommended practices.	Kurt and Butkovic (2015).
Business	"The Economic Cost of Publicly Announced Information Security Breaches: Empirical Evidence from the Stock Market"	Information security breaches are ubiquitous therefore understanding public sentiment is crucial. Data breaches pose a large risk to all businesses.	Breaches result in no statistically significant loss for an entire sample. Breaches involving unauthorized access to customer personal data or firm proprietary data result in an average loss of firm value of 5.5%. The highly significant, negative market reaction for information security breaches with unauthorized access to data.	Campbell, Gordon, Loeb, and Zhou (2003).

Government	"Government Data Breaches"	Public sector data breaches are not as heavily researched or investigated as much as in the private sector. Government data breaches are important because they hold many different types of information.	They found that human/software incompetence was the largest cause of government data breaches. It is hard to say how often these occur because there is no unified reporting system for the government.	Froomkin (2009).
Business	"Is there a Cost to Privacy Breaches? An Event Study"	Data breaches can negatively impact many parties such as consumers and companies. It is beneficial for a company to understand the associated cost of a data breach to protect themselves and consumers.	Through empirical analyses and an event study, the research showed a negative and statistically significant impact of data breaches on the company's market value on the day the breach had been publicly exposed.	Acquisti, Friedman, and Telang (2006).
Business	"The Effect of Data Breaches on Shareholder Wealth"	Data breaches pose a large risk to all businesses, specifically due to the personal information they hold. Businesses have large associated costs with data breaches.	The impact of a data breach on shareholder wealth is negative and statistically significant. After 40 days, it appears market value returns to prebreach levels. Firms with higher market-to-book ratios have higher, negative returns with the data breach. Firm size and subsidiary status mitigate the negative effects of the data breach.	Gatzlaff and McCullough (2010).
Business	"The Effect of Internet Security Breach Announcements on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers"	The issue this study tries to answer is the difficulty of measuring the associated costs of data breaches. Businesses hold an immense amount of data and can potentially be severely damaged by a data breach.	This study found that a security breach announcement is negatively associated with the market value of the firm. On average the firms lost 2.1% of their market value within two days, post announcement. The market value of security developers is positively associated with a data breach	Cavusoglu, Mishra, and Raghunathan (2004).

			announcement. They received an average, abnormal return of 1.36% during the two-day, post-announcement period.	
Business	"Estimating the Market Impact of Security Breach Announcements on Firm Values"	Security breaches can significantly damage companies; their reputation can suffer, and they can pay heavy, government driven fines.	The results of this study found that the announcement of a security breach has a significant negative impact on the market value of the company. The impact on the firms was a loss of about 1% of the market value.	Goel and Shawky (2009).
Business	"Quantifying the Financial Impact of IT Security Breaches"	Data breaches continue to happen at a rapidly increasing rate and will always be a main concern for all companies.	This study found that on average, the loss to a company was $17-28 million per incident. The market reacted the most severely to credit card information theft. Denial-of-Service incidents had a larger negative impact on share prices compared to website defacements. Security companies also are positively impacted by security breaches.	Garg, Curtis, and Harper (2003).
Business	"The Impact of Denial-of-Service Attack Announcements on the Market Value of Firms"	Security breaches have been increasing in companies for years. Assessing the impact of security breaches is crucial for policymakers when making security policies.	This study found that in general, the market does not penalize companies for an attack. However, the market does react and penalize companies that are internet specific, more than other companies.	Hovav and D'Arcy (2003).
Healthcare	"What Caused the Breach? An Examination of Use of Information Technology and Health Data Breaches"	Data regarding a person's health information is highly sensitive. Thus, an increase in data breaches of health information is not good and the cause of these breaches should be considered to help prevent them in the future.	This study found that 47.5% of breaches affecting individuals were from theft and second was from loss 27.4%. For covered entities and business associates, 20.2% were from unauthorized access or disclosure. Hacking/IT represented 7.1% of total	Wikina (2014).

			individuals and 8.6% for covered entities and 13.1% for business associates.	
Healthcare/ Education	"Organizational Data Breaches 2005-2010: Applying SCP to the Healthcare and Education Sectors"	In the healthcare field, insurance policy information can be used to file claims and obtain prescriptions. Educational institutions have millions of records of student, faculty, and alumni data.	The passage of reporting legislation within the healthcare field increased the number of incidents reported. For educational institution data breaches, there is an overall decrease in incidents. This study suggests that since there is no centralized reporting database for all data breaches it prevents a definitive analysis.	Collins, Sainato, and Khey (2011).
Healthcare	Information Security and Privacy in Healthcare: Current State of Research"	The healthcare industry has formed an increasing need for the transfer of digital records which makes it susceptible to data breaches. The sensitivity of healthcare information makes it extremely important to be protected.	This paper summarized the current research in this area and found many papers proposing methodologies to combat privacy in the healthcare sector. For future research, the paper suggests considering internal factors such as by organization type. They also suggest researching limits to be placed on all types of users who interact with the data. This paper continues to go into detail to identify the gaps in current research.	Appari and Johnson (2010).

Hypothesis Development

There are many factors that determine whether an organization will experience a data breached. Based on the literature reviewed, specifically for data breaches in business institutions, most institutions were large, well-known firms. Perhaps they were breached due to their stature

or they were breached for the sheer amount of records they hold. Due to the extensive research on data breaches, it is predicted that universities with higher student enrollment are more likely to be breached because they contain more personal information about a larger number of students and employees. In essence, there is more data at larger universities thus, a higher chance of breaching more personal data records. Below is the first hypothesis explored in this research.

H_1 – Larger universities are more susceptible to a data breach.

Similar to larger organizations, the literature indicates that companies with strong financial backgrounds tend to get breached more often. The research shows there are monetary incentives to data breaches. Thus, it can be inferred that universities with financial prestige are more likely to be breached due to the higher financial gain to a person with access to breached data. For example, a university with higher average family income is more likely to be breached due to the indication of larger amounts of funds associated with their Social Security number. Below is the second hypothesis explored in this research.

H_2 – Universities with more financial resources are more susceptible to a data breach.

The previous research studies conducted in the healthcare industry tend to focus on solutions to data breaches due to the immense amount of private data held by these organizations. These studies propose many solutions for data management, inferring the better data management there is, the less likely a breach would occur. As a result, it can be predicted that universities with tighter data protection policies are less likely to be breached because they have more controls on student and faculty records. Below is the third hypothesis explored in this research.

H_3 – Universities with stricter data protection policies are less susceptible to a data breach.

Methodology

Secondary data was used to examine the aforementioned hypotheses. Specifically, data from Privacy Rights Clearinghouse was used to gain knowledge about all data breaches reported from 2005 through 2017. Privacy Rights Clearinghouse is a nonprofit, consumer education organization that seeks to bring attention to all privacy-related issues (Data Breaches n.d.). Data can be downloaded in the form of an Excel document based on the type of breach, organization type, and year. For this study, all types of data breaches were downloaded from 2005 through 2017 for the education sector. The second source of data came from College Scorecard, a data collection program run by the U.S. Department of Education from 1996 to 2016 for all undergraduate degree-granting institutions (College Scorecard Data n.d.). This organization reports all data collected via their website. The data collected contains attributes about all institutions. College Scorecard breaks up their attributes by the following, overarching identifiers, academics, admissions, costs, student body, financial aid, competition and retention, earnings, repayment, and school. Each overarching identifier then breaks down into descriptive measurements related to the broad identifier.

Once the above data sheets were downloaded, they were cleaned. Cleansing and preparation for analysis were all done via Excel. For major formatting issues for the Privacy Rights Clearinghouse data, VBA macros were recorded and looped through each record to prepare for analysis. Each record was then identified to ensure non-postsecondary institutions were deleted from the sample. Descriptions of each breach were reviewed to identify what type

of data was stolen. The descriptions were searched by the following keywords: Social Security, financial, medical, phone, email, address, driver's license, credit card, debit card, and password. Dummy variables were created for each keyword, with a one (1) indicating that type of data was exposed and a zero (0) indicating that type of data was not exposed. One record could have multiple exposures to the previous keywords. The College Scorecard data required less cleansing. However, each attribute on the College Scorecard data was identified as relevant to the topic, if deemed irrelevant, the attribute column was then deleted. After the datasets were cleaned and prepared for analysis, they were merged using a unique identifier; OPEID (Office of Postsecondary Education Identification. This eight-digit identifier is the OPEID number created and assigned by the U.S Department of Education (Department of Defense n.d.). Each branch of any university has their own unique OPEID, making it the best unique identifier for this research. Once these datasets were merged, a breached column was added. This column was a dummy variable indicating whether a university was breached (1) or not breached (0).

After merging and cleansing the data set, descriptive statistics were computed via Excel, to better understand the data. A visualization software, Tableau, was used to create graphic representations of how, what, where and when data breaches occurred. Once the dataset was better understood, IBM's statistical software, SPSS Statistics, was used to create a correlation matrix to understand the relationships between the variables. Thereafter, a logistic regression was conducted to test the hypotheses. The findings from the associated analysis are presented in the paragraphs that follow.

Findings

This section will describe the full sample of data as well as a subsample of the breached universities. The College Scorecard and Privacy Rights Clearinghouse datasets were merged, resulting in a total of 7,594 total records. Of these records, 604 were breached universities. To further understand the full sample, each university was categorized into small, medium, and large. Small universities include all universities with student enrollments below 5,000. Medium universities have student enrollments between 5,001 and 15,000. Large universities have student enrollments greater than 15,001. For the breached sample, there are the most instances of large universities (285) and a close second of medium-sized institutions (206). For the full sample of data, almost half of the records are small universities (49.92%). This could be due to the larger number of smaller universities in the United States than larger universities. The table below shows the number of instances, as well as percentages, for each size categorization of the breached universities, as well as the full sample of data.

Table 2: Data Breach Instances per University by Size

	Breach		**Full Sample**	
	Instances	**%**	**Instances**	**%**
Small	100	16.56%	3791	49.92%
Medium	206	34.11%	1095	14.42%
Large	285	47.19%	1972	26.42%
No Data	13	2.15%	736	9.69%

To continue understanding the full sample of data, the location of each university was explored. The data was examined first by state and then by region. The regions include the Northeast (CT, DC, DE, MA, MD, ME, NH, NJ, NY, PA, RI, VA, VT, WV) the Southeast (AL, FL, GA, KY, MS, NC, SC, TN) the Southwest (AR, LA, NM, OK, TX) the West (AK, AZ, CA, CO, HI, ID, MT, OR, NV, WA, WY, UT) the Midwest (IA, IL, IN, KS, MI, MN, MO, ND, NE, OH, SD, WI) . By state, California has the most breach instances as well as the most instances

for the full sample. Wyoming has no instances for the breach set of data, while Alaska has the least amount of instances for the full sample. For the breach instances, the West region has the largest number of breaches (160), while the Southwest has the least (48). For the full sample data, the Northeast has the most instances (1,868) while the Southwest has the least (901). The tables below show the location of breach instances as well as full sample data instances.

Table 3: Breach Instances by State

	Breach		Full Sample	
	Instance	**%**	**Instance**	**%**
Alabama	8	1.33%	96	1.26%
Alaska	2	0.33%	9	0.12%
Arizona	5	0.83%	133	1.75%
Arkansas	2	0.33%	92	1.21%
California	90	14.95%	770	10.14%
Colorado	17	2.82%	125	1.65%
Connecticut	18	2.99%	97	1.28%
Delaware	4	0.66%	19	0.25%
District of Columbia	2	0.33%	25	0.33%
Florida	25	4.15%	441	5.81%
Georgia	15	2.49%	182	2.40%
Hawaii	6	1.00%	25	0.33%
Idaho	4	0.66%	41	0.54%
Illinois	15	2.49%	289	3.81%
Indiana	23	3.82%	169	2.23%
Iowa	15	2.49%	90	1.19%
Kansas	8	1.33%	99	1.30%
Kentucky	11	1.83%	105	1.38%
Louisiana	4	0.66%	128	1.69%
Maine	4	0.66%	41	0.54%
Maryland	4	0.66%	96	1.26%
Massachusetts	20	3.32%	195	2.57%
Michigan	15	2.49%	210	2.77%
Minnesota	5	0.83%	155	2.04%
Mississippi	3	0.50%	65	0.86%
Missouri	13	2.16%	190	2.50%
Montana	7	1.16%	32	0.42%
Nebraska	5	0.83%	51	0.67%
Nevada	7	1.16%	45	0.59%
New Hampshire	4	0.66%	41	0.54%
New Jersey	8	1.33%	165	2.17%

New Mexico	7	1.16%	51	0.67%
New York	41	6.81%	468	6.16%
North Carolina	18	2.99%	205	2.70%
North Dakota	1	0.17%	30	0.40%
Ohio	32	5.32%	355	4.68%
Oklahoma	7	1.16%	149	1.96%
Oregon	10	1.66%	93	1.22%
Pennsylvania	21	3.49%	405	5.33%
Rhode Island	1	0.17%	26	0.34%
South Carolina	7	1.16%	110	1.45%
South Dakota	1	0.17%	31	0.41%
Tennessee	13	2.16%	185	2.44%
Texas	28	4.65%	481	6.33%
Utah	5	0.83%	80	1.05%
Vermont	3	0.50%	27	0.36%
Virginia	22	3.65%	188	2.48%
Washington	7	1.16%	127	1.67%
West Virginia	2	0.33%	75	0.99%
Wisconsin	7	1.16%	116	1.53%
Wyoming	0	0.00%	11	0.14%

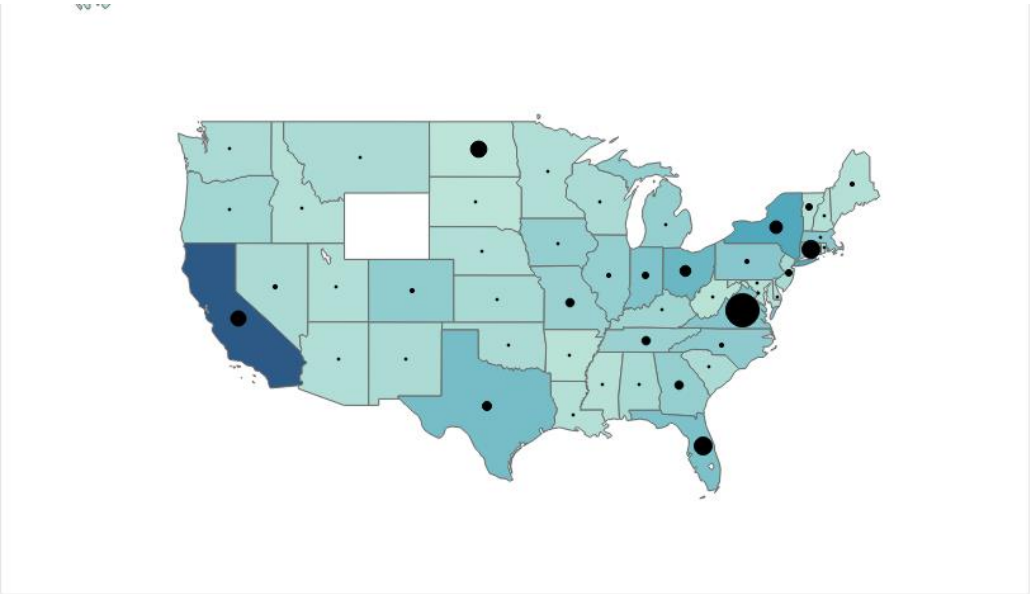
Table 4: Breach Instances by State

	Breach		**Full Sample**	
	Instance	**%**	**Instance**	**%**
The Northeast	154	25.58%	1,868	24.60%
The Southeast	100	16.61%	1,389	1829%
The Southwest	48	7.97%	901	11.87%
The West	160	26.58%	1,491	19.64%
The Midwest	132	21.93%	1,686	22.20%

To further understand the breached data, visualizations were constructed. As mentioned above, out of 604 breached observations, California had the largest number of data breaches (90) with New York just below California, at 41 data breaches. This can be explained due to the number of universities in each state, California has the most universities in the United States thus it would be expected that California has the most data breaches. It appears there is a direct correlation between the number of universities in each state and the number of data breaches in

each state. On the contrary, Virginia holds the most total records breached (353,923), meaning the most amount of data was breached there. Connecticut is behind Virginia at 112,761 records breached from 2005 to 2017. This is more difficult to explain given that there is no correlation between the number of universities in each state and the total number of records breached in each state. Below is a graphical representation of location with respect to data breaches, the darker the shading in a state represents a higher number of data breaches while a larger circle on a state represents a higher number of total records breached.

Figure 1: Number of Breaches and Total Number of Records Breached by State



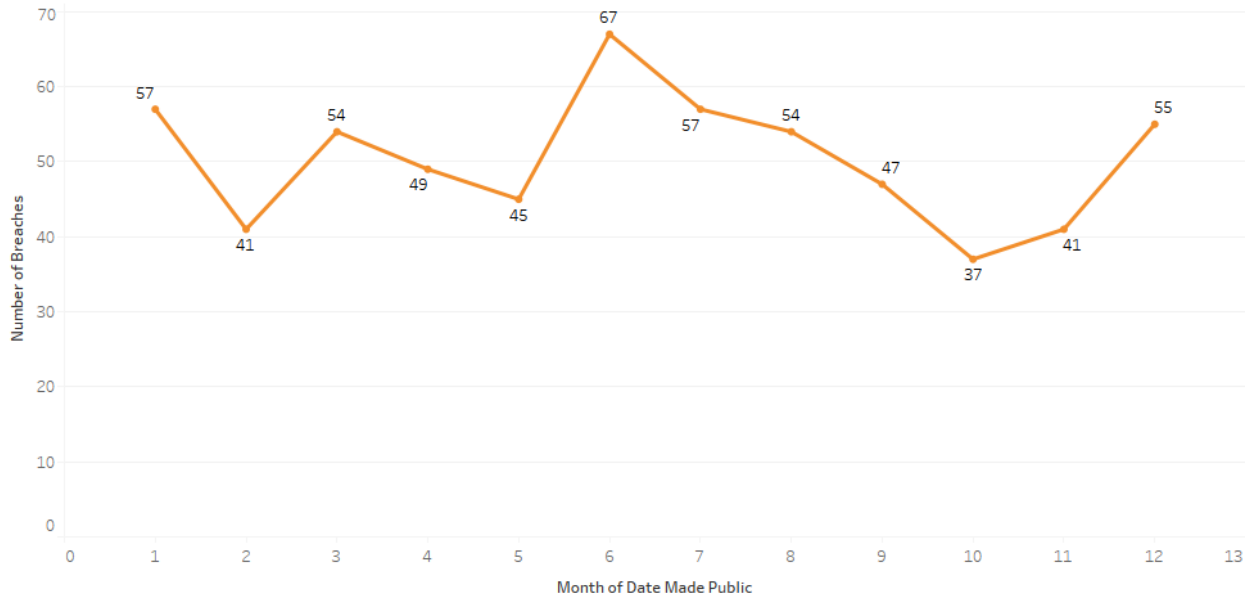
After understanding where data breaches tend to take place, the timing of breaches was explored. More specifically, the time the data breach was announced and made public. In terms of years, the occurrence of data breaches appears to be decreasing however, the total number of records breached per year does not have the same dramatic reduction as the occurrence of data

breaches. This indicates that although data breaches as a whole are decreasing, the number of records breached during a single breach is larger. After analysis of data breach announcements per month, most announcements were made in June (67) followed by January (57). Lastly, in terms of days, the company announcement days are typically announced on Friday's (144). Perhaps giving the weekend as a buffer from public scrutiny.

Table 5: Number of Data Breaches and Total Records by Year

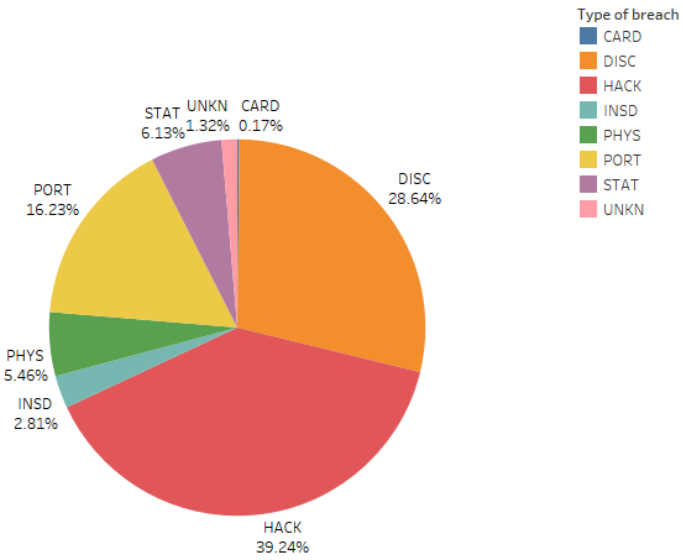
Year	Number of Data Breaches	Total Records Breaches
2005	62	62,578
2006	83	64,056
2007	82	48,247
2008	76	107,528
2009	51	100,005
2010	54	99,494
2011	46	244,990
2012	59	135,175
2013	32	160,090
2014	21	43,988
2015	10	1,013
2016	12	238
2017	11	51

Figure 2: Number of Breaches per Month



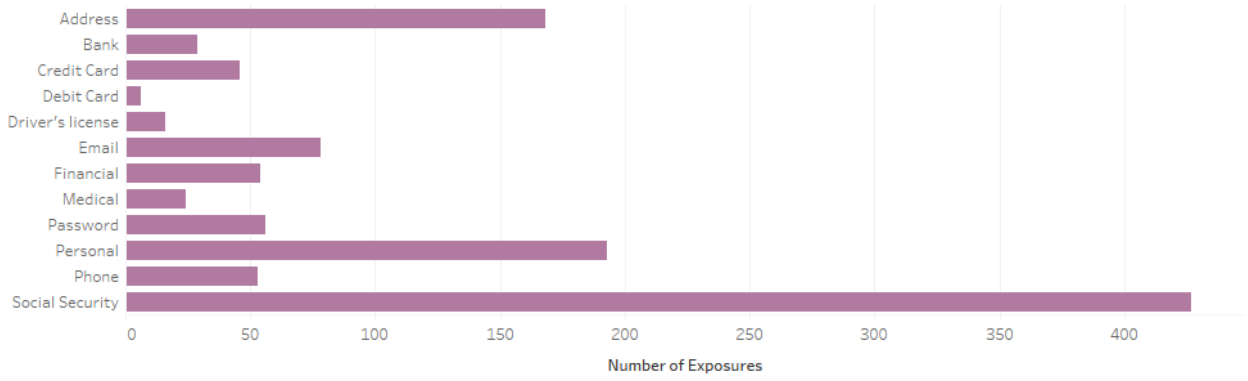
After understanding where the breaches happened and when it is important to examine how breaches occur and what type of data is stolen. Privacy Rights Clearinghouse breaks up data breach announcements by the type of breach. There are eight types of breach categories. They include, payment card fraud, hacking or malware, insider, physical loss, portable device, a stationary device, unintended disclosure and unknown (Data Breaches n.d.). In this dataset, most breaches originate from hacking or malware (39.24% or 237 breaches) followed by unintended disclosures (28.64% or 173 breaches). It is important for institutions to understand that 31.45% of all data breaches originate internally, whether through a malicious insider or an unintended disclosure situation.

Figure 3: Types of Breaches



At universities, faculty, students, and alumni are the most interested in what learning the type of their data that is prone to a breach. After conducting analysis, Social Security numbers are most likely to be stolen; out of 604 data breaches, 427 of them exposed Social Security numbers, followed by personal data (193) and addresses (168).

Figure 5: Types of Data Exposed



5.1 Correlation Analysis

The College Scorecard database allowed for the identification of multiple variables to test the hypothesized relationships. Table 6 shows the variables that were used in the study. The variables associated with university size in H_1 are Size and Main Campus, while those associated with data protection policies suggested in H_2 are Privacy Suppressed Instances and Privacy Suppressed. The remaining variables capture the financial dimension mentioned in H_3.

Table 6: Variable Definitions

Variable	Description	Used to Test Hypothesis
Breach	A binary variable that reflects whether or not a university has been breached. (1=True; 0=False).	n/a
Size	A variable indicating total enrollment of undergraduate, degree-seeking students.	H_1
Main Campus	A binary variable that reflects whether or not the campus is the main campus. (1=True, 0=False)	H_1
Privacy Suppressed	A binary variable that reflects whether or not a university suppresses data for privacy purposes. (1=True, 0=False).	H_2
Privacy Suppressed Instances	A variable indicating the number of data elements are suppressed by a university.	H_2
Faculty Salary	A variable indicating the median faculty salary of the university.	H_3
High Income Students	Number of Students from households earning $110,001 or higher.	H_3
Average Family Income	A variable indicating the average family income at the university.	H_3
Median Family Income	A variable indicating the median family income at the university.	H_3

The correlations confirm the hypotheses previously stated. As shown below, the size and main campus variables are positively correlated to the breach variable. Larger universities hold more faculty, student and alumni data proving that the more records a university holds, the more likely they are to be breached. In addition, monetary variables such as high-income students, faculty salary, and median and average family income show significance, which supports the second hypothesis (H_2) stated above. It is more valuable for the entity committing the breach to gain data from individuals with higher net worth, as the correlation shows, the higher amount of family income and faculty salary, the more likely a data breach will occur at that university. Lastly, the correlation matrix indicates a negative correlation between a data breach and a university that takes action protecting faculty and student records, this shows support for the third hypothesis stated (H_3). Similarly, the more data elements a university protects, the less likely a breach will occur. This indicates that universities should not only be taking actions to secure privacy but also to ensure the most amount of data possible is suppressed.

Table 7: Correlation Matrix

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**
(1) Breached	1								
(2)Size	.448**	1							
(3)Main Campus	.110**	.182**	1						
(4)Privacy Suppressed Instances	-.137**	-.261**	.308**	1					
(5)Privacy Suppressed	-.144**	-.316**	.317**	.376**	1				
(6)High Income Students	.217**	.130**	.337**	.282**	.219**	1			
(7)Average Family Income	.249**	.211**	.206**	-.199**	.039**	.983**	1		
(8)Median Family Income	.229**	.207**	.210**	-.190**	.049**	.966**	.977**	1	
(9) Faculty Salary	.342**	.404**	.229**	-.073**	-.051**	.569**	.562**	.519**	1

**. Correlation is significant at the 0.01 level (2-tailed).

5.2 Logistic Regression

To assess the influence of independent variables on breaches, a logistic regression model was created. Recall that H_1 predicted that the larger the university, the more susceptible it will be to data breach, while H_2 predicted that universities with more financial resources are more susceptible to a data breach, and that H_3 predicted that universities with stricter data protection policies are less susceptible to a data breach. In the preceding correlation analysis, multiple variables were used to capture the various dimensions (i.e., University Size, Data protection Policy Strictness and University Financial Resources) introduced in the hypotheses. However, given that multiple variables seek to explain the same dimension including each variable in the logistic regression is likely to cause parameter estimates to be inaccurate due to multicollinearity. Therefore, principal components analysis (PCA) was used as method of data reduction prior to creating the logistic regression model.

The PCA extracted 3 components with eigenvalues above 1. These three factors collectively account for 79.90% of the variance. Table 8 shows the component loadings and cross-loadings. Component 1 represents each of the financial resources of a university while component 2 and 3 represent Data Protection Policies and the university size respectively.

Table 8: Rotated Component Matrix

	Component		
	1	2	3
High Income Students	**0.970**	0.169	0.076
Average Family Income	**0.963**	0.199	0.110
Average Family Income	**0.952**	0.206	0.086
Faculty Salary	**0.623**	-0.078	0.510
Privacy Suppressed	0.079	**0.821**	-0.058
Privacy Suppressed Instances	0.251	**0.699**	0.029
Size	0.075	-0.443	**0.770**
Main Campus	0.136	0.440	**0.765**

Using the components, the following logistic regression model was created.

$$Breach = \beta_0 + \beta_1 * Financial\ Resources + \beta_1 * Data\ Protection\ Policies + \beta_1 * University\ Size$$

The proportion three predictor variables explain a considerable portion (Nagelkerke R Square = 36.3%) of the variance in the dependent variable. The results for the logistic regression are shown below. Based on these results, each of the hypotheses are supported. Specifically, increasing university size increases the odds of being breached. Thus, H_1 is supported. Increasing financial resources increases the odds of being breached. Thus, H_2 is supported. Finally, suppressing data or having stricter data protection policies decreases the odds of being breached. Thus, H_3 is supported.

Table 9: Logistic Regression Results

	B	S.E.	Wald	df	Sig.
Financial Resources	0.788	0.060	172.162	1	0.000
Data Protection Policies	-0.421	0.076	30.559	1	0.000
University Size	1.211	0.085	201.819	1	0.000
Constant	-3.015	0.094	1023.448	1	0.000

Discussion

After conducting analysis, there were many interesting observations. For example, there appears to be a level of "prestige" that data hackers are after. As the correlation matrix shows, the higher average faculty salary, the more likely a data breach will occur. It can be assumed that higher faculty salaries could indicate more internal resources the university has, thus the more

money a university has to pay salaries. Similarly, on the external side of a university, the higher median family income indicates higher net worth for the student. Meaning their personal information such as Social Security numbers are of more worth to a hacker. It is of more worth to a hacker to steal an identity because more money in bank accounts and higher credit scores make it easier to use monetary funds as well as apply for credit cards, etc. Overall, it appears that universities with more financial resources are more susceptible to a data breach.

Another interesting observation pertains to when universities announce a data breach. After analysis of the months in which data breaches are announced, it appears that they are typically announced when students are out of school. During Winter break (December and January), during Spring break (March) and during the first month of school after graduation (June) more data breach announcements are made. This could indicate two scenarios; the first that universities wait to announce data breaches to avoid any public or internal scrutiny from faculty and students. The next scenario could be that universities do not find the breach until students are away from the universities because they are typically busier during the times students are in session.

Implications

After analyzing the data, there are many key takeaways that universities and students should understand. For example, a university is more likely to be breached if it is a larger university. Perhaps due to their well-known image or the immense about of data held within large universities. It can be assumed the entity breaching the universities wants as many personal records as possible, therefore hacking a larger university is advantageous for them. The correlation analysis also showed the type of ownership plays a factor in a data breach. If it is a public university they are more likely to be breached. Larger, public universities are typically

more well-known than smaller ones, indicating that well-known universities are more likely to be breached. It is crucial for larger universities, more specifically large, public universities, to pay attention to data security because size is the most significant variable when determining the likelihood of a university data breach.

Many data breaches at universities stem from meagre data management practices. After analysis of how data breaches occur, 31.45% of data breaches originate internally. More specifically 28.64% originate from unintended disclosures or non-malicious internal employee/student error. To combat this, training of university employees and students, on best practices for data management, is essential. Employee's need to know how to properly handle student data to avoid accidentally sharing this information via internal or external servers. Similar to some companies, universities could provide employees and new students with a mandatory online data management course. This would provide them with knowledge on how to appropriately handle personal information as well as how to handle suspicious, external materials, such as phishing emails.

Social Security numbers are the most stolen piece of personal data. Inferring that identify theft is what malicious hackers are after. This could be detrimental to employees and students. Specifically, more for students as they are most likely in their late teens early twenties and just beginning their independent financial lives. For example, some students will be applying for credit cards, renting apartments, or buying a car, all of which Social Security numbers and credit score checks are imperative. Because of how private Social Security numbers are, universities should seek effective ways of protecting Social Security numbers both for students and employees. Perhaps they could suppress all data elements that hold Social Security numbers. They could also allow only the last four digits of a Social Security number show on the

employee or student record. In essence, all personal data needs to be carefully handled, however Social Security numbers must be handled with very extreme and confidential care because they are the most common data element hackers are looking for.

While breaches at universities seem to be decreasing, there are still multiple breaches that occur every year. As technology continuously improves, more records are being stolen in each single instance of a data breach. Universities must remain vigilant and continuously maintain internal security systems as well as data management practices. It is also important to always be aware of how data flows throughout an organization to be aware of who is seeing or handling the different types of personal data.

Future Research

Although this study gives an in-depth introduction to the causes of data breaches in higher education institutions, there is room for further research. For example, these are only announcements; therefore, all the breached may not be covered. There may be universities that have been breached and did not publicly disclose the breach that could have been excluded in the sample. There is currently no federal regulation or reporting standard to hold every university to the same reporting level. Therefore, finding alternative approaches to identifying breaches could ensure breached universities are included and thus, would give better insights into the research questions posed.

This study identifies the causes of a data breach however, it does not explore preventative measures. An area for further research could be identifying preventative measures universities currently have and looking at their associated data breaches to see if there are some preventative measures that combat data breaches better than others. Similarly, research to understand the most

essential preventative measures that need to be employed by different types of universities could be undertaken. For example, smaller universities could implement different preventative measures than larger universities, or after analysis, it could be found that university attributes do not affect the type of data security measures put in place. Generally, after identifying what causes data breaches, the next step would be to research how to prevent data breaches.

This study encompasses numerous types of data breaches. For further research, data breaches could be broken up into malicious hackers, both internally and externally, compared to unintended disclosures or breaches that occur unintentionally. This could help identify which data elements universities should specifically focus on managing.

References

Acquisti, A., Friedman, A., and Telang, R. (2006). Is There a Cost to Privacy Breaches? An Event Study. *Proceedings of the 27th International Conference on Information Systems*.

Appari, A., Johnson, E.M. (2010). Information security and privacy in healthcare: current state of research. *Internet and Enterprise Management*, 6(4) 279-314

Campbell, K., Gordon, L.A., Loeb, M.P., and Zhou, L. (2003). The economic cost of publicly announced information security breaches: Empirical evidence from the stock market. *Journal of Computer Security*, 11(3), 431-448.

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*, 9(1), 69-104.

College Scorecard Data. (n.d.). Retrieved March, 2018, from https://collegescorecard.ed.gov/data/

Collins, J.D., Sainato, V.A., Khey, D.N. (2011). Organizational Data Breaches 2005-2010: Applying SCP to the Healthcare and Education Sectors. *International Journal of Cyber Criminology*, 5(1), 794-810.

Data Breaches. (n.d.). Retrieved December 12, 2017, from https://www.privacyrights.org/data-breaches

Department of Defense. (n.d.). Retrieved May 6, 2018, from https://www.dodmou.com/Home/EnterOpeidBeforeCreateUserAccount

Froomkin, M.A. (2009). Government Data Breaches. Berkeley Technology Law Journal, 24(3), 1019-1059.

Garg, A., Curtis, J., and Haper, H. (2003). Quantifying the financial impact of IT security breaches. *Information Management & Computer* Security, 11(2), 74-83.

Garkoti, G., Peddoju, S. K., and Balasubramanian, R. (2014). Detection of Insider Attacks in Cloud Based e- Healthcare Environment. *2014 International Conference on Information Technology*.

Gatzkaff, K.M., and McCullough, K.A. (2010). The effect of data breaches on shareholder wealth. *Risk Management & Insurance Review*, 13(1), 61-83.

Goel, S., and Shawky, H.A. (2009). Estimating the market impact of security breach announcements on firm values. *Information & Management*, 46(7), 404-410.

Hovav, A., D'Arcy, J. (2003). The Impact of Denial-of-Service Attack Announcements on the Market Value of Firms. *Risk Management and Insurance Review*, 6(2), 97-121.

Kurt, Asim. (2015). Effectiveness of Cyber Security Regulations in the US Financial Sector: A Case Study.

NCES Fast Facts Tool (2017a). https://nces.ed.gov/fastfacts/display.asp?id=372 Date Accessed: December 12, 2017

NCES Fast Facts Tool (2017b). https://nces.ed.gov/fastfacts/display.asp?id=61 Date Accessed: December 12, 2017

Wikina, S.B. (2014). What Caused the Breach? An Examination of Use of Information Technology and Health Data Breaches. *Perspectives in Health Information Management*, 11(Fall).