Emmet Hall-Hoffarth
23081152

The principle data source used in this study is the database of publicly announced data breaches obtained from Privacy Rights Clearinghouse – a non profit organization that advocates for privacy protection. This study will consider data breaches which occurred in the years 2005-2017 in the United States. Data from the United States are particularly rich because there companies are required by law to announce when they have fallen victim to unauthorized data access (NCSL, 2018). The database contains information about the time, location, and magnitude (number of records leaked) of data breaches, as summarized in table 1C. Furthermore, the table includes short descriptions for each event which come primarily from news articles and press releases. From this I was able to construct a

**Table 1A: Summary Statistics for Fincial Variables**

| Statistic | N | Mean | St. Dev. |
|---|---|---|---|
| Net income (Millions USD) | 437,859 | 51.1 | 541.3 |
| Revenue (Millions USD) | 402,217 | 790.2 | 3,961.4 |
| Sales, General, and Administrative Expenses (Millions USD) | 353,249 | 301.0 | 1,640.2 |
| Market Value (Millions USD) | 339,657 | 3,033.3 | 16,026.5 |

**Table 1B: Types of Data Loss**

| Statistic | Mean |
|---|---|
| Customer | 334 |
| Employee | 224 |
| Credit Card | 154 |
| CVV | 6 |
| Social Security Number | 369 |
| Name | 442 |
| Address | 283 |
| Personal Information | 146 |
| Total Breaches | 791 |

**Table 1C: Magnitude of Data Loss**

| Statistic | N | Mean | Min | Max | St. Dev. |
|---|---|---|---|---|---|
| Records Leaked per Breach | 791 | 6,445,309 | 0 | 3,000,000,000 | 108,950,764 |

**Table 1D: Summary Stock Market Data**

| Statistic | N | Mean | St. Dev. |
|---|---|---|---|
| Daily Firm Return | 1,149,144 | 0.0005 | 0.03 |
| Value Weighted Market Return | 1,161,325 | 0.0003 | 0.01 |
| Risk Free Market Return | 1,161,468 | 0.04 | 1.20 |
| SMB factor | 1,161,468 | 0.004 | 0.57 |
| HML factor | 1,161,468 | 0.001 | 0.65 |

**Table 1E: Summary Statisitics for Control Variables**

| Statistic | N | Mean | St. Dev. |
|---|---|---|---|
| Nominal GDP (Billions USD) | 559,102 | 16,026.2 | 1,932.3 |
| Real GDP (index 2012) | 559,102 | 100.2 | 6.0 |
| Inflation (% yoy) | 548,418 | 1.8 | 1.0 |

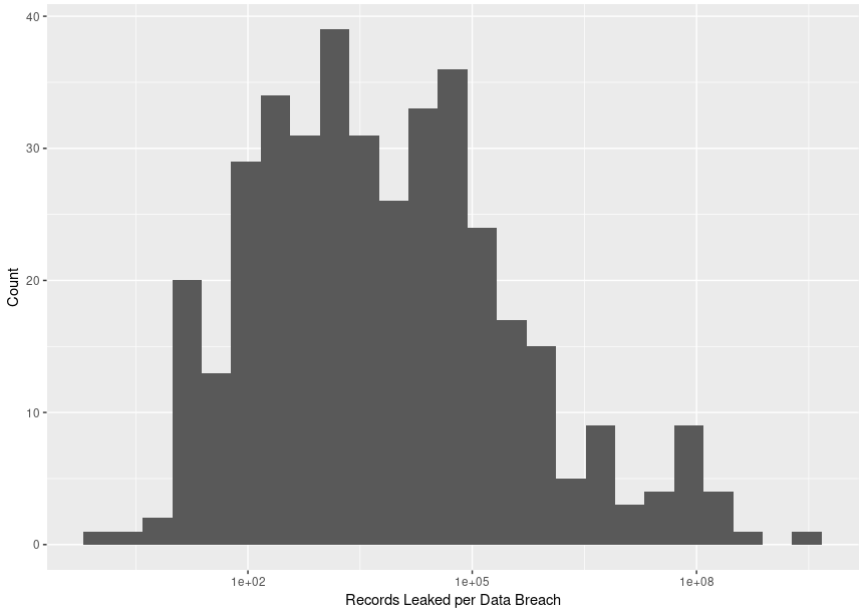
measure for the types of data leaked by creating dummies which are equal to one when the description contains certain keywords such as, “credit card.” In particular, since employees are often mentioned as having stolen customer data, the dummy for employee data is only equal to one if the description does not also contain, “customer.” While this measure is certainly not perfect, it will help to develop some idea about the types of private information that consumers care most about. This information is given in table 1B.

Unfortunately, there was no standard identifier given for companies listed in the Privacy Rights Clearinghouse database, and performing a fuzzy match on the given names provided unsatisfactory results. Therefore, the listed companies were matched to standard identifiers manually, by searching both Google and the COMPUSTAT database. Even though this process was thorough, the resulting match rate is still relatively low – 33% or 791 of the 2366 total data breaches targeting for profit enterprises. This is because the sample is restricted to publicly traded companies for which financial information is readily available. Companies which were private at the time of data breach, as well as the small number of those located outside the United States were excluded from the sample. During the matching process, 67 companies were identified as subsidiaries of publicly traded companies at the time of the breach. These can possibly be included in the sample, while controlling for the fact they are subsidiaries, which may lead to a very dampened effect on outcomes for the parent company. In addition to these, five companies were identified but excluded because they were acquired in the same quarter as the data breach – making the two effects on revenue impossible to distinguish.

As one might expect, the matched publicly traded companies are associated with larger data breaches on average, as they account for only 33% of the data breaches, yet 53% of total records leaked and 42% of breaches in which over 1 million records are leaked. Nevertheless, breaches of private

firms are also significant, and these firms may vary endogenously in many ways from public companies, so the results of this paper should not be generalized to all firms. Here it is worth noting that the distribution of number of records leaked – shown in Figure 1 – is very strongly skewed to the right[1]. Of the 791 matched data breaches, 403 resulted in no records being leaked, while the largest breach targeted Yahoo in 2016 and resulted in 3 billion records being exposed. Therefore, the mean number of records leaked of over 6 million does not give a good indication of the magnitude of a typical data breach.

Figure 1: Records Leaked per Data Breach (log_10 scale)




The primary dependent variable for the sales study is quarterly total revenues, as reported on the financial statements of publicly traded companies. These data were obtained from COMPUSTAT. Once companies in the data breach database were identified, data were queried from COMPUSTAT for all companies with an SIC code that matches the first three digits of any firm affected by a data breach

1 So skewed, in fact, that any attempt to plot it in a meaningful way without a log transformation would be fruitless.

over the study period. This is primarily to allow meaningful industry fixed effects to be calculated, as well as serving as the control group for the difference in differences regression. The data breach database was merged with this COMPUSTAT database over the GVKEY of institutions, as well as calendar quarter.

Summary statistics for the COMPUSTAT database are shown in Table 1A. Other than revenues, the COMPUSTAT database also contains Selling, General, and Administrative (SGA) expenses – a proxy for marketing expenditure. This variable is a useful because it is likely that firms may respond to data breaches by increasing advertising spending to counter the bad publicity. Whether or not this happens can be tested by using this variable as an outcome in the regression model. Whether or not it is effective can be tested by including an interaction term in the main revenue regression. Furthermore, controls for macroeconomic shocks (real GDP and inflation) were obtained from the BEA, and are summarized in Table 1E. Finally, Google Trends provides a monthly index for the frequency of searches for a given term, which in this study is used as a proxy for the media attention . These data was collected by searching both for the stock ticker of a company, as well as its official name – after stripping superfluous strings such as “CORP,” and “INC,” in order to programmatically define reproducible search terms. This variable can be used both as an outcome – to test whether publicity about a company increases in the wake of a data breach – and as an independent variable – to test whether potential lost revenues may be driven by media coverage.

For the stock market study the dependent variable is daily stock returns. These data were obtained from CRSP, and are summarized in Table 1D. This database also contains a wealth of other standard stock information such as trade volume, volatility, and dividend payments. Since an event study methodology will be employed, factors for the Fama-French model were obtained from Kenneth

French's website. The data breach database was merged to the CRSP database over company stock tickers and the date.