Data Descriptor

A Dataset Containing S&P500 Information Security Breaches and Related Financial Firm Performances

Nynke Voermans and Francesco Lelli *

Posted Date: 14 June 2024

doi: 10.20944/preprints202406.0975.v1

Keywords: dataset, information security breaches, stock measures, accounting measures, event data, S&P500, financial data, security

 doi:10.20944/preprints202406.0975.v1

Data Descriptor

A Dataset Containing S&P500 Information Security Breaches and Related Financial Firm Performances

Nynke Voermans and Francesco Lelli *

Tilburg University; n.voermans@tilburguniversity.edu
* Correspondence: f.lelli@tilburguniversity.edu

Abstract: In this paper, comprehensive datasets are presented to advance research on information security breaches. The datasets include data on disclosed information security breaches affecting S&P500 companies between 2020 and 2023, collected through manual search of the Internet. Overall, the datasets include 504 companies, with detailed information security breach and financial data available for 97 firms that experienced a disclosed information security breach. This document will describe the datasets in detail, explain the data collection procedure and shows the initial versions of the datasets.

Keywords: dataset; information security breaches; stock measures; accounting measures; event data; S&P500; financial data; security

Introduction

In the current digital world, information security breaches pose a significant threat to companies. Therefore, the datasets described herein contain data on information security breaches and related financial data of S&P500 companies. Initially, these datasets have been collected to obtain insight into the impact of information security breaches on overall firm performance and the moderating impacts of insider involvement and breach identification. Data on the information security breaches affecting S&P500 companies between 2020 and 2023 has been collected by manual search of the Internet. The financial data, including stock data and several accounting measures have been collected via Refinitiv (LSEG). With this document, a description of the datasets, the procedure for collecting the data and an initial version of the datasets are provided. While these datasets are useful for examining the impact of information security breaches on firm performance, they are also useful for other research purposes. For example, the identification of patterns and trend in information security breaches, the effect on other financial metrics or the development of predictive models.

The rest of the document is structured as follows. First, a specification table presenting a general overview of the characteristics of the dataset will be given. Then, the value of the data will be explained. The data description part describes in detail what the datasets contains. Lastly, the data collection process will be explained.

Specification Table

Table 1 present an overview of the datasets, facilitating easy access to all relevant characteristics of the data. It includes details such as the subject area, data type, the acquisition of the data and the data format. Additionally, it outlines the data collection process and its parameters.

Table 1. Tabular description of all information in the datasets.

Subject	Information security / economics
Specific subject area	Information security breaches and firm performance measures
Type of data	Six Microsoft Excel documents (.xls)

 doi:10.20944/preprints202406.0975.v1

How data were acquired	Manual search of Google Chrome, financial data from Refinitiv Workspace (LSEG)
Data	Raw
Parameters for data collection	Data has been collected on disclosed information security breaches between 2020 and 2023 to capture recent data and ensure relevance to recent trends and developments. This process involved sourcing from various highly visible media outlets, as these are likely to be followed by investors and consumers and therefore likely able to detect an effect. Specific key- words were used to identify relevant events. This research focused on S&P500 companies, as their security breaches are likely to attract attention and their high data availability. Several data collection and filtering steps were taken to ensure data integrity
Description of data collection	The event data collection process involved identifying disclosed information security breaches among S&P500 companies from 2020 and 2023. This was done through manually searching various highly visible media outlets, including technology platforms, news sites and firm's websites. 4 keywords were used to identify disclosed information security breaches. Initially, 205 breached firms were identified, with data collected on disclosure date, insider involvement and the identifier of the breach. Matching control firms were selected based on industry SIC codes and size similarity. Financial data, including stock data and several accounting measures were obtained from Refinitiv (LSEG) for both short- and long term impact assessment.

Value of the Data

The datasets records disclosed information security breaches of S&P500 companies between 2020 and 2023 and financial stock data and accounting measures. These comprehensive datasets are vital for examining the impact of disclosed information security breaches on performance measures, shedding light on the connection between security breaches and firm performance. As the dataset focuses on recent data, it is valuable for gaining new insights and identifying patterns and trends in

 doi:10.20944/preprints202406.0975.v1

information security breaches. Moreover, the availability of accounting metrics alongside stock data enables the examination and understanding of the overall economic consequences. Researchers across various sectors such as finance, cyber security and IT can leverage this data to enhance their understanding of information security breaches. These insights are not only valuable for researchers, but also inform investor decision making incorporating cyber security risks. In addition, companies can utilize this information to enhance their risk management strategies and information security policies. Beyond examining the impact of information security breaches on firm performance, this dataset can be used for analysis of trends over time, industry- or firm-specific comparisons and investigation into other moderating factors like firm characteristics. The events facilitate the analysis of impact on other metrics like market share and brand reputation. Additionally, the data enables the development of predictive models of the occurrence and impact of information security breaches. Furthermore, comparative analysis of the datasets with other datasets or research findings in the field of information security enables researchers to identify commonalities, discrepancies and trends. All are advancing research in information security and contribute to deeper understanding of information security breaches.

Data Description

Overall, three types of datasets have been collected. The first dataset centers on specific information security breach events. The second and third dataset focus on short term financial data and long term financial data, respectively. In the following subsections, each type of dataset will be described in detail

1. Information Security Breach Data

The event dataset includes rows corresponding to each individual S&P500 company and columns representing various variables of interest, capturing crucial information regarding information security breaches among S&P500 companies. Table 2 offers a comprehensive overview of the columns in the dataset, including their data types and descriptions. The primary key of the dataset is the ticker: a one, two, three or four lettered code which uniquely identifies each company. In addition to this ticker, the company name serves as another identifier for each company in the dataset ensuring clarity. The dataset includes each company's operating sector, categorized in one of 19 possible sectors. Focusing on the information security breaches, a boolean variable has been added, labeled yes when the firm experienced a disclosed information security breach between 2020 and 2023 and no if they did not. In case of a disclosed information security breach, the date of first disclosure has been added. If a firm did not experience a disclosed information security breach, the disclosure date is denoted as X. In some cases of breaches a short description is added, such as "multiple breaches" if the firms experienced multiple disclosed breaches in the period, "no open source" when the disclosure was not publicly available", or for example "MOVEit" or "Cl0P" to specificy the name of the breach. If the firm did not experience a disclosed information security breach or no description was added, this is again denoted as X. In the case of a security breach, the dataset includes the URL of the source where the first disclosure about the breach was found. For some breaches, an additional source has been added which has been used to obtain the URL of the initial disclosure. If no disclosed information security breach occurred and no additional source was utilized, it is denoted as X. In addition, a column has been added to indicate the level of insider involvement in case of a breach. Table 3 outlines the possible values of insider involvement, along with a description of each value. If no information was disclosed regarding insider involvement during a breach, it is denotes as "unknown". If there was no disclosed information security breach, it is denoted as X. Then, a column has been added to indicate who identified the information security breach. Table 4 outlines the possible values of the identifier with a description for each value. Again, if no information was disclosed about the identifier, it is denoted as "unknown" and if there was no security breach, it is denoted as X. Two additional columns have been included to capture firm characteristics: the four-digit SIC code representing its industry and the firm size in thousands relative to industry averages. If data on these characteristics were unavailable, they are marked as "unknown". In case of no information security breach, it is denoted as X. For each breached firm, an associated control firm has been identified. The unique ticker symbol of this control firm is provided in the control firm ticker column. In cases where no breach was disclosed, no control firm was needed,

 doi:10.20944/preprints202406.0975.v1

thus marked as X. If no suitable control firm could be determined, is it denoted as ”unknown. Finally, the dataset includes two columns specifying the range of stock data required, based on the date of disclosure. This range spans from 125 days before to 125 days after the first disclosure of the information security breach. In case of no security breach, they are both denoted as X. Initially, data was obtained for 503 companies. Among them, 205 companies experienced a disclosed information security breach within the specified period. After sample filtering, the final sample comprises 97 companies that faced a disclosed information security breach between 2020 and 2023 for which all information is available. The dataset includes all firms experiencing a disclosed information security breach, regardless of the completeness of information, and firms without disclosed information security breach, as all may provide valuable insights for further analysis

Table 2. Information security breach dataset overview.

Variable	**Type**	**Description**
Ticker	Categorical	Ticker symbol of the S&P500 company to uniquely identify the company
Company name	Text	Name of the S&P500 company
Sector	Categorical	Sector in which the S&P500 company operates
Disclosed information security breach between 2020 and 2023	Boolean	Indicates whether the company experienced a disclosed information security breaches between 2020 and 2023 (yes or no)
Date of disclosure	Date	Date of first disclosure of the information security breach
Short description	Text	Brief description of the information security breach
Source	Text	URL of the first disclosure of the information security breach
Additional source	Text	Additional URL used to obtain the URL of the first disclosure of the information security breach
Insider involvement	Categorical	Involvement of insiders and outsiders in the information security breach (insider error, malicious insider or outsider)
Identifier of the breach	Categorical	Identifier of the information security breach (company itself, hacker or third party

 doi:10.20944/preprints202406.0975.v1

SIC code	Categorical	Four-digits Standard Industrial Classification (SIC) code of the breached firm
Firm size (in thousands)	Numerical	Size of the breached company in proportion to industry averages (given that they have more than 100 million euros operational revenue, more than 200 million euros total assets and more than 1000 employees)
Control firm ticker	Categorical	Ticker symbol of the related S&P500 control firm
First stock date	Date	Starting date from which stock prices of the breached and control firm are necessary, 125 days before the event date
Last stock date	Date	End date from which stock prices of the breached and control firm are necessary, 125 days after the event date

Table 3. Insider involvement column specification.

Values	**Description**
Insider error	The information security breach was the result of an unintentional action or mistake by someone within the organization, for example through phishing
Malicious insider	The information security breach was caused by someone within the organization with malicious intent
Outsider	The information security breach was caused by an external hacker or hacker group who gained unauthorized access without the involvement of an insider

Table 4. Identifier of the breach column specification.

Values	**Description**
Organization Itself	T he information security breach was identified internally by the breached organization

 doi:10.20944/preprints202406.0975.v1

Hacker	The information security breach was identified by / diclosed to the breached firm by the external hacker who gained unauthorized access, for example through ransomware
Third party	The information security breach was identified by a benign third party, for example a cybersecurity firm or regulatory agency

2. Short Term Financial Data

The short term financial dataset consists of two files structured similarly: one for the breached firms and one for the control firms. Each row corresponds to a combination of a firm with a trading date for which stock data has been collected. The columns represent various stock-related variables of interest. Table 5 offers a comprehensive overview of the columns in the dataset, including their data types and descriptions. Similar to the event dataset, the ticker symbol and company name are included. In addition, the date variable corresponds to the date for which stock data is recorded, aligning with the range provided by the first and last stock date in the event dataset (125 days before to 125 days after the event). The unique combination of the ticker and the date is the primary key in this dataset. Focusing on the stock related variables, the closing price of the breached or control company's stock has been included. Additionally, it incorporates the net change in stock price compared to the previous trading day, indicating whether the difference is positive or negative. This variable is available for each day, except for the initial date data is collected for a company, where it is left blank. In addition, the percentage change in the company's stock compared to the previous trading day has been included, except for the initial date when data is collected. Moreover, the dataset includes the opening price of the company's stock on the given date, as well as the lowest and highest prices observed on that day. The dataset also includes the total number of traded shares for the company's stock on a certain date, along with the total value of those traded shares. Furthermore, the flow of the company's stock on a certain trading date is included, indicating the overall direction of trading activity by comparing inflow (stock bought) and outflow (stock sold) of the stock. A positive flow indicates a net buying activity, when more shares were bought than sold on that specific date. Conversely, a negative inflow indicates a net selling activity, when more shares were sold than bought. The flow has a value of 0 on each company's initial trading date. All values are denoted in USD to be able to easily compare them. The short term financial dataset is supplemented by an additional dataset that contains the S&P500 market index data. This data has been collected to control for industry specific influences later in this research, such as events affecting the entire industry like the Covid-19 pandemic. Similar to the collection of stock data for the breached and control firms spanning from October 1, 2019 to May 1, 2024, corresponding market data has been gathered. The variables collected are similar to those in the short term financial dataset: ticker, date, stock close, stock net, percentage change, stock open, low, high, volume, turnover and flow, allowing for direct comparison of the breached firms with the market index

Table 5. Short term financial dataset overview.

Variable	**Type**	**Description**
Ticker	Categorical	Ticker symbol of the S&P500 company to uniquely identify the company
Company name	Text	Name of the S&P500 company

 doi:10.20944/preprints202406.0975.v1

Date	Date	Date for which stock data is recorded
Stock close	Numerical	The closing price of the company' stock on the given date in USD
Stock net	Numerical	The net change in the company's stock price compared to the previous trading day in USD
Percentage change	Numerical	The percentage change in the company's stock price compared to the previous trading day
Stock open	Numerical	The opening price of the company's stock on the given date in USD
Low	Numerical	The lowest price of the company's stock during the given trading date in USD
High	Numerical	The highest price of the company's stock during the given trading date in USD
Volume	Numerical	The total number of shares traded for the company's stock on the given date, including both buying and selling
Turnover	Numerical	The total value of shares traded for the company's stock on the given date in USD
Flow	Numerical	The flow of the company's stock on the given date, indicating overall direction of trading activity through a comparison of inflow (stock bought) and outflow (stock sold)

3. Long Term Financial Data

The long term financial dataset also comprises two similarly structured files: one for the breached firms and one for the control firms. Every row corresponds to a breached or control firm,

 doi:10.20944/preprints202406.0975.v1

while the columns represent various accounting metrics of interest for each firm. Table 6 offers a comprehensive overview of the columns in the long term financial dataset, including their data types and descriptions. The ticker symbol of the firm serves
as the primary key, uniquely distinguishing each row. In addition, the company name has been included. For each ticker, the date of first disclosure of the information security breach, as recorded in the event dataset, has been included. Based on this date, the accounting metrics have been gathered. Specifically, data for each metric has been collected from financial statements spanning from 2 quarters prior to two quarters after the disclosure. Consequently, for each firm, data from differences quarters has been collected. Each metric, in combination with a quarter (-2, -1, 1 or 2), is represented in a column. This timeframe was chosen to enable to measurement of the long term impact of disclosed information security breaches while minimizing the risk of other significant events occurring that could affect the metrics. The long term financial dataset includes data on sales (S) and operating income (OI), reported in millions of USD. Additionally, it comprises return on assets (ROA) and return on equity (ROE), expressed in percentages and calculated pretax. Lastly, the dataset includes operating income divided by assets (OI/A) and operating income divided by sales (OI/S), given as ratios. These specific accounting measures are carefully selected as they are most commonly used as firm performance indicators, reflecting firm profitability and efficiency (Bharadwaj, 2000; Ko & Dorantes, 2006). Negative values are indicated with a ”-” sign and are shown in red text. In case no data was available for a specific metric in a specific quarter, it is left blank.

Variable	**Type**	**Description**
Ticker	Categorical	Ticker symbol of the S&P500 company to uniquely identify the company
Company name	Text	Name of the S&P500 company
Date of breach	Date	Date of the information security breach events to be able to determine the needed quarterly data
S – 2	Numerical	Sales in millions two quarters before the information security breach
S – 1	Numerical	Sales in millions one quarter before the information security breach
S + 1	Numerical	Sales in millions one quarter after the information security breach
S + 2	Numerical	Sales in millions two quarters after the information security breach
OI – 2	Numerical	Operating income in millions two quarters before the information security breach

 doi:10.20944/preprints202406.0975.v1

OI – 1	Numerical	Operating income in millions one quarter before the information security breach
OI + 1	Numerical	Operating income in millions one quarter after the information security breach
OI + 2	Numerical	Operating income in millions two quarters after the information security breach
ROA – 2	Numerical	Pretax return on assets two quarters before the information security breach
ROA – 1	Numerical	Pretax return on assets one quarter before the information security breach
ROA + 1	Numerical	Pretax return on assets one quarter after the information security breach
ROA + 2	Numerical	Pretax return on assets two quarters after the information security breach
ROE – 2	Numerical	Pretax return on equity two quarters before the information security breach
ROE – 1	Numerical	Pretax return on equity one quarter before the information security breach
ROE + 1	Numerical	Pretax return on equity one quarter after the information security breach
ROE + 2	Numerical	Pretax return on equity two quarters after the information security breach
OI/A – 2	Numerical	Operating income divided by total assets two quarters before the information security breach
OI/A – 1	Numerical	Operating income divided by total assets one quarter before the information security breach

 doi:10.20944/preprints202406.0975.v1

OI/A + 1	Numerical	Operating income divided by total assets one quarter after the information security breach
OI/A + 2	Numerical	Operating income divided by total assets two quarters after the information security breach
OI/S – 2	Numerical	Operating income divided by sales two quarters before the information security breach
OI/S – 1	Numerical	Operating income divided by sales one quarter before the information security breach
OI/S + 1	Numerical	Operating income divided by sales one quarter after the information security breach
OI/S + 2	Numerical	Operating income divided by sales two quarters after the information security breach

Experimental Design, Materials and Methods

To obtain the three types of datasets described above, several steps have been taken related to the collecting of this data, which will be described below.

1. Information Security Breach Dataset

The information security breaches are identified through manual search of the full text of various publicly accessible media outlets on the Internet. Examples of these public sources include information security and technology platforms like bleepingcomputer.com and securitymagazine.com, news sites like Forbes and CNN and public disclosures on the breached company's website. All used sources represent highly visible media outlets likely followed by investors and consumers. It focuses specifically on S&P500 companies, as they are among the largest and most well-known publicly traded companies. Therefore, they are likely to attract significant attention from the media, investors and consumers in case of a security breach, providing a sample robust to detect an effect. In addition, data about these companies is expected to be largely available and well-documented. Lastly, focusing on S&P500 companies enables the possibility to compare the impact of information security breaches across different firms and industries within the same index. Data on the ticker symbols of these companies, the company's names and the operating sectors are taken from liberatedstocktrader.com. This study covers security breaches spanning between 2020 and 2023 to capture recent data and ensure relevance to current trends and developments, generating current and valuable insights. After careful examination of prior research, the online search to identify information security breaches used the keywords "information security breach", "cyber security breach", "cyber attack" and "data breach" (Campbell et al., 2003; Gordon, Loeb & Zhou, 2011; Paul & Das, 2024). Based on these criteria, 205 S&P500 companies that experienced a disclosed information security breach between 2020 and 2023 were initially identified. For each breached company, the date of the first public disclosure has been gathered. Furthermore, data regarding insider involvement and the identifier of the breach has been collected by carefully reading either the initial source disclosing the breach or supplementary media outlets

 doi:10.20944/preprints202406.0975.v1

2. Financial Data

For the short term financial data, stock price data has been obtained from financial market data provider Refinitiv (LSEG). Data of the breached and control firms' stock have been collected for 125 days prior to 125 days after the disclosure of the information security breach, ensuring a reliable time frame for event studies (Campbell et al., 2003; Gordon, Loeb & Zhou, 2011; Tripathi & Mukhopadhyay, 2022). All has been merged into two XLS files, one for the treatment firms and one for the control firms. In addition, the stock price data from the general market index (S&P500) has been collected from October 1, 2019 to May 1, 2024, similar to the range of data collected for the breached and control firms.

For the long term financial data, several accounting measures are employed. Using accounting measures of performance to assess firm performance is a widely adopted and accepted approach (Bharadwaj, 2000; Hunton, Lippincott & Reck, 2003; Ko & Dorantes, 2006). Consequently, data on four profit-based ratios has been gathered: Return on Assets (ROA), Return on Equity (ROE), Operating Income to Assets (OI/A) and Operating Income to Sales (OI/S). ROA and ROE are indicators of the profitability of a firm, while OI/A and OI/S serve as measures of the direct effect (Bharadwaj, 2000). Hence, these profit ratios are considered reliable measures of firm performance and are therefore collected. In addition, data on sales and operating income has been gathered. Data regarding these performance measures has been collected from Refinitiv (LSEG). Quarterly financial performance data has been collected for both the treatment and control samples, focusing on two distinct timeframes: before and after the disclosed information security breach. For each treatment sample, the performance metrics spanning two quarters before to two quarters after the security breach have been collected based on the financial statements of the firms. Correspondingly, the performance metrics have been collected for each related control firm for the same applicable quarters included in the treatment firm. The decision to limit the analysis to two periods before and after the disclosed security breach was made to minimize the potential impact of other significant events unrelated to this research. Again, all has been merged into two XLS files, one for the treatment firms and one for the control firms

Conclusions

In this document, a collecting of datasets has been presented to advance research in information security breaches. Data regarding disclosed information security breaches has been collected through manual search of Internet. In addition, short term and long term financial data has been retrieved from Refinitiv (LSEG). The breach dataset includes data on 504 companies, of which 97 experiences an disclosed information security breach and had all necessary data available. For these 97 firms, both short term and long term financial data has been collected. The datasets spans data from 2020 to 2023 affecting S&P500 companies.

References

Bharadwaj, A. S. (2000). A Resource-Based Perspective on Information Technology Capa- bility and Firm Performance: An Empirical Investigation. MIS Quarterly, 24 (1), 169-196. https://doi.org/10.2307/3250983

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. https://doi.org/10.3233/JCS-2003-11308

Gordon, L. A., Loeb, M. P. and Zhou, L. (2011). The Impact of Information Security Breaches: Has There Been a Downward Shift in Costs? Journal of Computer Security, 19 (1), 33-56. https://doi.org/10.3233/JCS-2009-0398

Hunton, J. E., Lippincott, B. and Reck, J. L. (2003). Enterprise Resource Planning Sys- tems: Comparing Firm Performance of Adopters and Nonadopters. International Jour- nal of Accounting Information Systems, 4 (3), 165-184. https://doi.org/10.1016/S1467-0895(03)00008-3

Ko, M. and Dorantes, C. (2006). The Impact of Information Security Breaches on Fi- nancial Performance of the Breached Firms: An Empirical Investigation. Journal of Information Technology Management, 17 (2), 13-22.

Paul, S. and Das, S. (2024). Public Disclosure of Information Security Breach Incidents:Short-term Stock Market Reaction on Indian Listed Firms. Journal of Organizational Computing and Electronic Commerce, 1-29. https://doi.org/10.1080/10919392.2024.2335689

 doi:10.20944/preprints202406.0975.v1

Tripathi, M. and Mukhopadhyay, A. (2022). Does Privacy Breach Effect Firm Performance? An Analysis Incorporating Event-Induced Changes and Event Clustering. Infor-mation & Management, 59 (8). https://doi.org/10.1016/j.im.2022.103707