REVISITING THE COST OF DATA BREACH DISCLOSURES:

A DECADE LATER - EXTENSION

By

MARTON ISTVAN SZEP

_____

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree
With Honors in

Management Information Systems

THE UNIVERSITY OF ARIZONA

M A Y 2 0 1 9

Approved by:

_____

Dr. Matthew Hashim

Revisiting the Cost of Data Breach Disclosures: A Decade Later - Extension

A short paper

Abstract

This article is an extension of the article, "Revisiting the Cost of Data Breach Disclosures: A Decade Later." This extension takes the first step towards the identification and classification of underlying factors affecting the shift in stock market backlash discovered in the paper. The goal is to get an insight into several factors and determine whether or not said factors are correlated to the market response shift corresponding to the studied events. The primary concern regarding this research is a lack of available information concerning the majority of the events. While stock market archives are well kept and a majority of news sources have historic archives, more often than not such archives are not complete and often have access limitations. As such, this study was to make do with readily available news articles of the events. The project relied on the utilization of sentiment analysis to determine the sentiment of articles around a filtered random sample of the 198 events analyzed in the paper. It was found that there was no significant correlation between the general article sentiment and the shift in market backlash. However, in terms of a specific, majority subcategory (Tech), correlation can be found and statistically shown.

Introduction

The primary goal of this project is to better understand the underlying reasons behind the shift in stock market backlashes of data breach announcements. The problem is relevant because it may contribute significantly to forecasting stock market reactionary behavior in the future, not only in terms of data breach announcements but any isolated, semi-frequent, negative company events. As such there are several hypotheses which give a reasonable explanation to the shift in attitude. First and foremost the argument also mentioned in the paper [6], that public sentiment has shifted, and users simply care less nowadays about data loss as they did in the past. This would logically explain why the reaction to such incidents to which humanity has no doubt gotten more used to over the decade has become milder on the stock market. This is the hypothesis this project is focusing on evaluating. Further, alternate hypotheses will be mentioned in the conclusion and future remarks section.

The main issue governing the project is the fact that we have to analyze broad public sentiment on a scarce data problem. Since the events are in the past, surveying is not an option. Twitter, another widely used method to generally assess public sentiment, is also not ideal for this case, as the behemoth social networking site was still in rather early ages when the targeted period of events started. As such an alternate route had to be found to study public sentiment in this case.

Theory

Given the underlying assumption that press adjusts to public opinion and sentimentality to optimize revenue, estimating general population sentiment of events can be accomplished via the evaluation of news coverage sentiment for the given event. Please note that this is an estimation method and not an exact indicator of general sentiment, which is rather difficult to measure (and could be potentially further explored in the future by looking at (for example) Twitter historical data, which however requires a Twitter Premium API).

In order to optimize sentiment analysis efficiency, it is optimal to look at news articles primarily from general newspapers, as these appeal to a wider audience and thus give a better reflection of public sentiment towards the given event. [4] In terms of highly technical blogs and articles, the focus is rather on technical detail and accuracy, as well as often conciseness, which eliminates sentiment from the article. Nonetheless, due to the scarcity of available data in the case of this study, we cannot allow to completely ignore data from sub-optimal sources, thus we have included them in our collection and labeled them accordingly under the umbrella term "security papers."

Methodology

A random selection of events was selected from amongst the 198 events researched within the paper. Three articles released within a week of the event were selected for each chosen event. The articles were prioritized the following way. First, the primary responses on Google News were evaluated. If not enough news articles were found for the given event on Google, as a secondary database, we researched first Softpedia's data-breach article collection, and then the University of Arizona Library's news article database. For an article to be qualified as the relevant article it had to focus mainly on the incident, and not surrounding legal dealings or world events. If an article did not qualify, simply the next article was taken. When finding articles, the primary concern was to reduce bias, however, given the very scarce nature of older articles some selection bias was unavoidable. As such, throughout the selection process, if not enough news articles were found for a given event, rather than selecting strictly technical event descriptions in historical data breach disclosure archives, we have opted to skip these events. This was due to the nature of such archives, which typically include the disclosure letter from the company as well as numerical data and statistics of the breach event. Since the primary goal of this research was to find sentiment data on the events, such articles would heavily affect the relatively small sample and introduce significant bias into sentiment data. The articles selected in the end were screened for content to maintain the primary criteria that these articles are only focusing on the incidents themselves, and upon approval were entered into the sample data set.

Further important notice on article selection is the rare availability of news articles related to older events, especially before the year of 2009. This has unfortunately introduced some unavoidable sampling bias to the small random sample, as we had to drop a considerable number of events for the simple reason that there was no available news coverage of the event online. Albeit also present at later events, similar issues affected newer events at a much lesser scale.

After article selection, events were categorized in two ways. Firstly, by the nature of the company represented in the event. The categories here included tech companies, financial institutions, and stores. While there weren't enough 'stores' to create a viable subcategory, they were separated since they did not fit into either other category. Every event was fit into one of these categories based on the closeness to the categories of the company's primary services and products offered. As an example, Apple Inc. was fit into the Tech category, and Macy's Inc. was fit into the Store category. The second form of categorization was based on the source of the news article. This categorization, rather than one for each event was assigned on an individual basis to each article, which were labeled as "News" or "Security." If the article came from a source which could be categorized as a general news source, and as such the source would have different sections outside the software and cyber security columns (under which these articles were most generally published), the article was labeled "news." If the journal or news source was generally only focusing on software or cyber security related issues, then they were labeled as "security." This distinction was made due to the fact, that in general, the more specific the news source was, the less sentimental the articles presented by the news source. This is also the reason we have stayed away in the article finding process from data-breach historical archives, as these have generally no public sentiment reflected in the articles.

The Natural Language Processing (NLP) component of the project relied on the utilization of the TextBlob (Figure 1.) Python library. This library was built on the foundations of the parsing solutions developed for the Natural Language Toolkit (NLTK), and the "pattern.web" library/database. [10] TextBlob is an ideal choice for simple NLP tasks since it has many otherwise complex features integrated and as such is much simpler to use in straightforward applications than its more customizable counterparts.

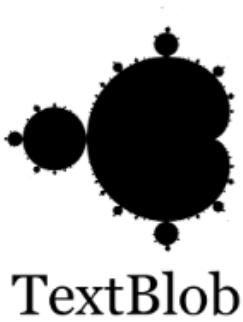


Figure 1. TextBlob Logo [11]

In terms of sentiment analysis, TextBlob relies on a lexicon file which has several, pre-assigned sentiment polarity (-1.0, 1.0), subjectivity (0.0, 1.0), and intensity (0.5, 2.0) values for each English language word. [1] When processing a text, TextBlob separates the entry into sentences (default option), and after identifying phrases and word stems, it replaces words within the sentence with their respective values for calculation. This part becomes more complex when a word is identified as a modifier word, in which case the Polarity and Subjectivity values are ignored, and the word simply multiplies the next word based on its only remaining intensity value. There are also some exceptions, on which more information can be found in the TextBlob GitHub library. [11]

However, to maintain the simplicity and error-free operation of the code, TextBlob does have some compromises, the primary of which is that the algorithm simply ignores any words not found in its lexicon. This in the case of a data breach study looking at non-preprocessed articles is very useful, as jargon may vary based on affected industry, breach specifics, and also article source. [1]

When using TextBlob, each sentence in an article was evaluated and in the final processing averaged to give a singular polarity and subjectivity value to the article. The individual sentence values were kept for data verification purposes.

Along with the polarity values for the articles, we have also evaluated a combination of polarity and subjectivity to reduce the effect of objective articles in the process. This was done to give us a secondary perspective with an increased utilization of the results given the scarce amount of data available.

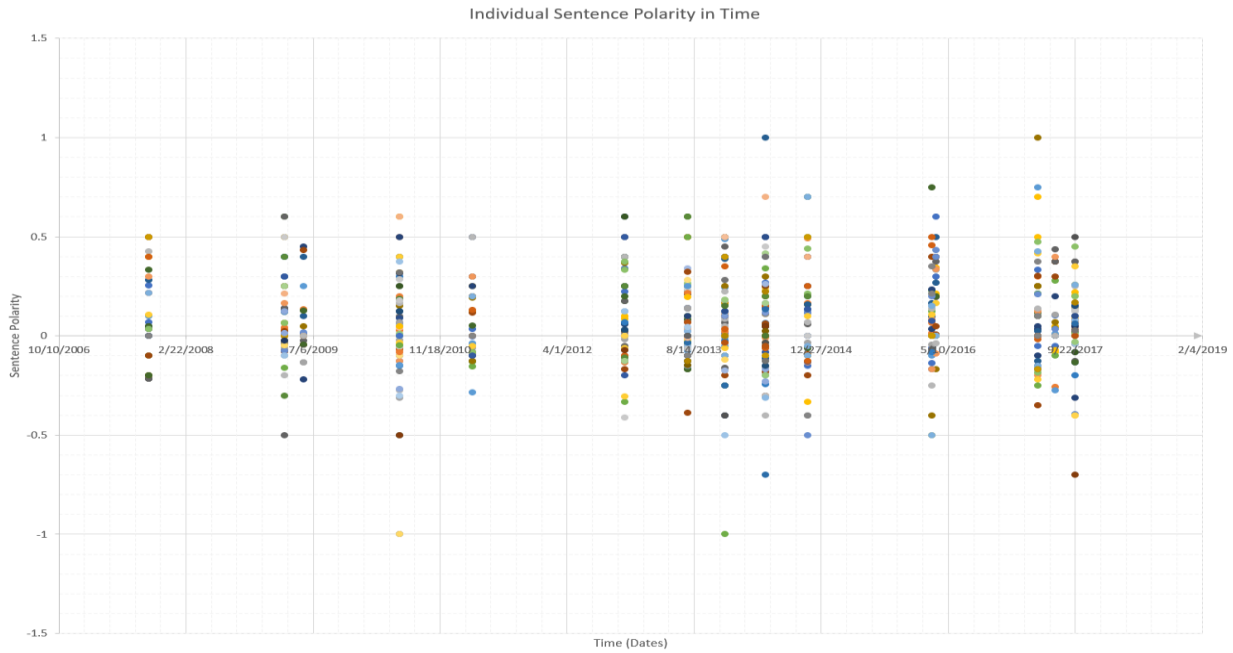
There was a total of 15 events and 45 articles sampled and evaluated.Scores were entered into a database and evaluated through MATLAB. Correlation analysis was completed for both general results (including all articles and categories), and categorized results. The possibility of correlation within the data was determined and evaluated.

Outlier articles were looked at specifically, to identify the cause of such a phenomenon. These examinations were done manually since the format and availability of these articles differ vastly; hence, automation of such tasks was not possible under the confines of the project. Outlier data was primarily identified to fall under two categories, the first of which were shorter articles, where headline phrasing would heavily affect the polarity of the article. (Due to this concern articles shorter than 5 sentences were not included in the sample.) Secondly, due to each event being unique there were cases where the situation in which the leak/breach happened prompted a more positive response than expected. (As an example of such a case, in terms of the Comcast data breach the initially thought to be exposed information was much more than what was actually lost, hence changes to the articles were made reflecting a more positive tone.) To adequately understand all that plays into such variation a much more in-depth analysis is required, as such we have kept these values, since there was no objective way to eliminate such events.

Results

Graph 2. shows the individual sentence polarity results with respect to time. As discussed above, each sentence is rated on a scale of -1 to 1, and per article, there were up to 64 sentences. Also, no article below 5 sentences was accepted into the sample due to otherwise resulting in strongly skewed averages.

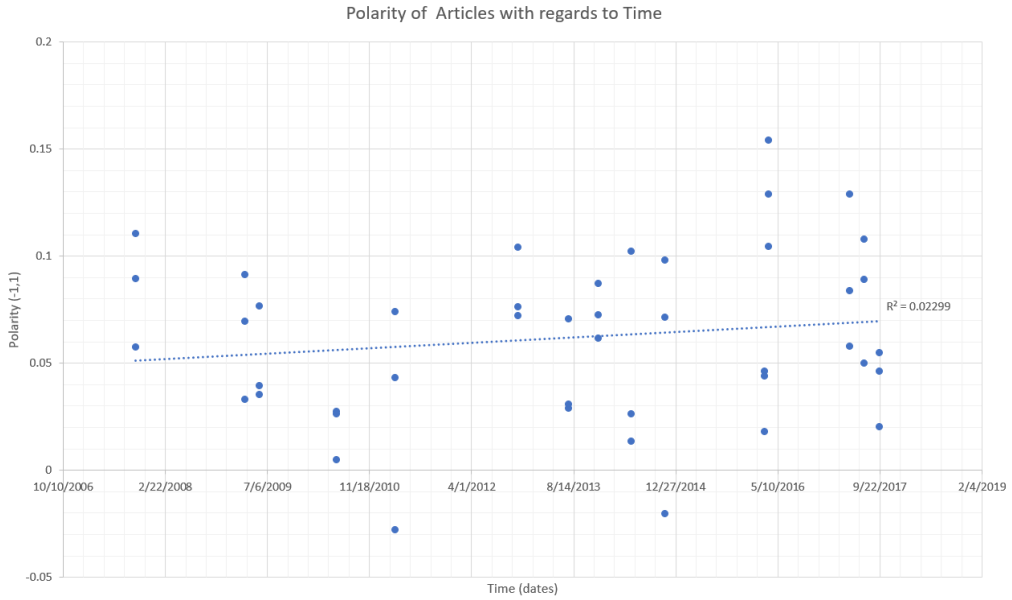
Graph 1. Individual sentence polarity as a function of time, raw data graph



This graph provides a good visualization of the sentiment spread within the articles, but generally only serves as a sentiment data verification tool.

As already noticeable on Graph 1. there is no statistically significant correlation between general article sentiment and the market backlash shift over the studied 10-year time period. A more accurate representation of this is shown in Graph 2.

Graph 2. Average polarity per article as a function of time



While the fitted line on the graph appears positive, the graph still portrays no correlation between time passing and article sentiment, as shown quantitatively by the p-value = 0.32. More details on the fitted line are shown in Table 1.

Table 1

mdlPolarity.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	-0.1475	0.2083	-0.7082	0.4826
2 x1	5.0512e-06	5.0222e-06	1.0058	0.3201

The Polarity times Subjectivity as a function of time graph, albeit being significantly condensed, still bears no correlation with regards to time. As shown in Graph 3. and Table 2.

Graph 3. Polarity multiplied by subjectivity as a function of time

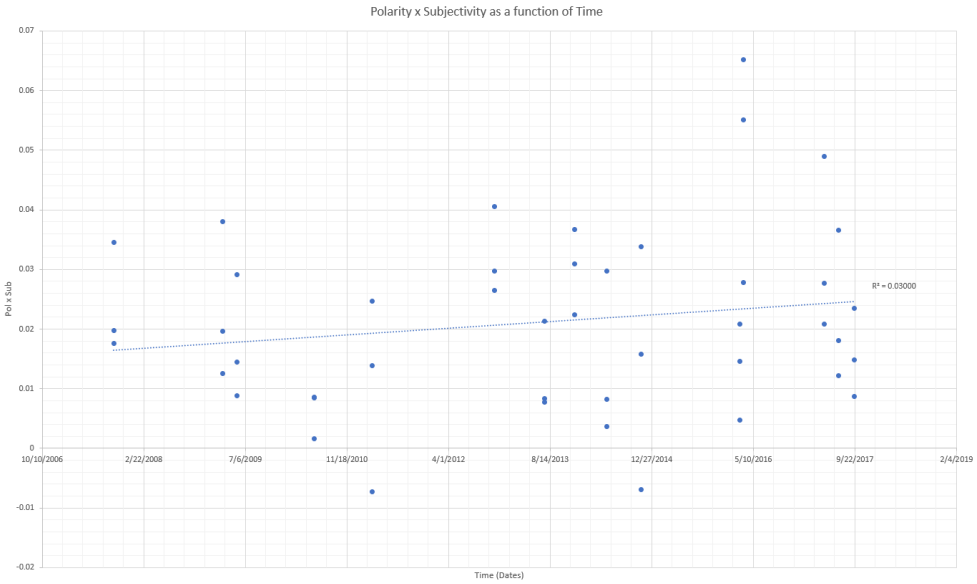


Table 2

mdlPolSub.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	-0.0711	0.0801	-0.8882	0.3794
2 x1	2.2267e-06	1.9309e-06	1.1532	0.2552

By looking at the magnitude of the values shown on the graph it becomes apparent that the graph is condensed. This is to be expected as Subjectivity values are smaller than 1, hence reducing all values (closer to 0). Given that the rate of reduction, however, is not constant, the idea behind the method is that more objective articles will contain less sentimentality hence will be generally higher on the polarity graph. However, if they were to be multiplied by their own Subjectivity measure, which measures 0 if perfectly objective and 1 if subjective, we could theoretically get closer to real sentiment.

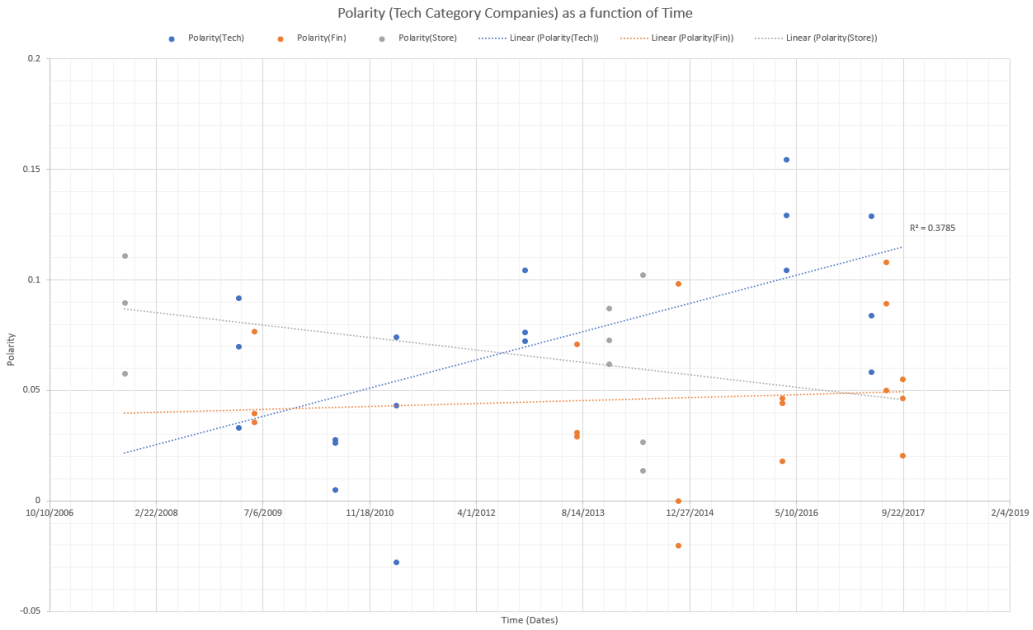
Table 3

mdlTechPol.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	-0.9868	0.3386	-2.9148	0.0101
2 x1	2.5624e-05	8.2084e-06	3.1217	0.0066

Subcategories of the polarity results are graphed on Graph 4. The description of these subcategories can be found in the methodology section's first paragraph. None but the Technological category was found to have a correlation to the shift in market backlash. The values for the equation of the Tech subcategory are in Table 3. By convention, any correlation with a p-value below 0.05 is considered numerically and statistically significant correlation.

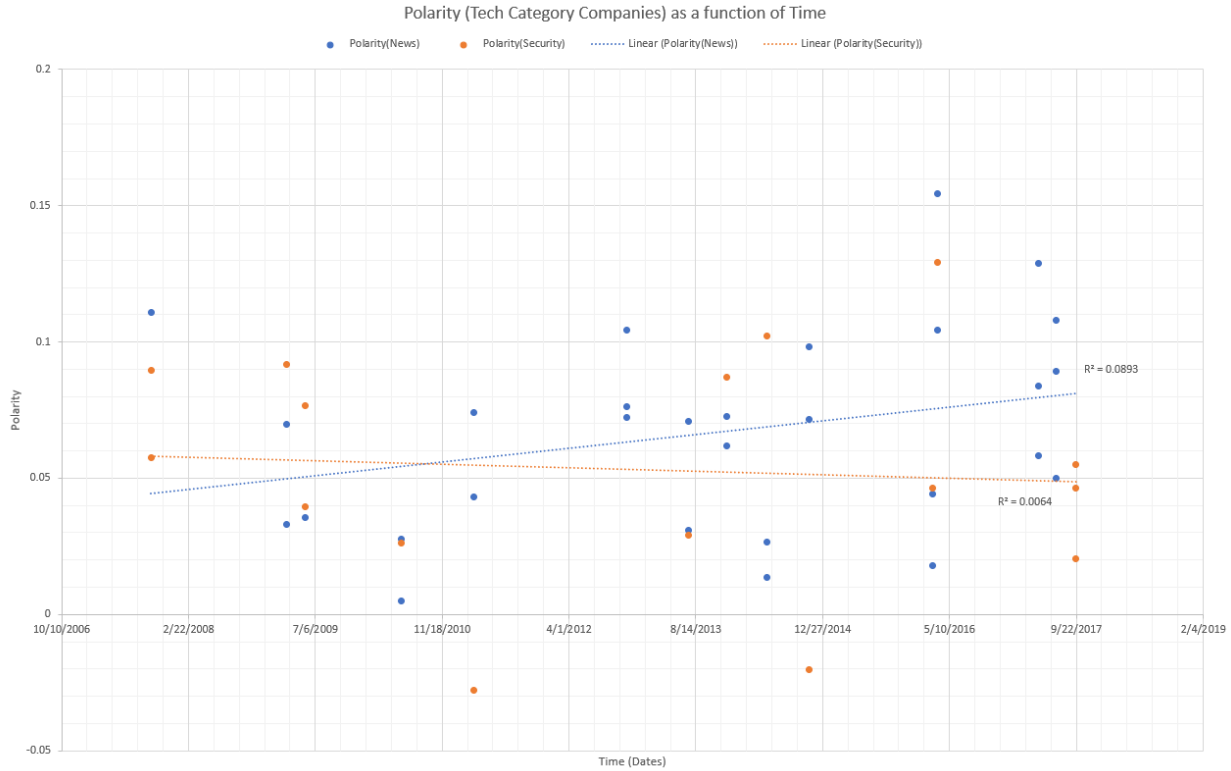
Graph 4. The three selected sub-categories' polarity as a function of time.



As can be seen in Graph 4. While the Financial institutions are also seeming to have a positive correlation, nonetheless, based on a p-value of 0.73 there is no statistically significant correlation.

Finally Graph 5. Showcases the difference between News sources and security blogs/journals/websites in terms of correlation with the financial market.

Graph 5. The Polarity of News articles and Security journal articles as a function of time.



While there is no correlation between polarity and time for either of the two represented distributions, it can be noted that the news distribution portrayed more characteristics of a positive correlation relative to passing of the decade. The difference between them is apparent also when looking at the fitted lines on the graph. The values of the best fit line are shown in Table 4 & 5.

Table 5

mdlNewsPol.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	-0.3546	0.2591	-1.3689	0.1823
2 x1	1.0138e-05	6.2304e-06	1.6272	0.1153

Table 4

mdlSecurityPol.Coefficients

	1 Estimate	2 SE	3 tStat	4 pValue
1 (Intercept)	0.1585	0.3525	0.4497	0.6598
2 x1	-2.5565e-06	8.5394e-06	-0.2994	0.7690

The p-value of News articles is much closer, however still not close enough to have a numerically and statistically justifiable correlation.

Analysis

In terms of the general analysis, no correlation was found between article sentimentality over the years and the decrease of stock backlash on data-breach disclosures. This does not necessarily mean that there is absolutely no correlation between public sentiment and the recorded stock market backlash for data breach incidents. It may very well be that with more data available, the indicated slight slope becomes numerically significant. However, based on the available information so far, this does not seem to be the case. Based on currently available information there is thus no general correlation between public sentiment and data breach stock fluctuations.

This indicates that there have to be other determinants affecting the size of such fluctuations. Further detail on this is discussed in the Conclusion section.

On the other hand, looking at a specific category (Tech), correlation can be found and statistically proven. The results and relatively large variation between categories indicate that public sentiment adapts differently for data-breaches in different field of business. This is a fairly logical assumption, and one which has been shown in the original paper as well, albeit the difference in categories.

Conclusion / Future Steps

As a conclusion of the findings, it should be noted that the lack of general correlation indicates the involvement of other factors besides general public sentiment towards data breaches to be influencing the effect of market backlash.

Such other influential factors could be the increased rate and effectiveness of closing data breaches and isolating (non-isolated) incidents. In the past decade, there has been a significant decrease in company response times and increase in effectiveness of handling such events. [8-9] Furthermore, it should be noted that based on past precedent, companies and stock values have bounced back from such incidents at an overwhelming rate indicating the delayed sell of stock to be futile and unprofitable. [2,3,5,6,7] Hence, investors who miss the instantaneous sell opportunity at the time of the breach, simply rather opt towards keeping their stocks and awaiting the short re-bounce of the stock to avoid unnecessary losses. These possibilities were not evaluated in this paper yet, however, may yield good results in identifying the cause of the shift.

Understanding such underlying causes of stock market behavior is key in long term success in the financial industry and could also help in predicting future market behavior regarding data breaches and other similar short time-span events.

Future steps in the realm of data-breach / stock market event study explanation projects could involve further news article analysis, Twitter sentiment analysis, potential retrieval and analysis of televised news footage, comparison of breach events to similar incidents affecting company stock, etc.

References

[1] Aaron Schumacher. TextBlob Sentiment: Calculating Polarity and Subjectivity. (2015). Retrieved from https://planspace.org/20150607-textblob_sentiment/

[2] Acquisti, A., Friedman, A., and Telang, R. Is There a Cost to Privacy Breaches? An Event Study. (2006.). ICIS 2006 Proceedings. p 94.

[3] Cavusoglu, H., Mishra, B., and Raghunathan, S. The Effect of Internet Security Breach Announcements on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers. (2004.). International Journal of Electronic Commerce (9:1). pp. 70-104.

[4] Devin McGinley. Newspaper Publishing in the US. IBISWorld. (2018). Retrieved from http://clients1.ibisworld.com.ezproxy4.library.arizona.edu/reports/us/industry/default.aspx?entid=1231

[5] Goel, S., and Shawky, H. A. Estimating the Market Impact of Security Breach Announcements on Firm Values. (2009.). Information & Management (46:7). pp. 404-410.

[6] Hashim, M., Khern-am-nuai, W., & Renno, W. Revisiting the Cost of Data Breach Disclosures: A Decade Later. (n.d.).

[7] Kannan, K., Rees, J., and Sridhar, S. Market Reactions to Information Security Breach Announcements: An Empirical Analysis. (2007.), International Journal of Electronic Commerce (12:1), pp. 69-91

[8] Ponemon Institute. 2017 Cost of Data Breach Study. (2017). Retrieved from https://www.ibm.com/downloads/cas/ZYKLN2E3

[9] Ponemon Institute LLC. 2018 Cost of a Data Breach Study: Global Overview. (2018). Retrieved from https://www.ibm.com/downloads/cas/ZYKLN2E3

[10] Shubham Jain. Natural Language Processing for Beginners: Using TextBlob. (2018). Retrieved from https://www.analyticsvidhya.com/blog/2018/02/natural-language-processing-for-beginners-using-textblob/

[11] Sloria. TextBlob: Simplified Text Processing. (n.d.). Retrieved from https://github.com/sloria/TextBlob/tree/eb08c120d364e908646731d60b4e4c6c1712ff63