University Degree in Business Administration
Academic Year 2020-2021

Bachelor Thesis

"Corporate Data Breaches and Narrative Disclosures"

María Aguilera García

Tutor: Encarnación Guillamon Saorin

Getafe 2020-2021

ABSTRACT

In a digitally connected world, data breaches create risks and threats to capital markets and companies. Due to the unregulated structure of the narratives of corporate reports, this thesis intends to analyze public U.S. company's response to data breaches and, in particular, the extent to which these events are reflected in the narratives included in company public disclosures. We focus on the concept of opportunistic managerial discretionary disclosure behaviour that results in biased reporting or "cheap talk". We explore whether there is a change in the subsequent linguistic cues of the 10-K annual accounts. Specifically, we examine the change in tone (optimistic and abnormally optimistic), complexity (readability and length of 10-K) and vague language (use of uncertain and weak modal words). We use data from the Loughran and McDonald (2011) dictionary lists and the Bog Index to construct our linguistic cues. We obtain our sample of data breaches reported to the Privacy Rights Clearinghouse (PRC) organization over the period 2005-2019 and match by name and fiscal year to the Compustat and CRSP databases. We predict that firms use linguistic cues such as optimism to mislead corporate report users about their intentions to address the deficiencies that allowed the data breaches or to distract attention from the data breach itself. We find that breached companies exhibit a change in corporate communications of their annual accounts the year of the breach and the following years of the breach, supporting our hypothesis of breached firms strategically using language by altering the tone and complexity of the narratives of the reports in an attempt to conceal the consequences of data breaches.

Keywords: discretionary accounting narratives, opportunistic managerial behaviour, data breaches, linguistic cues, optimistic tone, vague language, complexity

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The purpose of this project is to focus on the occurrence of data breaches as potential drivers of substantial losses for the company and its stakeholders and examine the way in which they influence corporate communication by investigating whether managers employ opportunistic managerial discretionary disclosure behaviour in the narratives of the 10-K annual reports or whether they provide incremental helpful information aimed at enhancing decision-making by bridging information asymmetries between managers and company outsiders.

"Data breaches have become so pervasive and concerning in recent years that the SEC chair considers them the biggest systemic risk facing U.S. corporations."(Ackerman, Wall Street Journal, 2015).

The extended use of computerised technologies has stressed the importance of cybersecurity as a source of corporate risk (AICPA, 2018). As corporations' exposure to and dependence on interconnected structures have grown, the resultant threats and number of cybersecurity incidents also have grown. We define a data breach incident as "a security violation in which sensitive, protected or confidential data is copied, transmitted, viewed, stolen or used by an authorised individual." (Privacy Rights Clearinghouse 2018).

In the last years, the frequency and severity of breaches have been unprecedented, and breaches have been identified regularly (Islam et al., 2018). Although the fourth industrial revolution has led to an information highway and drastically altered the way in which individuals interact and carry out business (Gordon et al., 2003), this "did not come out without a price- and this price is vulnerability" (Ganslier & Lucyshin, 2005, p.2). The digital age has introduced corporations to a host of novel threats resulting from attacks through automated systems. (Agrafiotis, 2018). For both breached firms and those individuals impacted, the consequences of data breaches are substantial (Kamhoua, 2015). "Firms affected by cyber-attacks tend to suffer economic and reputational losses" (Haapamäki & Sihvonen, 2019, p. 808), and for this reason, they may have incentives to conceal this information.

For example, Uber, the market-leading taxi services app, experienced a massive cyber-attack in 2016 that exposed the data of 57 million customers. The former CEO of Uber, Travis Kalanick, covered it up. A year after it happened, the breach was revealed due to an investigation conducted by media company *Bloomberg*[1]. Equifax, one of the world's largest bureaus, reported a data breach that involved 143 million customers. Equifax found the intrusion on July 29, 2017. Considering its capabilities and the apparent severity of the incident, Equifax made no announcements to the press or its clients until more than forty days after the breach (Rapoport & Andriotis, 2017).

Tangible as well as intangible costs are both possible repercussions of a data breach. These potential costs include (tangible) the costs of investigating and identifying the causes of the data breach, the costs to hire staff to plan and design the new measures to be taken, the costs of restoring /improving information systems, legal costs, and increased cost of borrowing due to reduced credit ratings as well as (intangible) loss of reputation, staff and customer turnover, and increased risk of future attacks (Layton & Watters, 2014). Consequently, "cybersecurity is the top concern for both executives and accounting professionals*"* (AICPA 2015; Protiviti 2016).

In this vein, the COVID-19 pandemic has dramatically modified the way we work, learn, shop and bank. With the governments' attempt to hinder the virus's spread, many people were instructed to stay at home. The change in our day-to-day routines has triggered a sharp decrease

[1] https://www.bloomberg.com/news/articles/2017-11-21/uber-concealed-cyberattack-that-exposed-57-million-people-s-data

in street crimes (Hawdon et al., 2020). At the same time, the transition to the unprecedented digital world created a larger array of opportunities and prospects for cybercriminals. “Risk to sensitive data from both insider and outsider threats has spiked since the WHO declared the COVID-19 outbreak a global pandemic on March 11, 2020”(*The DG Data Trends Report*, 2020). Companies were forced to adapt to the “new normal," and many were unprepared for this transition. Although the global average cost of a data breach declined from 3.92$ (2019) to 3.86$ million in 2020, costs increased for many organisations. Remote work has led to an uptick in the average cost of a U.S. data breach by $136,974, and the “worldwide average” is only $8.64million (Ponemon Institute, 2020).

Unlike the European Union’s General Data Protection Regulation (GDPR), the U.S. does not have a comprehensive federal law addressing the issue of data breach notification. Nevertheless, cybersecurity and its related disclosure have become a significant concern for regulators. In an effort to improve cyber risk disclosures among companies, the U.S. Securities Exchange Commission (SEC) issued guidelines in 2011 and then again in 2018[2]. It also formed a Cyber Unit that pays particular attention to investigating cybersecurity-related delinquencies in 2017 (Securities and Exchange Commission, 2018). Indeed, the SEC has warned investors and other users of corporate information of the risks of cybersecurity[3]:

“As markets grow more global and complex, so too are the threats through cyber intrusion, denial of service attacks, manipulation, misuse by insiders and other cyber misconduct. In the United States, aspects of cybersecurity are the responsibilities of multiple government agencies, including the SEC. Cybersecurity is also a responsibility of every market participant. The SEC is committed to working with federal and local partners, market participants and others to monitor developments and effectively respond to cyber threats."
(https://www.sec.gov/spotlight/cybersecurity)

Beyond regulatory efforts that are intended to increase transparency on the *occurrence* of incidents, it is relevant to investigate the extent to which companies communicate or obfuscate the *significance* and *possible implications* of data breach incidents. Within the framework of accounting, people usually view financial reports as figures. Interestingly, recent research is increasingly paying attention to accounting language “as the medium through which companies communicate to their externalities” (Hales et al., 2010). As the narratives of corporate accounts have become lengthier and more sophisticated, their content and unregulated structure have grabbed the attention of many recent studies.

It is well known that corporate narratives are effective means of disseminating knowledge (Merkley, 2014) with economic ramifications for the company (Huang et al., 2014; Tetlock et al.,2017; Tetlock et al.,2018). Narratives allow preparers of annual reports to communicate firm-specific knowledge to market participants; as stated by Smith and Taffler (2000), the narratives of the annual report supply “almost twice the amount of… information as do the basic financial statements” (Smith & Taffler, 2000, p.624). The increasing prominence of informative parts of corporate documentation enables businesses to address knowledge asymmetries by including

[2] SEC 2018. Commission Statement and Guidance on Public Company Cybersecurity Disclosures. Release Nos. 33-10459; 34-82746. https://www.sec.gov/rules/interp/2018/33-10459.pdf,

[3] While there are no current notification requirements for cybersecurity accidents, this guideline argues that are implicit in existing disclosure requirements. Firms must recognize the significance of these events when filing the disclosure provided under the Securities Act of 1933 and the Securities Exchange Act of 1934, as well as annual and ongoing reporting under the Exchange Act. When a corporation is expected to conform a report statement with the SEC, the required form normally y applies to the disclosure provision of Regulation S-K. Regulation S-K described the form and content of financial statements filed with the SEC. While these transparency provisions to not expressly mention cybersecurity threats and accidents, a variety of them place a duty to report certain risks based on a corporation’s unique circumstances.

more comprehensive information and interpretation, thus increasing their decision judgement. In this respect, (Merkley, 2014) demonstrates that narratives provide accurate evidence material. These disclosures may be insightful into the actual state of a firm's current state. (Grossman, 1981; Milgrom, 1981). Nevertheless, managerial disclosure is a discretional managerial decision, as it typically emphasises positive news in order to boost market values and offers less weight to bad news (Verrecchia, 1983; Dye, 1985).

This work is inspired by the overwhelming prominence and relevance of narrative disclosures in 10-K studies. In fact, in a standard financial statement, the textual narrative accounts for the vast majority of the disclosure- about 80% of the annual report (Li, 2010; Lo et al., 2017), and this qualitative material is critical for understanding a company's present success, historical operations and prospective prospects (Clarkson et al., 1994). Instead of analyzing the MD&A (Management Discussion and Analysis) section of the 10-K annual reports, we analyze the narratives of the entire 10-K reports. As Kothari et al. (2009) point out, the 10-K reports include various sections in which managers may reveal self-serving details (e.g., in the footnotes) in a consistent format throughout the whole report. These are written in a structured debate style of communication that differs significantly from the substance of more informal and interactive discourses such as conference calls. Therefore, rather than investigating disclosures of data breaches on corporate reports, we investigate the effect on linguistic cues in the narratives of 10-K annual accounts, such as optimistic tone, complexity, and vague language, as communication strategies with potential beneficial or harmful effects for shareholders and other financial statements users.

In doing this, we follow and build on the approach taken in prior research, which analyses the annual report to link the linguistic characteristics of their narratives to financial performance or financial constraints (Law & Mills, 2015; Loughran & McDonald, 2016). Three prominent studies in this area are Li (2010) and Huang et al. (2014), and Li (2008). Li (2010) uses a Naïve Bayesian machine learning technique to explore the information content of forward-looking statements in the MD&A section of annual and quarterly filings (10-K and 10-Q), while Huang et al. (2014) demonstrate that optimistic abnormal tone is used strategically in earnings press releases to conceal weak future realizations using Loughran and McDonald Dictionary List. Li (2008) investigates the effect between annual report readability and company financial profitability and earnings persistence and discovers that companies with lower profits have lower annual report readability, while annual reports from firms with better readability are more likely to have consistently decent earnings.

Just as Li (2010), Li (2008) or Huang et al. (2014), we study the association of data breaches with the use of optimistic (abnormal) tone, complexity (length of 10-K) and vague language (uncertain and weak modal words) of 10-K filings of corporate communication. Specifically, this project intends to do the following:

1) Describe the extent of data breach incidents, the characteristics of companies that suffer them, and their effect on firm value;
2) Describe the association between the linguistic cues and financial performance;
3) Describe whether the association between financial performance and qualitative characteristics of the narratives is modified when considering the occurrences of data breaches.

Opportunistic firms may use linguistic cues of narratives such as optimism to mislead corporate report users about their intentions to address the deficiencies that allowed the data breaches or to distract attention from the data breach itself. However, the tone of the corporate narratives may convey private and hard-to-quantify information to truly signal intention to improve the communication with interested parties (Davis et al.,2012; Tetlock et al., 2018). If the

truthful disclosure motive prevails, we expect the narratives' tone not to change significantly before and after the data breach event or even more negative tone after the data breach. However, if the intention is to mislead or distract attention from the data breach, we may observe a significant increase in optimistic (abnormal) tone, a decrease in the vague language (uncertain and weak modal words) and an increase in the complexity (length) of the corporate narratives compared to the year before the occurrence of the data breach.

To assess the disclosure behaviour of U.S. companies that are victims of these attacks caused by cyberattacks, we construct our sample of data breaches disclosed by the Privacy Rights Clearinghouse (PRC) between 2005 and 2019. We use this database because, as opposed to other cybersecurity incidents, the disclosure requirements imposed by breaches that involve the loss of personal information are expected to report in a timely manner in compliance with Security State Breach Notification Laws. We obtain historical financial data, and stock returns data from the Compustat and CRSP database. We then construct our linguistic cues variables using Loughran and McDonald 2011 dictionary list and the Bog Index (Bonsall et al.,2017).

The findings presented in this paper suggest that U.S. firms that experience a data breach use linguistic cues opportunistically to cover up the negative consequences of data breaches. Even though data breaches are usually regarded as largely idiosyncratic events, and with the exception of some cases, unrelated to firm's products and financial condition (Akey et al., 2018), it is always conceivable that latent firm characteristics could be responsible for both the occurrence of data breaches and subsequent losses thereby affecting the disclosure reactions of firms. We, therefore, compare firm and industry characteristics between our sample of data breaches and Compustat. We then turn to a more direct examination by constructing a logistic regression. We find that our sample of breached firms tends to be more visible and growing. While we find a significant effect is in our logistic regressions examining the likelihood of data breaches, we discover that firm characteristics do a very weak job at forecasting data breaches since our R-squared without fixed effects is low, 0.03.

Similarly, one might be concerned that "data-breachers" select companies on the basis of other (possibly unobservable) variables that are correlated with the variables that we study. In untabulated analysis, we perform nearest-neighbour matching (within industry, within time) for each affected firm in un tabulated analysis and confirm that our findings persist even after restricting the sample to only include affected firms with their closest comparisons. The PRC database does not include all data breach reports[4]; therefore, it is plausible that selection bias exists in our sample.

The rest of the paper is organized as follows. Section 2 discusses background on data breaches and financial reporting strategic management literature, culminating in the research hypotheses. In section 3, we describe our sample construction and present the distribution of data breaches. In section 4, we construct tone management proxies and explore their determinants following Huang et al. (2014) and Li (2010). In section 5, we analyze the impact of data breaches on linguistic cues. In section 6, we conclude.

[4] This is because even though the majority of the states enacted the cyberattack notification regulations by 2009, requiring companies in the state to warn impacted citizens, three states (i.e., Alabama, New Mexico, and South Dakota) lacked such legislation during the study span. The U.S. Security Breach Notification Laws and Regulations force publicly traded firms in the U.S. to warn affected individuals about data breaches and disclose the breaches to state governments and other governmental bodies. These are: (i) the "State Security Breach Notification Laws"; (ii) "SEC Cybersecurity Disclosure Guidance"; (iii) "HIPAA Privacy Rule".

2. LITERATURE REVIEW

In order to construct out theoretical predictors, we tap into two strands of literature. First, we review the literature on cybersecurity and data breaches and second, we cover the literature on financial reporting quality and the analysis of narrative linguistic cues.

2.1 Cybersecurity

"Every minute, we are seeing about half a million attack attempts that are happening in cyberspace."(Taylor, 2015)

"There are only two types of companies: Those that have been hacked and those that don't know they have been hacked."(Barnes, 2018)

A few years back, a data breach that compromised the privacy of a hundred million users may have been made of considerable public interest. Now, breaches that concern hundreds of millions of people are all too popular. The above quotes epitomize the modern market climate where "the assumption of a beach is a new norm" (Hayden, 2014) and data breaches are growing "larger in number and impact" (De Groot, 2019).

Recent studies suggest "cybersecurity has grown into one of the most significant risk challenges facing every type of organization and society" (Haapamäki & Sihvonen, 2019, p.808). The American Institute of Certified Public Accountants (AICPA, 2018) stated, "Cybersecurity is one of the top issues on the minds of management and boards in nearly every company in the world- large and small, public and private" (AICPA, 2018, p.1). Gordon et al. (2010) argued that a cybersecurity incident could bring down an entire vital infrastructure sector, jeopardising a country's economy and national defence.

Empirical work has examined corporate disclosure on cyber-attacks and data breaches and has found that firms in critical sectors such as banking, industrial services, insurance, telecommunications, financial services, and health care tend to be more proactive in providing voluntary disclosure of security-related activities (Gordon et al. 2006). Likewise, they argue that regulation impacts a firm's incentives to engage in security technology. For instance, the Sarbanes-Oxley Act of 2002[5]imposed stringent organizations' conditions (Hausken, 2006). Gordon et al. (2006) investigated the effects of voluntary corporate notification of information security practices. The scientific research demonstrated unequivocally that the SOX has a beneficial effect on voluntary transparency.

In this vein, Gordon et al. (2010) examine voluntary disclosures related to cybersecurity. They argue that voluntary disclosures in the annual report on cybersecurity provide signals to the markets that "the firm is actively engaged in preventing, detecting and correcting security breaches" (Gordon et al., 2010, p. 568). Overall, Gordon et al. (2010) provide empirical evidence for the claim that voluntary disclosures about cybersecurity are firmly and substantially linked to stock price.

Although Gordon et al. (2003) suggested that sharing information "has been promoted an important tool in enhancing welfare", they concluded that in the absence of sufficient economic incentives, companies would take advantage of the security expenditure of others. According to Hausken (2007), weighing the costs and advantages of information sharing for a defence scheme is inextricably tied to other tactics for strategic advantage.

[5] The Sarbanes-Oxley Act's primary objective was to reform financial company auditing in the United States, consistent with its complete, legal name: the Public Company Accounting Reform and Investor Protection Act of 2002.

Wang et al. (2013) argue that reporting information on corporate cybersecurity could be positive or negative. On the one hand, Wang et al. (2013) established that companies that share risk-reducing details are less likely to be involved with security incidents. The results suggest that firms taking proactive action have an incentive to disclose their information security truthfully. On the other hand, disclosing information on data security could provide information for competitors and to those interested in attacking the company. For example, Ettredge et al. (2018) examined the relationship between companies disclosing the presence of trade secrets in company reporting and cyber theft of corporate data. The study contributes to the literature by reflecting on breaches that threaten trade secrets and demonstrating that companies that mention the presence of trade secrets have a considerably greater risk of being compromised than firms that do not. Lainhart (2000) asserted that "for many organizations, information and the technology that supports it represent their most valuable assets" (Haapamäki & Sihvonen, 2019, p.808), arguing that successful control of information is crucial in this global information environment in which information flows through cyberspace.

Presently, a wide range of studies have examined the consequences of cybersecurity incidents. For example, Campbell et al. (2003) and Spanos and Angelis (2016) report meaningful negative short-term stock market reactions to corporate data breaches. Kamiya et al. (2018) also examine the cross-sectional effects of unexpected cyberattacks, using a wide range of sorting variables. While numerous papers examine the negative impact of cyberattacks on firm fundamentals, there is a growing interest to study the qualitative aspects of various firm communications with investors as the occurrence and concern of data breaches is continually increasing.

2.2 Financial Reporting Quality and Linguistic Ques

Analogous to mass communication, corporate reporting must satisfy a variety of heterogeneous audience knowledge requirements (Parker, 1982). Accounting is essentially concerned with information. Besides numerical data, company disclosures comprise a substantial number of unstructured textual data. Quantitative information alone provides investors with a partial, imperfect view of a firm's current financial and economic circumstances. However, recent research and reviews have emphasized accounting language as a means by which businesses express their externalities.

Given the increasing relevance and growing length of the narratives of the annual reports, preparers of these are permitted to disclose further detailed information and explanation of events, reducing the knowledge asymmetry that may arise due to weaknesses in existing accounting standards (Clatworthy and Jones 2003; Merkl-Davies and Brennan, 2007). Textual disclosures offer a valuable contest for interpreting financial data and testing economic hypothesis. They contain knowledge about the data generation role of the numeric financial data; therefore, understanding the textual details used in public disclosures is critical for financial accounting analysis (Li, 2010). Examining managers' communication, textual analysis can unveil a variety of managerial traits (Li 2010). As a result, accounting numbers supplemented by accounting narratives can provide a complete picture of a firm's fundamentals.

Two potential factors have driven the heightened awareness of research analysis using textual analysis of qualitative information. First, the availability of unstructured textual data has recently been electronically usable online and public. Second, major advancements in science in the fields of computational linguistic text mining and machine learning over the last two decades have supplied researchers with valuable methods to better interpret corporate disclosures.

Among the wide range of methodologies to quantifying qualitative data are manual-based and computer-based content analysis. The manual approach may result in more precise and

tailored results; however, the sample sizes and costs limit the scope of the empirical results. The computer-based methodology allows for better reproducibility of the empirical findings, and hence longer follow-up studies. Additionally, raising the sample size improves the statistical power for empirical findings. Among the computer-based content analysis methodologies, there are two major approaches: a rule-based ("dictionary") approach and a statistical approach. The dictionary technique makes use of a "mapping algorithm", in which words (or phrases) are classified according to any predefined rules or category (i.e., the dictionary), and the statistical approach is based upon statistical techniques. Examples include the naïve Bayesian algorithm (Li 2010; Huang et al. 2011), psychological dictionaries such as General Inquirer and Diction (Kothari et al.,2009), and financial-customized word lists (Loughran & McDonald 2011; Henry 2008).

Prevailing textual analysis research varies in terms of the means[6] of disclosure, the measure of the qualitative characteristic and the outcomes examined. This paper examines three qualitative characteristics: optimistic tone, complexity, vague language.

"The natural question with respect to corporate textual disclosures is whether they have information content. The alternative hypothesis is that these disclosures are boilerplate generic disclosures and not informative" (SEC, 2003: Bloomfield, 2008).

2.2.1 Tone

It has long been recognised that 'style' plays a critical role in facilitating efficient and convincing dialogue. By word usage, tone may be used to impart a desirable connotation or impact on readers of narratives. As with other stylistic features, the tone may be used to promote clarity and transparency to help in the distribution of incrementally valuable knowledge or the reader's impressions of a narrative's subject matter. Frequently, words with the same literal or dictionary definition will have a distinct connotative interpretation. "Just like brush strokes collectively contribute to the mood of a painting" (Brill, 1992, p. 32), the word choice of narratives can combine to produce one or more dominant tones. Hart et al. (2013) described tone as a "tool" employed to "create distinct social impressions via word choice" (Hart et al., 2013, p. 9).

Thematic deception usually entails changing the tone of narratives to hide negative news and emphasise good news. Numerous previous studies have discovered signs of a negative but mainly positive bias. Instead of complementing accounting numbers, firms could take advantage to obfuscate perceptions using tone, i.e., tone management. Tetlock et al. (2008) investigate the impact of derogatory terms in firm-related news and finds that companies with more pessimistic terms are more likely to report lower profits. Huang et al. (2014) demonstrate that an optimistic abnormal tone is used strategically in earnings press releases to conceal weak future realizations. Huang et al. (2014) define tone management as "the choice of tone level in the qualitative text that is incommensurate with the concurrent quantitative information" (Huang et al., 2014, p. 1083). He deconstructs tone into two dimensions. The normal dimension represents a neutral tone consistent with concurrent knowledge regarding the firm's financial performance's actual and anticipated potential quantitative success. He argues that, although narrative rhetoric is necessary for comprehending quantitative data, where agency motivations are present, the rhetoric can be used opportunistically rather than informatively and thus mislead investors (Huang et al., 2014).

Opportunistic firms may use linguistic cues of narratives such as the employment of positive words, i.e., optimism, to mislead corporate report users about their intentions to address the deficiencies that allowed the data breaches or to distract attention from the data breach itself.

[6] Mandatory filings and disclosures, earning updates and other news releases, conference calls, financial media articles, analyst reviews and research comments, legal announcements, and social networks are among the unstructured data sources analysed.

However, the tone of the corporate narratives may convey private and hard-to-quantify information to truly signal intention to improve the communication with interested parties (Davis et al., 2012; Tetlock et al., 2008). If the truthful disclosure motive prevails, we expect the narratives' tone not to change significantly before and after the data breach event or even more negative tone after the data breach. However, if the intention is to mislead or distract attention from the data breach, we may observe a significant change in the tone (increase in optimism) compared to the year before the occurrence of the data breach. Following this opportunistic use of the tone, we pose the first hypothesis as follows.

H1. *Companies facing a data breach use a more optimistic (abnormal) tone in the narratives of corporate reports.*

2.2.2 Vague Language

Aside from the dichotomy of positive vs negative words, vague language in 10-K filings may well reflect the increase in firm risk due to the data breach occurrence. According to latest studies, the use of weak modal words such as *"might, could, maybe, depending and possible"* indicate a lack of confidence and the list of uncertain words such as "*approximate, assume, contingent, depend, and indefinite"*, expresses imprecision. Recent findings have shown that vague company disclosure texts influence assessment confusion. Guo et al. (2017) reveal that companies facing increased competition use a more ambiguous tone in their corporate reports as a possible countermeasure against hostile takeover attempts.

The use of uncertainty in financial statements may affect whether users of such reports fully understand them and whether they can make informed decisions. Prior research has investigated the uncertainty embedded in the narratives of corporate reports (Law & Mills, 2015; Loughran & McDonald, 2011; Loughran & McDonald, 2016). In this line, Guo et al. (2017) find that firms with risk of competition include more vagueness in their annual reports. In this study, we investigate whether companies that face a data breach reflect this negative event in the narratives of the 10-K providing additional information to users or, on the contrary, companies hide this uncertainty related to the data breach and to the potential negative consequences. We pose our second hypothesis as follows.

H2. *Companies facing a data breach use more vague language (uncertainty and weak modal words) in their annual corporate reports.*

2.2.3 Readability

The study of textual complexity is inspired by its direct association with communicative efficacy. Since the Securities Act of 1933 was enacted, regulators such as the SEC has made continuous attempts to improve the readability of public company information statements (Firtel, 1999) to safeguard stakeholders' interests and have also issued recommendations on improving the readability of public disclosures). The SEC's simple English transparency guidelines, enacted on January 22, 1998, was the most current of these attempts. The fundamental rationale for plain English transparency legislation is that (1) a company may use ambiguous terminology and style in disclosure to conceal negative facts, and (2) ordinary investors may not comprehend complex papers, resulting in stock market inefficiency.

Following the incomplete revelation hypothesis, derived from Bloomfield (2002), management opportunism theory states that managers have a greater thrust to conceal facts, while current success is poor (Bloomfield, 2002). Notably, research indicates that the narrative disclosures in 10-K filings are nuanced and challenging to interpret and comprehend, and managers have strategically employed disclosure readability to cover up financial information, placing knowledge processing costs on consumers and contributing to markets responding less fully to information found in narratives (Lehavy et al., 2011; Li, 2010)

According to several studies(e.g., Bonsall et al., 2017; Li, 2008), U.S. companies have low readable narratives in their 10-K. Others claim that companies intentionally have less readable 10-K records in order to conceal negative facts and results. Lo et al. (2017) investigate how annual report readability is viewed in relation to earnings management and also documents that firms with greater propensity to manage profits in order to compensate earnings from the previous year have less readable financial disclosures. Li (2008) explores annual report readability in the MD&A section and discovers that companies with lower and less persistent earnings report fillings are more difficult to read. Remarkably, he proposes a positive correlation between linguistic features of annual reports and company financial results. Previous research has shown the capacity to manipulate readability to obscure "negative news" through strategic changes in textual sophistication and promote "positive news" through simple and straightforward vocabulary.

We use two measures of text readability. The first one focuses on the premise that longer reports are more deterrent and necessitate higher knowledge costs (Bonsall & Miller, 2017; Bonsall, Leone et al., 2017). The second one relates to the fact that companies may use the complexity of their corporate reports to obfuscate or hide information about data breaches from their intended audience. Overall, prior research finds that companies use the complexity of corporate reports to “camouflage” bad news. We expect the narratives of the 10-K to reflect the intention of the company to obfuscate the negative consequences of data breaches. We pose our third hypotheses as follows.

H3. *Companies facing a data breach use more complex language (longer documents) in their annual corporate reports.*

These arguments and evidence found in previous research led us to wonder if companies consider that providing information about data breaches is positive for the company and that stakeholders appreciate this information, learn and value it. If companies report information about data breaches, this reflects their awareness of the vulnerabilities and that they are taking actions to prevent potential attacks. We are interested in assessing whether the content of the narratives disclosed by companies who have faced a data breach change after the event, which will indicate that these narratives are informative and contain information about the data breach. To this end, we investigate companies that suffer these attacks and compare their corporate reports with those of companies that did not suffer the attacks to assess the differences in disclosure behaviour. We also intend to compare the information content of corporate narratives before and after the data breach event occur for the same company to find out how the characteristics of these narratives are associated with the occurrence of a data breach.

3. DATA SOURCES AND DESCRIPTIVE STATISTICS

3.1 Corporate Data Breaches

To analyze public U.S. company's response to data breaches, we obtain our sample of data breaches from the Privacy Rights Clearinghouse (PRC)[7]. We use this database because the disclosure requirements imposed by breaches that involve the loss of personal information are expected to report in a timely manner in compliance with the Security State Breach Notification Laws. Even though it is probable that a breached firm in our sample withheld details and postponed public disclosure, the disclosure requirements help mitigate potential sample underreporting biases that may arise in other studies without such reporting requirements.

We begin with a total of 9,015 data breaches that were reported to the PRC database between 2005 and 2019. We then restrict the sample to incidents involving the compromise of more than 1,000 records (Akey et al., 2020), totalling 4,400 incidents. While this restriction reduces our sample size, it ensures that we have a representative sample of data breaches, thereby eliminating the risk that our results are influenced by outliers that are not representative of the population. Additionally, we exclude private businesses such as government, non-profit or educational institutions. We then match organization names with names listed in Compustat using Excel Fuzzy Lookup Add-in. We face several limitations, including:

i) *Name changes:* while company names in the PRC database contain the company's name at the time of publication, companies in Compustat appear by their most current name with no records of previous names[8].

ii) *Ownership structure:* a majority-owned subsidiary and a holding company may comprise numerous identification numbers and records within Compustat.

iii) *Changes in ownership:* ownership of a firm may change throughout our sample due to mergers, acquisitions, and spinoffs[9].

iv) *Compustat unique company identifier over time:* Compustat uses GVKEY to track companies over time. A single GVKEY may correspond to multiple GVKEYs within the Compustat database.

Considering that about 40% of firms in Compustat change their names at least once, extensive manual checks were required. First, we searched all companies on the "Securities Exchange Commission" website. Second, to avoid including years of data submitted by the focal company before the firm became publicly available, we defined an active record as a year with positive common shares traded. Third, when breached corporations are unlisted subsidiaries of listed corporations, we matched these as having occurred in their listed parent corporations. Lastly, although Compustat contains the most current name, we identified that we could match our data to the CRSP database to check proper historical name matching. Our final sample yielded 645 breach incidents and 511 unique firm-year breaches. If a firm experience more than one attack in a given year, we treat them as a single incident. Our sample includes 359 unique firms attacked in 15 years. Table 1 summarizes those that matched with Compustat, and we present breach statistics of our initial sample distribution. Our sample of breached firms varies over our subsequent analysis.

[7] Privacy Rights Clearinghouse | Privacy Rights Clearinghouse

[8] Therefore, many studies would consider ("CONM") the name that appears for each record in the most recent Compustat file.

[9] After an M&A, companies typically stop being traded independently, and therefore should be searched by their existing owner.

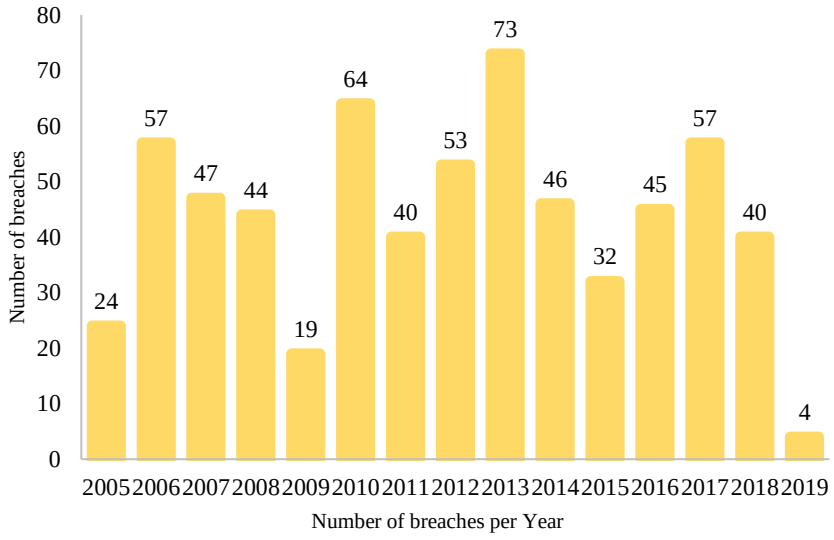
Table 1. Summary Breach Sample Construction

	Observations
Total breaches reported by PRC 2015-2019	9,015
Excluded observations:	
Breaches <1,000 records compromised	4,401
Private companies	3,837
Breaches matched with Compustat-CRSP	652
Firm-year breaches in Compustat-CRSP	511

Figure 1. presents the frequency of breaches over time. There are only four data breaches in 2019 because the PRC database only reports data breaches in the year's first month. We do not find a generally increasing trend in our sample. In 2009, according to the Digital Guardian, "the number of total data breaches exposed in the U.S. dropped to 498, from 656 in 2008.However, the total number of exposed individual records increased sharply[10]", signifying that although there were fewer breaches, individual breaches were larger. We also see a sharp decline from 2010 to 2011 and then from 2014 to 2015, demonstrating that our sample selection should be expanded to reflect better the last years of our sample. As mentioned earlier, there are 642 data breach incidents and 511 firm-year breached firms. Figure 2 shows that 242 companies had only one breach in a fiscal year, while the rest were involved in multiple breaches in the same fiscal year.

Table 2 presents the distribution of the 642 data breach incidents in our sample by industry (Fama French 12-industry classification). Prior analysis has shown that data breaches occur more frequently in banks, retailing and internet services company. In our sample, we find that data breaches occur most frequently in finance (38%), business equipment (16%) and wholesale & retail (11%), suggesting that businesses with a huge number of clients are more vulnerable to data breaches.

Figure 1. Frequency of breaches over the years



[10] The History of Data Breaches | Digital Guardian

Figure 2. Breaches by firm in a fiscal year

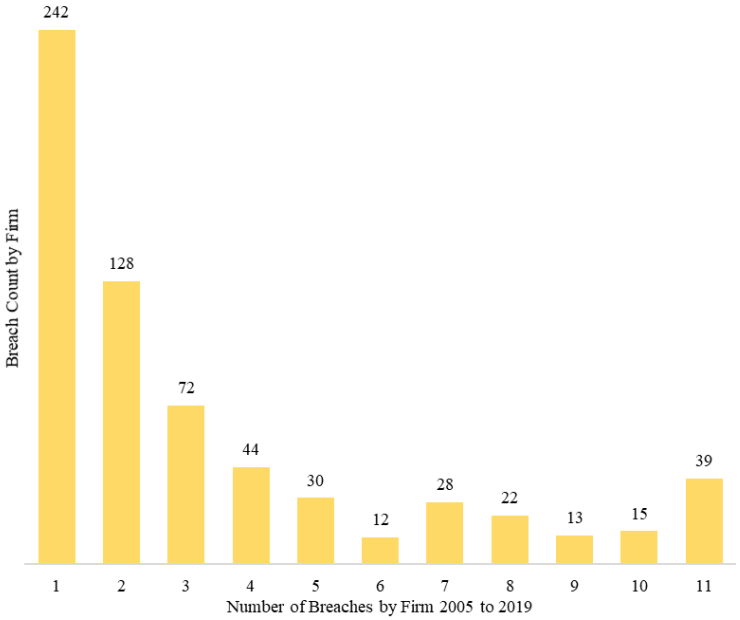


Table 2. Breaches by Industry

Fama French Industry Categories		Number of breaches	Per cent of Total
1.	Consumer Nondurables	23	4%
2.	Consumer Durables	6	1%
3.	Manufacturing	22	3%
4.	Oil, Gas, Coal Extraction and Products	2	0%
5.	Chemicals and Allied Products	6	1%
6.	Business Equipment	105	16%
7.	Telephone and Television Transmission	31	5%
8.	Utilities	8	1%
9.	Wholesale, Retail, and Some Services	72	11%
10.	Healthcare, Medical Equipment, and Drugs	58	9%
11.	Finance	243	38%
12.	Other	66	10%
		642	100%

3.2 Data Breach Characteristics

To study firm characteristics that drive data breaches[11,] we begin by comparing the characteristics of breached firms, which we refer to as "*Breached*", with those that have not been breached, which we refer to as "*Control*". When a firm undergoes several data breaches in a fiscal year, these will be treated as a single breach, so the sample reduces from 645 data breaches to 511 data breaches. We then eliminate those without sufficient available information. Table 3 presents summary statistics for 372 firm-year observations compared to 29,190 firm-year observations of Compustat that were not breached the following year over the period 2005-2019. We reduced our control sample by eliminating companies not in the same industry (4-digit SIC code) as those in our sample of breached firms. This eliminated 42% of our control sample. Each year, all continuous variables are winsorized at the 1% and 99% percentiles.

By examining firm-level characteristics, we discover that breached firms are generally bigger, older, higher valued, and have a larger presence among Fortune 500 companies. These results indicate that compromised firms in our study are more noticeable than non-compromised firms. We also find that our *breached* sample generally have greater current performance *(ROA)*,

[11] We follow (Kamiya et al., 2021, p. 744) to choose our variables to compare our breached and control sample.

are less likely to incur negative earnings *(LOSS)*, have higher growth opportunities (*Tobin's Q* and *Growth*), are less financially constrained *(WWindex and Leverage)*, invest less in R&D, and have a greater number of business and geographical segments *(BUSSEG and GEOSEG)*.

Table 3. Summary Statistics before Data Breach

Characteristic	Control Firm-years no breach following year, N = 29,109		Breached Firm-years followed by a breach, N = 372		Test indifference Breached - Control	
Variable	Median	Mean	Median	Mean	Median	Mean
ln(assets)	2.939	2.961	3.987	4.0140	1.048***	1.053***
Firm Age	13.000	17.608	19.000	24.470	6.00***	6.862***
BIG4 =1 (%)	67.00%	67.00%	95.00%	95.00%	28.00%***	0.280***
Foreign =1 (%)	21.00%	21.00%	24.00%	24.00%	3.00%	3.00%
TobinsQ	1.309	1.884	1.452	1.811	0.143***	-0.073**
$TobinsQ_{t-1}$	1.333	1.960	1.455	1.844	0.122**	-0.116**
ROA	0.013	-0.013	0.040	0.040	0.027***	0.053***
LOSS =1 (%)	27.00%	27.00%	11.00%	11.00%	-16.00%***	-0.160***
Growth	0.068	0.127	0.075	0.139	0.007**	0.012**
RET	0.060	-0.002	0.097	0.008	0.037	0.0100
STD_RET	0.326	0.392	0.245	0.288	-0.081***	-0.104**
Leverage	0.071	0.149	0.169	0.215	0.098***	0.066***
R&D/assets	0.000	0.045	0.000	0.014	0.00***	-0.031***
CAPX	0.016	0.030	0.023	0.030	0.007***	0.00***
Asset Intangibility	-0.051	-0.146	-0.076	-0.159	-0.025***	-0.013***
Fortune500 =1 (%)	2.90%	2.90%	14.00%	14.00%	11.10%***	0.111***
GEOSEG	0.000	0.276	0.301	0.351	0.301***	0.075***
BUSSEG	0.000	0.267	0.477	0.452	0.477***	0.185***
Extraordinary items =1 (%)	1.00%	1.00%	2.40%	2.40%	1.4%**	0.014***
BTM	0.771	0.722	0.693	0.690	-0.078***	-0.032***
SIZE	2.705	2.763	3.947	3.877	1.242***	1.114***
TACC	-0.037	-0.055	-0.047	-0.055	-0.01**	0.00**
M&A =1 (%)	6.60%	6.60%	6.70%	6.70%	0.10%	0.10%
Special Items	-0.000	-0.012	-0.002	-0.009	-0.002***	-0.003***
Wwindex	-0.328	-0.333	-0.442	-0.437	-0.114***	0.219***
Credit Rating						
C	25.00%	25.00%	10.00%	10.00%	-15.00%	-15.00%
B	51.00%	51.00%	57.00%	57.00%	6.00%	6.00%
A	12.00%	12.00%	26.00%	26.00%	14.00%	14.00%
Missing	12.00%	12.00%	6.70%	6.70%	-5.00%	-5.00%

Notes: Mean and median of 372 (251 unique firms) firm-year observations that undergo a breach in the next fiscal year and 29,109 (4,168 unique firms) firm-year observations that did not suffer a breach the following year during the time frame of 2005-2019. The control group consists of all firms in industries that at some point suffered a data breach. The construction of the variables is described in detail in Appendix A Table A.1. ***,**,* denote the t-tests (Wilcoxon z-tests) for the mean (median) differences in firm and industry characteristics between attacked and non-attacked firms at the 1%,5% and 10% levels, respectively.

We then proceed by conducting a logit regression to analyse the likelihood of firms being victims of a breach that compromises more than 1,000 records (*Large Breach*). We use the same sample used for our analysis in Table 3. Table 4 presents the results of our logistic regressions. The dependant variable *"Large Breach"* indicates whether a firm was breached with over 1,000 records in a given fiscal year and 0 otherwise. The independent variables are measured one year before the occurrence of the data breach incidents except for Tobin's Q, which is calculated two years before a data breach due to its high association with historical stock returns. Standard errors are clustered at the firm level and adjusted for heteroskedasticity using the Huber-White procedure. Controls of columns (1), (2), (3) were selected following Kamiya et al. (2021, p. 744), and in columns (4), (5), (6), we add additional controls to capture other factors used in other research papers (Xu et al., 2019, p. 271), (Higgs et al., 2016, p. 86).

On the whole, all of our regressions indicate that firms with greater recognition and market presence *(In(assets), Fortune500, BUSSEG)*, greater profile and current profitability (*Tobin's Q, ROA, Growth)*, greater reputation (*Asset intangibility)*, and higher financially constrained firms (*Leverage, WWindex)* that invest less in R&D and have higher capital expenditures *(CAPX)* are more likely to be victims of a "*Large Breach"*.

In column (1) and (4), results indicate that firms with lower volatility of stock returns have a higher chance of being breached. In columns (2) and (5), we include year fixed effects to control for unobserved variables that evolver over time and are constant across firms. Results seem to corroborate our overall conclusions. In regressions (3) and (6), we include industry-by-year fixed effects to study the differences within a given industry within a given year. The loss of significance in some coefficients is due to the inclusion of the fixed effects. In columns (1) and (4), where we include models without fixed effects, the adjusted R-squared is only 0.03, suggesting that observable firm characteristics are somehow ineffective at forecasting data breaches.

Table 4. Likelihood of a Data Breach

	(1) LARGE BREACH	(1) LARGE BREACH	(3) LARGE BREACH	(4) LARGE BREACH	(5) LARGE BREACH	(6) LARGE BREACH
(Intercept)	-8.933***			-8.787***		
	(0.3348)			(0.3687)		
In(assets)	2.122***	2.146***	2.160***	1.775***	1.807***	1.905***
	(0.1693)	(0.1682)	(0.2145)	(0.1956)	(0.1959)	(0.2482)
In(firm age)	-0.0093	-0.0056	0.0257	-0.0770	-0.0809	-0.0432
	(0.0585)	(0.0594)	(0.0682)	(0.0574)	(0.0582)	(0.0674)
$TobinsQ_{t-1}$	0.1163**	0.1473***	0.1063*			
	(0.0428)	(0.0414)	(0.0542)			
ROA	1.940*	1.901*	0.8860	2.305*	2.526*	1.437
	(1.053)	(1.026)	(1.088)	(1.049)	(1.092)	(1.183)
Growth	0.6459**	0.6723***	0.8097***			
	(0.1972)	(0.1871)	(0.2041)			
RET	-0.1527	-0.1535	-0.1410	-0.2118.	-0.2014	-0.1634
	(0.1249)	(0.1435)	(0.1589)	(0.1270)	(0.1477)	(0.1638)
STD_RET	-0.6534*	-0.5202	-0.7469.	-0.6380*	-0.4883	-0.5524
	(0.2999)	(0.3381)	(0.3937)	(0.3132)	(0.3489)	(0.3901)
Leverage	1.023***	1.126***	0.7247.	0.7436*	0.8605**	0.7080.
	(0.2709)	(0.2761)	(0.4043)	(0.2972)	(0.3020)	(0.4120)
Wwindex	7.799***	7.872***	7.000***	6.394***	6.483***	5.301*
	(1.502)	(1.483)	(1.937)	(1.703)	(1.693)	(2.300)
R&D/assets	-5.443***	-5.568***	-5.380*	-5.020**	-4.838**	-3.392
	(1.653)	(1.627)	(2.223)	(1.636)	(1.657)	(2.095)
CAPX/assets	4.940**	4.939**	2.324	5.394**	5.753**	3.700.
	(1.702)	(1.749)	(2.207)	(1.755)	(1.790)	(2.181)
Asset	1.508***	1.568***	0.4338	1.694***	1.774***	0.6260
intangibility	(0.3828)	(0.3937)	(0.6453)	(0.3864)	(0.3984)	(0.6667)
Fortune500	0.4274**	0.6288**	0.5391*	0.5304**	0.6345**	0.5795*
	(0.1646)	(0.2372)	(0.2744)	(0.1636)	(0.2360)	(0.2756)
BIG4				1.065***	0.9519***	0.4160
				(0.2512)	(0.2547)	(0.2781)
BUSSEG				0.3829**	0.5198***	0.3018.
				(0.1284)	(0.1373)	(0.1711)
M&A				0.1287	0.1443	-0.0489
				(0.2189)	(0.2206)	(0.2452)
Fyear FE	No	Yes	No	No	Yes	No
Fyear x SIC2 FE	No	No	Yes	No	No	Yes
Observations	29,481	29,481	17,047	29,481	29,481	17,047
Squared Cor.	0.03316	0.03693	0.10527	0.03400	0.03798	0.10540
Pseudo R2	0.14900	0.16733	0.25439	0.15364	0.17067	0.25053
BIC	3,541.7	3,612.6	4,754.0	3,533.5	3,609.6	4,777.5

Notes: Table presents our logistic regressions to analyse the likelihood of firms being breached, where the dependant variable, "Large Breach", is an indicator that takes the value of one if a firm experiences a breach in a given year and the breach compromised over 1,000 records. The control group consists of all firms in industries that at some point suffered a data breach. The sample consists of 29,481 firm-year observations covered in Compustat over the period 2005 to 2019. All explanatory variables are measured one year before the breach except for Tobin's q that is measured two years before the breach. The construction of the variables is described in detail in Appendix A Table A.1. Standard errors are reported in parenthesis and are adjusted for heteroskedasticity and clustering at the firm level using the Huber-White procedure. ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

3.3 Data Breaches and Firm Value

We analyze the extent to which data breaches affect firm value by analysing how a firm's market-to-book ratio (MTB) change in the years following a data breach. To better understand the key drivers behind firm value changes, we decompose the market-to-book (M/B) ratio into

return on equity (ROE) and the price-to-earnings ratio (P/E). Changes in ROE capture how the data breach impacts the firm's current performance, while changes in P/E ratios capture how market participants view the impact of the data breaches on the firm's longer-term growth, opportunities, and growth options. Hence, this decomposition allows us to examine whether value changes are primarily driven by changes in short-run firm profitability or changes in long-term market sentiment and expectations. Details on the variable construction are included in Appendix A, Table A.2.

We build an annual panel of all firms from 2000 to 2020. As discussed earlier, we restrict our sample to firms in 4 digit-SIC industries that have at some point experienced a data breach. This restriction eliminates 42% of firms of our control sample Compustat. In order to study the firms' reactions following data breaches, we follow Akey et al., 2018's methodology. We include two specifications.

$$Y_{ijt} = \alpha + \gamma Post_{ijt} + \beta x_{ijt} + f_{it} + f_i + \varepsilon_{ijt} \quad (1)$$

$$Y_{ijt} = \alpha + \gamma Post_{ijt} + \delta Treated + \beta x_{ijt} + f_{it} + \varepsilon_{ijt} \quad (2)$$

We construct an indicator variable, *Post*, that identifies firm-year observations following the disclosure of a breach. We use three different definitions of *Post* to capture responses over different time periods. Our first definition, *Post0-1*, includes the year of the data breach, together with the subsequent year. Hence, *Post0-1* is equal to 0 for all firms that were never subject to a data breach as well as firms that did not experience a data breach within the previous two years. Our second and third definitions include the breach year plus two years *(Post0-2)* and three years following the breach *(Post0-3)*.

In our first specification, the inclusion of firm fixed effects ensures that our identification controls for any time-invariant characteristics that may differ across affected and unaffected firms. However, "by using fixed-effects models, researchers make not only a methodological choice but a substantive one" (Bell and Jones 2015). This is because we are estimating extra N parameters by the inclusion of the firm fixed effect, thereby reducing the variation and finally the precision in our estimation. Therefore, our main specification only includes industry-by-year fixed effects. The inclusion of industry-by-year fixed effects ensures that our comparisons are within industry, within a given year, between affected and unaffected firms. In this model, we include the variable *Treated*, which identifies whether a firm has even been subject to our data breach in our sample. All other variables are defined as stated earlier.

Our identifying assumption is that data breaches constitute an exogenous negative shock to a firm's reputation (Akey et al., 2018). Specifically, it can be said that a considerable majority are associated with a firm's products or services. Conversely, they may lure negative attention to the firm and may influence the firm's stakeholders. For example, consumers' credit card numbers, passwords or personal information are exceptionally valuable to hackers but do not directly affect the products or services.

Additionally, another concern could be that those firms subject to data breaches differ from unaffected firms, potentially in unobservable ways. For example, firms that under-invest in research and development expenses may also under-invest in other areas. However, this rationale is unlikely, as explained by Malcolm Marshall, KPMG's Global Head of Cyber Security, July 2015: "Any CEO who understands risk knows that cyber is possibly the most unpredictable risk there is. It's more unpredictable than a flood or tornado"; or Erik Avakian, Chief Information Security Officer, Commonwealth of Pennsylvania, USA: "JP Morgan is a company that has 2,000 people dedicated to cybersecurity. They have spent $250 million dedicated to cybersecurity. They did everything right, and they still got hacked."

Nevertheless, we handle this concern by employing a within-firm variation approach. Including year fixed effects, we ensure that we compare years following a data breach to the same firm at a different point in time. We do not exclude firms that were never hacked so that we can better assess the year-by-industry fixed effects. In this specification, the necessary identifying assumption is that there are no omitted-time varying firm characteristics that covary with the probability of a data breach. Some firms may be more vulnerable to data breaches, but we would not expect their data vulnerabilities to vary over time in a predictable way. We believe this assumption is plausible, particularly given that a firm's information technology infrastructure is difficult to change and requires long-term investment.

Lastly, we restrict our sample of data breaches to those that have at least compromised 1,000 records. Many data breaches do not have the records compromised recorded; therefore, this limitation reduces our sample size by approximately 50%. However, this allows us to focus on breaches that are arguably the most similar in nature.

Table 5 presents the results of the analysis. Panel A examines how a firm's M/B, ROE, and P/E change in the two years following a data breach, while Panel B presents the results for the three years following a data breach and Panel C following a data breach. Columns (1) – (2) of the three panels study changes in M/B, columns (3)- (4) study changes in ROE and columns (5)-(6) study changes in P/E. The control group in all tests consists of all firms in industries that at some point suffered a data breach. All columns contain industry-by-year fixed effects, and columns (2), (4), and (6) contain firm fixed effects. Finally, all columns include controls for *In(assets)*, *In(assets²)*, and the firm's *Market Leverage*.

We retrieve clear evidence that *long-term* value declines following unexpected data breaches. In the two years following a data breach, breached firm's M/B declines by nearly 0.54 units (which is nearly 10% of our sample standard deviation) relative to unaffected firms in the same industries. We also find that data breaches negatively impact both firms' current profitability and firm's expected growth opportunities. For example, in the three years following the event, P/E ratios decline by -2.6 to 3.1. ROE declines 1% to 4% in the four years following an event. Although the ROE results are statistically mixed. The economic magnitudes of these results are substantial: the sample standard deviation of ROE is 0.57, suggesting that the across-firm, four-year coefficient is 2.2%-8% of a standard deviation.

Table 5. Effect of Data Breaches on Firm Performance

	Panel A: Years 0-1					
	(1)	(2)	(5)	(6)	(7)	(8)
Dependent Var.:	M/B	M/B	ROE	ROE	P/E	P/E
Post0	-0.5794***	-0.5367***	0.0034	-0.0235.	-1.496.	-1.999*
	(0.1416)	(0.1590)	(0.0136)	(0.0133)	(0.8913)	(0.9851)
ln(assets)	1.742***	1.574***	-0.1049***	-0.1580***	3.328***	9.066***
	(0.1997)	(0.0927)	(0.0252)	(0.0108)	(0.6044)	(0.3485)
In(assets) ^{2}	-0.1679***	-0.2000***	0.0241***	0.0412***	-0.0220	-0.6216***
	(0.0402)	(0.0180)	(0.0044)	(0.0020)	(0.1461)	(0.0775)
Market Leverage	-3.788***	-4.232***	-0.2113***	-0.1866***	-10.06***	-10.95***
	(0.1604)	(0.1270)	(0.0211)	(0.0142)	(0.6948)	(0.5116)
Treated		0.4975***		0.0109		1.850*
		(0.1330)		(0.0110)		(0.7194)
Firm FE	Yes	No	Yes	No	Yes	No
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	81,670	81,670	81,548	81,548	83,319	83,319
R2	0.38382	0.09407	0.35301	0.06562	0.38149	0.13012
Within R2	81,670	81,670	81,548	81,548	83,319	83,319

Panel B: Years 0-2						
	(1)	(2)	(5)	(6)	(7)	(8)
Dependent Var.:	M/B	M/B	ROE	ROE	P/E	P/E
Post0-2	-0.5294***	-0.5062**	-0.0015	-0.0315*	-2.602**	-3.114**
	(0.1596)	(0.1674)	(0.0126)	(0.0124)	(0.9513)	(1.007)
ln(assets)	1.740***	1.573***	-0.1049***	-0.1580***	3.312***	9.058***
	(0.1997)	(0.0927)	(0.0252)	(0.0108)	(0.6035)	(0.3485)
$\ln(\text{assets})^2$	-0.1671***	-0.1998***	0.0242***	0.0413***	-0.0142	-0.6194***
	(0.0402)	(0.0180)	(0.0044)	(0.0020)	(0.1459)	(0.0775)
Market Leverage	-3.787***	-4.232***	-0.2113***	-0.1865***	-10.05***	-10.94***
	(0.1604)	(0.1270)	(0.0211)	(0.0142)	(0.6945)	(0.5115)
Treated		0.5242***		0.0141		2.213**
		(0.1342)		(0.0112)		(0.7394)
Firm FE	Yes	No	Yes	No	Yes	No
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	81,670	81,670	81,548	81,548	83,319	83,319
R2	0.38383	0.09408	0.35301	0.06564	0.38155	0.13021
Within R2	0.01565	0.04684	0.00439	0.01979	0.00474	0.06524

Panel C: Years 0-3						
	(1)	(2)	(5)	(6)	(7)	(8)
Dependent Var.:	M/B	M/B	ROE	ROE	P/E	P/E
Post0-3	-0.4767**	-0.4742**	-0.0124	-0.0433***	-3.097***	-3.527***
	(0.1593)	(0.1664)	(0.0129)	(0.0125)	(0.9237)	(0.9679)
ln(assets)	1.739***	1.572***	-0.1050***	-0.1582***	3.295***	9.049***
	(0.1997)	(0.0927)	(0.0252)	(0.0108)	(0.6030)	(0.3485)
$\ln(\text{assets})^2$	-0.1666***	-0.1996***	0.0242***	0.0413***	-0.0070	-0.6172***
	(0.0402)	(0.0180)	(0.0044)	(0.0020)	(0.1457)	(0.0775)
Market Leverage	-3.787***	-4.231***	-0.2112***	-0.1864***	-10.05***	-10.93***
	(0.1604)	(0.1270)	(0.0211)	(0.0142)	(0.6944)	(0.5115)
Treated		0.5417***		0.0188.		2.483**
		(0.1345)		(0.0114)		(0.7587)
Firm FE	Yes	No	Yes	No	Yes	No
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	81,670	81,670	81,548	81,548	83,319	83,319
R2	0.38382	0.09408	0.35302	0.06567	0.38160	0.13027
Within R2	0.01564	0.04684	0.00440	0.01983	0.00482	0.06531

Panel D: Summary Statistics for Firm Value Analysis									
Variable	N	Mean	SD	MIN	MAX	SE	Q0.25	Q0.5	Q0.75
Market Leverage	83697	0.2623	0.2627	0	0.9183	0.0009	0.0125	0.1823	0.4529
ROE	90107	0.0286	0.5721	-2.4688	2.9139	0.0019	-0.0687	0.0753	0.1693
P/E	83646	8.9127	27.2739	-121.5945	104.2136	0.0943	-3.3581	10.1856	20.1869
M/B	81993	2.5085	5.6098	-38.2182	35.6835	0.0196	0.7655	1.5767	3.2716
ln(assets)	95643	2.4329	1.1937	-0.288	4.8443	0.0039	1.6373	2.5736	3.3077
$\ln(\text{assets})^2$	95643	7.42	5.4433	0.0888	23.4674	0.0176	2.8153	6.6566	10.9411

Notes: Panel A "Years 0-1 Post" indicates whether a firm has disclosed a data breach in the current or previous year. Panel B "Years 0-2 Post" indicates whether a firm has disclosed a data breach event within the current or past two years, and Panel C "Years 0-3" indicates whether a firm has disclosed a data breach event within the current or past three years. Data breaches are included if the number of affected records is known, and it is at least 1,000. The control group consists of all firms in industries that at some point suffered a data breach. Controls include ln(Assets), ln(assets²) and market leverage. The construction of the variables is described in detail in Appendix A Table A.2. Standard errors are reported in parenthesis and adjusted for heteroskedasticity and clustering at the firm level, and ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

4. LINGUISTIC CUES & FINANCIALS

We use Professor McDonald's 10-K filing summary file freely available online to construct our main sample variables related to linguistic cues[12]. The summary is based on the U.S. SEC online electronic EDGAR filings. The summary file has several advantages; first, it is factual and reproducible. The variables in Loughran and McDonald (2011) rely on an automated parsing algorithm that seeks to eliminate human misinterpretations and errors. Second, unlike traditional measures that rely on labour-intensive processes for limited tests, the data is accessible to every company that file a 10-K. The usage of these steps dependant on a bag of terms that are already incorporated into an integrated database significantly expands the availability of public corporations to practically the whole cosmos of these. Third, the word list is comparatively exhaustive compared to Henry (2008) word list: there are no frequently occurring negative or positive word terms omitted. The word list contains 254 positive and 2,329 negative words. As mentioned above, this data is available online and free to use for researchers. Additionally, Loughran and McDonald's dictionary was prepared expressly for accounting reports, as opposed to other word lists used in the accounting and finance literature such as Henry (2008), Harvard's General Inquirer (GI) and Diction.

We define our set of narrative characteristics of the 10-K annual accounts based on a sequence of sentiment counts following Loughran and Mc Donald (2011):

Table 6. Linguistic Cues Definitions

Variable	Description
Tone	The number of positive words – the number of negative words divided by the total number of words.
Use of uncertainty words	The number of uncertainty words divided by the total number of words. Examples are approximate, contingency, depend, fluctuate and uncertain.
Use of possibility words	The number of possibility (weak modal) words divided by the total number of words. Examples are could, may suggest, possibly, and possible.
ln(Net file Size)	Natural logarithm of the net file size of the 10-K report.
Bog Index	Proxy of narratives complexity following Bonsall, Leone, Miller, and Rennekamp (2017) and based on the formula Bog Index = Sentence Bog + Word Bog
Abnormal Tone	The residual of tone following Huang et al. (2014).

Further, to analyze complexity, instead of using the highly criticized Fog Index, we use two metrics. The natural log of the net file size of the 10-K document (in megabytes) Loughran and McDonald (2014) recommend, and the Bog Index[13]. The Bog Index is a multifaceted predictor of plain English readability. (Bonsall & Miller,2017; Bonsall, Leone, et al., 2017).

4.1 Constructing Abnormal Tone

[14]To construct abnormal tone, we follow Huang et al.(2014). The concept of abnormal tone is built on the premise that both economic fundamentals and management incentives have an effect on it. In other terms, both honest and strategic disclosures coexist. He, therefore, "decomposed tone into non-discretionary component based on economic fundamentals and a discretionary component that could reflect managerial incentives, managers' private information about future firm fundamentals, manager's biased estimation of fundamentals, or noise" and "find that

[12] Available at https://www3.nd.edu/~mcdonald/

[13] Brian P. Miller: Kelley School of Business: Indiana University (iu.edu).

[14] We follow Huang et al. (2014) and Li (2010) to construct our tone models. Huang et al. (2014) measure whether managers strategically use tone in earning press releases and uses Loughran and Dictionary, while Li (2010) measures the information content of the Management's Discussion and Analysis of 10-Q and 10-K corporate fillings using a Naïve Bayesian algorithm. We do both as neither are exactly comparable to our analysis.

abnormal positive tone is used to facilitate managerial incentives to mask weak future fundamentals" (Huang et al.2014, p.1087).

There are many explanations why managers can enhance a positive tone. It may simply be an expression of positive actual and anticipated financial results. Alternatively, tone may be skewed upwards for a variety of purposes. Managers may use an optimistic tone bias to signal investors private knowledge regarding positive potential results that current quantitative disclosures do not disclose, perhaps due to GAAP restrictions. Positive prejudice can often arise from managers' manipulative efforts to conceal bad current results or to hype investors' perceptions of potential performance in order to confuse investors.

Huang et al. (2014) perform annual cross-sectional regressions. Abnormal positive tone as the residual U_{jt} of the model and defines it as "tone management". Specifically, the regression is:

$$TONE_{jt} = \beta_0 EARN_{jt} + \beta_1 RET_{jt} + \beta_2 \Delta EARN_{jt} + \beta_3 SIZE_{jt} + \beta_4 BTM_{jt} + \beta_5 STD\_RET_{jt} + \beta_6 STD\_EARN_{jt} + \beta_7 AGE_{jt} + \beta_8 BUSSEG_{jt} + \beta_{10} GEOSSEG_{jt} + \beta_{11} LOSS_{jt} + U_{jt} \quad (3)$$

"The determinants are measures for current available fundamental information, growth opportunities, operating risks, and complexity."
(Huang et al., 2014 p.1091)

Current financial and market performance: captured by profitability (*EARN*) and two performance benchmarks (*LOSS*) and change in earnings ($\Delta EARN$). We expect the level of tone to increase with current profitability *(EARN)*. *LOSS* is an indicator variable set to 1 when EARN is negative. We expect a negative coefficient. $\Delta EARN$ is the change in earnings from the prior year. The hypothesis is unclear as firms may use tone to deliver a different image to readers or not.

Future performance: captured by stock return (*RET*) and book-to-market ratio (*BTM*). *RET* is the contemporaneous annual stock returns. We are unsure of the relationship with tone. For example, lawsuits may lead better-performing firms to be more prudent. BTM controls for *growth opportunity*. The investment incentive set and development prospects of low BTM firms vary from those of high BTM ratio firms. Growth firms (firms with low BTM ratios) are exposed to increasingly unpredictable potential economic factors; therefore, a negative relationship is anticipated.

Price and return operating risks: captured by the volatility of stock returns (*STD_RET*), volatility of earnings (*STD_EARN*). Firms with more volatile market settings may be more careful to address upcoming events due to uncertainties about future results. Managerial and investor knowledge asymmetry is often more common in these companies. Lastly, profitability instability may have an impact on a company's exposure to litigation proceedings. We hypothesize a negative correlation.

Operational complexity: Geographic segments (GEOSEG) and business segments (BUSSEG), firm age (AGE) and firm size (SIZE) proxy for operating complexity of the firm. Firms with a greater number of business and geographic environments may be more cautious of information uncertainty regarding future performance. Additionally, age measures the *lifecycle stage of a company*. The younger the firm, the more uncertainties they face. Young executives are expected to be more vigilant when considering potential opportunities. Larger firms may employ cautiousness because of their higher political and legal costs due to their visibility.

Table 7. Constructing Abnormal Tone

Dependent Var.:	TONE
(Intercept)	-0.005624***
	(0.000105)
EARN	0.001011***
	(0.000135)
STD_EARN	-0.003100***
	(0.000184)
Δ EARN	-0.002685***
	(0.000178)
LOSS	-0.001257***
	(0.000041)
BTM	-0.002861***
	(0.000057)
SIZE	-0.000349***
	(0.000010)
BUSEG	-0.000064**
	(0.000020)
GEOSSEG	-0.000164***
	(0.000019)
AGE	0.000325***
	(0.000020)
RET	-0.000126***
	(0.000027)
STD_RET	0.000112*
	(0.000061)
Observations	61,155
R2	0.07670
Adj. R2	0.07653

Notes: Table 7 shows the model estimated using a pooled OLS regression. The sample comprises 61,155 observations for the period 2000-2018. The construction of the variables is described in detail in Appendix A Table A.3. Standard errors reported in parenthesis and are clustered at the firm level and adjusted for heteroskedasticity using the Huber-White procedure. All continuous variables have been winsorized at 1% and 99% level to mitigate the effect of outliers ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

We find that a more positive level of tone is associated with higher current earnings (*EARN)*, lower stock returns (*RET)*, smaller firms (*SIZE)*, lower *BTM* ratio firms (i.e., growth firms) the lower variability of earnings (*STD_EARN*), and the lower number of business and geographic segments. Even though Huang et al. (2014) have positive coefficients for earnings volatility, we expected a negative relation. Our findings vary from Huang et al. (2014) because some aspects of our research are different. First, we concentrate on 10-K records because they provide formal information and are structured in such a way that administrators can disclose information to a variety of different audiences (Kothari et al., 2009). A current study by Davis and Tama-Sweet (2012) reveals that disclosure styles vary greatly between earning press releases and 10-K annual reports. Second, time intervals for the samples are different. Huang et al. (2014) used a sample period of 1997 to 2007 while ours is from 2000 to 2018. Lastly, it is worth noting that our adjusted R-squared obtained in our model is 0.077 and that we obtain statistical significance for every variable. As a result, we conclude that our model adequately accounts for *optimistic tone* and that the residuals from this model may be used to assess *optimistic abnormal tone.*

Each year, we calculate the mean, median, standard deviation, and 1st, 25th, 75th, and 99th percentiles for each variable in our study. The annual average of the cross-sectional figures for the variables in Table 8 is then recorded. The mean (median) tone is -0.0095% (-0.0098%), confirming previous literature that annual 10-K reports tend to incorporate greater negative words.[15]

[15] E.g. . Huang. et al. (2014) report a higher positive tone in earning press releases

Table 9 summarises the Spearman correlations between 15 firm characteristics and (1) TONE, (2) Δ TONE and (3) ABTONE, Spearman correlation of 15 firm characteristics. The findings indicate that ABTONE has a significantly lower association with firm fundamentals than TONE and Δ TONE. These findings validate Huang et al. (2014) argument that ABTONE is a better proxy for discretionary tone than the other two tests.

Table 8. Descriptive Statistics

	Mean	SD	Median	Q0.01	Q0.25	Q0.75	Q0.99
TONE	-0.0095	0.0008	-0.0098	-0.0105	-0.0101	-0.0089	-0.0080
ABTONE	-0.0001	0.0008	-0.0004	-0.0010	-0.0008	0.0004	0.0017
ADACC	0.0745	0.0067	0.0732	0.0653	0.0699	0.0758	0.0905
DACC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EARN	-0.0852	0.0152	-0.0874	-0.1126	-0.0955	-0.0717	-0.0591
CFO	-0.0042	0.0115	-0.0022	-0.0235	-0.0133	0.0036	0.0178
RET	-0.0082	0.2687	0.0245	-0.6712	-0.1273	0.1631	0.4463
SIZE	5.0920	0.3747	5.1380	4.3727	4.9717	5.3487	5.7406
BTM	0.6748	0.0494	0.6682	0.6146	0.6355	0.7098	0.7882
STD_RET	0.4495	0.1171	0.3956	0.3418	0.3703	0.4849	0.7355
STD_EARN	0.0996	0.0107	0.1022	0.0825	0.0913	0.1082	0.1149
AGE	2.2564	0.0783	2.2585	2.0478	2.2365	2.2884	2.4047
BUSSEG	0.4000	0.2791	0.2348	0.1491	0.1989	0.6401	0.8538
GEOSEG	0.3728	0.2237	0.2619	0.1470	0.2121	0.5742	0.7336
LOSS	0.4085	0.0387	0.3997	0.3531	0.3796	0.4408	0.4408
Δ EARN	-0.0041	0.0110	-0.0055	-0.0193	-0.0133	0.0006	0.0161
UNC	0.0129	0.0015	0.0133	0.0102	0.0118	0.0141	0.0145
POSS	0.0055	0.0010	0.0056	0.0040	0.0047	0.0064	0.0070
ln(Net File Size)	5.4776	0.0571	5.4964	5.3566	5.4436	5.5083	5.5564
BOG	85.0975	3.3907	85.0423	79.2199	82.9550	87.3120	91.6639
Δ TONE	-0.0003	0.0004	-0.0002	-0.0010	-0.0003	-0.0001	0.0002
R&D	0.0392	0.0035	0.0385	0.0343	0.0367	0.0406	0.0460
CAPEX	0.0357	0.0053	0.0359	0.0262	0.0323	0.0387	0.0488
SPI	-0.0053	0.0009	-0.0050	-0.0077	-0.0054	-0.0047	-0.0043
MTB	2.0672	0.1547	2.0978	1.6886	1.9832	2.1754	2.2782
DLW	0.5444	0.0194	0.5347	0.5265	0.5310	0.5564	0.5852

Notes: Table 8 presents summary statistics for our sample. Each year, we obtain the median, median, standard deviation,1st, 25th, 75th, and 99th percentile of the variables in our sample. We then report the annual average of the cross sectional-statistics. The sample comprises 61,155 observations for the period 2000-2018. The construction of the variables is described in detail in Appendix A Table A.3. All continuous variables have been winsorized at 1% and 99% level to mitigate the effect of outliers.

Table 9. Tone Correlations and Firm Characteristics

	TONE	Δ TONE	ABTONE
EARN	0.13****	0.07****	0.06****
RET	-0.01***	0.00	0.00
SIZE	-0.05****	0.02****	0.01*
BTM	-0.12****	-0.05****	-0.01**
STD_RET	-0.06****	-0.07****	-0.01
STD_EARN	-0.05****	0.01*	-0.01*
AGE	0.07****	0.03****	0.01*
BUSSEG	-0.04****	0.03****	0.01*
GEOSEG	-0.04****	0.03****	0.01**
LOSS	-0.12****	-0.08****	-0.01*
Δ EARN	0.02****	0.09****	0.01*
CFO	0.09****	0.03****	0.06****
R&D	0.01****	0.01	0.02****
CAPEX	0.11****	0.01****	0.16****
TACC	0.01***	0.06****	-0.02****

Notes: Table 9 presents spearman correlation for *Tone*, *Change in Tone* and *Abnormal* Tone with firm industry characteristics. The sample comprises 61,155 observations for the period 2000-2018. The construction of the variables is described in detail in Appendix A Table A.3. All continuous variables have been winsorized at 1% and 99% level to mitigate the effect of outliers ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

Abnormal Tone *(ABTONE)* is constructed in such a way that it is irrelevant to recent financial performance or other firm fundamentals. We then assess its capacity to forecast future financial performance incremental to the published financial numbers and controls to see how it can detect the consequences of strategic managerial action. If abnormal positive tone forecasts optimistic potential profits and cash flows, it includes additional managerial private knowledge that, due to GAAP restrictions, cannot be transmitted through financial numbers. If abnormal tone

forecasts no/negative future profits and cash flows, managers are likely to use tone to merely hype or disguise mediocre future results to deceive investors.

We follow Huang et al. (2014) to ascertain the impact of strategic managerial behaviour and analyze the relationship between one-to-three year ahead financial results in Table 9. Our controls are discretionary accruals[16] *(DA)*, earnings *(EARN)*, stock returns *(RET)*,size, *(SIZE)*, book-to-market ratio *(BTM)*, volatility of stocks *(STD_RET)* and earnings *(STD_EARN)*.

We control for two-digit SIC industry[17] and year dummies akin to Huang et al. (2014). Standard errors are clustered at the firm level and adjusted for heteroskedasticity using the Huber-White procedure to account for cross-sectional and time-series error. Specifically, our regressions are:

$$EARN_{jt+n} = \beta_0 ABTONE_{jt} + \beta_1 DA_{jt} + \beta_2 EARN_{jt} + \beta_3 SIZE_{jt} + \beta_4 BTM_{jt} + \beta_5 STD\_RET_{jt} + \beta_6 STD\_EARN_{jt} + \beta_7 RET_{jt} + U_{jt} \quad (4)$$

$$CFO_{jt+n} = \beta_0 ABTONE_{jt} + \beta_1 DA_{jt} + \beta_2 EARN_{jt} + \beta_3 SIZE_{jt} + \beta_4 BTM_{jt} + \beta_5 STD\_RET_{jt} + \beta_6 STD\_EARN_{jt} + \beta_7 RET_{jt} + U_{jt} \quad (5)$$

where n= (1,2,3).

If *ABTONE* forecasts optimistic future earnings and cash flows, we can confirm it is correctly calculated, and it provides additional managerial private knowledge. Table 10 presents our results. Panel A presents the relationship between abnormal positive tone future cash flows, and Panel B presents the relationship between future earnings and abnormal positive tone. We find ABTONE has a negative and significant coefficient across all regressions and time horizons. Therefore, as abnormal positive tone is more prevalent in companies with unfavourable future fundaments, we can confirm that abnormal tone is correctly calculated and provides additional managerial private knowledge.

Table 10. Abnormal Positive Tone and Future Financial Performance

Panel A: Abnormal Positive Tone Future Cash Flows

	(1)	(2)	(3)
Dependent Var.:	CFO_{t1}	CFO_{t2}	CFO_{t3}
ABTONE	-0.470406***	-0.666636***	-0.783843***
	(0.133656)	(0.145282)	(0.153983)
DACC	-0.362729***	-0.310367***	-0.276409***
	(0.007796)	(0.008441)	(0.008555)
EARN	0.600692***	0.510818***	0.441075***
	(0.005390)	(0.005812)	(0.005880)
SIZE	0.004066***	0.005755***	0.006758***
	(0.000332)	(0.000365)	(0.000382)
RET	0.046320***	0.045007***	0.041957***
	(0.001186)	(0.001340)	(0.001394)
BTM	-0.012891***	-0.019161***	-0.021235***
	(0.001961)	(0.002136)	(0.002261)
STD_RET	-0.010060***	-0.014980***	-0.016848***
	(0.002552)	(0.002838)	(0.002956)
STD_EARN	-0.073689***	-0.077706***	-0.080374***
	(0.007119)	(0.007702)	(0.008086)
Fyear x SIC2 FE	Yes	Yes	Yes
Observations	44,036	41,317	38,288
R2	0.66173	0.57955	0.51904
Within R2	0.60996	0.51899	0.45544

[16] We measure discretionary accruals similarly to Huang et al. (2014). Details of the construction are in Appendix A, Table A.3 and an example of a model is in Appendix B, Table B.1.

[17] We also control for industry-by-year fixed effects as we explore the effect of data breaches using an industry-by-year approach and supply the results in Appendix B, Table B.2.

Panel B: Abnormal Positive Tone Future Earnings			
Dependent Var.:	$EARN_{t1}$	$EARN_{t2}$	$EARN_{t3}$
ABTONE	-0.351161*	-0.915370***	-1.247844***
	(0.164208)	(0.186650)	(0.199168)
DACC	-0.215492***	-0.217122***	-0.201641***
	(0.009378)	(0.010411)	(0.011120)
EARN	0.659712***	0.549125***	0.482392***
	(0.006333)	(0.007147)	(0.007642)
SIZE	0.003655***	0.006333***	0.008457***
	(0.000420)	(0.000469)	(0.000501)
RET	0.092959***	0.088110***	0.060943***
	(0.001682)	(0.001831)	(0.001903)
BTM	-0.033380***	-0.012713***	0.000442
	(0.002582)	(0.002856)	(0.003051)
STD_RET	-0.054492***	-0.036355***	-0.030962***
	(0.003495)	(0.003763)	(0.003926)
STD_EARN	-0.116876***	-0.150744***	-0.161707***
	(0.008824)	(0.010138)	(0.010867)
Fyear x SIC2 FE	Yes	Yes	Yes
Observations	44,075	41,366	38,375
R2	0.65998	0.55574	0.47648
Within R2	0.61720	0.50116	0.41649

Notes: Table 10 presents OLS regression for examining the relationship between Abnormal Positive Tone and Future Financial Performance. The dependant variables are cash flows one to three years ahead in Panel A and earnings one to three year ahead in Panel B. The construction of variables is defined in detail in Appendix A, Table B.3. We include 2-digit SIC industry and year fixed effects. Standard errors are reported in parenthesis, clustered at the firm level, and adjusted for heteroskedasticity using the Huber-White procedure. All continuous variables have been winsorized at 1% and 99% level to mitigate the effect of outliers. ***,**, and * represent significance at the 1%, 5% and 10% levels, respectively.

4.2 Linguistic Cues Determinants

We follow by constructing our baseline regressions. We select our control variables following Li (2010). We control for year fixed effects[18] and cluster our standard errors at the firm and year level, adjusting for heteroskedasticity using the Huber-White procedure to account for cross-sectional and time-series error. Specifically, our specification is:

$$LINGUISTIC\ CUE_{jt} = \beta_0 EARN_{jt} + \beta_1 RET_{jt} + \beta_2 TACC_{jt} + \beta_3 SIZE_{jt} + \beta_4 BTM_{jt} + \beta_5 STD\_RET_{jt} + \beta_6 STD\_EARN_{jt} + \beta_7 AGE_{jt} + \beta_8 BUSSEG_{jt} + \beta_9 GEOSSEG_{jt} + \beta_{10} M\&A_{jt+1} + \beta_{11} SEO_{jt+1} + \beta_{12} SPI_{jt} + \beta_{13} DLW_{jt} + U_{jt} \quad (6)$$

4.2.1 Determinants of Tone and Abnormal Tone

Table 11 reports the OLS regression results for *optimistic tone* in columns (1) and (2) and *abnormal optimistic tone* in columns (3) and (4). In columns (1) and (3), we include year fixed effect as Li (2010), and in columns (2) and (4), we include industry-by-year fixed effects to compare our results with our subsequent analysis of Section 5. We focus on columns (1) and (3).

In column (1), we find a positive relationship between *optimistic tone* and current performance (*EARN* and *RET),* indicating that managers of high-performing companies address their potential prospects in a more optimistic tone. Accruals *(TACC)* are significantly negatively correlated to *optimistic tone,* implying that while actual accruals are very high, narratives with 10-K annual reports of the firm's potential prospects are more pessimistic. Given that accruals are adversely linked to potential performance, administrators should be aware of the implications of accruals on future performance. Moreover, as shown by the negative coefficient *(SIZE),* larger

[18] In untabulated results we have also calculated the models with no fixed effects, year fixed effect, year and 2-digit SIC industry fixed effect and industry-by-year fixed effect. to be able to compare our results properly to Li (2008), Li (2010) and Kim (2018) as they each use a different fixed effect. Even though, neither control for industry-by-year fixed effects, we include we include industry-by-year fixed effects because our analysis in section 5 controls for industry-by-year fixed effects.

companies tend to have a greater negative tone aligned with the theory that large companies are more cautious in their disclosures due to political and legal concerns. Furthermore, companies with low BTM (book-to-market) ratios are growth companies, and these face more uncertain information conditions and are therefore more conservative in forecasting potential events.

Additionally, firms with greater volatile and unpredictable environments (*STD_EARN* and *STD_RET)* appear to use a lower optimistic tone when considering their potential prospects. We also find that older firms with a greater number of business segments (*BUSSEG*) and geographic segments (*GEOSEG*) that are involved in seasoned equity offerings (SEO_{t+1}) are also positively related to optimistic tone. Additionally, firms incorporated in Delaware *(DLW)* or have a merger and acquisition ($M\&A_{t+1}$) the following year exhibit a negative relationship. In contrast, special items *(SPI)* exhibit a positive relationship.

Li (2010) finds a negative relationship between *GEOSEG*, *SPI* and *Tone*, while we find positive relationships. Our results differ from Li (2010) because our research approaches are different. First, Li (2010) uses a Bayesian algorithm, while we use Loughran and McDonald Dictionary List. Second, our period ranges are different: Li (2010) use a sample period between 1994 and 2007 while ours is 2000 to 2018. Third, Li (2010) uses 10-K annual and 10-Q quarterly reports and includes three reporting quarter dummies while we analyze only annual 10-K filings. Finally, Li (2010) focuses on the MD&A section of the corporate reports while we analyze the entire document.

In column (3), we present the determinants of abnormal tone. We do not expect a good fit as an abnormal tone was created exactly for this reason. Overall, our findings indicate that bigger firms *(SIZE)*, older *(AGE)*, with lower current earnings *(EARN)*, lower accruals *(TACC)*, higher book-to-market ratios *(BTM)*, with lower stock returns volatility *(STD_RET)*, and higher earnings volatility (*STD_EARN)* with a higher number of geographic *(GEOSEG)* and business segments *(BUSSEG)*, with an M&A and without an SEO the following year generally have a more optimistic abnormal tone.

Table 11. Determinants of Tone and Abnormal Tone

	(1) TONE	(2) TONE	(3) ABTONE	(4) ABTONE
EARN	0.001074***	0.000404**	-0.001308***	-0.002047***
	(0.000124)	(0.000130)	(0.000124)	(0.000129)
TACC	-0.000644***	0.000396*	-0.000459*	0.000768***
	(0.000188)	(0.000199)	(0.000188)	(0.000198)
SIZE	-0.000290***	-0.000243***	0.000020*	0.000070***
	(0.000010)	(0.000010)	(0.000010)	(0.000010)
RET	0.000048*	0.000065**	0.000109***	0.000148***
	(0.000023)	(0.000023)	(0.000031)	(0.000031)
BTM	-0.002642***	-0.001749***	0.000266***	0.001204***
	(0.000058)	(0.000063)	(0.000058)	(0.000062)
STD_RET	-0.000998***	-0.001394***	-0.000762***	-0.001183***
	(0.000065)	(0.000066)	(0.000065)	(0.000065)
STD_EARN	-0.003119***	-0.004275***	0.000657***	-0.000630***
	(0.000181)	(0.000186)	(0.000181)	(0.000186)
GEOSEG	0.000269***	0.000007	0.000457***	0.000180***
	(0.000020)	(0.000022)	(0.000020)	(0.000022)
BUSSEG	0.000391***	0.000143***	0.000456***	0.000185***
	(0.000021)	(0.000022)	(0.000021)	(0.000022)
FIRM AGE	0.000022***	0.000011***	0.000009***	-0.000002*
	(0.000001)	(0.000001)	(0.000001)	(0.000001)
$M\&A_{t+1}$	-0.000110*	-0.000110*	-0.000152**	-0.000159**
	(0.000050)	(0.000050)	(0.000050)	(0.000050)
SEO_{t+1}	0.000381***	0.000348***	0.000364***	0.000332***
	(0.000058)	(0.000058)	(0.000058)	(0.000057)
DLW	-0.000170***	-0.000359***	-0.000100***	-0.000309***
	(0.000028)	(0.000028)	(0.000028)	(0.000028)
SPI	0.032115***	0.040543***	0.026297***	0.034987***
	(0.001253)	(0.001257)	(0.001251)	(0.001253)
Fyear FE	Yes	No	Yes	No
Fyear x SIC2 FE	No	Yes	No	Yes

Observations	59,564	59,564	59,564	59,564
R2	0.15929	0.26943	0.09327	0.21649
Within R2	0.09827	0.08860	0.03923	0.03820

Notes: Table 11 presents OLS regression for the determinants of tone and abnormal tone following Li (2010). The dependant variable, "Tone" in column (1) and (2), is the difference between the number of positive and negative words over total words of 10-K annual accounts. The dependant variable, "Abnormal Tone", in column (3) and (4) is the residual of Table 7. following Huang et al. (2014). . The construction of the variables is described in detail in Appendix A Table A.3. Column (1) and (3) includes year fixed effects following Li (2010); column (2) and (4) includes industry-by-year fixed effects so that we can compare our results in Section 5. Standard errors are reported in parenthesis and are clustered at the firm level and adjusted for heteroskedasticity using the Huber-White procedure. All continuous variables have been winsorized at a 1% level to mitigate the effect of outliers. ***, **, and * represent significance at the 1%, 5% and 10% levels, respectively.

4.2.2 Determinants of Vague Language

Table 12 reports the OLS regression results for the determinants of vague language. Columns (1) and (2) capture our first proxy, *use of weak modal words* and columns (3) and (4) capture our third proxy: *uncertainty words*. In columns (1) and (3), we include year fixed effect as Li (2010). In columns (2) and (4), we include industry-by-year fixed effects to compare our results with our subsequent analysis in Section 5.We compare our results Kim (2018). Again, we focus on columns (1) and (3).

We find that the *use of weak modal words (POSS)*, column (1) *and* is associated positively with firms with lower current earnings *(EARN)*, with a higher number of accruals *(TACC)*, of bigger size *(SIZE)*, higher stock returns *(RET)*, lower book-to-market ratio *(BTM)*, higher volatile business environments *(STD_RET; STD_EARN)*, a higher number of geographic segments *(GEOSEG)*, lower number of business segments *(BUSSEG)*, younger *(AGE)*, with a greater number of special items *(SPI)*, if a firm is incorporated in Delaware *(DLW)* or has a seasoned equity offering (SEO_{t+1}) or a merger and acquisition *(*$M\&A_{t+1}$*)* the following year. Results are identical for the use of *uncertainty words (column 3)*, but controls for firm events are not statistically significant.

These results are congruent to prior research (Loughran and McDonald, 2011; Kim, 2018). Firm size captures a firm's operational and business environment; therefore, larger firms with more volatile environments face a more dynamic and unpredictable market climate should be expected to employ words of a higher ambiguous nature in their corporate accounts. Growth firms (*low BTM firms)* may also face a more uncertain and business environment. Firms with a greater number of geographic segments also face more uncertainty. Younger firms tend to have greater information asymmetry. Delaware firms follow different laws and regulations, have more takeover bids, and are valued higher than similar firms in other states, therefore more prone to employing vague language. Firms with firm events such as mergers and acquisitions and seasoned equity offerings with a greater number of special items also face a more volatile, unpredictable environment.

Our results differ from Kim (2018), possibly because our research approaches are slightly different. First, he excludes regulated utilities and financial firms. Second, our sample periods are different. Third, our control variables are slightly different: he retrieves M&A data from the SDC Platinum M&A database and SEO data from SDC Global New database and does not control for accruals and current performance.

Table 12. Determinants of Vague Language

	(1) POSS	(2) POSS	(3) UNC	(4) UNC
EARN	-0.000762***	-0.000698***	-0.000475***	-0.000201*
	(0.000049)	(0.000051)	(0.000088)	(0.000093)
TACC	0.000154*	0.000186*	0.000663***	0.000266.
	(0.000074)	(0.000078)	(0.000134)	(0.000143)
SIZE	0.000067***	0.000100***	0.000103***	0.000132***
	(0.000004)	(0.000004)	(0.000007)	(0.000007)
RET	0.000024*	0.000015	0.000040.	0.000022
	(0.000012)	(0.000012)	(0.000022)	(0.000022)
BTM	-0.000515***	-0.000196***	-0.000110**	0.000017
	(0.000023)	(0.000025)	(0.000042)	(0.000046)
STD_RET	0.000374***	0.000347***	0.000133**	0.000206***
	(0.000025)	(0.000026)	(0.000045)	(0.000047)
STD_EARN	0.002445***	0.001909***	0.001962***	0.001502***
	(0.000072)	(0.000074)	(0.000130)	(0.000135)
GEOSEG	0.000064***	-0.000037***	0.000135***	0.000035*
	(0.000008)	(0.000009)	(0.000014)	(0.000016)
BUSSEG	-0.000105***	-0.000122***	-0.000193***	-0.000139***
	(0.000008)	(0.000009)	(0.000015)	(0.000016)
FIRM AGE	-0.000017***	-0.000016***	-0.000031***	-0.000030***
	(0.000000)	(0.000000)	(0.000001)	(0.000001)
$M\&A_{t+1}$	0.000084***	0.000056**	0.000066.	0.000046
	(0.000020)	(0.000020)	(0.000036)	(0.000036)
SEO_{t+1}	0.000072**	0.000065**	0.000032	0.000023
	(0.000022)	(0.000022)	(0.000040)	(0.000041)
DLW	0.000200***	0.000131***	0.000118***	0.000091***
	(0.000011)	(0.000011)	(0.000020)	(0.000020)
SPI	0.001521**	0.003170***	0.003965***	0.003894***
	(0.000485)	(0.000490)	(0.000892)	(0.000903)
Fyear FE	Yes	No	Yes	No
SIC2 FE	No	No	No	No
Fyear x SIC2 FE	No	Yes	No	Yes
Observations	59,564	59,564	59,564	59,564
R2	0.39086	0.44041	0.28421	0.33971
Within R2	0.19966	0.12376	0.04439	0.03205

Notes: The table presents OLS regression for the determinants of vague language following Li (2010). The dependant variable, "POSS" in column (1) and (2), is the number of weak modal words over the total number of words. The dependant variable, "UNC", in column (3) and (4), is the number of uncertainty words over the total number of words. The construction of the variables is described in detail in Appendix A Table A.3. Column (1) and (3) includes year fixed effects following Li (2010); column (2) and (4) includes industry-by-year fixed effects so that we can compare our results in Section 5. Standard errors are reported in parenthesis and are clustered at the firm level and adjusted for heteroskedasticity using the Huber-White procedure. All continuous variables have been winsorized at 1% level to mitigate the effect of outliers. ***, **, and * represent significance at the 1%, 5% and 10% levels, respectively.

4.2.3 Determinants of Annual Report Readability

Table 13 reports the OLS regression results for the determinants of annual report readability. In columns (1) and (3), we include year fixed effect as Li (2010), and in columns (2) and (4), we include industry-by-year fixed effects to compare our results with our subsequent analysis in Section 5. We focus on columns (1) and (3).

We use two metrics to assess the readability of 10-K papers. Columns (1) and (2) present the results using our first metric: the natural log of the net file size of a 10-K document (in megabytes) (Loughran and Mc Donald 2014) as Loughran and McDonald suggest file size is a significant and robust indicator of readability of financial statements. It is based on the premise that longer reports are more deterrent and require higher knowledge costs. Columns (3) and (4) presents the determinants of our second metric: the Bog Index. We use the Bog Index as our main readability metric since it predicts plain English readability in a variety of ways (Bonsall & Miller, 2017; Bonsall, Leone et al., 2017). A higher bog index indicates worse text readability. To compare our results, we compare our findings to Li (2008).

In column (1) and (3), we find that firm current performance *(EARN)* is negatively associated with the length and complexity of the 10-K document following the managerial

obfuscation hypothesis, which states that poorly performing companies employ textual complexity strategically to obfuscate facts in corporate transparency. We find that bigger firms *(SIZE)*, with greater volume of accruals *(TACC)*, growth firms *(lower BTM)*, greater volatility of operations (*STD_RET; STD_EARN)*, with more complex operations *(GEOSEG, BUSSEG)*, younger (*FIRM AGE)*, that have a seasoned equity offering *(SEO_{t+1})* or a merger acquisition *($M\&A_{t+1}$)* the following year and are incorporated in Delaware *(DLW)* and fewer special items *(SPI)* are associated with less readable results. We find two controversies in our results. First, in column (3), analyzing the determinants of the *Bog Index*, we find a negative correlation with current earnings *(EARN)* but a positive effect with current stock returns *(RET)*. Second, in column (1), measuring the determinants of *ln(NETFILESIZE)*, we find a significant negative instead of positive between the length of the 10-K document and the number of geographic segments *(GEOSEG*). However, Li (2008) finds the same result and notes, "The counter-intuitive result is the negative coefficient on geographic segments, suggesting firms with more geographic segments tend to have less complicated annual reports" (Li 2008, page 21). We can confirm our results are similar to Li (2008).

Li (2008) finds that larger firms, firms with more volatile business, firms with merger and acquisition (M&A) transactions, and firms incorporated in Delaware are positively associated with Fog Index (listed items are associated with less readable reports). However, his finding suggests that firm age, firms with special items, firms with geographic segments, and firms issuing new equity are negatively associated with 10-K reports (i.e., listed items are associated with more readable 10-K reports).

Table 13. Determinants of Annual Report Readability

	(1) ln(NETFILESIZE)	(2) ln(NETFILESIZE)	(3) BOG INDEX	(3) BOG INDEX
EARN	-0.232322***	-0.173615***	-9.641453***	-5.805163***
	(0.006060)	(0.006380)	(0.220005)	(0.209713)
TACC	0.077857***	0.040949***	7.889513***	3.378174***
	(0.009528)	(0.010048)	(0.338663)	(0.329398)
SIZE	0.068602***	0.064280***	0.998801***	1.096261***
	(0.000474)	(0.000501)	(0.017332)	(0.016825)
RET	-0.001512	-0.004051*	0.400462***	0.132542*
	(0.001657)	(0.001700)	(0.055006)	(0.051981)
BTM	0.186219***	0.157537***	0.557310***	2.516443***
	(0.002912)	(0.003227)	(0.106249)	(0.105561)
STD_RET	0.067775***	0.093778***	2.076110***	2.131321***
	(0.003346)	(0.003432)	(0.113703)	(0.109188)
STD_EARN	0.093053***	0.107676***	9.144828***	4.586813***
	(0.008743)	(0.009045)	(0.319943)	(0.299205)
GEOSEG	-0.011596***	0.006886***	0.170470***	-0.243015***
	(0.000931)	(0.001039)	(0.035695)	(0.037159)
BUSSEG	0.003851***	0.006361***	0.737366***	0.408124***
	(0.001011)	(0.001092)	(0.037904)	(0.038320)
FIRM AGE	-0.001727***	-0.001427***	-0.006227***	-0.015016***
	(0.000074)	(0.000075)	(0.001669)	(0.001578)
$M\&A_{t+1}$	0.007435**	0.009352***	0.827063***	0.527049***
	(0.002538)	(0.002540)	(0.088220)	(0.081379)
SEO_{t+1}	0.020500***	0.013746***	0.302354**	0.083599
	(0.002821)	(0.002819)	(0.102464)	(0.089962)
DLW	0.018565***	0.025438***	1.114900***	0.852312***
	(0.001430)	(0.001439)	(0.050509)	(0.047250)
SPI	-1.002248***	-1.354389***	-34.601831***	-22.843736***
	(0.062557)	(0.062611)	(2.234454)	(2.070004)
Fyear FE	Yes	No	Yes	No
SIC2 FE	No	No	No	No
Fyear x SIC2 FE	No	Yes	No	Yes
Observations	59,571	59,571	62,736	62,736
R2	0.34444	0.40788	0.28189	0.43688
Within R2	0.28845	0.27871	0.16863	0.10714

Notes: The table presents OLS regression for the determinants of readability following Li (2010). The dependant variable, "ln(NETFILESIZE)" in column (1) and (2), is the natural log of the net file size of the 10-K document. The dependant variable, "Bog Index", in column (3) and (4) is the Bog Index. The construction of the variables is described in detail in Appendix A Table A.3. Column (1) and (3) includes year fixed effects following Li (2010); column (2) and (4) includes industry-by-year fixed effects so that we can compare our results in Section 5. Standard errors are reported in parenthesis and are clustered at the firm level and adjusted for heteroskedasticity using the Huber-White procedure. All continuous variables have been winsorized at 1% and 99% level to mitigate the effect of outliers. ***, **, and * represent significance at the 1%, 5% and 10% levels, respectively.

5. LINGUISTIC CUES, FINANCIALS & DATA BREACHES

To analyze the impact of data breaches in the linguistic cues of the corporate narratives, we build an annual panel of all firms from the period 2000-2018. As before, our control sample consists of all firms in industries (by 4-digit SIC) that at some point suffered a data breach. This restriction eliminates 47% of firms in Compustat. Again, to construct our difference-in-difference model, we follow Akey et al., 2018's methodology. Our main specification is the same model (2) of Section 3.3:

$$Y_{ijt} = \alpha + \gamma Post_{ijt} + \delta Treated + \beta x_{ijt} + f_{it} + \varepsilon_{ijt} \qquad (7)=(2)$$

As in Section 3.3, we construct an indicator variable *Post* that identifies firm-year observations over different time frames. We add two other definitions of Post to analyze the impact of data breaches on linguistic cues: *Post0* and *Post0-4. Post0* is an indicator variable that takes the value of one if a firm was breached the current year; *Post0*-4 takes the value of 1 if a firm was breached the current year of the previous four years. Table14, Table 15 and Table 16 report the OLS regression results for the effect of data breaches on our linguistic cues. Column (1) reports the effect of data breach incidents on the year of the breach *(Post0),* column (2) the year of the breach and the previous year *Post0-1*, and so on.

We include industry-by-year fixed effects to ensure our comparisons are within industry, within a given year, between affected and unaffected firms. We do not exclude firms that were never hacked so that we can better assess the year-by-industry fixed effects. We include the variable *Treated,* which identifies whether a firm has even been subject to a data breach in our sample period. Each year, all continuous variables are winsorized at the 1% and 99% percentiles to diminish the effects of outliers in our analysis. Our identifying assumptions are those explained in Section 3.3.

We use two different sets of controls. To analyze the effect of optimistic tone and abnormal optimistic tone, we include the controls Huang et al. (2014) selected to explore abnormal positive tone in strategic settings. These controls are discretionary accruals *(DACC),* earnings *(EARN),* size, stock return *(RET),* and stock return volatility *(STD_RET)* and earnings volatility *(STD_EARN).* To analyze the effect of readability and vague language, our second set of controls are those used in Section 4 to study the determinants of linguistic cues following Li (2010).

5.1 Effect of Data Breaches on Tone & Abnormal Tone

Tables 14 presents the results of the effect of data breaches on optimistic tone (Panel A) and optimistic abnormal tone (Panel B) using the controls Huang et al. (2014).

We find strong evidence that companies that suffer a data breach employ a greater *optimistic tone* the years following the data breach. The optimistic tone of the corporate reports of firms that suffer a data breach the current year increases by 13% *(Column 1)*; the current or the previous year by 9% *(Column 2)*; the current or the previous two years by 7% *(Column 3)*; the current or the previous three years *(Column 4)* and the current or previous four years *(Column 5)* by 6%.

Again, we find substantial evidence that managers manage *optimistic abnormal tone* strategically the year of the breach by 8% *(Post0),* the year of the breach and the following year by 5% *(Post0-1),* the year of the breach and the following two years by 3% *(Post0-2)*. Contrary to *optimistic tone*, we do not find a significant effect for *Post0-3* and *Post0-4*, but still a positive coefficient. Our results for optimistic tone and optimistic abnormal tone show that the effect is more pronounced the closer it is to the year of the breach and then diminishes. Therefore, we can

say that our results confirm our first hypothesis that firms that suffer a data breach manage optimistic tone strategically.

Table 14. Effect of Data Breaches on Tone and Abnormal Tone

Panel A: Effect of Data Breaches on Tone

	(1) TONE	(2) TONE	(3) TONE	(4) TONE	(5) TONE
DACC	0.0031***	0.0031***	0.0031***	0.0031***	0.0031***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
EARN	0.0011***	0.0011***	0.0011***	0.0011***	0.0011***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
SIZE	-0.0004***	-0.0004***	-0.0004***	-0.0004***	-0.0004***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
RET	-0.0001	-0.0001	-0.0001	-0.0001	-0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
BTM	-0.0026***	-0.0026***	-0.0026***	-0.0026***	-0.0026***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
STD_RET	-0.0018***	-0.0018***	-0.0018***	-0.0018***	-0.0018***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
STD_EARN	-0.0061***	-0.0061***	-0.0061***	-0.0061***	-0.0061***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
Treated	0.0002.	0.0001	0.0001	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Post0	0.0013***				
	(0.0003)				
Post0-1		0.0009***			
		(0.0002)			
Post0-2			0.0007***		
			(0.0002)		
Post0-3				0.0006**	
				(0.0002)	
Post0-4					0.0006**
					(0.0002)
Fyear x SIC2 FE	Yes	Yes	Yes	No	No
Observations	28,021	28,021	28,021	28,021	28,021
R2	0.19630	0.19621	0.19605	0.19595	0.19597
Within R2	0.06329	0.06320	0.06301	0.06290	0.06291

Panel B: Effect of Data Breaches on Abnormal Tone

	(1) ABTONE	(2) ABTONE	(3) ABTONE	(4) ABTONE	(5) ABTONE
DACC	0.0025***	0.0025***	0.0025***	0.0025***	0.0025***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
EARN	-0.0016***	-0.0016***	-0.0016***	-0.0016***	-0.0016***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
SIZE	-0.0000*	-0.0000*	-0.0000*	-0.0000*	-0.0000*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
RET	0.0001*	0.0001*	0.0001*	0.0001*	0.0001*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
BTM	0.0009***	0.0009***	0.0009***	0.0009***	0.0009***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
STD_RET	-0.0012***	-0.0012***	-0.0012***	-0.0012***	-0.0012***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
STD_EARN	-0.0010***	-0.0010***	-0.0010***	-0.0010***	-0.0010***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Treated	0.0002*	0.0002*	0.0002*	0.0002**	0.0002*
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Post0	0.0008***				
	(0.0002)				
Post0-1		0.0005**			
		(0.0002)			
Post0-2			0.0003*		
			(0.0002)		
Post0-3				0.0002	
				(0.0001)	
Post0-4					0.0001
					(0.0001)
Fyear x SIC2 FE	Yes	Yes	Yes	No	No
Observations	28,021	28,021	28,021	28,021	28,021
R2	0.19630	0.19621	0.19605	0.19595	0.19597
Within R2	0.06329	0.06320	0.06301	0.06290	0.06291

Notes: Table 14 presents OLS regressions for the effect of data breaches on optimistic tone (Panel A) and optimistic abnormal tone (Panel B) with control variables Huang et al. (2014) selected for analyzing abnormal tone in strategic settings. Column(1) Post0, Column (2), Column (3) Post0-2, Column (4) Post0-3, Column (5) Post0-4 are indicator variables that take the value of one if a firm disclosed a data breach in the current year, the current year and the previous year, the current year and the previous two, three, four years. "Treated" takes the value of one if a firm was ever affected by a data breach and zero otherwise. Data breaches are included if the number of affected records is known and at least 1,000. The construction of the variables is described in detail in Appendix A Table A.3. We include industry-by-year fixed effects. Standard errors are reported in parenthesis and adjusted for heteroskedasticity and clustering at the firm level, and ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

5.2 Effect of Data Breaches on Vague Language

Tables 15 present the results of data breaches on the use of vague language using the controls selected by Li (2010) as in Section 4. Table 15, Panel A presents the results for our first vague language proxy: *use of uncertainty words (UNC)*. We find breached firms employ a fewer proportion of uncertainty words the year of the breach and the following two, three and four years *(Post0-2, Post0-3, Post0-4)*. Table 15, Panel B presents the results for our second proxy: the *use of weak modal words (POSS)*. We find firms employ a fewer proportion of possibility words across all our time periods. We can therefore confirm our second hypothesis as our sample breached firms use less vague language as a tool to obfuscate negative news.

Table 15. Effect of Data Breaches on Vague Language

Panel A: Effect of Data Breaches on the use of Uncertainty words					
	(1)	(2)	(3)	(4)	(5)
	UNC	UNC	UNC	UNC	UNC
EARN	-0.000337*	-0.000338*	-0.000339*	-0.000340*	-0.000340*
	(0.000147)	(0.000147)	(0.000147)	(0.000147)	(0.000147)
TACC	0.000179	0.000180	0.000182	0.000183	0.000184
	(0.000227)	(0.000227)	(0.000227)	(0.000227)	(0.000227)
SIZE	0.000161***	0.000161***	0.000161***	0.000161***	0.000162***
	(0.000011)	(0.000011)	(0.000011)	(0.000011)	(0.000011)
RET	0.000051	0.000051	0.000051	0.000052	0.000052
	(0.000035)	(0.000035)	(0.000035)	(0.000035)	(0.000035)
BTM	-0.000092	-0.000091	-0.000090	-0.000089	-0.000088
	(0.000073)	(0.000073)	(0.000073)	(0.000073)	(0.000073)
STD_RET	0.000148*	0.000148*	0.000149*	0.000149*	0.000150*
	(0.000071)	(0.000071)	(0.000071)	(0.000071)	(0.000071)
STD_EARN	0.001863***	0.001862***	0.001863***	0.001864***	0.001865***
	(0.000201)	(0.000201)	(0.000201)	(0.000201)	(0.000201)
GEOSEG	0.000077**	0.000077**	0.000077**	0.000076**	0.000077**
	(0.000026)	(0.000026)	(0.000026)	(0.000026)	(0.000026)
BUSSEG	-0.000161***	-0.000161***	-0.000161***	-0.000162***	-0.000162***
	(0.000026)	(0.000026)	(0.000026)	(0.000026)	(0.000026)
FIRM AGE	-0.000022***	-0.000022***	-0.000022***	-0.000022***	-0.000022***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)	(0.000001)
$M\&A_{t1}$	0.000088	0.000087	0.000087	0.000087	0.000086
	(0.000057)	(0.000057)	(0.000057)	(0.000057)	(0.000057)
SEO_{t+1}	0.000020	0.000020	0.000020	0.000020	0.000020
	(0.000067)	(0.000067)	(0.000067)	(0.000067)	(0.000067)
DLW	0.000107***	0.000107***	0.000107***	0.000107***	0.000108***
	(0.000031)	(0.000031)	(0.000031)	(0.000031)	(0.000031)
SPI	0.004282**	0.004280**	0.004277**	0.004276**	0.004279**
	(0.001414)	(0.001413)	(0.001413)	(0.001413)	(0.001413)
Treated	-0.000189***	-0.000171**	-0.000151**	-0.000131*	-0.000117*
	(0.000053)	(0.000054)	(0.000056)	(0.000058)	(0.000059)
Post0	-0.000132				
	(0.000161)				
Post0-1		-0.000191			
		(0.000127)			
Post0-2			-0.000231*		
			(0.000111)		
Post0-3				-0.000260*	
				(0.000105)	
Post0-4					-0.000270**
					(0.000100)
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes
Observations	37,714	37,714	37,714	37,714	37,714
R2	0.31767	0.31770	0.31774	0.31777	0.31779
Within R2	0.02690	0.02695	0.02700	0.02706	0.02709

.

Panel B: Effect of Data Breaches on the Use of Weak Modal Words					
	(1)	(2)	(3)	(4)	(5)
	POSS	POSS	POSS	POSS	POSS
EARN	-0.000956***	-0.000956***	-0.000956***	-0.000957***	-0.000958***
	(0.000087)	(0.000087)	(0.000087)	(0.000087)	(0.000087)
TACC	0.000078	0.000080	0.000080	0.000082	0.000083
	(0.000130)	(0.000130)	(0.000130)	(0.000130)	(0.000130)
SIZE	0.000113***	0.000113***	0.000113***	0.000113***	0.000113***
	(0.000006)	(0.000006)	(0.000006)	(0.000006)	(0.000006)
RET	0.000040*	0.000040*	0.000040*	0.000040*	0.000040*
	(0.000019)	(0.000019)	(0.000019)	(0.000019)	(0.000019)
BTM	-0.000420***	-0.000419***	-0.000418***	-0.000417***	-0.000416***
	(0.000043)	(0.000043)	(0.000043)	(0.000043)	(0.000043)
STD_RET	0.000320***	0.000320***	0.000321***	0.000321***	0.000322***
	(0.000040)	(0.000040)	(0.000040)	(0.000040)	(0.000040)
STD_EARN	0.002095***	0.002094***	0.002095***	0.002096***	0.002097***
	(0.000118)	(0.000118)	(0.000118)	(0.000118)	(0.000118)
GEOSEG	-0.000035*	-0.000035*	-0.000035*	-0.000035*	-0.000035*
	(0.000015)	(0.000015)	(0.000015)	(0.000015)	(0.000015)
BUSSEG	-0.000128***	-0.000128***	-0.000128***	-0.000128***	-0.000129***
	(0.000014)	(0.000014)	(0.000014)	(0.000014)	(0.000014)
FIRM AGE	-0.000023***	-0.000023***	-0.000023***	-0.000023***	-0.000023***
	(0.000001)	(0.000001)	(0.000001)	(0.000001)	(0.000001)
$M\&A_{t1}$	0.000061.	0.000060.	0.000060.	0.000060.	0.000060.
	(0.000032)	(0.000032)	(0.000032)	(0.000032)	(0.000032)
SEO_{t+1}	0.000066.	0.000066.	0.000066.	0.000066.	0.000065.
	(0.000040)	(0.000040)	(0.000040)	(0.000040)	(0.000040)
DLW	0.000187***	0.000187***	0.000187***	0.000188***	0.000188***
	(0.000017)	(0.000017)	(0.000017)	(0.000017)	(0.000017)
SPI	0.005481***	0.005483***	0.005483***	0.005480***	0.005482***
	(0.000801)	(0.000801)	(0.000801)	(0.000801)	(0.000800)
Treated	-0.000120***	-0.000111***	-0.000103***	-0.000084**	-0.000067*
	(0.000026)	(0.000027)	(0.000028)	(0.000029)	(0.000029)
Post0	-0.000177*				
	(0.000074)				
Post0-1		-0.000159**			
		(0.000059)			
Post0-2			-0.000150**		
			(0.000052)		
Post0-3				-0.000192***	
				(0.000049)	
Post0-4					-0.000222***
					(0.000047)
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes
Observations	37,714	37,714	37,714	37,714	37,714
R2	0.31767	0.31770	0.31774	0.31777	0.31779
Within R2	0.02690	0.02695	0.02700	0.02706	0.02709

Notes: The table presents OLS regressions for the effect of data breaches on vague language: Panel A: the use of uncertainty words, "UNC"; Panel B: the use of weak modal words, "POSS," with control variables selected by Li (2010). Column(1) Post0, Column (2), Column (3) Post0-2, Column (4) Post0-3, Column (5) Post0-4 are indicator variables that take the value of one if a firm disclosed a data breach in the current year, the current year and the previous year, the current year and the previous two, three, four years. "Treated" takes the value of one if a firm was ever affected by a data breach and zero otherwise. Data breaches are included if the number of affected records is known and at least 1,000. The construction of the variables is described in detail in Appendix A Table A.3. We include industry-by-year fixed effects. Standard errors are reported in parenthesis and adjusted for heteroskedasticity and clustering at the firm level, and ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

5.3 Effect of Data Breaches on Annual Report Readability

Tables 16 present the results for the effect of a breached firm on the readability of the 10-K annual accounts using the controls Li (2010) used as in Section 4. Table 16, Panel A uses our first readability proxy *ln(NETFILESIZE)*, which is the natural log of the net file size of the 10-K corporate report. We find that breached firms increase the complexity of their 10-K annual reports the year of the breach and the following year *(Post0-1)* and the year of the breach and the following two, three, and four years. We find the effect is more pronounced for years after the year of the incident. Table 17 presents the findings using our second readability proxy, *Bog Index*. Again, we find breached firms use more complex language in all of our regression models and that the effect is gradually more significant. We can, therefore, accept our third hypothesis stating our sample of breached firms exploit readability as a way of obscuring "bad news," i.e., suffering a data breach by deliberate changes in textual sophistication.

Table 16. Effect of Data Breaches on Annual Report Readability

Panel A: Effect of Data Breaches on Document Size					
	(1)	(2)	(3)	(4)	(5)
	ln(NETFILESIZE)	ln(NETFILESIZE)	ln(NETFILESIZE)	ln(NETFILESIZE)	ln(NETFILESIZE)
	-0.167350***	-0.167303***	-0.167256***	-0.167196***	-0.167188***
EARN	(0.008080)	(0.008079)	(0.008079)	(0.008079)	(0.008079)
	0.043109***	0.043004***	0.042868***	0.042826***	0.042765***
TACC	(0.012914)	(0.012912)	(0.012911)	(0.012912)	(0.012912)
	0.060947***	0.060940***	0.060934***	0.060932***	0.060933***
SIZE	(0.000634)	(0.000634)	(0.000634)	(0.000634)	(0.000634)
	-0.008550***	-0.008548***	-0.008544***	-0.008575***	-0.008570***
RET	(0.002057)	(0.002056)	(0.002056)	(0.002056)	(0.002056)
	0.153338***	0.153271***	0.153174***	0.153132***	0.153095***
BTM	(0.004127)	(0.004127)	(0.004127)	(0.004127)	(0.004127)
	0.093945***	0.093908***	0.093863***	0.093832***	0.093816***
STD_RET	(0.004247)	(0.004247)	(0.004247)	(0.004248)	(0.004247)
	0.086770***	0.086803***	0.086752***	0.086678***	0.086607***
STD_EARN	(0.010968)	(0.010968)	(0.010966)	(0.010966)	(0.010966)
	0.006191***	0.006213***	0.006232***	0.006228***	0.006216***
GEOSEG	(0.001337)	(0.001337)	(0.001337)	(0.001337)	(0.001337)
	0.004379**	0.004396**	0.004420**	0.004435**	0.004453**
BUSSEG	(0.001415)	(0.001415)	(0.001415)	(0.001415)	(0.001416)
	-0.000319***	-0.000320***	-0.000321***	-0.000322***	-0.000322***
FIRM AGE	(0.000063)	(0.000063)	(0.000063)	(0.000063)	(0.000063)
	0.004041	0.004084	0.004107	0.004106	0.004136
$M\&A_{t1}$	(0.003229)	(0.003229)	(0.003228)	(0.003228)	(0.003228)
	0.013923***	0.013911***	0.013916***	0.013935***	0.013946***
SEO_{t+1}	(0.003683)	(0.003683)	(0.003683)	(0.003683)	(0.003683)
	0.026871***	0.026857***	0.026838***	0.026826***	0.026822***
DLW	(0.001786)	(0.001786)	(0.001786)	(0.001786)	(0.001786)
	-1.229919***	-1.229729***	-1.229498***	-1.229537***	-1.229809***
SPI	(0.078988)	(0.078982)	(0.078977)	(0.078979)	(0.078978)
	0.030645***	0.029290***	0.027684***	0.026994***	0.026445***
Treated	(0.002849)	(0.002958)	(0.003083)	(0.003208)	(0.003335)
	0.009461				
Post0	(0.007059)				
		0.014030*			
Post0-1		(0.005771)			
			0.017581***		
Post0-2			(0.005171)		
				0.016738***	
Post0-3				(0.004909)	
					0.016201***
Post0-4					(0.004759)
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes
Observations	37,714	37,714	37,714	37,714,	37,714
R2	0.41377	0.41382	0.41389	0.41389	0.41389
Within R2	0.27374	0.27380	0.27388	0.27388	0.27389

Panel B: Effect of Data Breaches on Bog Index					
	(1)	(2)	(3)	(4)	(5)
	Bog Index	Bog Index	Bog Index	Bog Index	Bog Index
EARN	-5.414445***	-5.413995***	-5.412715***	-5.409892***	-5.408584***
	(0.256114)	(0.256121)	(0.256117)	(0.256128)	(0.256121)
TACC	2.991059***	2.987665***	2.983036***	2.979628***	2.975297***
	(0.410893)	(0.410902)	(0.410901)	(0.410895)	(0.410870)
SIZE	1.031011***	1.030955***	1.030762***	1.030650***	1.030510***
	(0.021378)	(0.021379)	(0.021378)	(0.021377)	(0.021375)
RET	0.118244.	0.118358.	0.118493.	0.117239.	0.117209.
	(0.065092)	(0.065094)	(0.065091)	(0.065093)	(0.065089)
BTM	2.433951***	2.433134***	2.430805***	2.428440***	2.424791***
	(0.136587)	(0.136596)	(0.136607)	(0.136608)	(0.136599)
STD_RET	1.781015***	1.780780***	1.779117***	1.777126***	1.775213***
	(0.135678)	(0.135682)	(0.135668)	(0.135679)	(0.135661)
STD_EARN	2.278205***	2.278262***	2.277690***	2.275445***	2.271664***
	(0.361535)	(0.361532)	(0.361519)	(0.361513)	(0.361487)
GEOSEG	-0.246509***	-0.246572***	-0.245977***	-0.245799***	-0.245825***
	(0.048869)	(0.048874)	(0.048878)	(0.048883)	(0.048885)
BUSSEG	0.476471***	0.476798***	0.477574***	0.478319***	0.479280***
	(0.050580)	(0.050583)	(0.050590)	(0.050597)	(0.050604)
FIRM AGE	-0.004254*	-0.004271*	-0.004289*	-0.004332*	-0.004358*
	(0.002152)	(0.002152)	(0.002151)	(0.002151)	(0.002150)
$M\&A_{t1}$	0.502510***	0.504100***	0.504811***	0.504459***	0.506377***
	(0.100410)	(0.100409)	(0.100412)	(0.100373)	(0.100364)
SEO_{t+1}	0.100126	0.099742	0.099909	0.100577	0.101167
	(0.113383)	(0.113394)	(0.113400)	(0.113405)	(0.113403)
DLW	0.814164***	0.813935***	0.813432***	0.812798***	0.812037***
	(0.059725)	(0.059723)	(0.059714)	(0.059704)	(0.059693)
SPI	-18.311854***	-18.310482***	-18.305186***	-18.291780***	-18.305939***
	(2.621773)	(2.621904)	(2.621563)	(2.621018)	(2.620763)
Treated	-0.164586.	-0.177864.	-0.231286*	-0.286506**	-0.352693**
	(0.098314)	(0.101722)	(0.105372)	(0.108608)	(0.111183)
Post0	0.525571.				
	(0.272001)				
Post0-1		0.382486.			
		(0.219944)			
Post0-2			0.522621**		
			(0.196364)		
Post0-3				0.627054***	
				(0.185273)	
Post0-4					0.751785***
					(0.179784)
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes
Observations	37,714	37,714	37,714	37,714	37,714
R2	0.41377	0.41382	0.41389	0.41389	0.41389
Within R2	0.27374	0.27380	0.27388	0.27388	0.27389

Notes: Table 17 presents OLS regressions for the effect of data breaches on Annual Report Readability: Panel A: on the natural log of net file size, "ln(NETFILESIZE)"; Panel B: on the Bog Index with control variables selected by Li (2010). Column(1) Post0, Column (2) Post0-1, Column (3) Post0-2, Column (4) Post0-3, Column (5) Post0-4 are indicator variables that take the value of one if a firm disclosed a data breach in the current year, the current year and the previous year, the current year and the previous two, three, four years. "Treated" takes the value of one if a firm was ever affected by a data breach and zero otherwise. Data breaches are included if the number of affected records is known and at least 1,000. The construction of the variables is described in detail in Appendix A Table A.3. We include industry-by-year fixed effects. Standard errors are reported in parenthesis and adjusted for heteroskedasticity and clustering at the firm level, and ***,**,* denote the significance levels at the 1%,5% and 10% levels, respectively.

6. DISCUSSION & CONCLUSION

We examine public US companies' responses to data breaches, specifically the extent to which these events are reflected in narratives included in company public disclosures. We focus on the concept of opportunistic managerial discretionary disclosure behaviour that results in biased reporting or "cheap talk". We explore whether there is a change in the subsequent linguistic cues of the 10-K annual accounts. Specifically, we investigate whether breached firms use optimistic (abnormal) tone, vague language (use of uncertainty and weak modal words) and complexity/readability (length of net file size and bog index).

We first examine the characteristics of breached firms, and we find that compared to our control sample, breached firms are generally bigger, older, higher valued and have greater current performance. We then conduct logistic regressions and again confirm that those firms with greater recognition, market presence and profitability are more likely of being breached. However, our R-squared is low, indicating that observable firm characteristics are somehow ineffective at forecasting data breaches. We then explore the effect of data breaches on financial performance and find that corporate data breaches have a significant long-lasting impact on firm value and firm profitability. Firms M/B ratio, for example, is impacted for three to four years after the announcement of a data breach. Third, we construct our linguistic cues, including abnormal tone, following Huang et al. (2014). We find, just as Huang et al. (2014), that an abnormal tone predicts future negative future earnings and cash flows. We then explore the determinants of our linguistic cues and find similar results to previous research (include reference). Finally, we examine whether our linguistic cues are affected by data breaches and find significant coefficients in nearly all our regressions. Specifically, positive tone and complexity increases, and uncertainty decrease in corporate narratives in line with the opportunistic behaviour advanced in prior research (Guo et al., 2017; Hart et al., 2013; Lo et al., 2017). Overall, these results confirm our expectations that companies facing a data breach try to obfuscate the language in their annual reports to alleviate the negative impact of this incident.

Our results are of interest to market participants who can better evaluate the firm's performance and future uncertainty regarding information security by closely examining the linguistic cues that incorporate narratives. Our analysis is also useful to managers involved with disclosure decisions in that narratives and their decisions on how to treat their quality are relevant for users.

This paper has limitations. First, in addition to a binary indicator of breach announcement, we also considered using the actual contents of the breach announcements to get more granular information and details of the breach. This could be analysed in future research.

Other potential extensions are as follows. First, in our paper, we do not directly consider the market reactions given the opportunistic use of linguistic cues in the narratives. The text-mining analysis of business risk factors related to the breaches can also provide insights into how these risks affect different businesses. Last, as different types of media, such as social media and blogs, becomes more popular information sources for investors, we could extend our analysis to investigate the relation among different information sources, information security incidents, and stock price reactions.

REFERENCES

Abrahamson, E., & Park, C. (1994). Concealment of Negative Organizational Outcomes: An Agency Theory Perspective. *Academy of Management Journal*, *37*(5), 1302–1334. https://doi.org/10.5465/256674

Ackerman, A. (2015, May 8). *Cyberattacks Represent Top Risk, SEC Chief Says* [Press release]. https://www.wsj.com/articles/cyberattacks-represent-top-risk-sec-chief-says-1431097038

Agrafiotis, I., Nurse, J. R. C., Goldsmith, M., Creese, S., & Upton, D. (2018). A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity*, *4*(1), 1–15. https://doi.org/10.1093/cybsec/tyy006

AICPA. (2018). *Cybersecurity risk management reporting fact sheet*. American Institute of Certified Public Accountants. https://www.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/cybersecurity-fact-sheet.pdf

Akey, P., Lewellen, S., & Liskovich, I. (2018). Hacking Corporate Reputations. *SSRN Electronic Journal*, 1–47. https://doi.org/10.2139/ssrn.3143740

Barnes, S. (2018, March 29). *There are two types of companies: Those who know they've been hacked & those who don't* [Press release]. https://dynamicbusiness.com.au/topics/technology/there-are-two-types-of-companies-those-who-know-theyve-been-hacked-those-who-dont.html

Bell, A., & Jones, K. (2014). Explaining Fixed Effects: Random Effects Modelling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Bloomfield, R. J. (2002). The "Incomplete Revelation Hypothesis" and Financial Reporting. *Accounting Horizons*, *16*(3), 233–243. https://doi.org/10.2308/acch.2002.16.3.233

Bonsall, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, *63*(2–3), 329–357. https://doi.org/10.1016/j.jacceco.2017.03.002

Bonsall, S. B., & Miller, B. P. (2017). The impact of narrative disclosure readability on bond ratings and the cost of debt. *Review of Accounting Studies*, *22*(2), 608–643. https://doi.org/10.1007/s11142-017-9388-0

Bozkus Kahyaoglu, S., & Caliyurt, K. (2018). Cybersecurity assurance process from the internal audit perspective. *Managerial Auditing Journal*, *33*(4), 360–376. https://doi.org/10.1108/maj-02-2018-1804

Brill, D. (1992). Setting the right tone. Writer's Digest, 72(9), 32-36.

Campbell, K., Gordon, L. A., Loeb, M. P., & 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

Clarkson, P. M., Kao, J. L., & Richardson, G. D. (1994). The Voluntary Inclusion of Forecasts in the MD&A Section of Annual Reports. *Contemporary Accounting Research*, *11*(1), 423–450. https://doi.org/10.1111/j.1911-3846.1994.tb00450.x

Clatworthy, M., & Jones, M. J. (2001). The effect of thematic structure on the variability of annual report readability. *Accounting, Auditing & Accountability Journal*, *14*(3), 311–326. https://doi.org/10.1108/09513570110399890

Clatworthy, M., & Jones, M. J. (2003). Financial reporting of good news and bad news: evidence from accounting narratives. *Accounting and Business Research*, *33*(3), 171–185. https://doi.org/10.1080/00014788.2003.9729645

Core, J. E. (2001). A Review of the Empirical Disclosure Literature: Discussion. *SSRN Electronic Journal*, 31(1/3): 441–456. https://doi.org/10.2139/ssrn.258513

Courtis, J. K. (1986). An Investigation into Annual Report Readability and Corporate Risk-Return Relationships. *Accounting and Business Research*, *16*(64), 285–294. https://doi.org/10.1080/00014788.1986.9729329

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012a). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language*. *Contemporary Accounting Research*, *29*(3), 845–868. https://doi.org/10.1111/j.1911-3846.2011.01130.x

Davis, A. K., Piger, J. M., & Sedor, L. M. (2012b). Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language*. *Contemporary Accounting Research*, *29*(3), 845–868. https://doi.org/10.1111/j.1911-3846.2011.01130.x

De Franco, G., Hope, O.-K., Vyas, D., & Zhou, Y. (2014). Analyst Report Readability. Contemporary Accounting Research, 32(1), 76–104. https://doi.org/10.1111/1911-3846.12062

Dye, R. A. (1985). Disclosure of Nonproprietary Information. *Journal of Accounting Research*, *23*(1), 123. https://doi.org/10.2307/2490910

Eshbaugh-Soha, M. (2013). Political Tone: How Leaders Talk and Why, by Roderick P. Hart, Jay P. Childers, and Colene J. Lind. *Political Communication*, *30*(4), 658–660. https://doi.org/10.1080/10584609.2013.835690

Ettredge, M., Guo, F., & Li, Y. (2018). Trade secrets and cybersecurity breaches. *Journal of Accounting and Public Policy*, *37*(6), 564–585. https://doi.org/10.1016/j.jaccpubpol.2018.10.006

Frazier, L., Taft, L., Roeper, T., Clifton, C., & Ehrlich, K. (1984). Parallel structure: A source of facilitation in sentence comprehension. *Memory & Cognition*, *12*(5), 421–430. https://doi.org/10.3758/bf03198303

Firtel, K.B., 1999, Plain English: A reappraisal of the intended audience of disclosure under the securities act of 1933, *Southern California Law Review* 72, 851–897.

Gansler, J. S., & Lucyshyn, W. (2005). Improving the security of financial management systems: What are we to do? *Journal of Accounting and Public Policy*, *24*(1), 1–9. https://doi.org/10.1016/j.jaccpubpol.2004.12.001

Gibbins, M., Richardson, A., & Waterhouse, J. (1990). The Management of Corporate Financial Disclosure: Opportunism, Ritualism, Policies, and Processes. *Journal of Accounting Research*, *28*(1), 121. https://doi.org/10.2307/2491219

Gordon, L. A., Loeb, M. P., & Lucyshyn, W. (2003). Sharing information on computer systems security: An economic analysis. *Journal of Accounting and Public Policy*, *22*(6), 461–485. https://doi.org/10.1016/j.jaccpubpol.2003.09.001

Gordon, L. A., Loeb, M. P., Lucyshyn, W., & Sohail, T. (2006). The impact of the Sarbanes-Oxley Act on the corporate disclosures of information security activities. *Journal of Accounting and Public Policy*, *25*(5), 503–530. https://doi.org/10.1016/j.jaccpubpol.2006.07.005

Gordon, L. A., Loeb, M. P., Lucyshyn, W., & Zhou, L. (2015). Externalities and the Magnitude of Cyber Security Underinvestment by Private Sector Firms: A Modification of the Gordon-Loeb Model. *Journal of Information Security*, *06*(01), 24–30. https://doi.org/10.4236/jis.2015.61003

Gordon, Loeb, & Sohail. (2010). Market Value of Voluntary Disclosures Concerning Information Security. *MIS Quarterly*, *34*(3), 567. https://doi.org/10.2307/25750692

Gow, I. D., Ormazabal, G., & Taylor, D. J. (2010). Correcting for Cross-Sectional and Time-Series Dependence in Accounting Research. *The Accounting Review*, *85*(2), 483–512. https://doi.org/10.2308/accr.2010.85.2.483

Grossman, S. J. (1981). The Informational Role of Warranties and Private Disclosure about Product Quality. *The Journal of Law and Economics*, *24*(3), 461–483. https://doi.org/10.1086/466995

Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine learning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, *2*(3), 153–170. https://doi.org/10.1016/j.jfds.2017.02.001

Guo, W., Yu, T., & Gimeno, J. (2017). Language and Competition: Communication Vagueness, Interpretation Difficulties, and Market Entry. *Academy of Management Journal*, *60*(6), 2073–2098. https://doi.org/10.5465/amj.2014.1150

Haapamäki, E., & Sihvonen, J. (2019). Cybersecurity in accounting research. *Managerial Auditing Journal*, *34*(7), 808–834. https://doi.org/10.1108/maj-09-2018-2004

Hales, J., Kuang, X. J., & Venkataraman, S. (2010). Who Believes the Hype? An Experimental Examination of How Language Affects Investor Judgments. *SSRN Electronic Journal, 49(1), 233–255.* https://doi.org/10.2139/ssrn.1667244

Hart, R. P., Childers, J. P., & Lind, C. J. (2013). Political Tone: How Leaders Talk and Why. *Political Communication*, *30*(4), 658–660. https://doi.org/10.1080/10584609.2013.835690

Hasan, M. M. (2018). Readability of Narrative Disclosures in 10-K Reports: Does Managerial Ability Matter? *European Accounting Review*, *29*(1), 147–168. https://doi.org/10.1080/09638180.2018.1528169

Hausken, K. (2006). Income, interdependence, and substitution effects affecting incentives for security investment. *Journal of Accounting and Public Policy*, *25*(6), 629–665. https://doi.org/10.1016/j.jaccpubpol.2006.09.001

Hausken, K. (2007). Information sharing among firms and cyber attacks. *Journal of Accounting and Public Policy*, *26*(6), 639–688. https://doi.org/10.1016/j.jaccpubpol.2007.10.001

Hawdon, J., Parti, K., & Dearden, T. E. (2020). Cybercrime in America amid COVID-19: the Initial Results from a Natural Experiment. American Journal of Criminal Justice, 45(4), 546–562https://doi.org/10.1007/s12103-020-09534-4

Hayden, E. (2013, May 1). *Data breach protection requires new barriers*. SearchSecurity. https://searchsecurity.techtarget.com/feature/Data-breach-protection-requires-new-barriers

Henry, E. (2008). Are Investors Influenced By How Earnings Press Releases Are Written? *Journal of Business Communication*, *45*(4), 363–407. https://doi.org/10.1177/0021943608319388

Higgs, J. L., Pinsker, R. E., Smith, T. J., & Young, G. R. (2016). The Relationship between Board-Level Technology Committees and Reported Security Breaches. *Journal of Information Systems*, *30*(3), 79–98. https://doi.org/10.2308/isys-51402

Huang, X., Teoh, S. H., & Zhang, Y. (2011). Tone Management. *SSRN Electronic Journal*, 1083–1113. https://doi.org/10.2139/ssrn.2024674

Impression management in financial reporting. Evidence from the UK and Spain (Unpublished doctoral dissertation, University College, Dublin). (2006). Guillamon-Saorin, E.

Institute of Internal Auditors (IIA). (2018). *"The future of cybersecurity in internal audit. A joint research report by the internal audit foundation and Crowe Horwath."* https://www.crowe.com/-/media/Crowe/LLP/folio-pdf/The-Future-of-Cybersecurity-in-IA-RISK-18000-002A-update.pdf

Islam, M. S., Farah, N., & Stafford, T. F. (2018). Factors associated with security/cybersecurity audit by internal audit function. *Managerial Auditing Journal*, *33*(4), 377–409. https://doi.org/10.1108/maj-07-2017-1595

Jasinski, J. L. (2001). Sourcebook on Rhetoric: Key Concepts in Contemporary Rhetorical Studies. Sage Publications, Inc.

Jones, M. J. (1988). A Longitudinal Study of the Readability of the Chairman's Narratives in the Corporate Reports of a UK Company. *Accounting and Business Research*, *18*(72), 297–305. https://doi.org/10.1080/00014788.1988.9729377

Juliana De Groot, D. G. J. (2020, December 1). *The History of Data Breaches*. Digital Guardian. https://digitalguardian.com/blog/history-data-breaches

Kamhoua, C. (2015, June 1). *An evolutionary game-theoretic framework for cyber-threat information sharing*. IEEE Conference Publication. https://ieeexplore.ieee.org/document/7249499/;jsessionid=zzTaDdS8GlWVpFsCIQosQcS0FxL7TK-FDk1XPTtqUsNzYMTvziKJ!696014846?tp=&arnumber=7249499

Kamiya, S., Kang, J. K., Kim, J., Milidonis, A., & Stulz, R. M. (2021). Risk management, firm reputation, and the impact of successful cyberattacks on target firms. *Journal of Financial Economics*, *139*(3), 719–749. https://doi.org/10.1016/j.jfineco.2019.05.019

Kamiya, S., Kang, J.-K., Kim, J., Milidonis, A., & Stulz, R. M. (2018). What is the Impact of Successful Cyberattacks on Target Firms? *SSRN Electronic Journal*, 1. https://doi.org/10.2139/ssrn.3135514

Kim, M. S. (2018). The Effect of Uncertain and Weak Modal Words in 10-K Filings on Analyst Forecast Attributes. *FIU Electronic Theses and Dissertations. 3786.*, 3786. https://doi.org/10.25148/etd.fidc006848

Kothari, S. P., Li, X., & Short, J. E. (2009). The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. *The Accounting Review*, *84*(5), 1639–1670. https://doi.org/10.2308/accr.2009.84.5.1639

Lainhart, J. W. (2000). COBITTM: A Methodology for Managing and Controlling Information and Information Technology Risks and Vulnerabilities. *Journal of Information Systems*, *14*(s-1), 21–25. https://doi.org/10.2308/jis.2000.14.s-1.21

Law, K. K. F., & Mills, L. F. (2015). Taxes and Financial Constraints: Evidence from Linguistic Cues. *Journal of Accounting Research*, *53*(4), 777–819. https://doi.org/10.1111/1475-679x.12081

Layton, R., & Watters, P. A. (2014). A methodology for estimating the tangible cost of data breaches. Journal of Information Security and Applications, 19(6), 321–330. https://doi.org/10.1016/j.jisa.2014.10.012

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45(2–3), 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003

LI, F. (2010a). The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach. Journal of Accounting Research, 48(5), 1049–1102. https://doi.org/10.1111/j.1475-679x.2010.00382.x

Li, F. (2010b). Textual analysis of corporate disclosures: A survey of the literature. Journal of Accounting Literature, 29: 143-165

Lo, K., Ramos, F., & Rogo, R. (2017). Earnings management and annual report readability. Journal of Accounting and Economics, 63(1), 1–25. https://doi.org/10.1016/j.jacceco.2016.09.002

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2014). Measuring Readability in Financial Disclosures. *The Journal of Finance*, *69*(4), 1643–1671. https://doi.org/10.1111/jofi.12162

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. Journal of Accounting Research, 54(4), 1187–1230. https://doi.org/10.1111/1475-679x.12123

Lehavy, R., Li, F., & Merkley, K. (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. *The Accounting Review*, *86*(3), 1087–1115. https://doi.org/10.2308/accr.00000043

Merkley, K. J. (2013). Narrative Disclosure and Earnings Performance: Evidence from R&D Disclosures. *The Accounting Review*, *89*(2), 725–757. https://doi.org/10.2308/accr-50649

Merkl-Davies, Doris M. and Brennan, Niamh M., Discretionary Disclosure Strategies in Corporate Narratives: Incremental Information or Impression Management?. *Journal of Accounting Literature, Vol. 27, 2007*, pp. 116-196, Available at SSRN: https://ssrn.com/abstract=1089447

Milgrom, P. R. (1981). Good News and Bad News: Representation Theorems and Applications. *The Bell Journal of Economics*, *12*(2), 380. https://doi.org/10.2307/3003562

Miller, B. P. (2010). The Effects of Reporting Complexity on Small and Large Investor Trading. The Accounting Review, 85(6), 2107–2143. https://doi.org/10.2308/accr.00000001

Mitts, J. (2020). Insider Trading and Strategic Disclosure. SSRN Electronic Journal, 1–10. https://doi.org/10.2139/ssrn.3741464

Monitor, I. S. (2006). Managing Cybersecurity Resources: A Cost-Benefit Analysis. Information & Security: An International Journal, 18, 137–147. https://doi.org/10.11610/isij.1808

Parker, L. D. (1982). Corporate Annual Reporting: A Mass Communication Perspective. Accounting and Business Research, 12(48), 279–286. https://doi.org/10.1080/00014788.1982.9728820

Rapoport, M., & Andriotis, A. (2017, September 12). Equifax Lobbied for Easier Regulation Before Data Breach [Press release]. https://www.wsj.com/articles/equifax-lobbied-for-easier-regulation-before-data-breach-1505169330

Ponemon Institute. (2020). Cost of Data Breach Report 2020. https://www.ibm.com/account/reg/us-en/signup?formid=urx-46542

Privacy Rights Clearinghouse | Privacy Rights Clearinghouse. (n.d.). Privacy Rights Clearinghouse. Retrieved May 25, 2020, from https://privacyrights.org/

Protiviti. (2016). *Executive perspectives on top risks for 2016.* https://erm.ncsu.edu/az/erm/i/chan/library/NC-State-Protiviti-Survey-Top-Risks-2016.pdf

Rogers, J. L., Van Buskirk, A., & Zechman, S. L. C. (2011). Disclosure Tone and Shareholder Litigation. The Accounting Review, 86(6), 2155–2183. https://doi.org/10.2308/accr-10137

Securities and Exchange Commission. (2018). Commission Statement and Guidance on Public Company Cybersecurity Disclosures. Release Nos. 33–10459; 34–82746. https://www.sec.gov/rules/interp/2018/33-10459.pdf

Smith, M., & Taffler, R. (1992). THE CHAIRMAN'S STATEMENT AND CORPORATE FINANCIAL PERFORMANCE. Accounting & Finance, 32(2), 75–90. https://doi.org/10.1111/j.1467-629x.1992.tb00187.x

Smith, M., & Taffler, R. J. (2000). The chairman's statement - A content analysis of discretionary narrative disclosures. *Accounting, Auditing & Accountability Journal*, *13*(5), 624–647. https://doi.org/10.1108/09513570010353738

Spanos, G., & Angelis, L. (2016). The impact of information security events to the stock market: A systematic literature review. *Computers & Security*, *58*, 216–229. https://doi.org/10.1016/j.cose.2015.12.006

TAN, H. U. N.-T. O. N. G., YING WANG, E. L. A. I. N. E., & ZHOU, B. O. (2014). When the Use of Positive Language Backfires: The Joint Effect of Tone, Readability, and Investor Sophistication on Earnings Judgments. Journal of Accounting Research, 52(1), 273–302. https://doi.org/10.1111/1475-679x.12039

Taylor, H. (2015, December 28). *Biggest cybersecurity threats in 2016* [Press release]. https://www.cnbc.com/2015/12/28/biggest-cybersecurity-threats-in-2016.html

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, *62*(3), 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. The Journal of Finance, 63(3), 1437–1467. https://doi.org/10.1111/j.1540-6261.2008.01362.x

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics*, *5*, 179–194. https://doi.org/10.1016/0165-4101(83)90011-3

The DG Data Trends Report. (2020, May). https://info.digitalguardian.com/rs/768-OQW-145/images/digital-guardian-data-trends-report-may-2020.pdf?

Wang, T., Kannan, K. N., & Ulmer, J. R. (2013). The Association Between the Disclosure and the Realization of Information Security Risk Factors. Information Systems Research, 24(2), 201–218. https://doi.org/10.1287/isre.1120.0437

Xu, H., Guo, S. Y., Haislip, J. Z., & Pinsker, R. E. (2019). Earnings Management in Firms with Data Security Breaches. Journal of Information Systems, 33(3), 267–284. https://doi.org/10.2308/isys-52480

APPENDIX A: VARIABLE DEFINITIONS

Table. A.1 Variable Definitions for Data Breach Characteristics

Variable	Description	Source
Asset Intangibility	1-total property, plant and equipment (ppent)/ total assets (at)	Compustat
In(assets)	Natural logarithm of total assets (at)	Compustat
BIG4	Indicator variable takes the value of 1 if the firm is audited by BIG4 audit firm (au)	Compustat
Foreign	Indicator variable takes the value of 1 if the firm has foreign operations different than 0 (fca)	Compustat
TobinsQ	(Total assets (at)- stockholders equity (ceq) + market value of equity (prcc_f*csho))/total assets (at)	Compustat
EARN	Earnings before income and extraordinary items scaled by total assets	Compustat
LOSS	Indicator variable equals 1 if EARN <0	Compustat
Growth	$Sales_t$-$sales_{t-1}$/$sales_{t-1}$	Compustat
RET	contemporaneous annual stock returns calculated using CRSP monthly return data	CRSP
STD_RET	standard deviation of monthly stock returns over the fiscal year	CRSP
Leverage	Short term and long-term debt scaled by total assets(dltt+dlc)/at	Compustat
R&D/assets	Max(0, R&D expenditures (xrd))/total assets (at)	Compustat
CAPEX/assets	Capital expenditures (capx)/total assets	Compustat
Fortune500	Indicator variable equals 1 if the firm is considered a Fortune company in a given year	Compustat
GEOSEG	Natural logarithm(1+ number of geographic segments)	Compustat
BUSSEG	Natural logarithm(1+ number of business segments)	Compustat
Extraordinary items	Indicator variable equals 1 if extraordinary items (xi) are not equal to 0	Compustat
BTM	Total assets (at) /(market value (prcc_f*csho) + total liabilities (lt))	Compustat
SIZE	Natural logarithm market value of equity at fiscal yearend (prcc_f *csho)	Compustat
TACC	Income before extraordinary items (ibc)–operating activities/net cash flow (oancf) scaled by lagged total assets	Compustat
M&A	Indicator variable equals 1 if the amount of acquisition (AQC) is greater than 10% of beginning total assets	Compustat
WWindex	Higher values of the WW index imply greater levels of financial constraint. Constructed following Whited Wu (2006). WW = -0-091CF -0.062DIVPOS +0.021TLTD – 0.044LNTA + 0.102ISG – 0.035SG Where: CF = [income before extraordinary items (ib) + depreciation (dp)]/total assets (at) DIVPOS = indicator set to 1 if dividends (dvc+dvp) are positive, and 0 otherwise TLTD= long term debt (dltt)/total assets (at) LNTA = ln(total assets (at)) SG = sale (sale)/lagged sale where; ISG = average industry SG for each 2-digit SIC Industry each year.	Compustat
Credit Rating	S&P Quality Ranking (spcsrc)	Compustat
Special Items	Number of special items scaled by lagged total assets	Compustat

Table.A2 Variable Definitions on Effect on Firm Value

Variable	Description	Source
In(assets)	Natural logarithm of total assets (at)	Compustat
$In(assets)^2$	Natural logarithm of total assets (at) squared	Compustat
Market Leverage	Total debt in current liabilities and long-term debt (dlc +dltt)/ (Total debt in current liabilities and long-term debt (dlc +dltt) + market value of equity (prcc_f*csho)	Compustat
MTB	Market equity (prcc_f*csho)/book value of equity	Compustat
ROE	Earnings before extraordinary items (ib)/lagged book value of equity	Compustat
P/E	Market value of equity / earnings before extraordinary items (ib)	Compustat

Table. A.3 Variable Definitions for linguistic cues models following Huang et al. (2014)

Variable	Description	Source
EARN	Earnings before extraordinary items (ib) scaled by total assets (at).	Compustat
RET	The contemporaneous annual stock returns measured using the CRSP monthly stock database, ending three months after fiscal year-end.	CRSP
∆EARN	Change in earnings (ib) from the prior year scaled by total assets (at).	Compustat
SIZE	The logarithm of the market value of equity at fiscal yearend (prcc_f*csho).	Compustat
BTM	Total assets (at) over the market value of equity (prcc_f*csho) and total liabilities (lt).	Compustat
STD_RET	The standard deviation of the monthly stock CRSP database over the fiscal year, ending three months after fiscal year-end.	CRSP
STD_EARN	The standard deviation of *EARN* over the preceding five years, with a minimum of three years of data required.	Compustat
AGE	Log of 1 plus the first year the firm entered the CRSP dataset.	CRSP
BUSSEG	Log of 1 plus number of business segments, or 1 if the item is missing in Compustat	Compustat
GEOSEG	Log of 1 plus number of business segments, or 1 if the item is missing in Compustat	Compustat
LOSS	1 if earnings are smaller than 0, 0 otherwise.	Compustat
DA	Discretionary accruals (*DA)* are constructed using the cross-sectional modified Jones model again following Huang et al. (2014). Our study span of 2000-2018 allows us to quantify accruals using SFAS No.95 statement of cash flow rather than the balance sheet approach, which Hribar and Collins (2002) suggest is less accurate. $TACC_{jt} = EBEI_{jt} - (CFO_{jt} - EIDO_{jt})$ TACC = Total accruals EBEI = income before extraordinary items CFO = cash flow from operations; and EIDO = extraordinary items and discontinued operations included in CFO for each firm j in year t. We run the following regression for each industry in the Fama French 48 Industry Classification and extract the regression residuals to construct discretionary accruals. $TACC_{jt} = \beta_0(1/Assets_{jt-1}) + \beta_1(\Delta SALES_{jt} - \Delta AR_{jt}) + \beta_2 PEE_{jt} + V_{jt}$	Compustat
ADACC	The absolute value of discretionary accruals	Compustat
SEO_{t+1}	1 when the Sale of Common and Pref. Stock (sstk) one year after the earnings press release is greater than 10% of beginning total assets (at)	Compustat
$M\&A_{t+1}$	1 if the amount of acquisition (aqc) in one year after the earnings press release is greater than 10 per cent of beginning total assets and is 0 otherwise	Compustat
TACC	Cash flow of income before extraordinary items (ibc)- operating cash flow (oancf) scaled by beginning total assets (at)	Compustat
R&D	R&D expenditure (xrd) scaled by beginning total assets (at)	Compustat
CAPEX	Capital expenditure (capx) scaled by beginning total assets (at)	Compustat
CFO	Operating cash flows (oancf)/beginning total assets (at)	Compustat
SPI	Special items (spi) /beginning total assets (at)	Compustat
Asset intangibility	1-property plant & equipment (ppent) over beginning total assets (at)	Compustat
DLW	1 if firm is incorporated in Delaware; 0 otherwise	Compustat

APPENDIX B: ADDITIONAL TABLES

Table B.1. Constructing Discretionary Accruals

SIC2	Fyear	Intercept	1/assets	DREV	PPE	Adj.R2
99	2019	0.03	-1.91	0.25	- 0.04	0.99
29	2000	0.04	-2.31	- 0.02	- 0.09	0.93
49	2011	-0.00	-1.87	0.03	-0.05	0.89
65	2020	- 0.09	0.11	-0.42	0.03	0.89
49	2013	0.02	- 2.07	-0.09	-0.05	0.88
65	2013	0.01	-1.61	-0.07	- 0.02	0.87
23	2014	-0.06	- 1.50	0.16	- 0.01	0.86
79	2019	-0.02	- 1.78	- 0.16	- 0.03	0.86
87	2019	-0.05	- 1.67	0.07	- 0.02	0.85
30	2012	-0.01	4.81	- 0.01	- 0.07	0.84

	TACC
1/ASSETS	-0.428**
	(0.205)
DREV	0.087***
	(0.022)
PPE	-0.185***
	(0.058)
Constant	0.050**
	(0.025)
Observations	144
R2	0.182
Adjusted R2	0.164

Table B.2. Abnormal Positive Tone and Future Financial Performance

	Panel A: Future Cash Flows and Abnormal Positive Tone		
	(1)	(2)	(3)
Dependent Var.:	CFO_{t1}	CFO_{t2}	CFO_{t3}
ABTONE	-0.455002***	-0.672002***	-0.794823***
	(0.137543)	(0.148298)	(0.156482)
DACC	-0.372881***	-0.321702***	-0.287185***
	(0.007955)	(0.008588)	(0.008704)
EARN	0.615905***	0.524829***	0.453408***
	(0.005485)	(0.005902)	(0.005976)
SIZE	0.004239***	0.005766***	0.006684***
	(0.000340)	(0.000372)	(0.000389)
RET	0.007962***	0.005932***	0.005689***
	(0.001175)	(0.001255)	(0.001306)
BTM	-0.005533**	-0.012587***	-0.014786***
	(0.002019)	(0.002210)	(0.002322)
STD_RET	-0.016205***	-0.019734***	-0.021050***
	(0.002594)	(0.002859)	(0.002982)
STD_EARN	-0.085395***	-0.090120***	-0.092357***
	(0.007298)	(0.007891)	(0.008252)
Fyear FE	Yes	Yes	Yes
SIC2 FE	Yes	Yes	Yes
Observations	44,045	41,326	38,296
R2	0.64393	0.56116	0.50165
Within R2	0.58939	0.49786	0.43567

	Panel B: Future Earnings and Abnormal Positive Tone		
	(1)	(2)	(3)
	$EARN_{t1}$	$EARN_{t2}$	$EARN_{t3}$
ABTONE	-0.309373***	-0.923558***	-1.257570***
	(0.174444)	(0.195477)	(0.203879)
DACC	-0.233190***	-0.237345***	-0.216934***
	(0.009850)	(0.010840)	(0.011364)
EARN	0.679849***	0.571735***	0.498544***
	(0.006563)	(0.007381)	(0.007787)
SIZE	0.004112***	0.006437***	0.008407***
	(0.000449)	(0.000489)	(0.000510)
RET	0.036350***	0.021546***	0.012859***
	(0.001458)	(0.001599)	(0.001697)
BTM	-0.010521***	0.003914	0.011697***
	(0.002712)	(0.002995)	(0.003109)
STD_RET	-0.065245***	-0.045719***	-0.036707***
	(0.003741)	(0.003892)	(0.003970)
STD_EARN	-0.140856***	-0.174212***	-0.179003***
	(0.009321)	(0.010612)	(0.011109)
Fyear FE	Yes	Yes	Yes
SIC2 FE	Yes	Yes	Yes
S.E.: Clustered	Firm & Fyear	Firm & Fyear	Firm & Fyear
Observations	44,083	41,374	38,382
R2	0.61969	0.51287	0.45429
Within R2	0.57181	0.45300	0.39166

Table B.3 Effect of Data Breaches on Tone using controls selected by Li (2010)

	(1)	(2)	(3)	(4)	(5)
	TONE	TONE	TONE	TONE	TONE
EARN	0.0005*	0.0005*	0.0005*	0.0005*	0.0005*
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
TACC	0.0003	0.0003	0.0003	0.0003	0.0003
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
SIZE	-0.0003***	-0.0003***	-0.0003***	-0.0003***	-0.0003***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
RET	0.0001**	0.0001**	0.0001**	0.0001**	0.0001**
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
BTM	-0.0021***	-0.0021***	-0.0021***	-0.0021***	-0.0021***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
STD_RET	-0.0014***	-0.0014***	-0.0014***	-0.0014***	-0.0014***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
STD_EARN	-0.0051***	-0.0051***	-0.0051***	-0.0051***	-0.0051***
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
GEOSEG	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
BUSSEG	0.0001.	0.0001.	0.0001.	0.0001.	0.0001.
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
FIRM AGE	0.0000***	0.0000***	0.0000***	0.0000***	0.0000***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$M\&A_{t+1}$	-0.0002*	-0.0002*	-0.0002*	-0.0002*	-0.0002*
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
SEO_{t+1}	0.0004***	0.0004***	0.0004***	0.0004***	0.0004***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
DLW	-0.0005***	-0.0005***	-0.0005***	-0.0005***	-0.0005***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
SPI	0.0499***	0.0499***	0.0499***	0.0499***	0.0499***
	(0.0021)	(0.0021)	(0.0021)	(0.0021)	(0.0021)
Treated	0.0001.	0.0001	0.0001	0.0001	0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Post0	0.0008***				
	(0.0002)				
Post01		0.0006**			
		(0.0002)			
Post012			0.0004**		
			(0.0002)		
				0.0003*	
				(0.0001)	
					0.0003*
					(0.0001)
Fyear x SIC2 FE	Yes	Yes	Yes	Yes	Yes
Observations	37,714	37,714	37,714	37,714	37,714
R2	0.29853	0.29847	0.29840	0.29835	0.29834
Within R2	0.07849	0.07841	0.07832	0.07826	0.07824