Journal of Corporate Finance 43 (2017) 58–85

Contents lists available at ScienceDirect

Journal of Corporate Finance

journal homepage: www.elsevier.com/locate/jcorpfin

Do fraudulent firms produce abnormal disclosure?

Gerard Hoberg[a], Craig Lewis[b,*]

[a] Marshall School of Business, University of Southern California, 3670 Trousdale Parkway, Los Angeles, CA 90089, United States
[b] Owen Graduate School of Management, Vanderbilt University, 401 21st Avenue South, Nashville, TN 37203, United States

ARTICLE INFO

Article history:
Received 20 September 2016
Received in revised form 2 November 2016
Accepted 15 December 2016
Available online 19 December 2016

JEL codes:
G30
G32
G38

Keywords:
Fraud
Disclosure
Text analytics
Cost of capital

ABSTRACT

Using text-based analysis of 10-K MD&A disclosures, we find that fraudulent firms produce verbal disclosure that is abnormal relative to strong counterfactuals. This abnormal text predicts fraud out of sample, has a verbal factor structure, and can be interpreted to reveal likely mechanisms that surround fraudulent behavior. Using a conservative difference-based approach, we find evidence that fraudulent managers discuss fewer details explaining the sources of the firm's performance, while disclosing more information about positive aspects of firm performance. They also provide less content relating the disclosure to the managerial team itself. We also find new interpretable verbal support for the well-known hypothesis that managers commit fraud in order to artificially lower their cost of capital.

1. Introduction

Many studies suggest that managers committing fraud likely do so to achieve various objectives such as getting access to low cost capital (Dechow et al., 1996; Povel et al., 2007 and Wang et al., 2010) or to conceal diminishing performance (Dechow et al., 2011a).[1] We examine the question of whether a firm's 10-K MD&A disclosure to the U. S. Securities and Exchange Commission (SEC) reflects the fact that a firm committed fraud. This issue should be particularly salient to managers committing fraud because the SEC is tasked with identifying and pursuing accounting, auditing and enforcement actions (AAERs) against such firms. These disclosures also are simultaneously submitted to the public, which uses them to inform its decisions to allocate capital among firms seeking financing and other resources.

Accounting fraud occurs when a firm intends to deceive shareholders by misrepresenting a *material* fact it is legally required to disclose. In most cases, there is active misrepresentation by management in the form of earnings management; however, it also is possible for misrepresentation to occur through silence on key issues. Materiality legally requires that the missing or incorrect information caused the defrauded parties to make materially different decisions (for example, they might have

* Corresponding author.
E-mail addresses: hoberg@marshall.usc.edu (G. Hoberg), craig.lewis@owen.vanderbilt.edu (C. Lewis).

[1] Dechow et al. (2010) provide a detailed review of fraud literature, and we summarize this literature in detail in Section 2 of this paper.

http://dx.doi.org/10.1016/j.jcorpfin.2016.12.007
0929-1199/

sold their shares instead of retaining them). The open-ended nature of this legal definition allows for a large number of content dimensions along which verbal disclosure can be both abnormal and material relative to counterfactuals. As a result, highly aggregated variables such as document tone are limited by their low dimensionality. For example, a fraudulent manager might have incentives to be abnormally positive on one topic, but abnormally negative on another, washing away any signal in aggregate tone. A higher dimensional data structure can identify both effects alone or as a set.

We propose that the use of verbal factor analysis specifically based on the MD&A section of the 10-K is a useful laboratory for three reasons. First, MD&A is required for all publicly traded firms, which eliminates any concern regarding selection bias, which can be an issue for alternative non-periodic disclosure platforms. Second, all firms are required to discuss the same set of required topics in MD&A, as outlined in regulation S-K and the SEC's guidance for MD&A disclosures. As a result, MD&A has a strong factor structure in verbal content. Advances in computational linguistics such as Latent Dirichlet Allocation should thus permit us to both identify the set of interpretable factors, and then identify strong non-fraudulent counterfactuals for potentially fraudulent firms. Third, we note that manipulations of revenues and expenses are the basis for most AAER fraud allegations. In MD&A, managers discuss both as part of their required discussion of annual performance, and a further requirement of MD&A is to provide detail and interpretation specifically from management's perspective. MD&A content is thus at the core of what should be most relevant in assessing the impact of fraud. Overall, MD&A should reliably contain relevant content that represents management's view, and these regularity features mitigate econometric biases inherent to other approaches.

MD&A legally represents the manager's view of the firm, but some have argued that MD&A involves not only internal management, but also outside counsel. Although the SEC cautions against such practices, the end result could be a relatively sanitized disclosure. Brown and Tucker (2011), for example, find that firms make fewer modifications to their MD&A disclosures over the 1998 to 2006 time period. While this might reduce the information content of MD&A, managers have incentives to counterbalance these effects by influencing the disclosure process either intentionally through internal firm discussions or indirectly through managerial influence on corporate culture. Overall, participation by non-management implies a conservative bias against our finding results making this an empirical question. We do find strong results despite any conservative bias, and hence the aforementioned advantages of MD&A are likely more important.

Our initial tests examine our paper's central hypothesis that MD&A text contains systematic abnormal components in the presence of fraud. We first examine whether fraudulent firms have abnormal verbal disclosure measured relative to two counterfactuals. The first is the disclosure of industry peers of similar size and age. The second applies a higher bar and focuses on within firm variation only. We compare the disclosure of each firm in the years it commits fraud to the disclosure of the same firm in the years before and after the alleged fraud. In both tests, we find strong and uniform support for our central hypothesis. This finding holds in both in-sample and out-of-sample tests, and cannot be explained by the disclosure of industry peers, various controls from the literature, or firm and year fixed effects.

Before we address the question of why fraudulent firms have a common component in their abnormal MD&A disclosures, we note that the results of our aforementioned test are important in their own right. The simple ability to improve the detection of accounting fraud is practically relevant to future researchers, investors and regulators alike. The presence of such a signal in disclosed verbal text also provides a foundation for future theoretical and empirical research to further examine related hypotheses in various firm corpora.

To understand why fraudulent firms have common abnormal MD&A disclosures, we consider three hypotheses that are derived from the Communications and Psychology literature on deception, and the associated legal requirements for successfully litigating accounting fraud cases.[2] These theories predict that managers, consciously or unconsciously, attempt to influence investor expectations by providing deceptive verbal disclosures. The underlying intuition is that, once management has chosen to commit accounting fraud, it has incentives to engage in a type of word "shell game" in which it grandstands certain aspects of performance to deflect attention away from the economic events that precipitated the fraud, and also to deflect attention away from themselves. They will also under-discuss *material* aspects of performance that would have allowed investors to make an informed choice. Based on the existing literature, which we summarize in the next section, our text-based hypotheses are as follows: fraudulent managers (1) conceal details that might expose their fraudulent accounting, (2) grandstand their good performance, and (3) disassociate themselves from the fraudulent disclosure by avoiding references to themselves in MD&A.

Because the number of topics discussed in the MD&A is large, it is necessary to categorize its content. With this in mind, we employ Latent Dirichlet Allocation (LDA) to identify interpretable verbal topics that firms involved in AAERs emphasize compared to peers not involved in AAERs. LDA is based on the idea that the corpus of MD&A discussions can be represented by a set of common topics (akin to factors) and that the content of a specific MD&A can be described by the weights that managers place on these topics. Intuitively, LDA is a textual analog to factor analysis. We discuss LDA in more detail in the next section.

One of the challenges associated with topic modeling is how to accurately interpret the identified topics. We rely on two distinct approaches that provide computer generated resources. The first is a list of the most frequent commongrams or key phrases that associate with each topic. The second is a "representative paragraph", which best represents the content that is typical among firms that use the topic. The representative paragraph is an example of the disclosure that most closely associates with the topic. It can be viewed as the "typical" non-fraudulent discussion. A fraudulent firm that underdiscusses such a topic would thus provide less of this verbal content.

[2] Humphreys et al. (2011) discuss the relevant Communications and Psychology literature on deception. The authors also conduct a small pilot study of verbal content for 202 disclosures that further motivates our study.

Both methods help to associate each topic with a specific thematic interpretation. These calculations require few inputs, are fully automated and replicable, and cannot be impacted by researcher prejudice. Limiting researcher choice is critical given the high dimensionality of MD&A data, as researchers would have to make too many arbitrary choices using other methods.

We find support for all three hypotheses from the communications literature. Specifically, we find that firms committing fraud abnormally under-disclose details regarding why the firm experienced its observed level of performance. This conclusion is based on specific topics identified by the LDA methodology. For example, one of the representative paragraphs associated with this topic states:

> *"The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense.... The decrease was due primarily to the increased costs of operating the company-owned restaurants."*

The fact that fraudulent firms use less of this LDA topic related to providing quantitative details is direct evidence that they disclose fewer details explaining their performance.

We also find evidence that managers tend to disassociate themselves from fraudulent disclosure during years the firm is engaged in fraud. For example, fraudulent firms disclose abnormally low levels of an LDA topic that touts the manager's overall participation in the firm's vision and strategy. The representative paragraph from this topic begins with

> *"Since joining the company in January 1998, the new chief executive officer, along with the rest of the company's management team has been developing a broad operational and financial restructuring plan."*

By separately examining instances of revenue fraud and expense fraud, we are further able to test the hypothesis that fraudulent firms grandstand the manipulated performance itself. For example, we find that firms engaged in revenue fraud abnormally disclose more of a topic that highlights the firm's revenue growth. The representative paragraph associated with this topic, for example, starts with the statement

> *"Revenues increased by $29.9 million, or approximately 27.4%, to $139.1 million in 1997 from $109.2 million in 1996."*

Regarding firms engaged in expense fraud, they disclose abnormally high levels of a topic that associates with a representative paragraph stating

> *"Research and development expenses increased 20.7% to $6,006,000 in 1996, and increased as a percentage of net sales to 10.0% in 1996 from 6.1% in 1995. The increases in research and development expenses were primarily due to the expansion of the research and development staff, and expenses associated with its research and development facility."*

Because an active R&D platform is highly valued in the stock market, this is consistent with managers inflating the firm's perceived growth options by committing expense fraud that is reinforced by effervescent disclosure.

Because LDA provides a full taxonomy categorizing information in MD&A (see Ball et al., 2013), it is also easy for us to test two additional hypotheses from the existing literature that strongly relate to MD&A content. The first is the hypothesis that managers initiate fraud to reduce their cost of capital or to alleviate financial constraints (Dechow et al., 1996; Povel et al., 2007, and Wang et al., 2010). The second is the hypothesis that firms initiate fraud to lower the cost of M&A transactions that use equity as a means of payment.

Consistent with the lowering the cost of capital hypothesis, we find that fraudulent firms under-disclose the following topic associated with the firm's required summary of liquidity challenges:

> *"The company believes that its current cash, cash equivalents and short-term investment balances and cash flow from operations, if any, will be sufficient to meet the company's working capital and capital expenditure requirements for at least the next twelve months. Thereafter, the company may require additional funds..."*.

Hence fraudulent managers disclose less content that might indicate the limitations of their firm's supply of liquidity. Following Hoberg and Maksimovic (2015), this can indicate to investors that the firm is less constrained.

Given the popularity of this hypothesis in the literature, we also consider a quasi-natural experiment based on exogenous forced mutual fund selling following Coval and Stafford (2007) and Edmans et al. (2012). We find that when firms face negative liquidity shocks, they increase their disclosure of text that correlates highly with the abnormal text of fraudulent firms. We also find that the observed level of ex-post AAERs increases. This suggests that firms increase manipulative text and are more likely to commit fraud when their liquidity exogenously deteriorates.

The remainder of this article is organized as follows. Section 2 presents our hypotheses, Section 3 describes our data and methodology, and Section 4 presents our data and methods. Section 5 presents our abnormal disclosure regressions and Section 6 presents content analysis. Section 7 presents our quasi-natural experiment based on equity market liquidity, and Section 8 concludes.

2. Literature and hypotheses

The U. S. Securities and Exchange Commission (SEC) is responsible for prosecuting firms that intentionally misstate earnings. When the SEC decides to formally investigate a firm, it issues an Accounting and Auditing Enforcement Release (AAER). AAERs

are notices that actions concerning civil lawsuits have been brought by the Commission in federal court. The study of earnings management is based on the premise that management exercises significant judgment when preparing financial statements in accordance with Generally Accepted Accounting Principles (GAAP). Although many earnings management strategies are within GAAP limits, others intentionally violate it which is inherently illegal. The SEC (SEC, 2003) has issued interpretive guidance for MD&A which states:

> *"The MD&A requirements are intended to satisfy three principal objectives:*
>
> - *To provide a narrative explanation of a company's financial statements that enables investors to see the company through the eyes of management;*
> - *To enhance the overall financial disclosure and provide the context within which financial information should be analyzed; and*
> - *To provide information about the quality of, and potential variability of, a company's earnings and cash flow, so that investors can ascertain the likelihood that past performance is indicative of future performance.*
>
> *MD&A should be a discussion and analysis of a company's business as seen through the eyes of those who manage that business. Management has a unique perspective on its business that only it can present. As such, MD&A should not be a recitation of financial statements in narrative form or an otherwise uninformative series of technical responses to MD&A requirements, neither of which provides this important management perspective. Through this release we encourage each company and its management to take a fresh look at MD&A with a view to enhancing its quality. We also encourage early top-level involvement by a company's management in identifying the key disclosure themes and items that should be included in a company's MD&A."*

A key premise of our hypotheses is that MD&A is relevant to the incentives to commit fraud and to improving fraud detection. We thus note that manipulations of revenues and expenses are the basis for most AAER fraud allegations. In MD&A, managers discuss both as part of their required discussion of annual performance, illustrating why MD&A is highly relevant to testing our hypotheses.

There exists an extensive literature that attempts to measure earnings quality and to better understand the consequences of earnings management. Dechow et al. (2010) provide a comprehensive review. Many of these efforts have focused on the estimation of discretionary accruals (Jones, 1991; Dechow et al., 1995; Dechow et al., 2011b; Holthausen et al., 1995; Kothari et al., 2005, and Dechow and Dichev, 2002).

A common test of the efficacy of accounting quality models is to estimate their power to predict accounting fraud using financial (see Beneish, 1997; Beneish, 1999a,b; Dechow et al., 1996; Dechow et al., 2011b) and non-financial (Brazel et al. (2009) metrics. In a related literature, Feroz et al. (1991), Karpoff et al. (2008a,b) examine the issues that motivate fraud and their consequences. A common theme is whether common characteristics can predict accounting fraud.

The literature has continued to evolve as researchers have addressed power concerns and the large number of false positives. Dechow et al. (2011b) illustrate that fraud prediction models produce an unacceptably high rate of false positives. This is partly attributable to the low rate of detected fraud, which averages 1.5% per year over our sample period. A second problem is that there may be a correspondingly large number of violators that escape detection. Our study also suggests that undetected fraud is likely quite high.

These concerns have motivated a related literature examining deception in conference calls. For instance, Hobson et al. (2012) document that vocal markers of cognitive dissonance are useful in detecting financial misreporting. Larcker and Zakolyunkina (2012) use linguistic-based classification models of deceptive discussion during conference calls to examine the relation between deceptive language and restatements.

2.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) was developed by Blei et al. (2003) and is based on the idea that a corpus of documents can be represented by a set of topics. LDA has been used extensively in computational linguistics, is replicable, and is automated so it cannot be influenced by researcher prejudice. This contrasts with existing textual approaches that rely on researcher-selected lists of words.[3] LDA uses likelihood analysis and discovers clusters of text ("topics") that frequently appear in a corpus.

The approach is commonly referred to as a "bag-of-words" technique because the relative frequency of words in a document is centrally important, but not their specific ordering. It is helpful to define the common vocabulary as the union of all words that are contained in the MD&A corpus. LDA assumes that the document generation process arises from an underlying topic distribution, and not an individual word distribution. A particular topic can be characterized as a distribution over this common vocabulary of words where the relative probability weight assigned to each word indicates its relative importance to that topic. A topic is thus a word vector where each element is the weight associated with that word. For example, the words "oil" and "electricity" might be important to topics associated with Natural Resources and Manufacturing, but one might expect oil to

[3] Examples of recent work include Li (2006), Core et al. (2008), Larcker and Zakolyunkina (2012) and Loughran et al. (2009). While these studies are designed to test specific hypotheses, unlike our approach, they do not attempt to characterize the latent factors.

receive a higher weighting than electricity in the Natural Resources topic. The opposite might be true for the Manufacturing topic.

2.2. Hypotheses based on managerial disclosure incentives

We start with three hypotheses based on the incentives of fraudulent managers to alter their verbal disclosures. The first is derived from McCornack's (1992) Information Manipulation Theory (IMT). IMT argues that fraudulent firms have incentives to withhold relevant information so that investors will make incorrect inferences.

H1A (*Managerial detail concealment*)**.** Fraudulent managers under-report details explaining the firm's performance.

A second theory from the Communications literature is the Interpersonal Deception Theory (IDT) of Buller and Burgoon (1996). IDT assumes that managers manipulate verbal disclosures to achieve specific goals. We hypothesize that firms are more likely to grandstand good performance to create a deceptively positive impression.

H1B (*Managerial grandstanding*)**.** Fraudulent managers grandstand the firm's growth and the quantities they manipulated to increase the impact of the manipulation.

This notion of grandstanding is related to the finding in Kedia and Philippon (2009) that fraudulent firms invest and hire more workers than they might optimally need. Our hypothesis predicts that managers will complement this real investment inflation with disclosure grandstanding to portray a unified signal to investors about the firm's prospects. Grandstanding might also be likely when investors have high expectations for future growth (Dechow et al., 2011a and Skinner and Sloan, 2002).

Hypotheses H1A and H1B provide specific predictions about the magnitude of the verbal discussions that managers are expected to make conditional on their prior decision to commit accounting fraud. As a general rule, broad predictions about relative rates of discussion are difficult to make. However, our hypotheses have foundations in the existing literature and we also note that fraudulent firms tend to have poor operating performance (at least relative to their peers). Hence one can make directional predictions about topic discussions because there are powerful managerial incentives to avoid the ex post legal consequences of their actions.

The IDT theory also hypothesizes that managers avoid references to themselves because they wish to minimize personal responsibility if the fraud is discovered. The Four Factor Theory (FFT, Zuckerman et al., 1981) makes a similar prediction. The FFT links attempts to deceive to behavioral cues. Although this theory focuses on individual behavioral tendencies, Humphreys et al. (2011) argue that the "negative or unintended effects of deception may influence the deceivers to use *nonimmediate* language to disassociate themselves from the guilt induced by the deception." That is, managers attempt to shift the focus away from themselves. Larcker and Zakolyunkina (2012) examine conference call text and find evidence of fewer managerial self references among fraudulent firms. Together, these articles imply:

H1C (*Managerial disassociation*)**.** Fraudulent managers reduce the extent to which the management team itself is mentioned in the MD&A.

2.3. Hypotheses based on firm incentives

We now focus on two hypotheses from the existing literature that make additional predictions specifically regarding the content of MD&A. For example, MD&A discusses performance, investment, capital structure, liquidity needs, and growth strategies.

H2A (*Fraud to reduce the cost of capital*)**.** Managers commit fraud to increase the likelihood that they will be able to successfully raise capital at a low cost. This hypothesis is discussed by Dechow et al. (1996), Povel et al. (2007), and Wang et al. (2010).

H2B (*Fraud to reduce cost of M&A*)**.** Managers commit fraud to achieve more favorable stock exchange ratios in stock-based M&A transactions. This hypothesis is discussed by Erickson and Wang (1999) and Wang (2013).

H2A predicts that fraudulent firms will under-disclose information indicating liquidity problems[4]. H2B predicts a higher incidence of discussions of equity acquisitions during fraud-years. We also consider a quasi-natural experiment to further test H2A,

[4] We focus on liquidity discussions because firms are required to disclose liquidity problems in MD&A (see for example Hoberg and Maksimovic (2015)). The authors also note a high degree of heterogeneity in this liquidity disclosure, which indicates that there is likely adequate power to test the current hypothesis.

one of the most widely discussed motives for fraud. Here we consider exogenous forced mutual fund selling shocks following Coval and Stafford (2007) and Edmans et al. (2012). Hypothesis H2A predicts that exogenous shocks to equity market liquidity will increase the firm's use of text that associates with fraud. We also note that, although we are able to test many hypotheses using MD&A, we cannot test every hypothesis. For example, MD&A is unsuitable for testing the link between executive compensation motives and fraud (see for example Johnson et al., 2009; Goldman and Slezak, 2006, and Burns and Kedia, 2006). The reason is that executive compensation is usually discussed in Item 11 of the 10-K, which is distinct from the MD&A (Item 7 of the 10-K). Analogous tests using other sections of the 10-K offer excellent potential for future research.

3. Data and methodology

We create our sample and our key variables using two primary data sources: COMPUSTAT and the text in the Management's Discussion and Analysis section of annual firm 10-Ks (extracted using software provided by metaHeuristica LLC).

We first extract COMPUSTAT observations from 1997 to 2010 and apply a number of basic screens to ensure our examination covers firms that are non-trivial publicly traded firms in the given year. We start with a sample of 98,878 observations with positive sales, at least $1 million in assets, and non-missing operating income. We also discard firms with a missing SIC code or a SIC code in the range 6000 to 6999 to exclude financials, which have unique disclosures (especially because MD&A covers financial market liquidity and capital structure). This leaves us with 80,488 observations. After requiring that observations are in the CRSP database, we have 68,389 observations. Our sample begins in 1997 because this is the first year of full electronic coverage of 10-K filings in the Edgar database. Our sample ends in 2010 as this is the final year our sample has non-zero overlap with our AAER dataset, which we obtain from the Center for Financial Reporting and Management at the University of California at Berkeley.

We also require that each observation has a machine readable MD&A section with a valid central index key (CIK) link to the Compustat database.[5] We use software provided by metaHeuristica to web crawl and to extract the MD&A section from each 10-K. MetaHeuristica uses natural language processing to parse and organize textual data, and its pipeline employs "Chained Context Discovery" (See Cimiano, 2010 for details). The majority of 10-Ks (over 90%) have a machine readable MD&A section. The primary reason why a firm might not have a machine readable MD&A is when it is "incorporated by reference," and is not in the body of the 10-K itself.[6] These requirements leave us with a final sample of 55,666 firm-year observations having adequate data.

3.1. Accounting and auditing enforcement releases

We obtain data on Accounting and Auditing Enforcement Releases (AAERs) from the Center for Financial Reporting and Management at the University of California at Berkeley. Data on AAERs can also be collected manually from the Securities and Exchange Commission website[7] . Our sample includes AAERs indicating fraudulent behavior from 1997 to 2010. In addition to firm identifying data, which is needed to link AAER firms to our Compustat universe, we also obtain from the Berkeley database the beginning and ending dates each AAER alleges fraudulent activity. Our AAER dummy is set to one for firm fiscal years that overlap with these begin and end dates. This is our primary variable of interest, and we focus on how disclosure varies during these AAER years. For an example of how it is calculated, consider a firm that has a June 30 fiscal year end and committed a fraud that began in March 2009 and ended May 2011. The AAER dummy variable would be set equal to one for fiscal year ends 2009, 2010, and 2011. If this same firm initiated the fraud on August 2009 instead, the AAER dummy variable would only be coded one for fiscal years 2010 and 2011.[8]

For each AAER, we also identify a year that is definitively prior to the alleged fraudulent activity, and a year that is definitively subsequent to the public release of the AAER by the SEC. We refer to these as the pre-AAER year and the post-AAER year. Our assessing disclosure in three critical periods (prior to, during, and after the alleged fraud) serves two purposes. First, this serves as a placebo test, as we expect a strong signal only during the years of fraudulent activity, and not in the years prior to or after the alleged fraud. Second, this allows us to understand the disclosure life cycle of fraudulent firms.

Due to the approximate nature of stated fraud periods, we take a conservative approach when identifying the pre-AAER year and the post-AAER year. We define the pre-AAER year as the fiscal year preceding the first full calendar year that precedes the alleged fraud period. This ensures that, even with 10-K reporting delays and potential approximate identification of the fraudulent period, the pre-AAER year has disclosure that is unlikely to be contaminated by disclosure associated with the fraud. We identify the post-AAER year as the fiscal year end in the calendar year that is subsequent to the calendar year in which the AAER is announced to the public on the SEC website. This ensures that the firm had adequate time to update its disclosure subsequent to the alleged fraud.

[5] We use the WRDS SEC Analytics package, which has a backward compatible correspondence table linking Central Index Keys (from SEC Edgar) to gvkeys (Compustat).

[6] The typical scenario under which a MD&A section is incorporated by reference is when the annual report is submitted along with or referenced by the 10-K, and thus MD&A is not in the 10-K itself.

[7] http://www.sec.gov/divisions/enforce/friactions.shtml

[8] If a firm has more than one AAER, and the periods of alleged fraud overlap, we set the AAER dummy to one for any fiscal period where at least one fraud event is alleged.

3.2. Disclosure industry similarity

In this section, we focus on identifying the disclosure similarity between a firm and its size-age-industry matched peers. We refer to this as our "Industry Similarity" measure. Our approach of identifying common industry disclosure is related to Hanley and Hoberg (2010), who examine IPO pricing.

We first group all firms into bins based on industry (two-digit SIC codes), size and age. In particular, for each industry group in each year, we create a small firm and a large firm bin based on the median size of firms in each industry bin. We then divide bins once again based on median age (listing vintage). We thus have four bins for each SIC-2 industry, and each of the four bins has nearly the same number of firms. If a given bin has less than two firms, we exclude it from the rest of our analysis. Given that our two-digit SIC categories are rather coarse, this requirement affects less than one percent of our sample. We also note that our findings are robust to only using industry bins rather than these industry-size-age bins. We use these more refined bins because we expect material systematic differences in disclosure across firms of different size and age. We refer to a firm's peers in its industry, size, and age bin as its "ISA peers".[9]

Following standard practice in text analytics, we first discard stop-words and then convert the text in each firm's MD&A into vectors of common length across all firms. We define a "stop word" as any word appearing in more than 25% of all MD&A filings in the first year of our sample (1997). The length of the vectors we create is based on the universe of remaining words. Because our calculations are computationally intensive, we restrict attention to words appearing in the MD&A of at least 100 firms in the first year of our sample (1997).[10] The resulting list of words is stable over time, as 99.1% of randomly drawn words using our 1997-based screen would, for example, be included using an analogous screen based on 2008. Each firm-year's MD&A is thus represented by its word distribution vector $W_{i,t}$. This vector sums to one, and each element indicates the relative frequency of the given word in the given MD&A. Our use of 1997 data to determine the word universe is meant to be conservative, as we avoid any look ahead bias in our later regressions that are based on an out of sample predictive framework.

To quantify disclosure similarity with ISA peers, we next compute the average word usage vector for a given firm's ISA peers excluding itself ($ISA_{i,t}$). It is important that this average excludes the firm itself, as skipping this step would create a mechanistic degree of similarity for firms in less populous bins. Our measure of industry disclosure similarity (H_{it}) is the cosine similarity between $W_{i,t}$ and $ISA_{i,t}$.

$$H_{i,t} = \frac{W_{i,t}}{\sqrt{(W_{i,t} \cdot W_{i,t})}} \cdot \frac{ISA_{i,t}}{\sqrt{(ISA_{i,t} \cdot ISA_{i,t})}} \tag{1}$$

We consider the similarity between each firm's disclosure and that of its ISA peers to examine if fraudulent firms are more likely to herd with their industry peers, perhaps as a mechanism to escape detection by "hiding in the crowd". The cosine similarity is a standard technique in computational linguistics (see Sebastiani, 2002 for example). It is also easy to interpret, as two documents with no overlap have a similarity of zero, whereas two identical documents have a cosine similarity of 1. Finally, by virtue of its normalization of vectors to unit length, this method also has the good property that it correlates only modestly with document length.

3.3. Disclosure fraud similarity

In this section, we construct measures of the extent to which firms engaged in fraudulent behavior produce common disclosure, while controlling for the disclosure of ISA peers. We first compute abnormal disclosure for each firm ($AW_{i,t}$) as follows:

$$AW_{i,t} = W_{i,t} - ISA_{i,t} \tag{2}$$

We note that we only include non-fraudulent ISA peers in this calculation. The resulting vector sums to zero, as W_{it} and ISA_{it} each sum to one. We next compute the average deviation from industry peers made by firms known to be involved in SEC AAER enforcement actions (where N_{AAER} is the number of AAER firm-years from 1997 to 2001):

$$AAER_{vocab} = \sum_{j=1,\ldots,N_{AAER}} \frac{AW_j}{N_{AAER}} \tag{3}$$

Note that the vector $AAER_{vocab}$ does not have a time subscript, as we are summing the unique disclosures over all AAERs in a given universe. We note here that we only tabulate this average over firms with an AAER dummy of one in the years 1997 to 2001. We do not use the years 2002 to 2010 for training as we wish to preserve these years for assessing the out of sample

[9] In unreported results, we examine if our results are robust to further excluding fraudulent firms from the group of ISA peers. This has little influence on our results because fraudulent firms are relatively rare in our sample.

[10] This results in a vector length of roughly 10,000 words. We also note that our findings are robust to instead using a stricter screen based on 5000 words. Because we also do not see a material degree of improvement in going from 5000 to 10,000 words, we thus conclude that our universe is sufficiently refined to provide a relevant signal for testing our key hypotheses.

performance of our fraud similarity variable in later tests. Our results are stronger if we instead use our entire sample for the computation of the $AAER_{vocab}$. Our approach ensures that results are not driven by look ahead bias. We then define the fraud profile similarity (we will also refer to this as the "fraud score") of a firm in a given year F_{it} as the cosine similarity between $AW_{i,t}$ and $AAER_{vocab}$ as follows (we also exclude firm i itself from the computation of $AAER_{vocab}$ to avoid any mechanistic correlations):

$$F_{i,t} = \frac{AW_{i,t}}{\sqrt{(AW_{i,t} \cdot AW_{i,t})}} \cdot \frac{AAER_{vocab}}{\sqrt{(AAER_{vocab} \cdot AAER_{vocab})}} \quad (4)$$

4. Data and summary statistics

Table 1 displays summary statistics for our panel of 55,666 firm-year observations from 1997 to 2010 for four subsamples: overall, the hold-out subsample from 2002 to 2010, the fraudulent firm sample and the ISA matched sample. The hold out sample reflects the out-of-sample tests we run later in this study, as our fraud score variable is fitted using ex-ante text from the 1997 to 2001 part of our sample, and hence we use the 2002 to 2010 subsample as an out-of-sample test. The matched sample is based on choosing a match for each fraudulent firm from its set of ISA peers (defined in the last section), in particular we select the firm from this set that is closest in size to the given fraudulent firm.

The table shows that for the full sample, 1.0% of firm year observations are AAER-years. In the hold-out sample, this figure drops slightly to 0.9%. As it is based on cosine similarities between positive and negative word vectors (see Eq. (4), note that abnormal disclosure $AW_{i,t}$ is a difference of two positive vectors), the Fraud Similarity Score has a distribution in the interval

Table 1
Summary Statistics. Summary statistics are reported for our full sample of 55,666 observations based on annual firm observations from 1997 to 2010. We display mean, median, and standard deviation statistics for four subsamples: the full sample of 55,666 obs (upper left), a hold-out sample that includes 32,553 observations from 2002 and later (upper right), the sample of all 552 firm-years where an AAER indicates fraud (lower left), and a matched sample of 552 non-fraudulent firms with same industry, size and age as the fraudulent firms (lower right). The matched sample is created by first identifying, for each fraudulent firm-year, the set of non-fraudulent firm-years that are in the same industry, are in the same size group (above versus below median), and in the same age group (above vs below median). From this set, the matched firm with the closest market capitalization is selected as the matched firm for each fraudulent firm. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year. The industry similarity score is the raw cosine similarity of the given firm's MD&A disclosure and that of its industry-size-age peers. A higher figure indicates that the given firm has disclosure that is highly similar to its industry peers. To compute the fraud similarity score, we first compute each firm's abnormal disclosure as its raw disclosure minus the average disclosure of its industry-size-age peers. The fraud similarity score is then the cosine similarity of the given firm's abnormal disclosure and the average abnormal disclosure of all firms involved in AAERs in the sample period 1997 to 2001. We use these earlier years of our sample to identify the vocabulary of firms allegedly committing fraud so that we can consider out of sample analysis for the later years in our sample 2002 to 2008. Log sales is the natural logarithm of Compustat sales. Operating income/sales is Compustat operating income before depreciation scaled by sales and any values of operating income/sales less than minus one are set to minus one. R&D/sales and CAPX/sales are Compustat values of R&D and capital expenditures scaled by sales. The market capitalization is the CRSP equity market capitalization, the fiscal year stock return is the stock return experienced in the fiscal year, and the book to market ratio is the book value of equity divided by the CRSP market value of equity. All ratios are winsorized at the 1% and 99% level.

Variable	Mean	Median	Std. dev.	Mean	Median	Std. dev.
	Full sample			Hold-out sample		
AAER dummy	0.010	0.000	0.099	0.009	0.000	0.093
Fraud similarity score	0.000	−0.012	0.116	0.001	−0.007	0.101
Industry similarity score	0.669	0.673	0.080	0.667	0.669	0.078
Log sales	5.007	4.967	2.147	5.412	5.454	2.148
Operating income/sales	0.000	0.084	0.349	0.028	0.092	0.328
R&D/sales	0.185	0.000	0.744	0.193	0.001	0.810
CAPX/sales	0.118	0.036	0.312	0.099	0.031	0.261
Stock market capitalization	1785.78	219.685	8674.64	2235.57	329.952	9360.75
Book to market ratio	0.697	0.491	0.871	0.705	0.504	0.894
Fiscal year stock return	0.114	−0.015	0.782	0.143	0.027	0.757
	Fraud sample			ISA-matched sample		
AAER dummy	1.000	1.000	0.000	0.000	0.000	0.000
Fraud similarity score	0.065	0.060	0.126	0.010	−0.011	0.122
Industry similarity score	0.684	0.691	0.082	0.678	0.679	0.076
Log sales	5.994	5.923	1.893	6.036	6.118	2.047
Operating income/sales	0.062	0.104	0.257	0.051	0.100	0.284
R&D/sales	0.107	0.009	0.215	0.159	0.012	0.641
CAPX/sales	0.069	0.038	0.110	0.107	0.035	0.308
Stock market capitalization	4713.59	667.622	17860.8	4530.27	593.345	14534.7
Book to market ratio	0.548	0.378	0.692	0.613	0.398	0.905
Fiscal year stock return	0.207	−0.003	0.888	0.076	−0.019	0.660

[−1, +1] and a mean that is close to zero in these two samples. The table also shows that firm characteristics in the hold-out sample are similar to those in the full sample as for example their investment intensity and book to market ratios are comparable. Nominal figures such as stock market capitalizations are somewhat larger in the hold-out sample likely due in part to inflation over the sample period. Nevertheless, we control for many variables including size, year fixed effects, and also firm fixed effects in order to ensure that our results can be attributed to fraudulent firms and not unobserved firm characteristics.

The lower part of Table 1 displays summary statistics for fraudulent firms and their matched ISA peers. Comparing these statistics to the full sample indicates that firms involved in AAERs are somewhat larger, more profitable, and engaged in less investment than firms in the full sample. These differences illustrate why it is important to control for these firm characteristics in our main disclosure regressions. These differences also illustrate why we consider a matched sample test. In particular, we later consider regressions where only the fraudulent firms and their closest size-matched ISA peers are included in the regression, allowing us to examine if the ability to relate disclosure to fraud is robust to a test where fraudulent and non-fraudulent firm characteristics are observationally very similar. The statistics in Table 1 reveal that the fraudulent firms and their matched ISA peers are indeed very similar on most dimensions. Finally, the table shows our expected result that the fraud score variable is the highest for the fraudulent firm subsample.

Table 2 displays Pearson correlation coefficients for the full sample from 1997 to 2010 (Panel A), and the hold out sample from 2002 to 2010 (Panel B). In Panel A, the positive 5.6% correlation between the AAER dummy and the fraud similarity score (significant at the 1% level) foreshadows our later multivariate results. This suggests that firms involved in potentially fraudulent activity have abnormal disclosure relative to ISA peers that is common among AAER firms. The correlation between the AAER dummy and industry similarity is much weaker at 1.9%. Remarkably, the fraud similarity score is more correlated with the AAER dummy than it is with any of the other displayed variables including firm size (4.6% correlation). Panel B shows that the correlation patterns are similar during the hold-out sample relative to the full sample.

The correlation between fraud similarity and industry similarity is modest at 1.9% in Panel A and is −5.0 % in Panel B. The modest result is by construction, as fraud similarity is a function of abnormal disclosure after controlling for ISA peers. We also note that fraud similarity correlates little with firm size (1.5% in both panels), which also relates to its construction based on size-adjusted peers (in addition to industry and age adjustments). These aspects of our variables help to ensure a clear interpretation in both univariate and multivariate settings. Finally, these modest correlations indicate that multicollinearity is unlikely to be a concern.

Table 3 displays time series summary statistics regarding AAER-year observations in our sample from 1997 to 2010. The table shows a peak in 2000 to 2003 following the internet bubble's collapse, and also a steady stream of AAER years throughout our sample with the exception of the last three to four years, where the incidence rate is lower. As our analysis controls for both industry and time effects, as well as other controls, these features of our data cannot explain our results. The relatively low rate of AAERs during the post financial crisis years does not necessarily point to a reduction in the rate of fraud but is more likely explained by a change in the SEC's priorities following the crisis.

4.1. Initial evidence of disclosure differences

In this section, we explore the distributional features of our industry similarity and fraud similarity measures, and their links to observed AAER Enforcement actions. In Table 4, we sort firms into deciles based on their fraud similarity and industry similarity measures. We then report the fraction of firms in each decile that are involved in AAERs.

Table 2
Pearson Correlation Coefficients. Pearson Correlation Coefficients are reported for our sample based on annual firm observations from 1997 to 2010. We display results for two subsamples: the full sample of 55,666 obs (Panel A), and a hold-out sample that includes 32,553 observations from 2002 and later (Panel B). See Table 1 for the description of our key variables.

Row	Variable	AAER dummy	Fraud similarity score	Industry similarity score	Log sales	Operating income/sales	R&D sales
Panel A: Correlation Coefficients: Full sample							
(1)	Fraud similarity score	0.056					
(2)	Industry similarity score	0.019	0.015				
(3)	Log sales	0.046	0.015	0.055			
(4)	Operating income/sales	0.018	0.001	−0.041	0.523		
(5)	R&D/sales	−0.010	0.015	0.086	−0.305	−0.524	
(6)	CAPX/sales	−0.016	−0.015	0.045	−0.144	−0.148	0.197
Panel B: Correlation Coefficients: Hold-out sample							
(7)	Fraud similarity score	0.039					
(8)	Industry similarity score	0.014	−0.050				
(9)	Log sales	0.032	0.015	0.036			
(10)	Operating income/sales	0.010	−0.001	−0.043	0.524		
(11)	R&D/sales	−0.010	0.014	0.095	−0.325	−0.551	
(12)	CAPX/sales	−0.015	−0.004	0.041	−0.123	−0.069	0.193

Table 3
AAER timeseries statistics. The table reports time series statistics for our sample of 552 fraudulent firm-years from our full sample of 55,666 observations based on annual firm observations from 1997 to 2010. The AAER dummy is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year.

Row	Year	Number AAER firm years	Number of firms in sample	Fraction AAER firm years
1	1997	36	4672	0.008
2	1998	42	4661	0.009
3	1999	48	4728	0.010
4	2000	65	4647	0.014
5	2001	77	4405	0.017
6	2002	61	4171	0.015
7	2003	68	4008	0.017
8	2004	54	3914	0.014
9	2005	38	3522	0.011
10	2006	27	3395	0.008
11	2007	19	3419	0.006
12	2008	10	3491	0.003
13	2009	5	3389	0.001
14	2010	2	3244	0.001

Panel A of Table 4 displays these results for our entire sample, and shows that the incidence rate of AAERs is strongly positively correlated with the fraud similarity decile or in the industry similarity decile in which a firm resides. The results are economically large and decile sorting is close to monotonic. Regarding fraud similarity, the incidence rate of AAERs in decile 10 is 2.4% compared to just 0.4% for decile 1. The positive link between industry similarity and AAER incidence is weaker with high to low decile range of 2.0% to 0.8%.

Panel B of Table 4 displays analogous results for the out of sample period from 2002 to 2008. We remind readers that the key vocabulary used to compute the vocabulary associated with fraudulent firms is computed only using data from 1997 to 2001 (see Section 3.3). Hence, our assessment of the link between AAERs and the fraud scores in 2002 to 2010 is an out of sample test. We continue to observe strong positive associations with AAER incidence rates for fraud similarity, and the inter-decile range is 0.6% to 1.9%. Our later tests will show that our results for fraud similarity are especially strong both statistically and economically, and are also robust to multivariate regressions including controls for firm and industry fixed effects. In contrast,

Table 4
AAERs versus fraud similarities and industry similarity deciles. The table displays decile statistics for our sample of 55,666 observations based on annual firm observations from 1997 to 2010. Within each year, firms are sorted into deciles based on their fraud similarities (first two columns) and based on their industry similarity scores (latter two columns). We display results for two subsamples: the full sample of 55,666 obs (Panel A), and a hold-out sample that includes 32,553 observations from 2002 and later (Panel B). The fraction of firms involved in AAERs is then reported for each decile group. See Table 1 for the description of our key variables.

Decile	Fraud similarity score	Fraction AAER firm years	Industry similarity score	Fraction AAER firm years
Panel A: Full sample (1997–2010)				
1	−0.176	0.004	0.516	0.008
2	−0.118	0.006	0.586	0.009
3	−0.084	0.006	0.618	0.009
4	−0.055	0.009	0.642	0.009
5	−0.028	0.010	0.663	0.010
6	−0.000	0.009	0.683	0.010
7	0.032	0.010	0.703	0.012
8	0.074	0.013	0.725	0.010
9	0.132	0.020	0.751	0.015
10	0.227	0.024	0.792	0.020
Panel B: Out of sample (2002–2010)				
1	−0.158	0.006	0.520	0.006
2	−0.101	0.005	0.589	0.006
3	−0.069	0.006	0.619	0.008
4	−0.043	0.009	0.641	0.008
5	−0.019	0.006	0.660	0.008
6	0.005	0.008	0.680	0.008
7	0.030	0.008	0.700	0.010
8	0.060	0.007	0.722	0.007
9	0.106	0.013	0.749	0.012
10	0.199	0.019	0.793	0.014

industry similarity plays a more passive role, and its correlation with AAERs is not robust to firm fixed effects. The results in this section indicate that the unique vocabulary used by AAER firms, that is distinct from industry-size-age peers, has persisted over time.

4.2. Fraud similarity distributions

In this section, we examine the distribution of fraud similarity. Fig. 1 shows the empirical density function of this variable over its domain [−1, 1] for both the entire sample and the subsample of firm years involved in AAERs. The distribution is centered near zero and has both a bell shaped component and a right-skewed component. This shape is akin to that of a mixed distribution, where some draws for example come from a non-fraudulent firm distribution centered near zero and others come from a right-shifted fraudulent-firm distribution centered to the right of zero. The solid line shows the reflection of the distribution around the y-axis and illustrates the extent of the right skewness. As the figure indicates, the amount of probability mass that differs from the reflection is 4.5% of the total mass. This is materially larger than the observed 1.0% AAER rate indicated in Table 1. These results suggest that the total rate of fraud (including undetected fraud) might be substantially larger than the observed rate of AAERs.

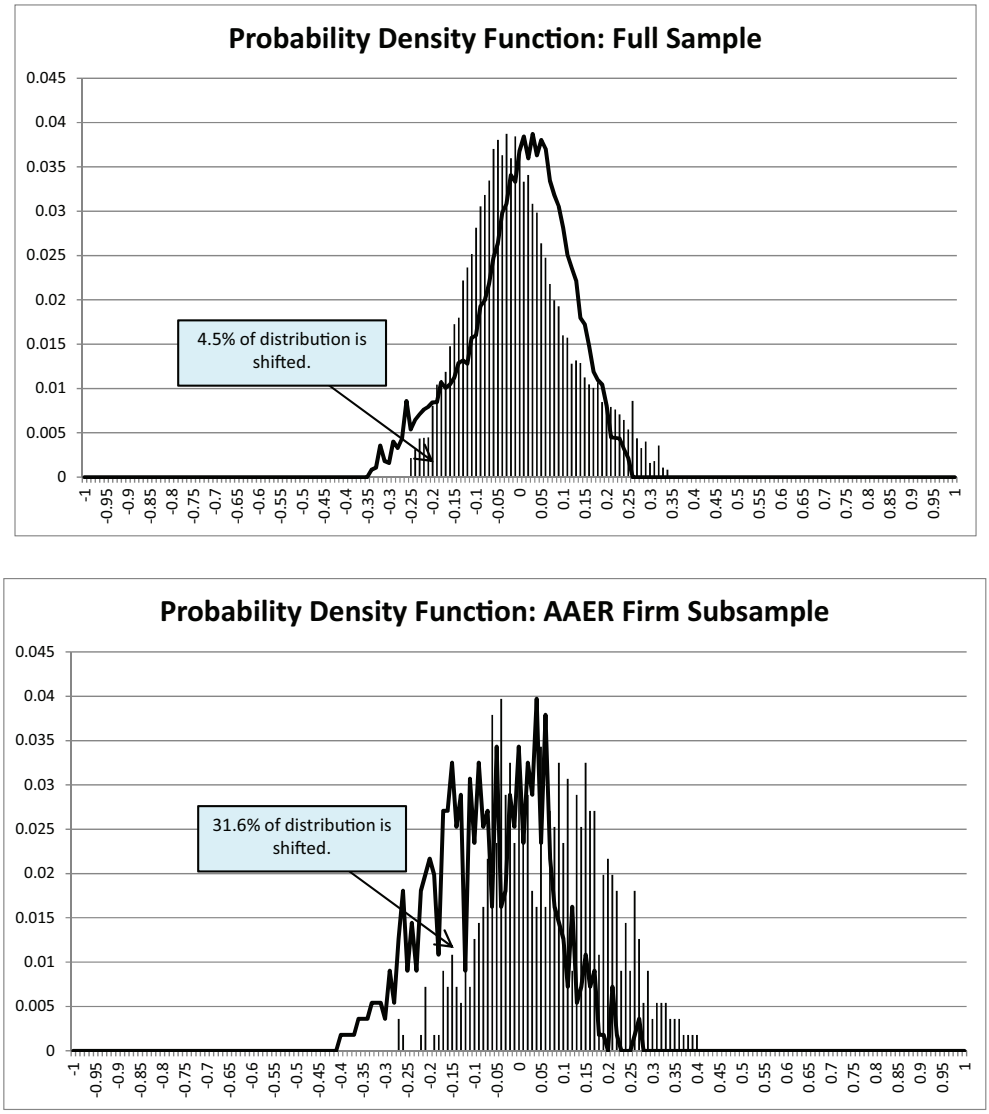

Fig. 1. Empirical distribution of firm fraud similarities for two subsamples. The upper figure's distribution is based on all firms in our sample excluding firm years involved in AAERs. The lower figure reports the fraud similarity distribution only for firms-years involved in AAERs. In both figures, the actual distribution is displayed using the bar chart format. To illustrate the degree of left-right asymmetry, the line plot displays the shape of the y-axis reflection of the actual distribution. The size of the asymmetric mass is then summarized.

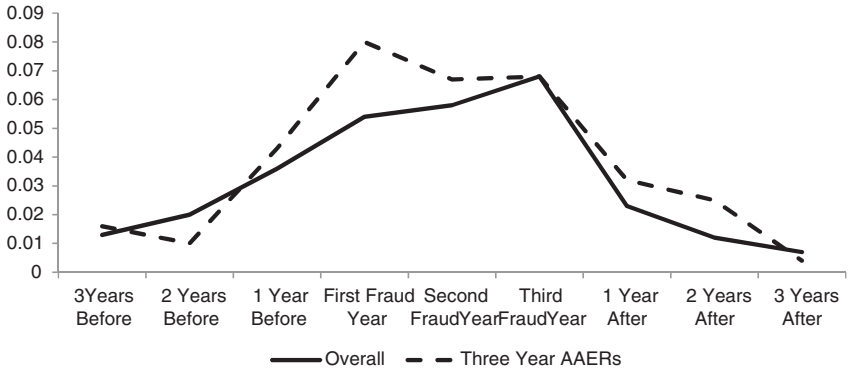

Fig. 2. Average fraud similarities over time for firms involved in AAERs. The figure displays the average fraud similarity score during the period of time that the AAER alleges fraud occurred, and also during the period of time preceding and after the period of the alleged fraud. Regardless of duration of the fraudulent period, we tag the three years prior to the fraud period as the ex-ante period and the three years after the fraud period as the ex-post period. For firms that had a fraud period of one or two years, they would be counted in the first fraud year and the second fraud year calculation, but not the third fraud year calculation. To ensure that fraud duration is not overly influencing our results, we also display results where we limit the sample to firms with alleged fraud that lasted at least three years.

The lower plot in Fig. 2 further illustrates why our approach might have good power for detecting fraud. The lower plot displays the density function of fraud similarity for firms that are known to be involved in AAERs. The figure shows a far higher degree of rightward shift than for the entire sample, indicating that fraud similarity is effective in separating AAER firms from non-AAER firms. The degree of asymmetric mass is 31.6%, which is far larger than the 4.5% in the upper figure.

Fig. 2 displays fraud similarity scores over time: before, during and after a firm is involved in an AAER. We also explore the extent to which fraud similarity varies when a firm is involved in an AAER alleging a longer duration of fraud. In particular, we tag the three years that are prior to the calendar year in which the AAER indicates that the fraud began as the pre-fraud period, and the three years after the calendar year in which the AAER indicates that the fraud ended as the post-fraud period. We then consider up to three years of time during which an alleged fraud occurred. If a firm's alleged fraud period is three or more years, it will enter the average fraud similarity calculation for the first three of these years. If the firm's alleged fraud lasted only one or two years, it will only be included in the first and second fraud year calculations, respectively. To ensure robustness, we also consider this calculation only for firms that experienced a fraud period of at least three years.

The figure shows a trapezoidal pattern for fraud similarity. During the three years preceding the alleged fraud, the average fraud similarity slowly increases from nearly zero to 0.04. During the period of alleged fraud, this score doubles to 0.08 and remains high during the years of alleged fraud. After the period of alleged fraud ends, fraud similarity drops sharply to less than 0.04 and dissipates to zero. Because the AAER is only announced after the fraud has occurred, these results provide strong time series evidence that we have identified a set of disclosure vocabularies that are used more by firms alleged to have committed fraud relative to those that have not. Because the figure reports scores for the same firms in all periods, these results are stark and control for built-in firm fixed effects.

5. Disclosure and fraud regressions

In this section, we use regression analysis to test our abnormal disclosure hypotheses using an unbalanced panel. As placebo tests, we consider not only disclosures in the year of an AAER, but also in the year prior and the year after the AAER. We expect a strong identifying signal only during the years of fraudulent activity, and not in the years prior to or after the alleged fraud periods. This approach allows us to fully understand the disclosure life cycle of fraudulent firms.

Table 5 displays the results of OLS regressions in which the dependent variable is the firm's disclosure strategy. As indicated in the first column, the dependent variable is either fraud similarity or the industry similarity score.[11] In Panels A to C, we report results for the entire sample, for larger firms, and for smaller firms, respectively. Firm size is identified using median assets in each year. These regressions are conservative in the sense that identification only is based on within-firm variation (they include controls for firm and year fixed effects). Standard errors are adjusted for clustering by firm. We also include several controls

[11] Readers interested in fraud detection might prefer an alternative specification where the AAER dummy is the dependent variable for convenience and the fraud similarity score is an independent variable. Although such a specification produces similar results and affirms the positive link between the fraud similarity score and AAER violations, we do not focus on this specification in our main tables due to potential endogeneity concerns. In particular, the sequencing of events in our framework is such that disclosure is created at the end of a fiscal year. Thus, the commission of fraud in the given fiscal year likely causes the ex-post disclosure to have an abnormal component when the managers later summarize their fiscal year's performance, and not vice-a-versa. This indicates that the use of the fraud similarity score as the dependent variable is the appropriate model.

Table 5
Disclosure outcome regressions (AAER-year). In Panels A to C, the table reports our baseline OLS regressions for our sample of 55,666 firm-year observations based on annual firm observations from 1997 to 2010. These baseline regressions are estimated with year and firm fixed effects and standard errors are clustered by firm. The dependent variable is based on a firm-year's disclosure in its 10-K and varies by row as indicated. For all Panels, standard errors are clustered by firm and *t*-statistics are in parentheses. See Table 1 for the description of our key variables. The AAER dummy is our primary variable of interest and is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year. To control for the economic conditions associated with the information in a given firm's MD&A, we additionally include controls for the average Tobins Q and Operating Income/Sales for the ten firms with MD&A Sections that are most similar to the given firm (these ten firms are those with the highest cosine similarity between their MD&A and that of the given firm in the given year). Panels D to F consider various robustness tests regarding the baseline model in Panel A. Panel D repeats the test in Panel A but only for our out of sample period including 32,553 annual firm observations from 2002 to 2010. These tests are out of sample because the base vocabulary used to compute fraud similarity is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. Panel E repeats the test in Panel A but adds three additional control variables aimed at challenging whether our results are due to narrower effects that have been documented in other studies but that might be better measured using text: a restatement variable, a litigation variable, a control for uncertainty, a control for inadequate CRSP data to compute uncertainty, and an acquisition dummy. Although we include these variables in the model, we do not report the additional coefficients to conserve space (see the Online Appendix for the presentation of those coefficients). Panel F repeats the test in Panel A but limits the test to the much smaller matched sample, which contains the 552 firm-years known to be associated with fraud, and for each such observation, the matched non-fraudulent firm-year in the same industry-size-age (ISA) matched grouping that is the closest match in terms of total assets. Panel F thus includes only year fixed effects, and standard errors are clustered by industry.

Row	Dependent variable	AAER dummy	Operating income/sales	R&D/sales	CAPX/sales	Log sales	MD&A peer implied Tobins Q	MD&A peer implied OI/sales	Obs.
Panel A: Entire sample									
(1)	Fraud profile sim.	0.034	−0.007	0.004	0.000	0.005	0.003	−0.000	55,666
		(6.26)	(−2.79)	(3.87)	(0.24)	(4.71)	(9.87)	(−1.25)	
(2)	Industry sim.	0.001	−0.012	0.004	0.003	0.008	−0.001	0.001	55,666
		(0.31)	(−5.95)	(4.96)	(2.01)	(9.83)	(−3.14)	(3.61)	
Panel B: Above median firm size only									
(3)	Fraud profile sim.	0.032	0.012	0.045	0.007	0.007	0.003	−0.000	28,754
		(5.11)	(1.86)	(2.64)	(2.00)	(3.46)	(7.68)	(−0.15)	
(4)	Industry sim.	−0.001	−0.005	−0.003	−0.000	0.004	−0.001	0.001	28,754
		(−0.10)	(−0.82)	(−0.24)	(−0.01)	(2.76)	(−3.31)	(2.14)	
Panel C: Below median firm size only									
(5)	Fraud profile sim.	0.024	−0.011	0.003	0.000	0.006	0.002	−0.000	26,912
		(2.27)	(−3.78)	(3.39)	(0.10)	(4.06)	(7.17)	(−1.29)	
(6)	Industry sim.	−0.003	−0.017	0.004	0.005	0.010	−0.000	0.001	26,912
		(−0.46)	(−7.30)	(5.21)	(2.48)	(9.34)	(−1.85)	(3.12)	
Panel D: Same as Panel A, but Out of sample years only									
(7)	Fraud profile sim.	0.015	−0.013	0.003	0.003	0.005	0.002	−0.001	32,553
		(2.10)	(−4.00)	(3.40)	(1.43)	(3.76)	(3.60)	(−2.44)	
(8)	Industry sim.	0.003	−0.014	0.003	0.006	0.007	−0.002	0.001	32,553
		(0.55)	(−4.80)	(3.14)	(2.31)	(6.22)	(−3.71)	(2.53)	
Panel E: Same as Panel A, but add additional controls									
(9)	Fraud profile sim.	0.032	−0.005	0.003	0.001	0.004	0.003	−0.000	49,014
		(6.10)	(−1.80)	(3.23)	(0.34)	(3.83)	(9.93)	(−1.09)	
(10)	Industry sim.	0.003	−0.010	0.004	0.004	0.008	−0.001	0.001	49,014
		(0.59)	(−4.64)	(4.69)	(2.40)	(9.41)	(−3.06)	(3.29)	
Panel F: Same as Panel A, but limit to matched sample									
(11)	Fraud profile sim.	0.108	0.019	0.011	−0.014	−0.002	0.013	0.009	1104
		(2.14)	(1.34)	(0.88)	(−1.76)	(−0.96)	(2.92)	(3.44)	
(12)	Industry sim.	−0.035	0.013	0.024	0.004	0.004	0.000	0.006	1104
		(−0.92)	(0.88)	(2.47)	(0.17)	(2.00)	(0.05)	(1.19)	

including the implied economic state of the firm (the average Tobins *q* and profitability of the ten firms in the given year having the most similar MD&A disclosure as the given firm based on cosine similarities).[12]

Panels D to F consider three robustness tests. Panel D considers the out of sample period (2002 and later). Panel E considers additional controls for restatements, litigation, mergers, and uncertainty. Panel F considers results based on a much smaller and more stringent matched-sample test. The matched sample is based on choosing one match for each fraudulent firm-year from

[12] The implied Tobins *q* and profitability of peers is particularly well-suited to control for economic conditions facing the firm in this setting as these are the conditions implied by the disclosure itself.

its set of ISA peers (defined in Section 3). We select the firm from this set that is closest in size to the given fraudulent firm. As a result, this sample only contains 1104 observations (552 fraudulent firms and 552 ISA-matched non-fraudulent firms).

Panel A of Table 5 shows that firms engaged in alleged fraud have significantly higher fraud profile similarities. This coefficient has a *t*-statistic of 6.26, and is significantly well beyond the 1% level. The results for industry similarity are not significant (a *t*-statistic of 0.31). We note again that these regressions are based on stringent within-firm identification. The results for fraud similarity confirm the intuition established in the discussion of Fig. 2, where we find that firms involved in fraud become more similar to other firms that committed fraud, but only in the years they are allegedly committing fraud. This suggests that these disclosures are likely related to the commitment of the fraud itself.

Panels B and C of Table 5 show that fraud profile similarity is robust at the 1% level for large firms and at the 5% level for small firms. We also continue to find that industry similarity is not significant. We thus focus our attention on fraud profile similarity for the remainder of our study and conclude that fraudulent firms have a strong common component that cannot be explained by ISA peers.

Panel D, shows that our results remain robust during the out of sample period from 2002 to 2010. This test is particularly stringent, as the combined effects of the smaller sample and the impact of firm fixed effects on the remaining degrees of freedom is more extreme. Nevertheless, the fraud similarity variable remains significant at the 5% level with a *t*-statistic of 2.10.

In Panel E, we further challenge our specification by including four additional control variables: restatements, litigation, uncertainty and mergers.[13] Although we do not display the coefficients for these variables in Panel E to conserve space, we do report the full set of coefficients in Table A1 of the Online Appendix of this study. The inclusion of these particular variables in Panel E raises the bar for our tests as it examines whether our results are potentially due to narrower effects that have been documented in other studies. The results in Panel E show that our results are highly robust, as the *t*-statistic for fraud profile similarity is roughly equal in Panels A and E.

Panel F shows that our results are also robust to using the much smaller matched sample. This indicates that our results continue to be robust even when fraudulent firms are rigidly compared to firms that are very similar regarding observable characteristics.

5.1. Existing fraud models

In this section, we stress-test the contribution of MD&A text when variables from existing models are included in our regressions. We consider our full sample model with firm fixed effects from Panel A of Table 5 and our more stringent matched sample model from Panel F of Table 5 as our baseline models. To each model, we then add variables used in the following studies: Beneish (1999a) (Panel A), Dechow et al. (2011a) (Panel B), absolute discretionary accruals from the modified Jones (1991) Model as shown in Dechow et al. (1995) (Panel C), and absolute discretionary accruals from the Ball and Shivakumar (2006) model (Panel D). The variables in each of these models are defined in the appendix. For both discretionary accruals models, we compute total accruals using the cash flow statement method in Eq. (3) of Collins and Hribar (2002).

Table 6 shows that our central finding cannot be explained by any of the existing models. In Panel A, when we control for the Beneish (1999a) model, we find that the AAER dummy remains a significant predictor of the fraud score with a *t*-statistic of 6.20, which is very similar to the 6.26 we found in Table 5 in our baseline model. Similarly, the *t*-statistic for the matched sample is 2.07, which is quite similar to the 2.14 we found in our baseline model. We also find similar results in Panel B for the model from Dechow et al. (2011a). These results suggest that the link between observed fraud and the fraud score textual disclosure variable cannot be explained by popular measures from financial statements.

Panels C and D consider popular discretionary accrual models. We find that both models also do not alter our inferences materially. In both cases, the full sample results are highly significant at the 1% level and the matched sample results are significant at the 5% level. We conclude that these controls have little influence on our study's conclusions.

We conclude this section with a formal analysis of fraud detection success rates using an approach similar to that in Beneish (1999a,b). We start with a benchmark regression model where the AAER dummy is the dependent variable and the RHS variables include (A) the eight variables from Beneish (1999a,b), (B) the six disclosure-motivated control variables we considered in Table 5, and (C) year and industry fixed effects. We then sort firms into deciles based on the predicted value from this regression and define the base model's success rate (P_{base}) as the fraction of firms in the highest predicted fraud decile that actually committed fraud. We then repeat this calculation after adding the 10-K based fraud score as an additional RHS variable. We define this augmented model's success rate as $P_{augment}$. Finally we also run a control model that only includes the year and industry fixed effects, and define this model's success rate as $P_{industry}$. The improved detection achieved by including the fraud score is then $\left[\frac{P_{augment}-Pindustry}{P_{base}-Pindustry}-1\right]$. This calculation intuitively measures relative improvement after acknowledging that industry and year fixed effects do not belong to any particular model. We repeat this procedure for all four existing models in Table 6.

[13] The restatement words variable is logarithm of one plus the number of times the word "restatement" appears in the firm's MD&A section of the firm's 10-K text. The litigation dummy is the logarithm of one plus the number of times the word "litigation" appears in the firm's MD&A section of the firm's 10-K text. These two controls are intended to maximize their ability to explain our results given that we report later that these particular words are significantly related to post-AAER firms. We control for uncertainty using the standard deviation of monthly stock returns from the previous year, and we also include a dummy that is equal to one if the given firm-year observation does not have adequate CRSP data to compute this variable. The acquisition dummy is one if the firm was an acquirer in a merger, or in an acquisition of assets transaction from SDC Platinum, in the previous year.

Table 6

Disclosure outcome regressions (control for various fraud models). The table reports our baseline OLS regressions for our sample of 55,666 firm-year observations based on annual firm observations from 1997 to 2010. We report results for two samples (noted in the first column). The full sample regressions are estimated with year and firm fixed effects, and standard errors are clustered by firm. The matched sample regressions limit the sample to the much smaller matched sample, which contains the 552 firm-years known to be associated with fraud, and for each such observation, the matched non-fraudulent firm-year in the same industry-size-age (ISA) matched grouping that is the closest match in terms of total assets. The matched sample tests thus include only year fixed effects, and standard errors are clustered by industry. In each panel, we consider the same regressions in Panel A and Panel F of Table 5 for the fraud similarity score variable as the dependent variable, and in each panel we add additional controls for an existing model of earnings management or fraud. In Panel A, we add the variables used in the model in Beneish (1999a). In Panel B, we add the variables used in the model in Dechow et al. (2011a). In Panel C, we add absolute discretionary accruals as used in the modified Jones (1991) Model (Modified in Dechow et al., 1995). In Panel D, we add absolute discretionary accruals as used in the Ball and Shivakumar (2006) model. All variables are defined in the appendix. For all accruals models, we compute total accruals using the cash flow statement method in Eq. (3) of Collins and Hribar (2002).

Row	Sample	AAER dummy	DSRI	GMI	AQI	SGI	DEPI	SGAI	LVGI	TATA	OI /sales	R&D /sales	CAPX /sales	Log sales	Peer Tob Q	Peer OI/sales	Obs.
Panel A: Baseline plus Beneish (1999a) Model																	
(1)	Full	0.033	0.001	−0.003	0.043	0.001	0.000	0.001	−0.000	0.002	−0.007	0.004	0.000	0.005	0.003	−0.000	55,666
		(6.20)	(1.45)	(−2.52)	(0.87)	(2.02)	(0.28)	(0.59)	(−0.19)	(0.85)	(−2.65)	(3.93)	(0.11)	(4.15)	(9.94)	(−1.24)	
(2)	Matched	0.107	0.005	0.002	0.114	0.002	0.017	−0.008	0.001	0.039	0.015	0.010	−0.017	−0.002	0.012	0.009	1,104
		(2.07)	(1.20)	(1.08)	(0.27)	(0.31)	(1.07)	(−0.49)	(0.10)	(1.11)	(0.80)	(0.73)	(−2.13)	(−0.87)	(2.53)	(3.44)	

Row	Sample	AAER Dummy	WC Accruals	RSST Accruals	CHG RECC	CHG INV	PCT SOFT	C-SALES	CHG ROA	Actual Issuance	OI /sales	R&D /sales	CAPX /sales	Log sales	Peer Tob Q	Peer OI/sales	Obs.
Panel B: Add Dechow et al. (2011a) Model																	
(3)	Full	0.032	−0.010	0.013	0.039	0.030	−0.002	−0.000	−0.009	0.001	−0.009	0.003	−0.001	0.005	0.002	−0.000	55,666
		(6.11)	(−1.61)	(5.57)	(5.16)	(2.81)	(−0.53)	(−1.75)	(−3.73)	(0.27)	(−3.47)	(3.57)	(−0.28)	(4.18)	(9.61)	(−1.18)	
(4)	Matched	0.108	−0.136	0.056	0.202	0.040	−0.059	−0.000	0.025	0.008	0.004	−0.001	−0.028	−0.002	0.011	0.009	1,104
		(2.11)	(−1.89)	(1.56)	(2.09)	(0.31)	(−2.44)	(−0.63)	(0.56)	(0.57)	(0.27)	(−0.08)	(−4.56)	(−0.73)	(2.36)	(3.34)	

Row	Sample	AAER dummy	Abs. discret. accruals	OI /sales	R&D /sales	CAPX /sales	Log sales	Peer Tob Q	Peer OI/sales	Obs. Obs.
Panel C: Add accruals from modified Jones (1991) Model										
(5)	Full	0.033	0.013	−0.006	0.004	0.000	0.004	0.003	−0.000	55,666
		(6.13)	(1.68)	(−2.48)	(3.67)	(0.22)	(3.92)	(10.12)	(−1.24)	
(6)	Matched	0.107	0.073	0.023	0.010	−0.016	−0.003	0.011	0.008	1104
		(2.10)	(0.87)	(1.55)	(0.81)	(−2.01)	(−1.11)	(2.83)	(3.47)	
Panel D: Add accruals from Ball and Shivakumar (2006) model										
(7)	Full	0.033	0.011	−0.006	0.004	0.000	0.004	0.003	−0.000	55,666
		(6.14)	(1.29)	(−2.47)	(3.65)	(0.22)	(3.92)	(10.13)	(−1.24)	
(8)	Matched	0.108	0.010	0.021	0.009	−0.016	−0.003	0.012	0.008	1104
		(2.14)	(0.11)	(1.42)	(0.76)	(−2.00)	(−1.18)	(2.85)	(3.33)	

These calculations indicate that $P_{base} = \{4.7\%, 5.2\%, 5.1\%, 4.9\%\}$ for the four models {Beneish, Dechow et al., Jones, Ball and Shivakumar}, respectively. After adding the fraud score, the success rates respectively increase to $P_{augment} = \{5.2\%, 5.8\%, 5.6\%, 5.6\%\}$. Because the industry and year fixed effects alone generate a success rate of 2.72%, it follows that the relative improvement achieved by the fraud score over and above the existing models is {22.2%, 26.5%, 22.6%, 34.1%}, respectively. We conclude that the fraud score variable adds, on average, a 25% improvement in success rates. This improvement is economically large and it is noteworthy that it is achieved simply by adding a single variable to the prediction models (the existing models include many variables). We also note that this approach is very conservative, and any signal that is shared by the fraud score and the various models is fully credited to the existing models. This understates the overall informativeness of the fraud score.

5.2. Placebo tests

Table 7 uses the same framework as Table 5, except that we consider the pre-AAER dummy (a dummy that is one if the firm will be involved in an AAER in the next fiscal year) and the post-AAER dummy (analogous AAER dummy based on the past fiscal year) as an explanatory variable instead of the actual AAER dummy. As a result, we are testing if fraud similarity is elevated in the year prior to and the year after the fraud period. This allows us to test hypotheses predicting that disclosure will strictly relate to the act of committing fraud and in the years it is committed, and not to passive long-term firm characteristics. We thus expect that the results should be substantially weaker than those in Table 5. As noted in the second column, we report separate regressions in each panel for the pre-AAER dummy and the post-AAER dummy.

Table 7 shows uniformly weak and statistically insignificant links between fraud profile similarity and both the pre-AAER dummy and the post-AAER dummy. This finding holds in all panels including the full sample, the out of sample period, and for large and small firms. This reinforces the graphical depiction of the average fraud score in Fig. 2, which shows that fraud scores quickly dissipate to zero outside the fraud period. We conclude that our evidence in Table 5 is strongly linked to the specific years that firms are allegedly engaged in fraud and our results cannot be explained by passive long-term firm characteristics.

In some specifications, including the full sample in Panel A and the full sample with added controls in Panel E, the post-AAER dummy is negative and marginally significant (opposite the baseline results where it is strongly positive). This finding suggests that, after they are caught, firms might adopt disclosures that distance themselves from prior bad behavior. One can think of this result as the "Repentant Manager" hypothesis. Overall, we conclude that our results are not related to passive firm characteristics, and are unique to firms committing fraud.

6. Content analysis

The results in the previous section support the conclusion that fraudulent firms have a strong common component to their disclosure that is unique to the specific years in which they commit fraud. However, these tests do not explain why MD&A disclosures have a strong abnormal component when firms commit fraud. In this section, we consider content analysis using the 75 verbal factors based on Latent Dirichlet Allocation (LDA) from Ball et al. (2013).[14]

LDA generates two detailed data structures. The first is the set of word-frequency distributions for each topic. For LDA with 75 topics, this data structure contains 75 word lists with corresponding word frequencies. The word lists also include commongrams, which are 2–3 word phrases that appear frequently in paragraphs that load highly on each topic. As do Ball et al. (2013), we fit the LDA model using only the first year of our sample (1997) to ensure there is no look ahead bias in our regressions. The second data structure quantifies the extent to which each of the 75 topics is discussed in individual MD&As. These firm-year variables are commonly referred to as "topic loadings". For each firm in each year, LDA provides a vector of length 75 stating the extent to which the given firm's MD&A discusses each of the 75 topics.

We use both data structures to provide two computer-generated resources for each topic that can be used to interpret each topic's content. The first is the list of the highest frequency commongrams, which is a list of key phrases that associate with each topic. The second is a complete "representative paragraph", which is a paragraph that best represents the content that is typical among firms that use the topic.

We compute representative paragraphs by first extracting the 1000 paragraphs that have the highest cosine similarity with the probability-weighted word list associated with each topic. We note that paragraphs are the standard unit of observation in the metaHeuristica software program and hence no additional steps are needed to parse or extract. We also note that the average paragraph in our sample contains 86 words. The second step is to sort these paragraphs by their document lengths, and extract the middle tercile, which contains the set of typical length representative paragraphs. Among these 333 candidate paragraphs, we then define the displayed "representative paragraph" as the paragraph with the highest total similarity to the other 332 paragraphs in this set. This calculation is akin to computing network centrality, and hence the chosen "representative paragraph" should indeed be representative of the content associated with the given topic (thus making the topic more easily interpretable).

[14] The number of topics is the only material input the researcher needs to specify when running LDA. We use 75 topics following Ball et al. (2013), who document that 75 topics best summarize the value relevant information in MD&As. These topics were generated using the metaHeuristica software program.

Table 7
Disclosure outcome regressions (pre-AAER and post-AAER disclosures). In Panels A to C, the table reports our baseline OLS regressions for our sample of 55,666 firm-year observations based on annual firm observations from 1997 to 2010. These regressions are the same as those in Table 5, except (A) we only report results for the fraud similarity score, and (B) we replace the AAER dummy with either the post-AAER dummy or the pre-AAER dummy (as indicated in the second column) to operationalize our placebo tests on both sides of the actual fraud dates. These baseline regressions are estimated with year and firm fixed effects. The dependent variable is based on a firm-year's disclosure in its 10-K and varies by row as indicated. For all Panels, standard errors are clustered by firm and *t*-statistics are in parentheses. See Table 1 for the description of our key variables. The AAER dummy is our primary variable of interest and is one if an AAER action indicates that the firm was involved in fraudulent activity in the current year. To control for the economic conditions associated with the information in a given firm's MD&A, we additionally include controls for the average Tobins Q and Operating Income/Sales for the ten firms with MD&A Sections that are most similar to the given firm (these ten firms are those with the highest cosine similarity between their MD&A and that of the given firm in the given year). Panels D to E consider various robustness tests regarding the baseline model in Panel A. Panel D repeats the test in Panel A but only for our out of sample period including 32,553 annual firm observations from 2002 to 2010. These tests are out of sample because the base vocabulary used to compute fraud similarity is fitted only using the earlier subsample from 1997 to 2001. One observation is one firm in one year. Panel E repeats the test in Panel A but adds three additional control variables aimed at challenging whether our results are due to narrower effects that have been documented in other studies but that might be better measured using text: a restatement variable, a litigation variable, a control for uncertainty, a control for inadequate CRSP data to compute uncertainty, and an acquisition dummy. Although we include these variables in the model, we do not report the additional coefficients to conserve space (see the Online Appendix for the presentation of those coefficients).

Row	Dependent variable	Pre- or post-AAER dummy	Operating income/sales	R&D/sales	CAPX/sales	Log sales	MD&A peer implied Tobins Q	MD&A peer implied OI/sales	Obs.
Panel A: Entire sample									
(1)	Pre-AAER dummy	−0.006	−0.007	0.004	0.001	0.005	0.003	−0.000	55,666
		(−0.64)	(−2.75)	(3.92)	(0.31)	(4.87)	(9.89)	(−1.26)	
(2)	Post-AAER dummy	−0.013	−0.007	0.004	0.001	0.005	0.003	−0.000	55,666
		(−1.82)	(−2.77)	(3.92)	(0.31)	(4.89)	(9.90)	(−1.26)	
Panel B: Above median firm size only									
(3)	Pre-AAER dummy	−0.008	0.012	0.047	0.008	0.008	0.003	−0.000	28,754
		(−0.80)	(1.97)	(2.69)	(2.06)	(3.56)	(7.74)	(−0.18)	
(4)	Post-AAER dummy	−0.012	0.012	0.047	0.008	0.008	0.003	−0.000	28,754
		(−1.45)	(1.95)	(2.69)	(2.06)	(3.57)	(7.75)	(−0.18)	
Panel C: Below median firm size only									
(5)	Pre-AAER dummy	−0.011	−0.011	0.003	0.000	0.006	0.002	−0.000	26,912
		(−0.77)	(−3.78)	(3.42)	(0.13)	(4.13)	(7.16)	(−1.29)	
(6)	Post-AAER dummy	−0.012	−0.011	0.003	0.000	0.006	0.002	−0.000	26,912
		(−1.00)	(−3.80)	(3.42)	(0.12)	(4.15)	(7.17)	(−1.29)	
Panel D: Entire sample (Out of sample years only)									
(7)	Pre-AAER dummy	−0.020	−0.013	0.003	0.003	0.005	0.002	−0.001	32,553
		(−1.23)	(−3.99)	(3.40)	(1.45)	(3.78)	(3.58)	(−2.46)	
(8)	Post-AAER dummy	−0.001	−0.013	0.003	0.004	0.005	0.002	−0.001	32,553
		(−0.16)	(−4.00)	(3.41)	(1.45)	(3.81)	(3.58)	(−2.46)	
Panel E: Same as Panel A, but add additional controls									
(9)	Pre-AAER dummy	−0.007	−0.005	0.003	0.001	0.004	0.003	−0.000	49,014
		(−0.87)	(−1.77)	(3.28)	(0.40)	(3.99)	(9.95)	(−1.10)	
(10)	Post-AAER dummy	−0.019	−0.005	0.003	0.001	0.004	0.003	−0.000	49,014
		(−2.55)	(−1.80)	(3.28)	(0.40)	(4.02)	(9.96)	(−1.10)	

We then use the panel database containing the 75 numeric topic loadings for each firm in each year, and estimate regressions to infer which of the 75 verbal topics are most related to abnormal disclosures during periods of fraud, relative to disclosures the same firms make during years they are not allegedly committing fraud. The topics that are significantly different can then be interpreted regarding their consistency with our hypotheses.

6.1. Abnormal content in AAER-years

Table 8 displays the results of 75 regressions that treat each of the LDA topic loadings as a dependent variable. We only report topics for which the AAER dummy is significantly different from zero at the 5% level or better.[15] We control for year

[15] To be conservative, we do not report 10% level significant results given the number of specifications.

Table 8
LDA topics driving fraud similarities. The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in AAER actions as compared to firms not involved in AAER actions (the last column). In addition to the topic commongrams and the representative paragraphs that describe each topic's content, the table displays coefficients and *t*-statistics for regressions where firm-year topic loadings are regressed on the AAER dummy. We only report results for the 5% level significant topics from the set in Table IX of Ball et al. (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported *t*-statistics are also adjusted for clustering by firm.

	Topic commongrams	Representative paragraph	Different in AAER years
1	Board directors, executive officers, officers directors, vice president, directors officers	Since joining the company in January 1998, the new chief executive officer, along with the rest of the company's management team has been developing a broad operational and financial restructuring plan. A broad outline of that plan has been presented to the company's board of directors in march 1998. the plan, which is designed to leverage the company's brand, distribution and technology strengths, includes reducing costs, outsourcing of certain components and products, disposition of certain assets and capitalizing on the company's patented digital television technologies. Restructuring costs must be incurred to implement the plan.	−0.250 (−3.36)
2	Acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	In august 1996, Prologic completed its acquisition of basis. all of the outstanding stock of basis was acquired for 337,349 shares of common stock of the company plus $500,000 in cash. the acquisition was accounted for as a purchase and accordingly the purchase price and all expenses directly associated with the acquisition were allocated to the assets acquired and the liabilities assumed based on their fair market values at the date of the acquisition determined by management estimates. goodwill in the amount of $1,459,661 was recorded in connection with the acquisition.	0.244 (3.50)
3	Gain sale, held sale, sale leaseback, gains sale, realized gains	The gain on sale of assets of $4.2 million, recognized in 1997, was associated with the sale of the company's interest in several blocks in India, the sale of an investment in a Philippines company and the sale of the gulf of Mexico properties. The 1996 gain on sale of assets of $1.0 million was from the sale of certain interests in India.	−0.204 (−3.58)
4	Legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	Assurance that any patent owned by the company will not be invalidated, circumvented or challenged, that the rights granted thereunder will provide competitive advantages to the company or that any of the company's pending or future patent applications will be issued with the scope of the claims sought by the company, if at all. Furthermore, there can be no assurance that others will not develop similar products or software, duplicate the company's products or software or design around the patents owned by the company or that third parties will not assert intellectual property infringement claims against the company. In addition, there can be no assurance that foreign intellectual property laws will adequately protect the company's intellectual property rights abroad. the failure of the company to protect its proprietary rights could have a material adverse effect on its business, financial condition and results of operations.	−0.202 (−2.61)
5	Partially offset, primarily due, offset decrease, due primarily, decreased decrease	Income from operations for 1997 totaled $974,549, an increase of $121,707 (14.3%) from 1996. The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense. Income from operations for 1996 totaled $852,842, a decrease of $40,532 (4.5%) from 1995. The decrease was due primarily to the increased costs of operating the company-owned restaurants.	−0.160 (−2.77)
6	Entered agreement, agreement dated, terms agreement, pursuant terms, agreement entered	During 1996, the company entered into a $55.0 million credit, security, guarantee and pledge agreement (the"1996 credit agreemen") with four financial institutions. permitted borrowings, as defined in the credit agreement, generally include acquisitions and capital expenditures. Borrowings under this agreement bear interest at either Libor plus 2.50%, 2.75% or 3.0%, or prime rate plus 1.0%, 1.25% or 1.50%, as defined in the agreement. At December 31, 1996, the company had borrowed the maximum of $55.0 million under the credit agreement. At June 30, 1997, the outstanding obligations under this agreement were paid in full in conjunction with the execution of the credit agreement described above.	−0.159 (−3.19)
7	Merrill lynch, market risk, balance sheet, hedging transactions, derivative instruments	The company uses derivative financial instruments to reduce certain types of financial risk. specifically, (1) foreign currency forward contracts, option contracts, and swap contracts are employed to reduce the risk of foreign currency fluctuations on future cash flows, and, (2) to a lesser extent, interest rate swaps are employed to effectively convert the company's outstanding floating rate debt to fixed rate debt. Hedging strategies and transactions are reviewed and approved by management before being implemented. Use of derivatives is limited to simple, non-leveraged instruments. Monthly market valuations and sensitivity analyses are performed to monitor the effectiveness of the company's risk management program.	0.159 (3.05)
8	Continued growth, business strategy, growth strategy, business opportunities, core business	The company's business has grown significantly since its inception, and the company anticipates future growth. The growth of the company's business and the expansion of its customer base have resulted in a corresponding growth in the demands on the company's management and personnel and its operating systems and internal controls. Any future growth may further strain existing management resources and operational, financial, human and management information systems and controls, which may not be adequate to support the company's operations.	0.155 (2.35)

(*continued on next page*)

Table 8 (*continued*)

	Topic commongrams	Representative paragraph	Different in AAER years
9	Adversely affected, results operations, operating results, negatively impacted, adversely impacted	Although management believes that inflation has not had a material effect on the results of its operations during the past three years, there can be no assurance that the company's results of operations will not be affected by inflation in the future.	0.151 (2.18)
10	Costs incurred, cost goods sold, labor costs, cost savings, costs related	The company incurred operating costs and expenses of approximately $6,571,000 for the year ended june 30, 1996 as compared to approximately $13,845,000 for the year ended june 30, 1995. Debt restructuring costs increased by $600,000 and represent fees owed to the company's financial advisors. the company realized significant cost decreases in development ($665,000), administrative costs including advertising ($1,514,000), payroll costs ($868,000), professional fees ($4,232,000), public company costs($452,000), travel costs ($570,000), and other operating costs ($336,000). some of these cost reductions were absorbed by the CDGC subsidiary.	0.139 (2.55)
11	Interest rate, interest rates, fixed rate, variable rate, prime rate	The company entered into two interest rate swaps with major financial institutions to exchange variable rate interest for fixed rate interest. The net result was to substitute a weighted average fixed interest rate of 7.81% for the variable libor rate on $13.0 million of the company's debt. the interest rate swaps expire in october and november of 2001. The company entered into an interest rate collar agreement with a major bank for $10.0 million. the agreement limits the net interest rate charged to 8.25%. the company receives no further interest rate benefit once the applicable interest rate falls below 6.55%. This agreement matures in june 1998.	0.127 (2.42)
12	Marketing expenses, professional fees, salaries benefits, expenses related, related expenses	Sales and marketing expenses increased to $835,000 for the year ended december 31, 1997 from $221,000 in 1996. the increase was due primarily to compensation and recruiting costs, product design costs, advertising expenses and other marketing expenses related to introduction of the company's infant jaundice product. Sales and marketing expenses are expected to increase in the future as the company begins to market this product.	−0.111 (−2.43)
13	Operating loss, operating losses, net operating, net operating loss, operating profits	The company recorded operating income of $392,566 for fiscal 1996 as compared to an operating loss of $1,346,003 in fiscal 1995. the operating income in fiscal 1996 is attributed to increased operating revenues, increased profit margins as well as a decrease in operating expenses. The operating loss in fiscal 1995 was primarily due to a restructuring charge of $800,765.	−0.111 (−2.17)
14	Clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	The company expects to incur substantial additional research and development expense including continued increases in personnel and costs related to preclinical testing and clinical trials. the company's future capital requirements will depend on many factors, including the rate of scientific progress in its research and development programs, the scope and results of preclinical testing and clinical trials, the time and costs involved in obtaining regulatory approvals, the costs involved in filing, prosecuting and enforcing patent claims, competing technological and market developments, the cost of manufacturing scale-up, commercialization activities and arrangements and other factors not within the company's control. the company intends to seek additional funding through research and development relationships with suitable potential corporate collaborators and/or through public or private financings. there can be no assurance that additional financing will be available on favorable terms, if at all.	−0.021 (−2.59)

and firm fixed effects, and standard errors are clustered by firm. Because we control for firm fixed effects, all reported links are conservative and based on within-firm identification. The first column reports the top 5 commongrams associated with each topic, and the second column displays the topic's representative paragraph. We focus on the AAER dummy as the independent variable of interest in the final column.

The table shows that 14 of the 75 topics are significantly linked to AAER years. These 14 topics are abnormally disclosed by firms involved in fraud relative to the same firms in non-AAER years. A positive and significant coefficient indicates that firms committing fraud disclose an abnormally long discussion of the given LDA topic relative to a strict within-firm counterfactual. A negative coefficient indicates that the firm discloses an abnormally short discussion of the given LDA topic, and hence this can be interpreted as under-disclosing the given topic. We note that each topic is well-described by its representative paragraph and list of frequent commongrams, and we remind the reader that both are generated automatically as described above. They are thus not subjected to researcher prejudice. We also note that finding 14 significant topics, many significant at the 1% level, is well beyond what one would expect by chance for 75 topics.

6.1.1. Hypotheses based on fraud verbal disclosure incentives

We next interpret the results in Table 8 through the lens of our three managerial incentive hypotheses (H1A, H1B, H1C). Regarding H1A, managerial incentives to conceal details of fraudulent accounting, row (5) is supportive. The representative paragraph shown in row (5) states for example:

> *"The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense. ... The decrease was due primarily to the increased costs of operating the company-owned restaurants."*

This paragraph explicitly explains details underlying the firm's performance, and Table 8 shows that it is negatively related to the fraud dummy at better than the 1% level. This suggests that firms committing fraud disclose less information regarding details that explain how their performance arises, which supports H1A directly.

Regarding H1B, managerial incentives to grandstand growth and strong performance, we see support in Table 8. Firms appear to grandstand their growth as shown in row (8) on page 2 of the table, as they disclose significantly more of an LDA topic with a representative paragraph that touts the firm's growth:

> *"The company's business has grown significantly since its inception..."*

Later in this section, we will report more refined evidence of hypothesis H1B when we examine subsamples of revenue fraud and expense fraud.

Regarding H1C, managerial incentives to avoid references to themselves in the presence fraudulent accounting, we observe support in row (1). The representative paragraph for this topic notes that managers often discuss their plans for the future in MD&A and associate the plans with references to themselves:

> *"Since joining the company in January 1998, the new chief executive officer, along with the rest of the company's management team has been developing a broad operational and financial restructuring plan..."*

Table 8 shows that this kind of self-reference is less likely to occur when the firm is committing fraud, which directly supports H1C as managers prefer to disassociate with the firm's disclosure when they are committing fraud, likely to insulate themselves from fallout should the fraud be discovered in the future.

We believe the above three results support our three managerial incentive hypotheses. We also note, however, that Table 8 reports 11 other significant topics. Many of these have further implications not only for our proposed hypotheses, but also can motivate future theoretical and empirical research regarding additional channels for fraud that have not yet been proposed. We discuss some of these results here, although we exercise caution in this section due to the speculative nature of discussing theories that do not yet exist. Yet the ability to produce results such as these are a strength of the LDA method because many important forces driving fraud might, in fact, have been overlooked by researchers.

The results in row (3) are also consistent with H1A as under-reporting details regarding asset sales could be a mechanism for concealing necessary details for detecting fraud due to the complexity of accounting for asset transactions. However, this result is more difficult to interpret. This discussion can also be linked to hypothesis H2B, which predicts that acquiring assets can create incentives to commit fraud, and hence fraudulent firms might be more focused on expanding their asset base (purchases) than shrinking it (sales).

The results in row (10) are potentially consistent with H1B as excessive reporting on cost savings is consistent with grandstanding outstanding expense management, and because manipulating expenses is one of the more common types of fraud. However, this result is also difficult to interpret. For example, it might reflect financially constrained firms cutting expenses in an attempt to produce internal liquidity.

Many additional results motivate potential new channels for fraud that have been discussed less in the literature. One example is row (4), which suggests that fraudulent firms under-discuss potential legal problems. One interpretation is that concealing legal problems can help the firm to better realize higher stock prices and improved liquidity consistent with an initiation to commit fraud motive. This result is particularly stark because, a priori, many would expect that firms involved in fraud might have more legal problems in general and would thus disclose more. Analogously, the results regarding the under-reporting of financial agreements as in row (6) are suggestive of new channels for understanding fraud incentives through the channel of the type and complexity of financial instruments issued by the firm. These latter results motivate future research on additional channels.

6.1.2. Hypotheses based on fraud initiation incentives

We now discuss the two fraud initiation hypotheses from the literature (H2A and H2B). Regarding H2A, that firms will under-report problems with their liquidity to maximize the likelihood of raising capital at an artificially low cost, we do not find direct support in Table 8 for our full sample of AAERs. However, we note that we do find supportive evidence when we specifically look at revenue frauds later in this section. Regarding H2B, the hypothesis that fraud is related to incentives surrounding stock-based mergers and acquisitions, finds strong support in row (2). The representative paragraph shown in row (2) states for example:

> *"Prologic completed its acquisition of basis. All of the outstanding stock of basis was acquired for 337,349 shares of common stock of the company plus $500,000 in cash."*

This representative paragraph is particularly direct in its support of our hypothesis because, not only does it directly mention acquisitions, but it also refers to the use of stock as a means of payment. Because this topic is positive and significant at the 1% level, it is consistent with managers being more likely to commit fraud when they are engaged in stock-based acquisitions. The incentives to commit fraud in this case likely relate to the incentive to temporarily inflate the stock price so that the target can be purchased at an artificially low price. These results are supportive of existing studies including Erickson and Wang (1999) and Wang (2013).

6.1.3. Placebo tests based on non-fraud periods

We next consider placebo tests and examine whether the above results are robust to replacing the AAER dummy with a pre-AAER dummy or a post-AAER dummy. If our results look materially the same in these placebo periods, that would indicate that our results would not be uniquely attributable to the periods during which firms actually commit fraud. We report the results in Table 9, where we report all topics that are significantly related to actual fraud years in the first column, pre-AAER years in the second column, and post-AAER years in the third column. The first column is thus from Table 8 for comparison.

Table 9 shows that all of our central results only exist in the actual AAER sample, and not in the pre-AAER or the post-AAER sample. Regarding the post-AAER sample, not a single coefficient is significant with the same sign. Regarding the pre-AAER sample, only the result for legal proceedings has the same sign and is significant. Given the number of topics that are related to AAERs in Table 8, these results suggest that our findings are indeed unique to the years during which firms are actually committing fraud.

6.2. Revenue and expense fraud

We next reexamine the tests in Table 8 specifically for revenue fraud and expense fraud. These tests are based on a separate hand-collected database containing information directly indicating whether AAERs are due to revenue and expense fraud. A fraud is identified as a revenue fraud if its 704 report code begins with the letter "A" indicating problems with revenue

Table 9
LDA topics driving fraud similarities (placebo tests). The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in AAER actions as compared to firms not involved in AAER actions (first column after Topic Descriptions). We also report significant topics for firms in the year after, and also the year before, they are alleged to be involved in fraud (last two columns). The table displays coefficients and *t*-statistics for regressions where firm-year topic loadings are regressed on the AAER dummy, the post-AAER dummy, and the pre-AAER dummy, respectively. We only report results for the 5% level significant topics from the set in Table IX of Ball et al. (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported *t*-statistics are also adjusted for clustering by firm.

	Topic commongrams	Different in AAER years	Different in pre-AAER years	Different in post-AAER years
1	Board directors, executive officers, officers directors, vice president, directors officers	−0.250 (−3.36)		
2	Acquisition, connection acquisition, acquired businesses, completed acquisition, acquisition accounted	0.244 (3.50)		−0.175 (−2.67)
3	Gain sale, held sale, sale leaseback, gains sale, realized gains	−0.204 (−3.58)		0.150 (2.47)
4	Legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	−0.202 (−2.61)	−0.151 (−2.05)	0.224 (2.45)
5	Partially offset, primarily due, offset decrease, due primarily, decreased decrease	−0.160 (−2.77)		
6	entered agreement, agreement dated, terms agreement, pursuant terms, agreement entered	−0.159 (−3.19)		0.136 (2.42)
7	Merrill lynch, market risk, balance sheet, hedging transactions, derivative instruments	0.159 (3.05)		−0.102 (−2.13)
8	Continued growth, business strategy, growth strategy, business opportunities, core business	0.155 (2.35)		
9	Adversely affected, results operations, operating results, negatively impacted, adversely impacted	0.151 (2.18)		
10	Costs incurred, cost goods sold, labor costs, cost savings, costs related	0.139 (2.55)	−0.152 (−2.05)	
11	Interest rate, interest rates, fixed rate, variable rate, prime rate	0.127 (2.42)		
12	Operating loss, operating losses, net operating, net operating loss, operating profits	−0.111 (−2.17)		
13	Marketing expenses, professional fees, salaries benefits, expenses related, related expenses	−0.111 (−2.43)		
14	Clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	−0.021 (−2.59)		
15	Investment securities, marketable securities, equity securities, investment portfolio, backed securities		0.177 (2.45)	
16	Foreign currency, foreign exchange, north america, currency exchange, domestic international		0.144 (2.15)	
17	Joint venture, joint ventures, limited partnership, minority interests, tci group		−0.128 (−2.61)	
18	License fees, consulting services, consulting fees, service fees, services provided		−0.104 (−2.22)	
19	Restructuring charge, restructuring charges, write downs, special charges, fourth quarter			0.181 (2.57)
20	Payments made, principal payments, payment dividends, pay dividends, dividends paid			0.128 (2.04)
21	Initial public offering, private placement, net proceeds, proceeds offering, public private			−0.127 (−3.09)
22	Net sales, sales sales, sales volume, sales marketing, sales force			−0.103 (−2.36)

recognition. A fraud is identified as an expense fraud if the first letter of this code begins with "B" indicating problems with expense recognition.[16]

These tests based on instances of revenue and expense fraud allow us to more precisely examine hypotheses H1A (incentives to conceal details), H1B (incentives to grandstand), and H2A (incentives to reduce the cost of capital for issuance). These hypotheses make unique predictions when managers commit revenue fraud or expense fraud. For example, given that a manager has committed revenue fraud, H1B predicts that managers will grandstand the strong revenue growth. In contrast, when the manager commits expense fraud, H1B predicts more grandstanding of the manager's cost management.

Table 10 reruns the same specifications in Table 8, which is based on all frauds, specifically for revenue and expense frauds. Regarding hypothesis H1B (grandstanding), row (4) provides direct evidence. The corresponding representative paragraph begins with:

> *"Revenues increased by $29.9 million, or approximately 27.4%, to $139.1 million in 1997 from $109.2 million in 1996."*

Managers committing revenue fraud abnormally disclose longer discussions touting their revenue performance. We also find that their MD&As under-disclose details regarding how their performance arises. This latter conclusion is supported by both row (1) and row (5), which directly indicate the disclosure of fewer details explaining the performance of the firm. For example, the representative paragraph associated with row (1) starts with:

> *"Income from operations for 1997 totaled $974,549, an increase of $121,707 (14.3%) from 1996. The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense."*

The results for expense fraud also support both H1A and H1B. Regarding H1B, row (8) shows that managers committing expense fraud disclose abnormally high levels of disclosure relating to R&D expense performance. The representative paragraph starts with:

> *"Research and development expenses increased 20.7% to $6,006,000 in 1996, and increased as a percentage of net sales to 10.0% in 1996 from 6.1% in 1995. The increases in research and development expenses were primarily due to the expansion of the research and development staff, and expenses associated with its research and development facility."*

This is consistent with grandstanding to convince investors that the firm has strong growth options consistent with a vibrant R&D program. These firms also disclose fewer details explaining their performance, as illustrated by the negative coefficient in row (1) regarding broad performance details and row (7) regarding specific cost reduction details.

Regarding H2A, row (3) shows that managers committing revenue fraud disclose abnormally low levels of disclosure relating to issues with firm liquidity. The representative paragraph starts with:

> *"The company believes that its current cash, cash equivalents and short-term investment balances and cashflow from operations, if any, will be sufficient to meet the company's working capital and capital expenditure requirements for at least the next twelve months. Thereafter, the company may require additional funds to support its working capital requirements or for other purposes and may seek to raise such additional funds through public or private equity financing or from other sources."*

These results suggest that firms under-disclose discussions of liquidity when they are committing revenue fraud. Indeed firms that are focused on revenue tend to be earlier-stage firms, and these firms likely find it difficult to raise equity capital. We also note that these regressions control for firm and year fixed effects. Hence, these results attribute the results specifically to the year the firm is committing fraud and not to surrounding years.

6.3. Robustness

As an additional robustness examination, we identify the individual words that are used more aggressively by AAER firms. These words are identified based on word-by-word tests of differences in each word's relative usage among AAER firms versus non-AAER firms. The details of this analysis are not reported here but are available in Table A2 of our online appendix. Table A2 shows that AAER years are often linked to individual words that might be indicative of governance issues including "duties", "investments", and "divert". We also find confirming evidence that acquisitions play a role, as the word "acquired" is the second most significant word that is linked to AAERs.

Our general conclusion, however, is that individual words are more difficult to interpret than are the results for LDA discussed previously. This comparison thus highlights how word-clustering methods like LDA can add clarity to content analysis. We also present a list of the top 25 most representative AAERs in Online Table A3, which lists the AAERs that have the highest fraud similarity scores.

[16] We only consider these two categories as other categories of fraud including those linked to business combinations or the "other" category for 704 report codes are not frequent enough to give us power to run analogous tests.

Table 10
LDA topics driving fraud similarities (revenue and expense fraud). The table lists the Topic Model Factors found to be statistically significant regarding their link to firms involved in specific types of AAER actions (revenue and expense fraud) as compared to firms not involved in these actions. We thus report significant topics separately for firms involved in revenue fraud and expense fraud. In addition to the topic commongrams and the representative paragraphs that describe each topic's content, the table displays coefficients and *t*-statistics for regressions where firm-year topic loadings are regressed on the Revenue-AAER dummy and the Expense-AAER dummy, respectively. We only report results for the 5% level significant topics from the set in Table IX of Ball et al. (2013). These regressions include controls for firm fixed effects and year fixed effects. Hence the significant topics listed here are statistically different for firms involved in AAERs even compared to the same firm prior to or after the AAER years. The reported *t*-statistics are also adjusted for clustering by firm.

	Topic commongrams	Representative paragraph	Revenue AAER years	Expense AAER years
1	Partially offset, primarily due, offset decrease, due primarily, decreased decrease	Income from operations for 1997 totaled $974,549, an increase of $121,707 (14.3%) from 1996. The increase was due primarily to the increase in sales, the decreased percentage of the general and administrative expenses and the decrease in depreciation and amortization expense. Income from operations for 1996 totaled $852,842, a decrease of $40,532 (4.5%) from 1995. The decrease was due primarily to the increased costs of operating the company-owned restaurants.	−0.221 (−2.57)	−0.239 (−2.42)
2	Legal proceedings, bankruptcy court, litigation settlement, settlement litigation, proprietary rights	Assurance that any patent owned by the company will not be invalidated, circumvented or challenged, that the rights granted thereunder will provide competitive advantages to the company or that any of the company's pending or future patent applications will be issued with the scope of the claims sought by the company, if at all. Furthermore, there can be no assurance that others will not develop similar products or software, duplicate the company's products or software or design around the patents owned by the company or that third parties will not assert intellectual property infringement claims against the company. In addition, there can be no assurance that foreign intellectual property laws will adequately protect the company's intellectual property rights abroad. The failure of the company to protect its proprietary rights could have a material adverse effect on its business, financial condition and results of operations.	−0.210 (−2.06)	
3	Sufficient meet, additional financing, sources liquidity, raise additional, additional funds	The company believes that its current cash, cash equivalents and short-term investment balances and cash flow from operations, if any, will be sufficient to meet the company's working capital and capital expenditure requirements for at least the next twelve months. Thereafter, the company may require additional funds to support its working capital requirements or for other purposes and may seek to raise such additional funds through public or private equity financing or from other sources. There can be no assurance that additional financing will be available at all or that if available, such financing will be obtainable on terms favorable to the company.	−0.162 (−2.60)	
4	Total revenues, revenues derived, percentage revenues, revenues revenues, revenues generated	Revenues increased by $29.9 million, or approximately 27.4%, to $139.1 million in 1997 from $109.2 million in 1996. This increase was primarily due to revenues generated by increased sales of commercial airtime inventory. acquisitions accounted for $5.7 million of this increase. excluding these revenues, same market revenues increased $24.2 million in 1997, or 22.2%. Revenues from reciprocal arrangements as a percentage of total revenues declined to 4.0% in 1997 from 8.0% in 1996.	0.162 (2.33)	
5	Partially offset, offset lower, partially offset lower, due higher, due lower	Earnings in 1996 were $34.3 million better than 1995. The improvement was mainly due to higher volumes in the company's major product lines, higher cement and U.S. ready-mixed concrete prices, lower operating costs in construction materials operations and lower imports to supplement production in the U.S. these increases were partially offset by lower divestments gains and lower clinker production in Canada.	−0.138 (−2.20)	
6	Clinical trials, research development, collaborative partners, collaborative arrangements, regulatory approvals	The company expects to incur substantial additional research and development expense including continued increases in personnel and costs related to preclinical testing and clinical trials. The company's future capital requirements will depend on many factors, including the rate of scientific progress in its research and development programs, the scope and results of preclinical testing and clinical trials, the time and costs involved in obtaining regulatory approvals, the costs involved in filing, prosecuting and enforcing patent claims, competing technological and market developments, the cost of manufacturing scale-up, commercialization activities and arrangements and other factors not within the company's control. The company intends to seek additional funding through research and development relationships with suitable potential corporate collaborators and/or through public or private financings. There can be no assurance that additional financing will be available on favorable terms, if at all.	−0.039 (−2.14)	

Table 10 (*continued*)

	Topic commongrams	Representative paragraph	Revenue AAER years	Expense AAER years
7	Efforts reduce, order reduce, effort reduce, economies scale, cutting measures	Factors which management believes may affect the future financial performance of the company include but are not limited to: the successful implementation of manufacturing and business processes which will reduce costs and improve efficiency; the investment in engineering and marketing activities which lead to improved sales growth; the successful integration of acquisitions into the company's operations; dealing with the external regulatory influences on the company's primary markets; and the effect on the competitive environment resulting in the consolidation of companies within the instrumentation industry.		−0.359 (−2.21)
8	Research development, research development expenses, product development, process research development, development stage	Research and development expenses increased 20.7% to $6,006,000 in 1996, and increased as a percentage of net sales to 10.0% in 1996 from 6.1% in 1995. The increases in research and development expenses were primarily due to the expansion of the research and development staff, and expenses associated with its research and development facility. The majority of these increased costs related to the support of the apex product development.		0.307 (2.49)
9	Management believes, management team, asset management, management aware, independent auditors	Pursuant to the management services agreement with standard management, premier life (luxembourg) paid standard management a management fee of $25,000 per quarter during 1997 and 1996 for certain management and administrative services. The agreement provides that it may be modified or terminated by either standard management or premier life (luxembourg).		0.206 (2.28)
10	Product line, product lines, product sales, distribution channels, product introductions	Historically, the company has introduced several new products each year. in prior years any increase in sales volume related to the new products was offset by discontinued products. In 1997, the company's product line was increased to 120 products from approximately 110 in 1996. Most of the new products were part of the new "Caribe Line" which replaced two colognes. As a result, sales of the fragrance product line increased approximately $342,000, or 163%, to $552,000 in 1997 compared to $210,000 in 1996. In 1998, the company plans to continue to expand its product line in an attempt to increase net product sales in North America.		0.103 (2.49)

In online appendix Table A4, we examine which topics are disclosed abnormally when SEC Comment Letters are issued. Because the comment letter process is the review of verbal disclosure by the SEC, this test thus examines if our results are driven by known SEC evaluation processes, or if they are instead linked to incentives as suggested by our hypotheses. We thus consider the same regression model as in Table 8, but we replace the AAER dummy with a dummy indicating whether the firm received a comment letter from the SEC relating to its MD&A disclosure in the given year from Audit Analytics. The table shows that there is little overlap between the disclosure of firms receiving comment letters and those committing fraud. The results thus support the conclusion that our results are not artifacts of the SEC review process itself.

Online appendix Table A5 tests whether managers use language that is difficult to read in order to obfuscate their disclosures. We compute the Gunning Fog Index for each firm's MD&A in each year, and consider regressions analogous to those in Table 5 where the Gunning Fog Index is the dependent variable. Under this hypothesis, we expect the AAER dummy to be a positive and significant predictor of the Gunning Fog Index. The formula for the Gunning Fog Index is $0.4\left[\frac{\#words}{\#sentences}+\frac{\#complexwords}{\#words}\right]$, where complex words are those with three or more syllables. We also consider the Automated Readability Index and the Flesch Kinkaid Index for robustness. The results do not support the hypothesis that managers use complex text when they are involved in AAERs. These results suggest that various fog indices might not be appropriate for measuring document complexity that might be associated with fraud. For example, firms that work in more technical product markets might use longer words to describe their operations.

7. Equity market liquidity

This section more deeply considers hypothesis H2A: the proposed link between fraudulent firm disclosure, equity market liquidity and equity issuance. We examine the more specific hypothesis that managers might commit fraud to get access to an artificially low cost of capital (see Dechow et al., 1996; Povel et al., 2007 and Wang et al., 2010). We consider whether, following exogenous negative shocks to equity market liquidity, managers are more likely to commit fraud and produce disclosure with a higher fraud similarity score, likely to inflate their odds of issuing equity.

We consider the Coval and Stafford (2007) and Edmans et al. (2012) forced mutual fund selling shock as an exogenous negative shock to equity market liquidity. As this measure of forced mutual fund selling is not sector-specific, and only affects equities, it is a direct shock to equity market liquidity. The authors also find that the effects of this shock can be long lasting, as much as two years. We examine regressions in which the dependent variable is the fraud profile similarity score or the AAER dummy, and the mutual fund selling shock is a key independent variable. If improving the odds of issuing equity is a strong motive for fraud that drives the common verbal disclosures made by fraudulent firms, the prediction is that negative shocks to equity market liquidity should result in increases in the fraud profile similarity score and the AAER dummy. This prediction arises from the assumption that the incentive to commit fraud increases when liquidity conditions deteriorate.

The results are presented in Table 11 Panel A (industry and year fixed effects) and Panel B (firm and year fixed effects). The results support our prediction that negative shocks to equity market liquidity lead firms to produce disclosure with higher fraud profile similarity scores. Moreover, the same firms are more likely to be involved in an AAER in these years as shown in rows (2) and (4).

In panel C, we examine regressions in which the dependent variable is equity issuance, and the key independent variable is the fraud profile similarity. We include firm and year fixed effects. As indicated in the first column, we consider equity issuance measured two ways: Compustat equity issuance/assets and SDC Platinum public SEO proceeds/assets. Our hypothesis is that if fraudulent disclosure is made to inflate the odds of issuing equity, and if the market is not fully aware of this link, then increased fraud profile similarity should predict more equity issuance.

We note, however, that these panel C regressions are only suggestive, as the link between disclosure and equity issuance is endogenous. We are not aware of any instruments for increased fraud profile similarity disclosure that are unrelated to liquidity. The results are consistent with the conclusion that firms with high fraud profile similarity issue more equity than firms with lower scores. Overall, our results in Panels A and B suggest a potential causal link between poor equity market liquidity and elevated levels of fraud profile similarity. Panel C is consistent with a non-causal link to equity issuance.

8. Conclusions

We first examine if firms committing fraud produce abnormal disclosure that is common among firms committing fraud. We define abnormal disclosure as that which cannot be explained by industry peers of similar size and age, or by firm fixed effects and various controls. We find such an abnormal component among fraudulent firms, and any firm's verbal similarity to this abnormal vocabulary predicts ex-post fraud both in sample and out of sample. The results are economically large. Firms in the lowest fraud vocabulary similarity decile commit fraud at a rate of 0.4%, while those in the highest decile commit fraud at a rate of 2.4%. These results are also robust to controlling for firm fixed effects, and we continue to find strong results even when disclosure is compared to the same firms before and after the AAER.

Having established the presence of abnormal disclosure, we consider mechanisms. Our tests reveal a link between fraudulent firms and the under-reporting of details explaining how the firm's accounting performance arises, and grandstanding the firm's growth potential and its strong performance. These results suggest that managers might respond to incentives to conceal details that might increase detection, and incentives to grandstand growth and performance to increase the positive impact the manipulation has on the firm's outcomes. Finally, we also find evidence that fraudulent managers disclose in such a way to disassociate their own names and reputations from the fraudulent performance.

We also find interpretable textual support for two financial hypotheses noted in the existing literature: managers commit fraud to improve their odds of raising capital, and managers commit fraud to improve acquisition terms. Further supporting the former, we find both an increased disclosure of verbal content that correlates with that of fraudulent firms, and higher rates of ex-post fraud following exogenous negative shocks to equity market liquidity. Overall, our results have implications for improving the ability to detect fraud and to further understanding its underlying mechanisms.

Acknowledgments

We thank Ken Ahern, Christopher Ball, Radha Gopalan, Kathleen Hanley, Yuan Li, Tim Loughran, Vojislav Maksimovic, Bill McDonald, Gordon Phillips, Scott Smart, and Harvey Westbrook for their excellent comments and suggestions. We also thank the conference participants at the AFA meetings in Boston, the European Financial Management Association meetings in Venice and the seminar participants at Boconni University, Columbia University, Duke University, Frankfurt School of Management and Finance, George Washington University, Goethe University, Hong Kong University, London Business School, London School of Economics, Rice University, U. S. Securities and Exchange Commission, University of Colorado, University of Hong Kong Science and Technology, University of Memphis, University of Oklahoma, University of North Carolina at Chapel Hill, University of Notre Dame, University of Southern California, University of Tennessee, Knoxville, University of Washington, and U.S. Department of Treasury, Office of Financial Research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Any remaining errors are ours alone.

Table 11
Equity market liquidity and issuance. The table reports OLS regressions for our sample of observations based on annual firm observations from 1997 to 2007. Our sample is limited to 2007 due to the availability of the mutual fund selling shock database. One observation is one firm in one year. The dependent variable is the fraud profile similarity, the fraud dummy, or Compustat equity issuance divided by assets as noted in the first column. All regressions include firm and year fixed effects. In Panel A, the dependent variable is either the Fraud Score or the AAER Dummy as noted in the first column. The fraud similarity score is the cosine similarity of the given firm's abnormal disclosure and the average abnormal disclosure of all firms involved in AAERs in the sample period 1997 to 2001. The AAER dummy is one in the year prior to a year in which a given firm was involved in an AAER. Equity issuance in Panel B is either Compustat equity issuance or SDC Platinum public SEO issuance. Both are in dollars and are scaled by assets. See Table 1 for the description of our key variables. All standard errors are clustered by firm. *t*-statistics are in parentheses.

Row	Dependent variable	Forced mutual fund selling	Operating income/sales	R&D/sales	CAPX/sales	Log sales	MD&A peer implied Tobins Q	MD&A peer implied OI/sales	Obs.
Panel A: Industry and year fixed effects									
(1)	Fraud score	0.008	−0.003	0.004	−0.013	−0.000	0.004	0.001	38,707
		(8.02)	(−1.08)	(3.97)	(−5.38)	(−1.17)	(1.40)	(2.81)	
(2)	AAER dummy	0.004	−0.002	0.001	−0.000	0.004	0.001	−0.000	38,707
		(4.43)	(−0.64)	(2.49)	(−0.13)	(4.34)	(3.63)	(−0.58)	
Panel B: Firm and year fixed effects									
(3)	Fraud score	0.003	−0.006	0.002	−0.001	0.005	0.003	−0.000	38,707
		(3.22)	(−1.88)	(1.74)	(−0.25)	(3.57)	(9.11)	(−1.20)	
(4)	AAER dummy	0.002	0.000	0.001	0.003	0.005	0.000	−0.000	38,707
		(1.78)	(0.13)	(2.51)	(2.63)	(3.31)	(0.94)	(−1.34)	
Row	Dependent variable	Fraud profile similarity	Operating income/sales	R&D/sales	CAPX/sales	Log sales	MD&A peer implied Tobins Q	MD&A peer implied OI/sales	Obs.
Panel C: Equity issuance: Firm and year fixed effects									
(5)	Compustat equity issuance	0.051	−0.084	0.007	0.050	−0.029	0.018	−0.003	38,707
		(4.56)	(−7.35)	(1.67)	(6.58)	(−9.57)	(8.88)	(−3.61)	
(6)	SDC equity issuance	0.024	0.012	0.006	0.011	0.001	0.004	−0.000	38,707
		(3.10)	(2.45)	(2.44)	(3.33)	(0.62)	(6.14)	(−0.36)	

Appendix A

Table A1
The table reports variable definitions for the variables used from four existing models from the literature in Table 6. All formulas are from the original studies, which are noted in the panel headers. The formulas are stated using the variable names as provided in the Compustat database. The character Δ refers to the given variable's first difference from time $t-1$ to time t. For both accruals models in Panel C and D, we compute total accruals using the cash flow statement method in Eq. (3) of Collins and Hribar (2002).

Variable name	Formula
Panel A: Beneish (1999a,b) *variables*	
Days Sales Receivables Index	$DSRI = \frac{RECT_t/SALE_t}{RECT_{t-1}/SALE_{t-1}}$
Gross Margin Index	$GMI = \frac{SALE_{t-1}-COGS_{t-1}}{SALE_{t-1}} / \frac{SALE_t-COGS_t}{SALE_t}$
Asset Quality Index	$AQI = \frac{1-ACT_t+PPENT_t}{AT_t} / \frac{1-ACT_{t-1}+PPENT_{t-1}}{AT_{t-1}}$
Sales Growth Index	$SGI = \frac{SALE_t}{SALE_{t-1}}$
Depreciation Index	$DEPI = \frac{DP_{t-1}-AM_{t-1}}{DP_{t-1}-AM_{t-1}+PPENT_{t-1}} / \frac{DP_t-AM_t}{DP_t-AM_t+PPENT_t}$
SG&A Index	$SGAI = \frac{XSGA_t}{SALE_t} / \frac{XSGA_{t-1}}{SALE_{t-1}}$
Leverage Index	$LVGI = \frac{DLTT_t+LCT_t}{AT_t} / \frac{DLTT_{t-1}+LCT_{t-1}}{AT_{t-1}}$
Total Accruals to Assets	$TATA = \frac{\Delta ACT-\Delta CHE-\Delta LCT-\Delta DD1-\Delta TXP-\Delta DP}{AT_t}$
Panel B: Dechow et al. (2011a) *Variables*	
WC Accruals	$WC_acc = \frac{[\Delta ACT-\Delta CHE]-[\Delta LCT-\Delta DLC-\Delta TXP]}{AvgTotalAssets_t}$
RSST Accruals	$RSST_acc = \frac{\Delta RSST_{WC}+\Delta RSST_{NCO}+\Delta RSST_{FIN}}{AvgTotalAssets_t}$
	$where\ RSST_{WC,t} = (ACT_t - CHE_t) - (LCT_t - DLC_t)$
	$and\ RSST_{NCO,t} = (AT_t - ACT_t - IVAO_t) - (LT_t - LCT_t - DLTT_t)$
	$and\ RSST_{FIN,t} = (IVST_t + IVAO_t) - (DLTT_t + DLC_t + UPSTK_t)$
Change in Receivables	$CHGRECC_t = \frac{RECT_t-RECT_{t-1}}{AvgTotalAssets_t}$
Change in Inventory	$CHGINV_t = \frac{INVT_t-INVT_{t-1}}{AvgTotalAssets_t}$
Percent Soft Assets	$PCTSOFT_t = \frac{AT_t-PPENT_t-CHE_t}{AT_t}$
Change in Cash Sales	$CHGCASHSAL_t = \frac{CSALES_t-CSALES_{t-1}}{CSALES_{t-1}}$
	$where\ CSALES_t = SALE_t - (RECT_t - RECT_{t-1})$
Change in ROA	$CHGROA = \frac{IB_t}{AvgTotalAssets_t} - \frac{IB_{t-1}}{AvgTotalAssets_{t-1}}$
Actual Issuance	$ACTISS = Indicator[SSTK >_{t-1} > 0 OR\ DLTIS >_{t-1} > 0]$
Panel C: Modified Jones (1991) *Model (modified in* Dechow et al. (1995)	
Discretionary Accruals	Residuals from $TA_t = \alpha_1 \frac{1}{TA_{t-1}} + \alpha_2 \frac{\Delta SALE-\Delta RECT}{AT_{t-1}} + \alpha_3 PPEGT_t + \epsilon_t$
Absolute Disc. Accruals	Absolute value of residuals from above model.
Panel D: Ball and Shivakumar (2006) *Model*	
Discretionary Accruals	Residuals from $TA_t = \alpha_1 \frac{1}{TA_{t-1}} + \alpha_2 \frac{\Delta SALE}{AT_{t-1}} + \alpha_3 \frac{PPEGT_t}{AT_{t-1}} +$
	$\alpha_4 CF_t + \alpha_5 Indicator[CF_t < 0] + \alpha_6 Indicator[CF_t > 0] CF_t$
	$\alpha_7 AR_t + \alpha_8 Indicator[AR_t < 0] + \alpha_9 Indicator[AR_t < 0] AR_t + \epsilon_t$
	$where\ AR_t = Firm\ \ Return_t - Market\ \ Return_t \ \ and \ \ CF_t = \frac{OANCF_t}{TA_{t-1}}$
Absolute Disc. Accruals	Absolute value of residuals from above model.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jcorpfin.2016.12.007.

References

Ball, C., Hoberg, G., Maksimovic, V., 2013. Disclosure, business change and earnings quality. University of Maryland Working Paper.
Ball, R., Shivakumar, L., 2006. The role of accruals in asymmetrically timely gain and loss recognition. J. Account. Res. 44, 207–242.
Beneish, M., 1997. Detecting GAAP violation: implications for assessing earnings management among firms with extreme financial performance. J. Account. Public Policy 16, 271–309.
Beneish, M., 1999a. The detection of earnings manipulation. Financ. Anal. J. 55, 24–36.
Beneish, M., 1999b. Incentives and penalties related to earnings overstatements that violate GAAP. Account. Rev. 74, 425–457.
Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1002.

Brazel, J., Jones, K., Zimbelman, M., 2009. Using nonfinancial measures to assess fraud risk. J. Account. Res. 47, 1135–1166.
Brown, S.V., Tucker, J.W., 2011. Large-sample evidence on firms' year-over-year MD&a modifications. J. Account. Res. 49, 309–346.
Buller, D., Burgoon, J., 1996. Interpersonal deception theory. Commun. Theory 6, 203–242.
Burns, N., Kedia, S., 2006. The impact of performance-based compensation on misreporting. J. Financ. Econ. 79, 35–67.
Cimiano, P., 2010. Ontology Learning and Population from Text: Algorithms Evaluation and Applications. Springer, New York.
Collins, D., Hribar, P., 2002. Errors in estimating accruals: implications for empirical research. J. Account. Res. 40, 105–134.
Core, J., Guay, W., Larcker, D., 2008. The power of the pen and executive compensation. J. Financ. Econ. 88, 1–25.
Coval, J., Stafford, E., 2007. Asset fire sales (and purchases) in equity markets. J. Financ. Econ. 86, 479–512.
Dechow, P., Dichev, I., 2002. The quality of accruals and earnings: the role of accrual estimation error. Account. Rev. 77, 35–59.
Dechow, P., Sloan, R., Sweeney, A., 1995. Detecting earnings management. Account. Rev. 70, 193–225.
Dechow, P., Sloan, R., Sweeney, A., 1996. Causes and consequences of earnings manipulation: an analysis of firms subject to enforcement actions by the SEC. Contemp. Account. Res. 13, 1–36.
Dechow, P., Ge, W., Schrand, C., 2010. Understanding earnings quality: a review of the proxies, their determinants and their consequences. J. Account. Econ. 2, 344–401.
Dechow, P., Ge, W., Larson, C., Sloan, R., 2011a. Predicting material accounting misstatements. Contemp. Account. Res. 28, 17–82.
Dechow, P., Hutton, A., Kim, J.H., Sloan, R., 2011b. Detecting earnings management, a new approach. J. Account. Res. 50, 275–334.
Erickson, M., Wang, S.W., 1999. Earnings management by acquiring firms in stock for stock mergers. J. Account. Econ. 27, 149–176.
Edmans, A., Goldstein, I., Jiang, W., 2012. The real effects of financial markets: the impact of prices on takeovers. J. Financ. 67, 933–971.
Feroz, E., Park, K., Pastena, V., 1991. The financial and market effects of the SEC's accounting and auditing enforcement releases. J. Account. Res. 29, 107–142.
Goldman, E., Slezak, S., 2006. An equilibrium model of incentive contracts in the presence of information manipulation. J. Financ. Econ. 80, 603–626.
Hanley, K., Hoberg, G., 2010. The information content of IPO prospectuses. Rev. Financ. Stud. 23, 2821–2864.
Hoberg, G., Maksimovic, V., 2015. Redefining financial constraints: a text-based analysis. Rev. Financ. Stud. 28, 1312–1352.
Hobson, J., Mayew, W., Venkatachalam, M., 2012. Analyzing speech to detect financial misreporting. J. Account. Res. 50, 349–392.
Holthausen, R., Larcker, D., Sloan, R., 1995. Annual bonus schemes and the manipulation of earnings. J. Account. Econ. 19, 29-74-392.
Humphreys, S., Moffitt, K., Burns, M., Burgoon, J., Felix, W., 2011. Identification of fraudulent financial statements using linguistic credibility analysis. Decis. Support. Syst. 50, 585–594.
Johnson, S., Ryan, H., Tian, Y., 2009. Managerial incentives and corporate fraud: the sources of incentives matter. Eur. Finan. Rev. 1, 115–145.
Jones, J., 1991. Earnings management during import relief investigations. J. Account. Res. 29, 193–228.
Karpoff, J., Lee, S., Martin, G., 2008a. The consequences to managers for financial misrepresentation. J. Financ. Econ. 88, 193–215.
Karpoff, J., Lee, S., Martin, G., 2008b. The cost to firms of cooking the books. J. Quant. Financ. Anal. 43, 581–612.
Kedia, S., Philippon, T., 2009. The economics of fraudulent accounting. Rev. Financ. Stud. 22, 2169–2199.
Kothari, S.P., Leone, A., Wasley, C., 2005. Performance matched discretionary accrual measures. J. Account. Econ. 39, 163–197.
Larcker, D., Zakolyunkina, A., 2012. Detecting deceptive discussions in conference calls. J. Account. Res. 50, 495–540.
Li, F., 2006. Do stock market investors understand the risk sentiment of corporate annual reports? University of Michigan Working Paper, SSRN Elibrary, Web Site.
Loughran, T., McDonald, B., Youn, H., 2009. A wolf in sheep's clothing: the use of ethics-related terms in 10-K reports. J. Bus. Ethics 89, 39–49.
McCornack, S., 1992. Information manipulation theory. Commun. Monogr. 59, 1–16.
Povel, P., Singh, R., Winton, A., 2007. Booms, busts, and fraud. Rev. Financ. Stud. 20, 1219–1254.
Sebastiani, F., 2002. Machine Learning in Automated Text Categorization. acmcs.
Securities and Exchange Commission, 2003. Interpretation: Commission Guidance Regarding Management's Discussion and Analysis of Financial Condition and Results of Operations,17 CFR Parts 211, 231 and 241,Release Nos. 33-8350; 34-48960; FR-72.
Skinner, D., Sloan, R., 2002. Earnings surprises, growth expectations, and stock returns or don't let an earnings torpedo sink your portfolio. Rev. Acc. Stud. 7, 289–312.
Wang, T., 2013. Corporate securities fraud: insights from a new empirical framework. J. Law Econ. 29, 535–568.
Wang, T., Winton, A., Xiaoyun, Y., 2010. Corporate fraud and business conditions: evidence from IPOs. J. Financ. 65, 2255–2292.
Zuckerman, M., DePaulo, B., Rosenthal, R.R., 1981. Verbal and nonverbal communication of deception. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology. vol. 14. Academic Press, New York, pp. 1–59.